Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

msii ÉCOLE DOCTORALE

DOCTORAL SCHOOL MSII

ICube Laboratory (UMR 7357)

Research Group CAMMA

Computational Analysis and Modeling of Medical Activities

THESIS presented by

Chinedu Innocent Nwoye

Defended publicly on: November 16, 2021

For obtaining the degree of **Doctor of Philosophy (P.hD.)**
From **the University of Strasbourg** delivered by the doctoral school MSII

Field: **Computer Science, Computer Vision, Deep Learning and Medical Robotics**

Deep Learning Methods for the Detection and Recognition of Surgical Tools and Activities in Laparoscopic Videos

THESIS DIRECTORS:	
Prof. Dr. Nicolas Padoy	Professor of Computer Science, Université de Strasbourg, IHU, Strasbourg, France
Prof. Dr. Didier Mutter	Professor of Surgery, Head of Hepato-Digestive Surgery, Hôpitaux Universitaires de Strasbourg, IRCAD, IHU Strasbourg, France
CHAIR OF THE JURY:	
Prof. Dr. Nassir Navab	Professor of Computer Science, Technische Universität München, Germany, Johns Hopkins University, USA
EXAMINERS:	
Prof. Dr. Stefanie Speidel	Professor of Translational Surgical Oncology, National Center for Tumor Diseases (NCT) Dresden, Germany
Prof. Dr. Raphael Sznitman	Professor of Computer Science, University of Bern, Switzerland

Wings are a constraint
that makes it possible to fly.
. . . Robert Bringhurst

Dedicated to my parents

Chief Chinwetoke Pius NWOYE & Mrs. Chinyere Mercy NWOYE (née Ọzọnze)

Abstract

Surgery, a core unit of the patient care system, is increasingly improving due to continuous technological innovations facilitating better patient outcomes and providing rich intraoperative data via information systems. This, however, increases the complexity of the workflow procedures, as well as, the surgeons' cognitive workload. Consequently, there is an increasing need to optimize surgical workflow via intelligent and analytical systems that can provide decision support and context-aware assistance to the surgeons. Despite the vast literature on activity recognition in medical computer vision, the coarse-grained nature of the tasks mostly tackled, e.g. recognizing surgical phases, are not detailed enough for a more helpful AI assistance in the operating room (OR). Modern high-tech surgery rooms require a more detailed activity recognition system: one that can meticulously capture finer actions, such as interactions between the instrument and tissue, and comprehensively describe the activities taking place.

In this thesis, we focus on the development of deep learning methods for the detection and recognition of surgical instruments and their fine-grained activities in laparoscopic videos. These activities are formalized as triplets of $\langle$*instrument, verb, target*$\rangle$ representing the tool-activity. We investigate, firstly, joint detection and tracking of surgical instruments in laparoscopic videos. To alleviate the difficulty of manually generating bounding box annotations for instruments in every video frame, we develop a novel localization method that is weakly supervised on binary presence labels, which are easier to generate. To leverage the temporal structure of surgical videos, we propose the use of a Recurrent Neural Network to track the motion of instruments, still without requiring any form of spatial training labels. Moreover, we create a large video dataset with spatial labels, which we use to validate the proposed method. Progressing to activity modeling, we generate a large-scale dataset of surgical action triplets and build several deep learning models for their recognition. First, we design a recognition pipeline that learns the individual components of the triplets using CNN features and establishes their association in a 3D feature space, as a frame can contain multiple triplets. Improving on the first method, we propose a new form of spatial attention to capture the individual triplet components more efficiently using activations resulting from the instruments. Furthermore, we introduce a new form of semantic attention, inspired by Transformer networks, to learn triplet components' association. Finally, we validate all the proposed approaches on the datasets introduced in this work, achieving state-of-the-art performance on each task.

Acknowledgments

Finally, brethren, whatever is true, whatever is honorable, whatever is right, whatever is pure, whatever is lovely, whatever is of good repute, if there is any excellence, and if anything worthy of praise, let your mind dwell on these things.
– Philippians 4:8

This thesis is the result of nearly four years of work at the *CAMMA, ICube Laboratory*. I return all the glory to Almighty God, whose grace and divine providence have been sufficient for me. My grandmother named me *Chukwunedumije* (abbreviated as *Chinedu*) - which translates to *God directs my paths.*

In the course of this thesis, I received exceptional teaching and support from numerous people which contributed immensely to the successful completion of my doctoral program. I dedicate this section to appreciating their kindness.

I would like to start by thanking my thesis supervisor, Prof. **Nicolas Padoy**, for welcoming and introducing me to the field of Computer-Assisted Interventions. I am incredibly grateful for his dedication, in both time, space, resources, and efforts, guiding me towards a painstaking and rigorous research culture. I appreciate his immeasurable patience and intimidating brilliance which motivate me to never settle for anything short of excellence. The most important take-away lesson that I learned from him is that "*if it is doable and worth doing, then I can do it and do it best*". His aura always pushes me to aspire for the highest pinnacle of glory. He is not only my thesis director, but also my discussant, friend, and mentor. I will forever remain indebted to him. I also appreciate my co-supervisor, Prof. **Didier Mutter**: knowing that I am also guided by a professor of Surgery builds in me the confidence needed in my field of research which intersects with surgery.

Special thanks go to my thesis committee: Prof. **Nassir Navab**, Prof. **Stefanie Spediel** and Prof. **Raphael Snzitmann**. I thank them for accepting to review my manuscript, and for their constructive comments, feedback, and august physical presence during my defense amidst *C*ovid-19 pandemic. I am most honored to be assessed by the very best and most esteemed members of the research community.

I appreciate the subject matter conversations and guidance that I received from Dr. **Pietro Mascagni**. I thank him for ensuring that my research is clinically sensible and surgically explainable. I appreciate his patience and tolerance especially in my request for repeated explanations and clinical significance reviews of my research findings over short notice and limited deadlines. In the same vein, I appreciate Dr. **Cristians Gonzalez** and Dr. **Barbara Seeliger** for their help with annotating the surgical data needed for research. Apart from being highly instrumental in the creation of CholecT40 and CholecT50 datasets, the clinical collaboration with them gave my research a beautiful coloration with a touch of precision. I thank them for always being available to answer all my questions on the surgical aspect of my research data.

My deepest gratitude goes to the awesome members of *CAMMA Research Group*. I thank especially my fellow Ph.D. Students, **Tong Yu** and **Vinkle Srivastav**, for sharing research ideas, scholarly views, and running around with me during administrative and academic endeavors of our doctoral program. Three of us started the same period and ended the same period. We saw all the years and struggles together and we conquered. I cannot forget the early positioning I received from the past Ph.D. student, Dr. **Nicolas Loy Rodas** and past post-doctoral fellow, Dr. **Rahim Kadkhodamohammadi**. They made settling and beginning a doctoral program in a new (French) environment and culture easier for me. I thank them for sharing their wisdom with all of us.

In sharing knowledge within the team, I received technical contributions from Research Engineers such as **Deepak Alapatt, Gaurav Yengera**, and **Armine Vardazaryan**. I had the opportunity to work with Interns: both those that I mentored: **Anvita Srinivas** and **Girinath Ravichandran**, and those I indirectly work with: **Siddharth Kannan, Fauzan Zaid, Taha Emre**, and **Cristian Rocha**. Collaborating with them helped me in keeping up to date with the basics, advances, and updates on Deep Learning concepts, frameworks, and libraries. Also, **Nicolas**'s effort at building the CAMMA group as a yardstick for professionalism brought me in contact with senior researchers: Dr. **Jean-Paul Mazellier** who is always on his toes to see that we are duly organized. I admire his selflessness a lot. Likewise, Dr. **Georgios Exarchakis** and Dr. **Alexandros Karargyris** for stepping in to provide further assistance and guidance. Their zeal, knowledge, and expertise encourage me to broaden my scope of learning.

My appreciation will be incomplete without mentioning Dr. **Isabella Bolognese**, the CAMMA administrative assistant. I am most humbled by her ultra-fine attitude to work and service. Within a short period, she already made an indelible impression that I will forever cherish, particularly helping me to prepare for my thesis defense. Did I forget to mention that I have a French tutor among my colleagues in the person of **Cindy Rolland**? I am incredibly grateful for the time and energy she put in teaching me some basics of the French Language and helping me to translate my documents and mails during these years. Indeed, I am surrounded by amazing teammates who, apart from offering suggestions on experiment tuning and troubleshooting, spent quality out-of-work time discussing news, football, stock market, books, movies, politics, economy, career advice, people's culture, and sometimes, funny discussions over a coffee or a beer. I thank them for helping me to balance work and life. The group which was initially led by **Luca Sestini** but now led by **Sanat Ramesh** comprises of, apart from the already mentioned teammates, the current Ph.D. students: **Idris Hamoud, Aditya Murali**, postdoctoral fellow: **Alexandre Krebs** current and past research engineers:

Antoine Fleurentin, Adrien Meyer, current and past interns: **Hasan Kassem, Thibaut Issenhuth, Jérémy Ramboasolo, Sarah, Vincent Roth, Emmanuelle Richer, Aakriti Agrawal, Rohit Arasanipalai, Philippe de Rooy, Pauline Guyot** and collaborating medical doctor Dr. **Maria Vannucci**.

I sincerely thank my colleagues at ICube's *Automatique Vision et Robotique* (**AVR**) group, as well as the personnel at **IRCAD** and **IHU** Strasbourg for their encouragement and inspiration. They have been so generous, friendly, and supportive. They made our workplace welcoming, vibrant, and lively. I thank them for all the refreshing seminars and get-together events over the years.

Finally, I would like to thank my family. I am particularly indebted to my parents for their limitless love and sacrifices in my upbringing. I appreciate the love and encouragement from my siblings: **Ikechukwu, Ifunanya, Chijioke, Chibuike,** and **Chinyere**. They have been supportive in my getting a higher education and pursuing my goals even if it is far away from them and my country of birth. I appreciate their understanding and letting me stay focused on my programs thereby staying away from home for five years now. My friends especially those back home have been of moral support, particularly **Isabella Iloka**, who immensely believed in me and constantly encourage me especially in times of despair. I cherish her genuine love and selfless concern for my well-being especially at the critical and terminal period of my program. Indeed, she has a special place in my heart.

Completing my doctoral program was possible because of financial support from French state funds managed within the Investissements d'Avenir program by BPI France (project CONDOR) and by the ANR under references ANR-11- LABX-0004 (Labex CAMI), ANR-16-CE33-0009 (DeepSurg), ANR-10-IAHU-02 (IHU Strasbourg) and ANR-20-CHIA-0029-01 (National AI Chair AI4ORSafety). I acknowledge the access privilege I received for the use of HPC resources of Unistra Mesocentre and GENCI-IDRIS (Grant 2021-AD011011638R1 and 2020-AD011011638).

May the Lord reward you for your kindness. *Ruth 1:8.*

Sincerely grateful,

Chinedu Nwoye

C.I.D. Nwoye
Strasbourg, 15 Nov. 2021

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

3D Three Dimensional. 7, 25, 35, 38, 50, 51

3Dis 3D Interaction Space. 85, 90, 92–94, 96, 98, 107, 110, 122, 123

AI Artificial Intelligence. 6, 9, 12, 17, 19, 23, 26, 80, 149, 150

AP Average Precision. 72, 73, 95, 96, 129

ASPP Atrous Spatial Pyramid Pooling. 172

BDI bile duct injury. 16

CAG Class Activation Guide. xxi, xxiii, 85, 90–92, 96, 97, 106, 109, 110, 113, 157

CAGAM Class Activation Guided Attention Mechanism. xviii, 106–108, 110, 112, 113, 117–119, 124, 131, 137, 157

CAI Computer Assisted Intervention. 5, 6, 9–12, 14–16, 30, 47, 49, 148–150, 152

CAM Class Activation Map. 92, 106–108

CholecTriplet2021 Surgical Action Triplet Recognition 2021. 27

CNN Convolutional Neural Networks. xvii, 14, 22, 31, 32, 34–36, 38, 44, 49, 51, 61, 67, 87, 88, 97, 113–115, 130, 132–134, 136, 169

ConvLSTM Convolutional Long Short Term Memory. xvi, xvii, 25, 40, 49, 51, 61, 62, 65–77, 169, 174, 175

CRF Conditional Random Field. 45

CVS Critical View of Safety. 16

DCNN Deep Convolutional Neural Networks. 60

FCN Fully Convolutional Network. 37, 61

FN False Negative. 72, 133

FP False Positive. 72, 133

GCN Graph Convolutional Networks. 32, 36, 51, 169

GMP Global Maximum Pooling. 107

GPNN Graph Parsing Neural Network. 45

GRU Gated Recurrent Unit. 33, 35, 40, 50

HAKE Human Activity Knowledge Engine. 44

HAR Human Activity Recognition. 43

HOI Human-Object Interaction. xvi, 27, 43–46, 49, 52, 53, 80, 169

IDSW Identity Switch. 72

IGS Image-Guided Surgery. 7

IoU Intersection over Union. 72

JDE Joint Detection and Embedding. 41

Lh-map Localization Heat Map. xvii, 63, 68–70, 75, 76, 148

LSTM Long Short Term Memory. 22, 33, 35, 45, 49–51, 61, 65, 67, 156, 168, 169

mAP Mean Average Precision. 96, 102, 169

MCMOT Multi-Class Multi-Object Tracking. 40, 74

MCT Multi-Class Tracking. 74

MHMA Multi-Head of Mixed Attention. 123, 125, 126, 158

MIS Minimally Invasive Surgery. 5, 7, 8

mL Mostly Lost. 40

MLP Multiple Layer Perceptron. 26

MOT Multiple Object Tracking. 40, 51, 72

MOTA Multiple Object Tracking Accuracy. 40, 72, 74, 77

MOTP Multiple Object Tracking Precision. 40, 72, 74

MRF Markov Random Field. 32

MRI Magnetic Resonance Imaging. 7

mT Mostly Tracked. 40

MTL Multi-Task Learning. 85, 88–92, 97–99, 112–116, 131–134, 136

NLP Natural Language Processing. 26

OR Operating Room. xv, 6, 10–12, 14–18, 21, 22, 24, 30, 46, 47, 51, 80, 148–150, 152, 156, 158, 166, 169, 170

RCNN Region-Based Convolutional Neural Network. 34, 44

RDV Rendezvous. xix, 123, 125, 128–132, 136, 137, 139–141, 158

RNN Recurrent Neural Network. 14, 25, 32, 33, 40, 47, 65, 67, 68, 76, 169

ROI Region of Interest. 37

SCMOT Single Class Multiple Object Tracking. 40, 51

SDE Separate Detection and Embedding. 41

SDS Surgical Data Science. 11, 20, 30

SGD Stochastic Gradient Descent. 71, 94, 172

SORT Simple Online and Realtime Tracking. 39

SOT Single Object Tracking. 40

SOTA state-of-the-art. 5, 21, 26, 43, 113, 114, 116, 122, 131, 137, 170

SQE Self Quality Evaluation. 40

TCN Temporal Convolution Network. xvi, 48, 113–115, 132–134, 136, 169

TP True Positive. 133

WSL Weakly-Supervised Learning. 107

Part I

Introduction, Motivation, and Related Work

1 Introduction

But biology and computer science - life and computation - are related...
I am confident that at their interface great discoveries await those who seek them
– Leonard Adleman

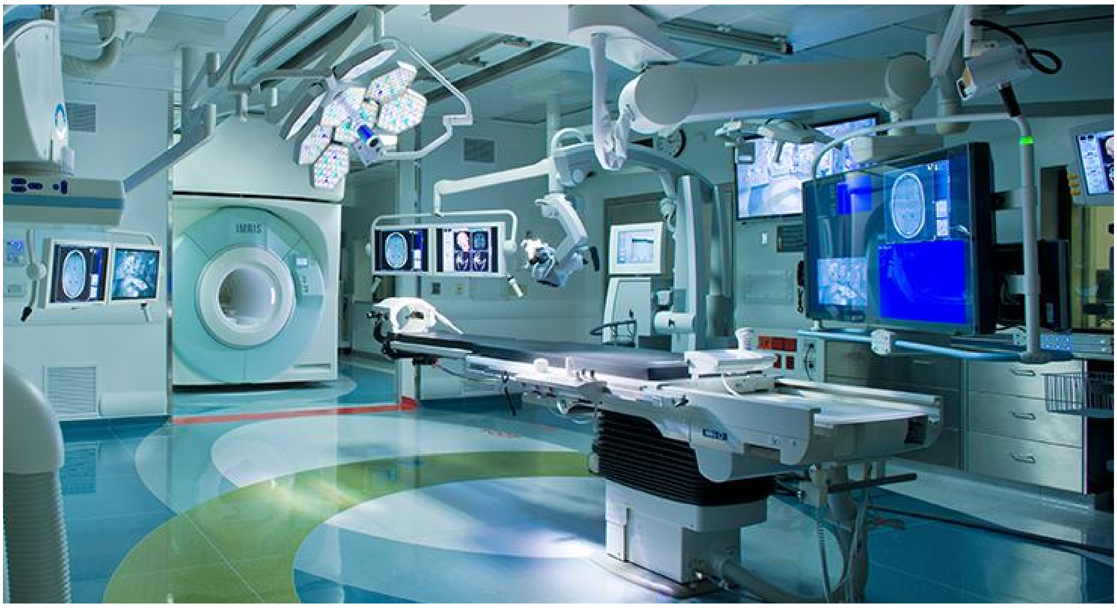


Figure 1.1 – An example of a hybrid operating room that combines a traditional operating room with an image guided interventional suite. AMIGO suite at Brigham and Women's Hospital, Boston, Massachusetts. *Image Credit: Copyright 2015 IMRIS, Inc.*

Chapter Summary

To remain alive, the human body must be kept in a healthy condition. When the body health deteriorates, medicine and medicinal foods are used to treat the ailing conditions. In some cases, the affected tissue or organ may be repaired or removed through a manual or instrumental technique known as surgery. According to Encyclopedia Britannica, surgery is a field and practice of medicine that involves the manipulation of bodily structures for diagnostic or therapeutic purposes. A number of medical conditions can only be effectively treated through surgery, especially those conditions that cannot be sustained over a long period without vital organ dysfunction. Post-surgery in such cases is usually characterized by relief, a surge in energy, and increased life expectancy. In the earlier days, most surgeries are carried out in open procedure which involves the cutting of skin and tissues large enough so that the surgeon can have a full view of the structures or organs to be operated upon. This type of procedure is usually associated with intense pain and long hospital stay due to the postoperative time required to heal the manipulated structures and incisions made on the body. Deciding to get surgery under such a scenario can feel overwhelming. Aside from the associated intraoperative pain and postoperative infection which can now be controlled by the patient's anesthesia and antiseptics (introduced in 19^{th} century), some surgical errors could also lead to life-threatening complications, re-admission, re-operation, and sometimes death [Nathan 2012, Birkmeyer 2013]. And so, many efforts are being made to improve surgery with lots of them focusing on the minimization of the risks associated with intraoperative errors. The expectation is to develop a surgical practice that is safe, effective, and efficient.

One of the most prominent improvements in surgery is the advent of Minimally Invasive Surgery (MIS): tiny holes are made on the body, and through the use of an endoscope and endoscopic devices, surgery is performed at a proxy. This type of surgery is enabled by advances in technology such as the introduction of cutting-edge surgical instruments, novel imaging technologies, control, monitoring, and support systems used during the procedure to provide greater control of the surgical procedure and reduced tissue trauma and disruption while granting better access to the anatomical structure. The minimally invasive technique creates a bedrock for the other technological innovations in surgery such as endoscopy, robot-assisted surgery, robotic surgery, etc., with improved patient outcomes. By being less traumatic and less invasive, it significantly alleviates some preoperative, intraoperative, and postoperative burden thereby leading to shorter hospital stay [Velanovich 2000]. However, this success comes at a price for the surgeon, who now deals with increased technical difficulty coming from the indirect vision and non-conventional handling of advanced surgical instruments [Ballantyne 2002], especially during complex surgical cases [Felli 2019].

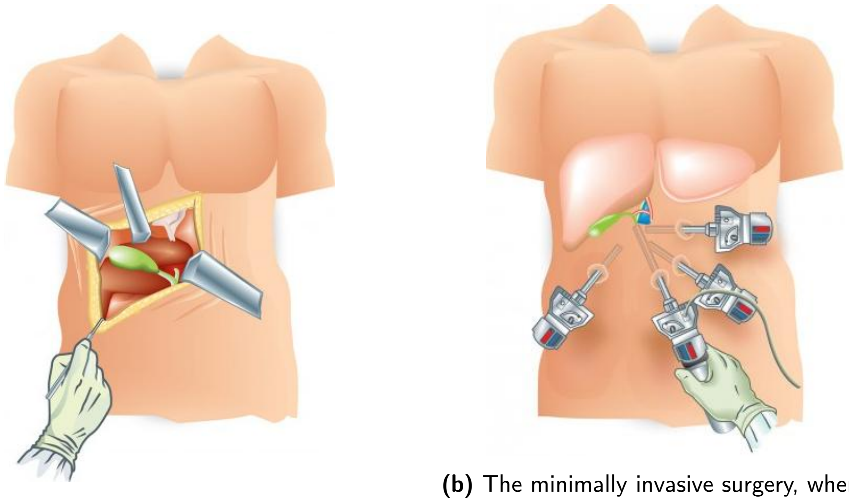
The elevated complexity of MIS is one of the motivations driving the development of context-aware support systems for surgery [Lemke 2005]; i.e. systems capable of assisting surgeons, for example via automated warnings [Vercauteren 2019], based on their dynamic perception and understanding of the surgical scene and workflow. And so, a copious amount of research in surgical data science [Maier-Hein 2017] provides data-driven computational models that are capable of extracting knowledge from medical data with reliability, accuracy, and speed [Vercauteren 2019]. Facilitated by the acquisition of a large amount of data using the novel imaging technologies, such as endoscopes, a growing number of state-of-the-art (SOTA) algorithms are being developed using Machine Learning and Artificial Intelligence methods to provide automated analysis of workflow [Simpson 2019, Gibson 2018]. These algorithms tackle several tasks including the classification of surgery type, detection of the used instruments, recognition of the surgical phases, prediction of activities within the phases, etc., providing information that is helpful for the development of Computer Assisted Intervention (CAI) in the OR. While there has been commendable progress in workflow analysis, the activity recognition is still very much at a coarse level, i.e., they are not detailed enough to give accurate information about the activities taking place. To obtain a comprehensive account of surgical activity, simultaneous recognition of the tools, the underlying tissues, and the relationship between them needs to be established.

The principal **aim of this thesis is to develop smart computational aids for the detection of surgical instruments and the recognition of their fine-grained activities in laparoscopic videos**. On the aspect of surgical instrument detection, the objective is to devise a new deep learning method that can learn from simple image-level labels the spatial positions and motion trajectories of surgical instruments - a complex task that would have ordinarily required expensive and difficult-to-generate spatial annotations for model training. This approach is known as weak supervision. On the activity recognition task, the objective is to build novel deep learning models that can recognize surgical actions as triplets of ⟨instrument, verb, target⟩ - a detailed level of granularity for activity recognition that is needed for a more helpful Artificial Intelligence (AI) assistance in the OR. This is known as surgical action triplet recognition [Neumuth 2006, Katić 2014, Katić 2015]. These objectives formed the central work carried out in this thesis. The benefit of the thesis is that decision support tools built from this research can be used to develop CAI solutions in the OR. Those tools will enable standardization of and objectivity in surgical care to provide better assessment, early detection of errors, safety monitoring, and guidance leading to improved patient care [Garrow 2021]. To facilitate this work, we also generate large fine-grained datasets to support research at this level of details.

In the following sections, we begin to dissect the clinical context for this thesis by throwing light into the background of the surgical procedure of our concern. There, we discuss the advancement in surgery that serves as the bedrock for the research and development of CAI systems in the OR. Going further, we present the problems being researched in this field while identifying a research gap that forms the central point in our research proposal. Thereafter, we highlight the motivations for our research, emphasizing the thesis interest in surgical tool and activity recognition, their challenges, research questions, and perspectives. This is followed by a presentation of the contributions of this work towards providing new deep learning solutions in this domain of research and concluded by outlining the structure of the thesis.

1.1 Clinical Context and Background

The growing popularity and acceptance of technology in our time is due to its perceived ability to enhance human performance in many spheres of life. Technology has also been utilized in surgery to improve surgeon's competence and provides access to relevant and complementary information that could help in reducing risks for patients. This has invariably increased the patients' confidence in getting surgery. Since technology and its methods are still emerging, there are increasing opportunities to advance interventional medicine. In this section, we present a brief overview of the surgical procedure of interest in this thesis. This will also include highlights on the innovative transformations that have created the enabling environment for the use of CAI systems in optimizing the surgical procedure. We will then conclude this overview section with the nature of ongoing studies modeling surgical workflow to provide the needed CAI solutions.



(a) The traditional open surgery, whereby a big incision is made to have a direct broad view of the anatomy being operated on.

(b) The minimally invasive surgery, where a tiny hole-like incision is made and the surgeons have indirect view of the anatomy via a computer screen.

Figure 1.2 – Two different procedures for cholecystectomy surgery (open vs laparoscopic) based on the size of incisions made during surgery. *Images obtained from* http://airnmed.com .

1.1.1 Image Guided Surgery

Image-Guided Surgery (IGS) is not only a direct product of surgical innovation but also one of the most notable enabler of other technological advancement in surgery. It is an umbrella term including all interventions performed looking at and relying on digital images [Bucholz 1995]. The idea dated back to the 1940s when Sir Victor Horsley and Robert Clarke developed the first stereotactic frame fitted over the head of a patient to undergo brain surgery [Galloway 2015]. At that time, IGS only refers to stereotactic surgery where preoperative images are registered to the surgical space through the use of reference markers called *fiducials* and a tracking device which displays the surgeon's anatomic position on Three Dimensional (3D) reconstructions of the preoperative films [Bucholz 1995]. IGS was originally developed for the treatment of brain tumors using stereotactic surgery and radiosurgery that are guided by intraoperative computed tomography (CT), Magnetic Resonance Imaging (MRI), and positron emission tomography (PET), all these targeted toward electro-physiological measures for refinement of position, optical localization, and image guidance [Galloway 2015]. A central aspect of IGS is creating accurate, detailed, patient-specific models from medical imagery. Using the model, surgical instruments are tracked relative to the patient, allowing the surgeon to effectively execute procedures while avoiding hidden, critical structures [Grimson 1999]. This also alleviates pains on the patient's side with fluoroscopy being the first image guidance technology adopted by pain specialists. IGS can be broadly classified into **minimally invasive surgery (MIS)**, interventional endoscopy, and interventional radiology (also called percutaneous surgery) [Mascagni 2018].

MIS refers to any surgical procedure performed through tiny incisions as shown in Fig. 1.2b instead of a large opening as in the traditional open surgery shown in Fig. 1.2a. This form of surgery is performed using miniaturized cutting-edge instruments supporting the drastic reduction of the size of incisions needed to gain access to the surgical site [Litynski 1999]. During MIS, some keyhole-like ports are made on the patient's body. The size of these ports is now dependent on the size of the instruments (e.g. trocar) [Fuchs 2002, Westebring-van der Putten 2008] rather than on the surgeon's holistic view of the surgical site as the case is in open procedures. The implication is that the patient generally experiences a decrease odds of nosocomial infection, less pain, less bleeding, and faster recovery times [Velanovich 2000, Olsen 1991] compared to the open surgery.

MIS is performed using a **rigid endoscope** held manually or tele-manipulated as opposed to interventional endoscopy, which uses flexible endoscopes intervening through natural orifices such as mouth, anus, etc. Note that MIS can be performed as well using surgical robots. In rigid endoscopy, a surgeon accesses the surgical site through trocar ports formed by minimally invasive incisions made on the patient's body [Litynski 1999]. Some specialized surgical instruments (such as electrified bipolar, hook, scissors, etc.) are passed via the trocar and the procedure is performed with the aid of the light and camera which allows the surgeon to see the inside of the patient's body via a monitor screen. Using the endoscopes, some of the hand motions, particularly the articulations, are transmitted to the instrument tips [Gaab 2013] for the dexterous manipulation of the anatomies. According to Healthline [1], rigid endoscopy falls into categories, based on the area of the body that they investigate. Some examples include: **laparoscopy (abdominal area)**, arthroscopy (joints), bronchoscopy (lungs), colonoscopy (colon), thoracoscopy (chest), ureteroscopy (ureter), etc. This thesis focus on a type of laparoscopic surgery known as *Laparoscopic Cholecystectomy*.

1.1.2 Laparoscopic Cholecystectomy

Laparoscopic cholecystectomy is a minimally invasive surgery that involves the removal of the gallbladder from the body [Olsen 1991]. Gallbladder removal is essentially carried out due to the presence of painful gallstones otherwise known as *Cholelithiasis* [Schirmer 1991] and the complications they cause. Meanwhile, a patient can live a normal life after the gallbladder is removed. The procedure is characterized by the dissection of the gallbladder from the surrounding anatomies, clipping, and cutting of structures (such as cystic-duct, cystic-artery, and other blood vessels) connecting to the gallbladder, and extraction of the gallbladder from the body using a specimen bag (see also Figure 1.3). Laparoscopic cholecystectomy is one of the most commonly performed surgical procedures in the world [Shaffer 2006]. It has become the gold standard approach to cholecystectomy [Pucher 2018] owing to its attributed low profile risk in removing the gallbladder. It is frequently used in research due to its high frequency of occurrence and well standardize protocol [Padoy 2012].

Being a type of minimally invasive surgery MIS, it equally enjoys a significant reduction of some preoperative, intraoperative, and postoperative issues such as pain and pain medication, invasiveness, blood loss, and recovery time [Olsen 1991, Velanovich 2000]. Nonetheless, its clinical outcome is only comparable to the traditional open surgeries [Ballantyne 2002]. Just

[1] https://www.healthline.com/health/endoscopy#types

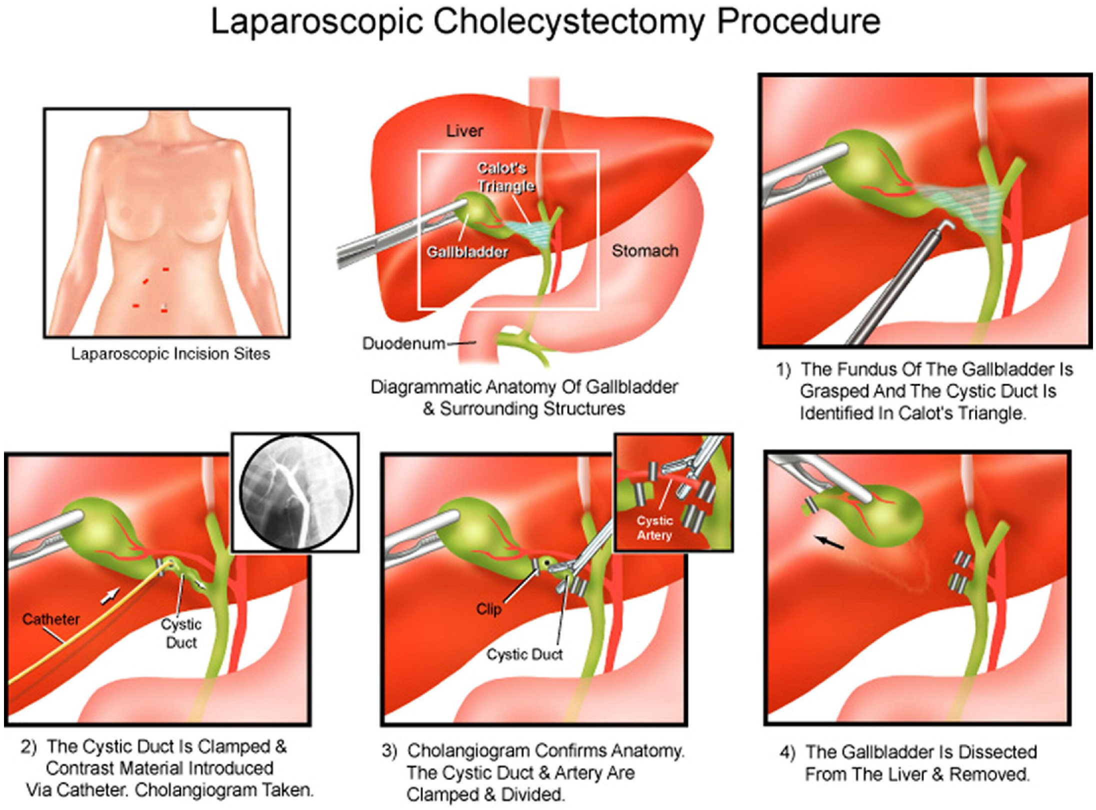

Laparoscopic Cholecystectomy Procedure
Laparoscopic Incision Sites
Liver
Calot's Triangle
Gallbladder
Stomach
Duodenum
Diagrammatic Anatomy Of Gallbladder & Surrounding Structures
1) The Fundus Of The Gallbladder Is Grasped And The Cystic Duct Is Identified In Calot's Triangle.
Catheter
Cystic Duct
2) The Cystic Duct Is Clamped & Contrast Material Introduced Via Catheter. Cholangiogram Taken.
Cystic Artery
Clip
Cystic Duct
3) Cholangiogram Confirms Anatomy. The Cystic Duct & Artery Are Clamped & Divided.
4) The Gallbladder Is Dissected From The Liver & Removed.

Figure 1.3 – A sequence of laparoscopic cholecystectomy. (source: [Massarweh 2007]) .

like open cholecystectomy, laparoscopic cholecystectomy is not immune to surgical errors. In fact, the non-conventional way of manipulating the laparoscopic instruments coupled with indirect observation of surgical scene via the screen of a monitor denies surgeons their conventional hand-eye coordination and direct hepatic feedback making it susceptible to visual illusion [Ballantyne 2002, Mascagni 2020, Mascagni 2021b].

Fortunately, laparoscopic cholecystectomy, compared to open cholecystectomy, has better vantage points to overcoming surgical risks. One of which, like other laparoscopic procedures, is that it enables a large amount of data to be acquired via the endoscope for their analysis. Leveraging this support, surgical workflow analysis on these data can help to develop intra- and post-operative context-aware decision support tools [Maier-Hein 2017] that can assist the surgeons to perceive, interpret, plan, and act on the digital visual data potentially fostering increased surgical safety and efficiency and decrease risk further. Also, the design of laparoscopy makes it easy for the integration of AI aids in surgery. This is because the use of computer systems and cameras in laparoscopy already provides imaging, deployment, and visualization platforms. The captured procedural data are readily available in digital form that AI systems can directly process online and in real-time. The analysis can be displayed on or next to the computer screen that the surgeons are already monitoring the procedure from.

In the next section, we will discuss the surgical workflow analysis in CAI in order to better understand the place and usefulness of the work done in this thesis.

1.1.3 Surgical Workflow Analysis

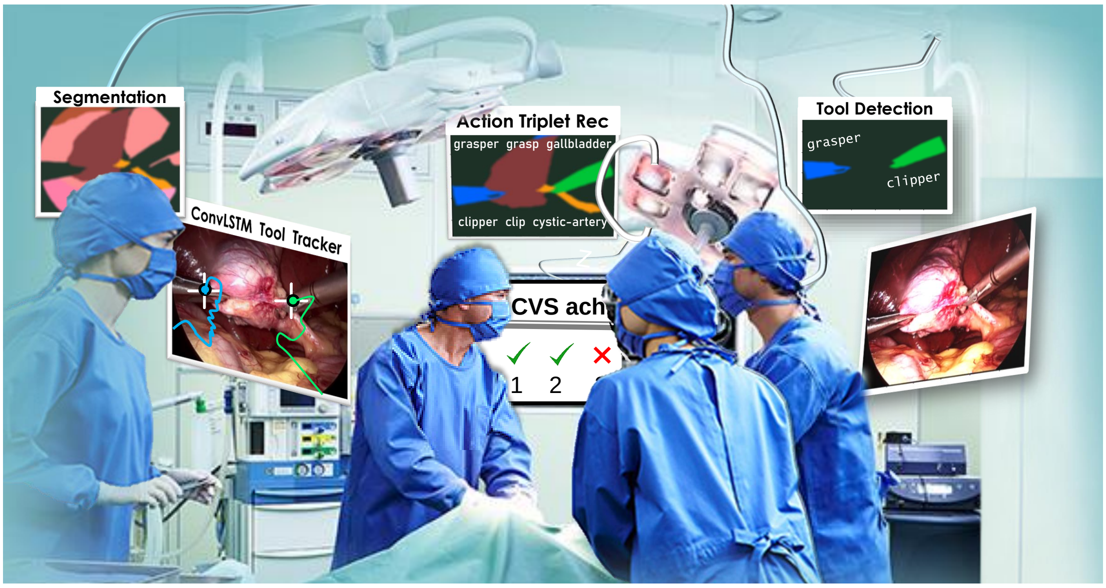


Figure 1.4 – An illustration of a context-aware assistance in the operating room using some of the systems developed in this thesis.

The technological advancement in medicine have not only improved medical practices, but also transformed surgery from a risky "art" into a scientific discipline capable of treating many diseases and conditions [Misra 2017, Twinanda 2017]. Some surgeries may require highly adaptive assistance systems [Kranzfelder 2013b, Jiang 2017, Liew 2018] which analyze surgical workflows and provide context-aware assistance [Lemke 2005] in the OR. Slowly, surgery is merging with technical disciplines and the procedure becoming more and more complex. There is now a growing need for documentation, computational model, and to analyze and support the surgical practice by means of computers and robots. This is one of the motivations driving the development of CAI systems to utilize the pre-operative and intra-operative patient-specific information from different sources, sensors, and imaging modalities to enhance the workflow, control, ergonomics, and navigation capabilities during surgery [Mirota 2011, Stoyanov 2012].

Computer-Assisted Intervention (CAI) is a field of research and practice, where medical interventions including clinical decisions are supported by computer systems and technology with the aim of augmenting the capability of clinicians to achieve a better clinical outcome. It encompasses medical robotics [Hager 1995, Hager 1996, Speidel 2014, Wagner 2021a, Vander Poorten 2020], interventional navigation [Navab 2002, Pfeiffer 2019b], intraoperative decision supports [Speidel 2006, Padoy 2008, Sznitman 2011], medical imaging [Navab 1999, Fitzek 2021], augmented reality and visualization [Navab 2007, Navab 2012, Rodas 2015], workflow and skill analysis [Speidel 2006, Speidel 2009, Jin 2018], surgical training [Stefan 2020], etc.. Among various functionalities of CAI, it is also expected to provide context-awareness assistance and intelligent decision support systems in the OR. Context-aware assistance entails knowledge representation and useful signals that can interactively update the surgeon on the state of the intraoperative procedure such as providing timely information through surgical phases [Maier-Hein 2017]. This type of information is useful in optimizing operating time, analyzing technical requirements, anticipating patient positioning, evaluating surgical skills, guiding against an unintentional use of instruments, and improving the pre-operative human-computer interface [Lalys 2014]. Decision support entails providing some complementary aids that can help surgeons to optimize their surgical decisions, for instance, retrieving of patient's history information, browsing decisions on similar cases from a surgical database, validating a tumor, quantifying blood loss, crosschecking safety checkpoints, measuring the length of structures in bypass, etc. Such informed context-aware and decision support systems could be developed from procedural data analysis in a scientific discipline known as Surgical Data Science (SDS).

SDS sets in to observe the preoperative, intraoperative, and postoperative activities with the aim of improving the value and quality of interventional healthcare [Maier-Hein 2017]. Its involvement is via data capturing, organization, analysis, and modeling [Maier-Hein 2017, Maier-Hein 2020]. While the advanced devices can provide increasingly more information from surgical procedures (e.g. surgical videos, instrument use, staff participation, instrument trajectories, etc.) [Kranzfelder 2014], it is SDS that is concerned with the analysis and understanding of the OR activities [Lalys 2014] using the data. This analysis can provide surgeons with quantitative support to aid decision-making and surgical actions, one of the fundamental needs in CAI.

Recently, surgical workflow analysis has become an active research area in surgical data science [Neumuth 2009, Maier-Hein 2017]. It is aimed at the automatic recognition of a predefined subset of tasks, or activities of interest by following surgical processes with real-time analysis of live video data acquired intraoperatively [Ahmadi 2006]. The surgical process of interest can be a set of one or more linked steps that collectively realize a surgical objective within the context of an organizational structure defining functional roles and relationships [Neumuth 2009]. The idea of describing the surgical procedure as a sequence of tasks was first introduced by [MacKenzie 2001] and formalized in [Jannin 2001]. The formalization allows surgical processes to be represented at the appropriate level of granularity (e.g.: activities, phases, steps, etc.) for the requisite decision making. Surgical Workflow Analysis identifies the stages of a procedure and gives guidance on what tool to use next, what the next step should be, or by displaying pertinent information at any given time. For example, if the operation is to remove a tumor, it can be relied on to determine if the growth is visible, or warn when an instrument is approaching a no-go zone in the body [Speidel 2008]. So, a more precise decision-making process would actually require filtering data and knowledge about surgical actions, instruments, anatomical structures, phases, and the workflow itself. And so there are numerous research focusing on different aspects of workflow analysis including the surgery type classification [Kannan 2019], surgery remaining time estimation [Aksamentov 2017], **tool detection** [Bouget 2017], phase recognition [Garrow 2021], **action/activity recognition** [Lalys 2014], clinicians pose estimation [Kadkhodamohammadi 2014], surgeon skill evaluation [Reiley 2011], etc. All these contribute in great amount towards the realization of CAI. For instance, surgical phase, action, and event recognition will mainly help in providing context-awareness. On the other hand, surgical instruments and anatomical structures recognition will play a role in improving safety by guiding against the use of the instrument, like a hook, in an unsafe dissection zone [Speidel 2008] or other safety checkpoints [Mascagni 2021b] such as anatomy validation and grading. The scrutiny from skill evaluation will help to improve surgical performance.

To wrap it up, surgical workflow analysis can be introduced into CAI systems and have a large impact on future surgical innovations, whether for planning, intra-operative or post-operative purposes. It would offer an additional layer of quality control and safety monitoring to surgical procedures. Furthermore, it would provide tools to keep surgeons and OR staff on track with every small detail.

1.2 Research Overview

1.2.1 Problem Statement

As the field of surgery is evolving with emerging technologies improving patient outcomes, the procedures are becoming more and more complex heightening pressure on surgeons who are now faced with complex handling of the sophisticated equipment. This equipment, especially the ones with sensors and imaging capabilities, such as endoscopes, captures a great deal of data from different sources and modalities. However, the general problem is that *"these large unprocessed surgical data are left unused"*. Analyzing these data would provide intelligent feedback, knowledge, communication signals, procedural and patient-specific model [Kranzfelder 2013b, Jiang 2017, Liew 2018] to the surgeons, helping to reduce their cognitive workload and improving their coordination and efficiency pre-operatively and intra-operatively [Lemke 2005]. The current information and communication technology in the OR cannot sufficiently extract useful information from these data or process the procedural data in a way that benefits the operational and clinical tasks without disrupting the surgical workflow.

Developing CAI solutions that can analyze the procedural data intraoperatively is one of the main focuses of surgical workflow analysis, mostly to provide interactive feedback to surgeons about the ongoing activities. Such a system should be able to understand the activities in every given surgical scene. Lots of research model information concerning the surgical intervention and its activities in their own way, such as performance time, instruments used, trajectories, or intervention phases. These exploits are encouragingly utilizing the available data to provide computer-based solutions to assist the intervention. But, the specific problem is that the **existing systems only describe surgical activities at a very coarse level, which are not detailed enough for more helpful AI-assistance in the OR**. Their coarse-grained predictions leave out substantial semantics, such as details about the tissue operated upon. For instance, the main task studied by the community, surgical phase recognition [Ahmadi 2006, Lo 2003], only describes scenes at a very coarse level. As an example the *clipping and cutting* phase [Twinanda 2016b] in cholecystectomy contains a multitude of important actions such as *graspers* holding anatomical landmarks, a *clipper* applying several clips, laparoscopic *scissors* cutting the *cystic-duct* and so on. The coarse phase labels can help navigate surgical videos and even help to provide selective documentation of critical events [Mascagni 2021a], but by itself, the phase information does not provide an accurate picture of the activities taking place. Such unaccounted finer details of the workflow activities are imperative for fostering improved safety in laparoscopic cholecystectomy. And so, it appears like the finer-grained the activity is modeled, the more value it gains in terms of clinical utility. Finer-grained workflow divisions such as step [Ramesh 2021], single verb action [Khatibi 2020, Rupprecht 2016] recognition made limited attempts at breaking the activities into finer units but still overlook interactions with the anatomy. Thus, the problem remains largely unresolved.

1.2.2 Research Questions

Our central research question is: **how can tool-tissue interactions be effectively modeled to infer fine-grained surgical actions from videos for the best clinical utility?** In an attempt to answer this research question, we are faced with disentangling activities into components entities involved in the interaction: the instrument, its role, and its target. It now looks like multiple recognition tasks are involved, but since the whole activities revolve around the instruments, localizing of these instruments becomes imperative as well for the recognition of the other interacting components that rely on the instrument position information.

However, there is a lack of spatially annotated datasets to train a deep learning model for instrument detection. But, since it is easier to generate binary labels indicating the presence or absence of surgical instruments, **how to exploit these easier-to-generate binary presence labels for tool localization and tracking?** becomes a complementary research question. Work done in this thesis is targeted at providing enough scope for investigation, practical answers, and insightful discussion to these research questions.

1.2.3 Purpose Statement

The purpose of this study is **to develop computational artificial intelligent systems for the detection and recognition of surgical instruments and fine-grained activities in laparoscopic videos**. Considering the interconnectedness of these two aspects of surgical workflow analysis: tool and activity, we propose, as our first task, to detect and track the surgical instruments in laparoscopic videos and thereafter, extend this proposal to also include the recognition of the actions performed using the detected instruments as would be introduced further and extensively discussed in the succeeding chapters.

1.2.4 Tasks and Methods

We present a concise overview of the tasks tackled in this thesis.

1.2.4.1 Surgical Tool Detection and Tracking

The surgical instrument recognition is an essential component for all the works done in this thesis. Its role in all our process modeling is crucial as tool information is a discriminative feature that enhances surgical activity recognition. Tool *recognition* entails the detection of the presence of tools and their classification. This recognition is a multi-label classification problem meaning that more than one tool can be present at the same time.

Our first task goes beyond multi-label recognition, to also include localization and tracking. By *localization*, the method predicts coordinate labels representing the spatial locations of the detected tools. And by *tracking*, the method associates and propagates the identities of the detected tools across video frames, including their re-identifications. Tool localization and tracking are respectively spatial and spatiotemporal tasks: they require a deep learning model to be trained on data in which the spatial coordinates of the tools have to be manually annotated. Owing to the lack of spatially annotated data, our tool detection method is designed to learn without requiring any form of spatial annotation, but by leveraging an approach termed *weak supervision*.

Taking everything into account, we build a new deep learning model, with both spatial and spatiotemporal considerations, that is trained end-to-end but weakly supervised on binary presence labels for joint detection and tracking of surgical tools in laparoscopic videos. Our model employs two main deep learning methods: (1) CNN for feature encoding and spatial localization; and (2) Recurrent Neural Network (RNN) for temporal refinement and tracking.

Building a joint detection and tracking pipeline for surgical instruments in the first instance creates a bedrock for the rest of the research work in this thesis.

1.2.4.2 Surgical Action Triplet Recognition

For its significance, especially in providing context-aware assistance in the OR, a copious amount of work models surgical activities recognition at different levels of granularity. Our activity recognition goes beyond the conventional recognition of surgical action as a single verb of the surgical instruments, to a deeper understanding of visual semantics that depicts the complex relationships between instruments and tissues. Specifically, we tackle surgical activity recognition at a more fine-grained and detailed level formalized as *triplet*, for comprehensive information to provide the best clinical utility in CAI systems.

Action triplet is here presented as ⟨*instrument, verb, target*⟩ combinations [Katić 2014]. Their recognition requires a lot of tasks, including simultaneous recognition of the three constituting components of the triplets, which is multi-label per component, and a semantic association of these components as there can be multiple triplets per frame. Characteristically, action triplets are instrument-centric, meaning that consideration of anatomy as part of a triplet is not by mere visibility, rather by its involvement in an interaction carried out using an instrument. Also, a verb is defined by the action of an instrument. This ultimately means that without an instrument, there is no verb, and also no anatomy will be marked as a target. The implication is that methods intending to recognize these two dependent components correctly would require the discriminative instrument information.

Taking everything into account, we build several new deep learning models, tackling both the components detections and their association. We still rely on binary presence labels to provide instruments' location information for guiding the detection of the other two components of the triplet. We explore two deep learning methods in this task: (1) CNN for feature encoding, spatial localization, triplet and triplet components recognition; (2) Attention Mechanism that models spatial and semantic reasoning for improved detection of the triplet components and their association.

The modeled task and methods stand out among other surgical workflow frameworks in providing truly fine-grained and comprehensive information on surgical activities.

1.2.5 Research Motivation

We discuss in this section, the motivation for the work done in this thesis.

1.2.5.1 Motivation for Surgical Tool Detection and Tracking

A key ingredient to developing CAI systems that can provide context-aware decision support for laparoscopic surgery is having a real-time knowledge of the presence of the surgical instruments, their locations, and their track of movement over time. Surgical instruments play a central role in the understanding of other aspects of surgical workflow analysis. For instance, the instruments' presence and their co-occurrence usage are some of the most important discriminating markers of the varying surgical phases and steps [Padoy 2012, Stauder 2014]. Also, the tools entry, exit, and substitution within the body determine the surgical phase transitions. Most surgical events are directly or indirectly tied to the instruments such as the presence of smoke resulting from the coagulating instruments (e.g.: bipolar) [Nwoye 2019].

Tool detection information could be useful in formulating and sending crucial signals in the OR. These signals could be for pre-operative or intra-operative needs. For instance, the detection of prolonged use of certain instruments such as irrigation and suction devices suggesting bleeding could be used to request a senior surgeon's intra-operative assistance. Whereas the detection of some instruments in combination with other foreign bodies such as a specimen bag could suggest the concluding part of a current surgery. This information would be useful in estimating the remaining surgery duration which is important for OR scheduling and pre-operative pain medication on the next patient in the waiting room. Furthermore, surgical instrument detection and tracking are essential in understanding the tool-tissue interaction in surgical videos [Nwoye 2019]. In this case, the instrument is central, meaning that such interactions revolve around the presence of an instrument [Nwoye 2020]. In robot-assisted surgery, the detection and tracking of the manually used instruments would provide information for the synchronization of kinematized and non-kinematizd instruments location [Nwoye 2019]. Since surgical skills are mostly accessed by the instrument usage pattern [Speidel 2006, Jin 2018], detection and tracking of instruments can be helpful in the evaluation of surgical skill and performance. Therefore, an instrument recognition model can be integrated into some learning systems to assist in surgical skill training and education.

Based on the significance, a lot of deep learning strategies have been explored on surgical instrument detection in laparoscopic videos and images including their classification, localization, segmentation, and tracking.

1.2.5.2 Motivation for Fine-Grained Activity Recognition

Having discussed the importance of surgical instrument detection in CAI, it will be necessary to also highlight that detected tool information alone is not sufficient to develop novel assistance systems that are reactive to the context, e.g. that can interpret tool-tissue interaction, provide timely instructions to OR staff, enforce safety checkpoints, or log automatically relevant information within the surgical report. This is because a particular instrument, such as a grasper, can occur in almost all the phases of the surgery but only its usage pattern could deliver distinguishing semantics for context-aware modeling. For instance, the clinical need of notifying the surgeons to observe CVS achievement may be overrun if a system fails to distinguish when an instrument like grasper is retracting gallbladder at the calot triangle dissection phase from when it is packing the specimen bag at the gallbladder retraction phase.

A step further in recognizing the surgical instruments is the understanding of the actions that they are performing at every point in time throughout the entire duration of the surgery. This is known as surgical activity or action recognition. The term *surgical tool-activity* as used in this thesis refers to those surgical actions/activities that are instrument-driven as opposed to other non-operative activities that can even happen even outside the patient's body without the use of surgical instruments. Modeling a tool activity encompasses the used instrument, the action of the instrument, and the targeted underlying anatomy. This is formalized as surgical action triplet in [Katić 2014]. These types of actions are instrument-centric. Their recognition usually starts from the point of instruments insertion into the body to their withdrawal from the body. A recognition model detects these instruments and recognizing their interactions with the tissues at every time interval in the procedure.

Surgical tool-activity recognition is highly essential towards the development of intra-operative and post-operative context-aware decision support systems since they provide additional information about the state of the detected instruments which are more relevant to the context-awareness of the procedure. Surgical tool-activity recognition is also essential in robotic surgery to keep track of the surgical actions controlled by humans. Surgical tool-activity recognition can be helpful in action anticipation. A sequence of predicted surgical actions could be used by a rule-based inference system to identify error-prone situations in complex cases, anticipate failures, and provide useful signals requesting assisted intervention.

At a fine-grained level of granularity, surgical activity recognition can help foster safety intraoperatively. For instance, in laparoscopic cholecystectomy, Critical View of Safety (CVS) achievement is a commonly advocated safety check to prevent bile duct injury (BDI). This medical error can lead to a major complication in surgery. According to [Strasberg 1995], CVS is defined by 3 criteria (1) the view of 2 and only 2 tubular structures, the cystic duct, and the cystic artery, connecting to the gallbladder, (2) that the hepatocystic triangle is cleared from fat and connective tissues, and (3) the lower part of the gallbladder is separated from the liver bed. These criteria can be achieved by careful dissection. Since CVS is assessed visually, this means that the assessment can be automated using computer vision [Mascagni 2020,Mascagni 2021b]. Hence, an activity recognition model which takes into consideration the detailed description of the instrument-tissue interaction would be of great benefit towards automating and giving feedback on CVS assessment. Furthermore, fine-grained activity recognition could potentially help to mitigate visual illusion by differentiating, via feedback, unrelated but easily mistaken tissue manipulations such as clipping cystic-duct vs clipping bile-duct. However, this level of granularity for surgical action detection, which is needed for a more helpful AI in the OR, is lacking in the existing recognition systems.

Surgical action recognition models can also be useful postoperatively via video captioning and report generation for post-surgery evaluation. In surgical education, they can be used for action-specific video indexing and retrieval. It can also help to improve the overall workflow of the hospital as the information can be made available to the administration, or computing overall statistics [Padoy 2008]. In general, surgical tool detection and tool-activity recognition will set a bridge for the development of many medical applications that could be useful pre-operatively, intra-operatively, and post-operatively.

1.3 Challenges

We broadly classify the challenges facing the modeling of surgical tool-activity recognition into two: data- and method-based challenges.

1.3.1 Challenges Related to Data

Over the years, interesting deep learning methods have been developed for visual recognition and language translation, however, these algorithms do not directly generalize on surgical data. Notwithstanding the endoscopic videos capturing most of the activities performed within the patient [Vercauteren 2019], automatic recognition of these activities is much more challenging than the classical human activity recognition for which most of the algorithms are benchmarked. Sometimes, these challenges are introduced by the constraints arising from surgical data acquisition protocol, annotations difficulty, and overlapping labels as discussed further.

1.3.1.1 Visual Challenges in Laparoscopic Images

Recognition models trained on natural images may not easily adapt on surgical data owing to a swift change in the task scene and visual ambiguity [Lalys 2014] affecting the image quality and visibility. In terms of coverage, the endoscopic videos are captured at a very close range which restricts the camera from obtaining sufficient contextual information. This also restricts the field-of-view and localization [Baumhauer 2008]. Since the videos were acquired in a controlled and constrained environment, the obtained images are typically similar to each other, resulting in low intra-class variability.

In the general vision tasks, shape, texture, and color are major discriminating properties learned by the recognition models, nevertheless, these features are not fairly representative in surgical data. The anatomies maintain fairly similar colors and textures, yet their shapes are mostly deformable. Worst still, these anatomies do not maintain clear boundaries from each other. The difficulty of recognizing and differentiating these anatomies affects learning the instruments based on their scene contextual characteristics. The surgical instruments are mostly similar except for the tips which are less than 20% of the whole body. There can

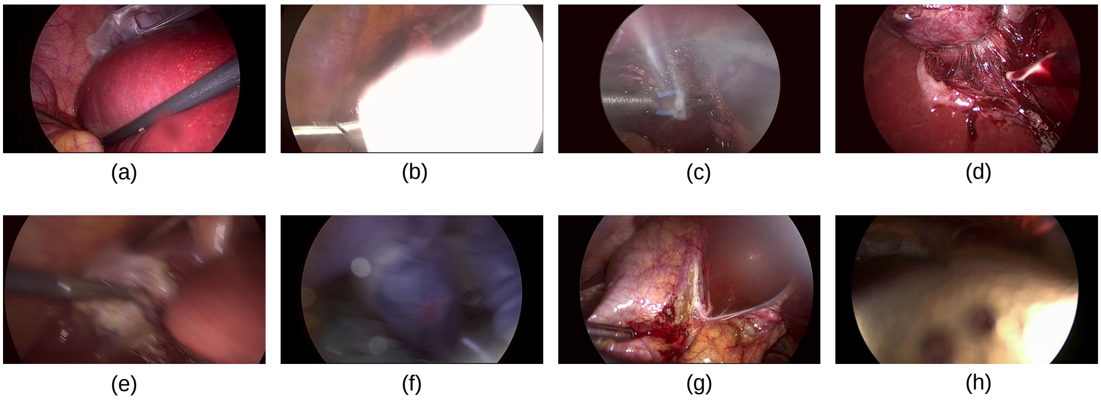


Figure 1.5 – Laparoscopic images illustrating several visual challenges: (a) occlusion, (b) specular reflection, (c) presence of smoke, (d) blood splatter, (e) rapid motion blurring, (f) out of body noise, (g) restricted field-of-view, (h) dirtiness of lens.

also be varying shapes for the instrument's tip resulting from their articulations. There can be occlusion of surgical instruments by the anatomies or other instruments [Speidel 2014]. And this becomes a bigger issue when the occlusion is on discriminative tips of the instruments. Even the camera lens can sometimes be occluded by the anatomies preventing it from capturing a clearer surgical scene. The orientation of the endoscopic camera leads to rapid appearance changes [Reiter 2010] and sometimes, can make the anatomies and instruments appear ambiguous when captured from different angles.

One peculiar visual challenge in surgical images is blood splatting [Haase 2013] on the instruments and the surrounding anatomies which can re-color them, thereby making their recognition more difficult. The camera lens can also be stained by blood and other fluids. At some point in the procedure, we witness the withdrawal and re-insertion of the camera to clean the stained lenses. Sometimes, when the camera is abruptly taken out of the patient's body, whether to clean them or not, it unintentionally captures other objects/persons in the OR which adds more noise to the dataset. More noise can be introduced from other sources such as poor resolution, some temporal blackouts, and lighting changes [Reiter 2010, Reiter 2012b, Sznitman 2012b]. This lighting can occur as specular reflection causing distorted brightness and contrast in the captured images. Another source of noise is the rapid camera motion leading to image blurring [Sznitman 2012a] which reduces the clarity of the instrument's and anatomy's boundaries. The quality of data obtained from the endoscopic camera can be also affected by the presence of smoke [Sznitman 2012b] caused by coagulating instruments.

In endoscopic videos, the instrument motions are backward, based on a fulcrum effect of the trocar insertion site which is antagonistic to the natural motion of the object in the real world scene. This affects the use of methods, such as object tracking, trained on natural vision datasets. There are also data variability across surgical teams, patient specification, and medical data centers [Vercauteren 2019]. These visual challenges, as illustrated in Figure 1.5, make it difficult to design discriminative visual features to represent the data.

1.3.1.2 Lack of Spatially Annotated Dataset

Having listed some of the visual challenges in surgical data preventing direct translation of vision deep learning algorithms in surgery, one would be tempted to ask why these algorithms are not directly benchmarked on the surgical datasets. A bigger data challenge in the field is the unavailability of large annotated datasets. Intelligent complementary aids for a complex procedure such as laparoscopy would require extensive analysis and model training on a large bank of surgical data. Before the advent of laparoscopy, most of the patient data are not digitalized and stored in a structured and standard manner [Maier-Hein 2017]. Even though the endoscopic camera captures a large amount of digital data, the bulk of them is not annotated.

A large chunk of the annotated ones such as Cholec80 [Twinanda 2016b], M2CAI-tool [Twinanda 2016a], etc., provides only binary labels. For the instrument recognition task, multi-label binaries of 0s and 1s are provided for each frame where the present instrument classes are annotated as 1s and the absent ones, labeled 0s. For the phase recognition task, a multi-class binary annotation is provided with the correct phase per frame labeled 1 and the rest marked 0s. This type of annotations is not designed for training AI models for complex tasks such as localization, segmentation, tracking, etc. And, creating spatial annotations such as region boundaries and pixel-wise masks is expensive, tedious, and time-consuming [Jia 2017, Vardazaryan 2018, Nwoye 2019]. This bottleneck has limited the exploitation of deep learning methods on only a very tiny fraction of the dataset that could be annotated spatially [Vardazaryan 2018, Nwoye 2019].

Since generating binary annotations just indicating the presence of the instruments requires less effort, it becomes an interesting research question to exploit these easier to generate binary labels for many complex tasks that would have ordinarily require spatial labels [Nwoye 2019]. Success in this direction would motivate increasing access and usage of large datasets [Nwoye 2019], which would, in turn, set the stage for a new generation of analytics that will support decision making, model benchmarking, and quality improvement in interventional medicine [Maier-Hein 2017].

1.3.1.3 Lack of Standardized Action Class Labels

One of the factors affecting research on surgical action recognition is the lack of standardized class labels for consistent benchmarking of recognition models. The impact of having standardized class labels for activity recognition is manifested in surgical phase recognition which has become one of the most researched workflow analyses in surgical data science. In laparoscopic cholecystectomy, surgical phase recognition is known for its seven common phase labels already established in the literature, namely: preparation, calot triangle dissection, clipping & cutting, gallbladder dissection, gallbladder packaging, cleaning & coagulation, and gallbladder extraction. Following this standardized labeling, several large public datasets have been generated for phase recognition which includes the famous Cholec80 [Twinanda 2016b], m2cai16-workflow [Twinanda 2016a], Endovis workflow challenge, and many other unpublished datasets.

Surgical action recognition could have also been as widely researched as the phase counterpart since it would be interesting to also detect the finer actions, such as interactions between the instrument and tissue, within the phases for a better understanding of the surgical activities. However, there exist no standardized action class labels across procedures even at a coarse-grained level. In laparoscopic surgery, work in [Lo 2003] recognizes four major action events as idle, retraction, cauterization, and suturing. The larger SDS challenges proposed different four verb classes (*cut, grasp, hold, and clip*) at MICCAI EndoVis challenge 2019 [Wagner 2021b]. Another challenge at MIDL 2020, the proposed action labels comprise 21 classes in EASD dataset [Bawa 2021]. A work on surgical image captioning [Xu 2021] generates semantic relationship classes from two different robotic surgery datasets. the first dataset comprises 11 action classes namely: manipulating, grasping, retracting, cutting, cauterizing, looping, suctioning, clipping, ultrasound sensing, stapling, and suturing, whereas the second dataset was annotated with 5 action classes comprising manipulating, grasping, cauterizing, suctioning, and clipping. The inconsistency in the label classes motivates their proposal for cross-domain adaptation across action labels in different surgical procedures.

Furthermore, works in gynecologic laparoscopy [Khatibi 2020, Petscharnig 2018b] recognize 8 action classes of suction and irrigation, cold cutting, blunt dissection, coagulation, suturing, high-frequency thermal cutting, sling, and injection. Another [Kletz 2017] proposed 11 actions classes for the same surgery. The lack of uniformity in the number and labels for the action classes across procedures hinders method comparison and incremental improvement of existing works for surgical action recognition. The inconsistency makes it even more difficult to combine several small datasets from several data centers since deep learning models are known to perform better when trained on a large dataset. A uniform, consensus, and standardized recommendations for annotating of surgical video data would enable assessment of algorithms and multi-institutional collaboration [Meireles 2021].

1.3.1.4 Lack of Fine-Grained Dataset for Detailed Workflow Analysis

Out of all existing frameworks for surgical workflow analysis in endoscopic videos, action triplet recognition stands out as the only one aiming to provide truly fine-grained and comprehensive information on surgical activities [Nwoye 2021]. However, there is a lack of public triplet datasets which can be attributed to the difficulty in generating a dataset of such detailed nature as rightfully pointed out in [Twinanda 2017]. The difficulty in generating this type of annotations can be connected to the particular need for precision in medicine. Most anatomical structures can not be easily differentiated without their texture information. Also, there are unclear boundaries between most anatomies making it difficult to precisely differentiate some actions when formulated as triplets. Additionally, the lack of triplet datasets can also be linked to the expert knowledge required for their labeling especially the anatomies, and sometimes, understanding the verb of the instruments on the anatomies is not straightforward. These difficulties have affected the generation of the dataset and in turn, hinders the design and training of recognition models for action triplet recognition.

Although generating a dataset of such magnitude is non-trivial, it is needed at this stage to drive the research forward, as shown by datasets such as Cholec80 [2], CATARACTS[3] and EndoVis [4], which have had tremendous impacts in the community.

Success stories in fine-grained activity recognition will motivate research in the field [Maier-Hein 2020], create a building block for onward development, model benchmarking, and offer direction for onward improvement.

1.3.2 Challenges Related to Methods

Vision-based approaches are very attractive in modeling surgical workflow since they do not require the redesign of the surgical instruments and/or OR. Also, they are equally achieving state-of-the-art (SOTA) performance in surgical workflow analysis. However, the difficulty of using deep learning for the detection and recognition tasks in surgery, for all its utility, is not to be overlooked. In this section, we discuss some of these challenges for a better perspective on the proposed tasks.

1.3.2.1 Huge Training Data Requirement

While human beings can learn abstract relationships in a few trials on a single or small data sample, deep learning algorithms need to be trained on large sets of labeled data over several iterations. In most cases, their performance scales with an increase in training examples [Sun 2017]. If the data is limited, deep learning tends to overfit the training samples [Horenko 2020]: this is because when a deep learning algorithm fits the variables, it also fits the noise specific to the given data. Several examples of similar cases are needed for the model to correctly concentrate on the deterministic features in the data. For instance, without training on large endoscopic data, a deep learning model may also include the shape of the camera to predict an instrument type.

Obtaining these data and their annotations is generally hard as previously discussed. It is also not straightforward to ascertain the size of a dataset needed for the effective training of deep learning algorithms. This may vary according to various factors. Firstly, the number of categories in a learning task. The more the categories, the more the overlap between their discriminating properties, and the more the training data needed by a deep learning model to accurately discriminate these categories.

Performance need is another factor affecting the data requirements for model training. While a small size dataset may be enough for a proof-of-concept study, a large dataset is needed for training a deep learning model for production, and an even larger dataset is needed when the model is intended to generalize across data sources or centers. Furthermore, class imbalance is another factor affecting the data requirements. In real-world examples, dataset categories usually differ in size. Deep learning models tend to have more false positives for the most frequently occurring classes and more false negatives for the less frequent classes.

There is no perfect way to deal with lack of data or missing data, but many efforts have been made to diminish the effects which include: data augmentation [Ding 2020], unsupervised

[2] http://camma.u-strasbg.fr/datasets

[3] https://cataracts.grand-challenge.org/

[4] http://endovissub-instrument.grand-challenge.org/

learning [Twinanda 2018], semi-supervised learning [Yu 2018, Shi 2021, Bodenstedt 2017], transfer learning [Neimark 2021, Dergachyova 2018], self-supervised learning [Funke 2018, da Costa Rocha 2019], synthetic dataset [Pfeiffer 2019a, Ding 2020], and of recent meta-learning [Dawoud 2020]. While all these approaches bring their own advantages to deep learning, they have not particularly removed the huge data requirements for training a deep learning model.

1.3.2.2 Computational Cost and Memory Requirement

Deep learning comes with a voracious appetite for computing power. While it has been shown that deep learning models perform better with deeper than shallower layers, unfortunately increasing the layers also means increasing the training parameters. This overparameterization of deep learning models which is intended to improve performance, however, increases the cost of training a deep learning model which scales with the product of the number of parameters and data points. Hence, many models require very high GPU computational power or even expensive TPU.

Apart from the computational demand, large memory is needed to fit large training data. Some networks are trained with smaller batch sizes to reduce the memory overhead. However, some tasks, such as activity recognition, tracking, etc., are better designed to capture longer temporal information even across a full video length. In this setup, training a CNN + Long Short Term Memory (LSTM) model in an end-to-end manner is usually impracticable as most of the time, the LSTMs most are intended to capture the temporal information of the full video. Unlike in general computer vision, where such tasks are modeled using very short video clips, laparoscopic videos are usually very long (avg. 1 hour in Cholec80 [Twinanda 2016b]). Hence, most LSTM-based models are not trained end-to-end since they relied on CNN extracted spatial features which would be stacked over a given temporal length. Instead, most of the algorithms are tailored for offline processing with pre-recorded videos [Ye 2016]. However, deep learning models would learn better representations as both the CNN and the LSTM components can benefits from each other when trained end-to-end [Yengera 2018], as some studies have shown using shorter videos [Ma 2016]. Also, since the same model training strategy is usually maintained during inference, it is hard to use an offline trained model for online inference as would be needed in the OR.

Another factor affecting the memory requirement for deep learning training on surgical data is image resolution. Due to the requirements for precision in medicine, images are usually preserved at high resolution. Compressing these images would lead to the loss of tiny structures in the image which might be contextually informative for feature extraction. Keeping surgical images at such high resolution, such as 1080x1920x3 as in Cholec80, leads to a huge memory bottleneck and affects model training especially for deeper layer models.

1.3.2.3 Time Constraint on Hyper-parameter Tuning

One of the most tedious efforts in the development and training of deep learning models is hyperparameter tuning. These hyperparameters such as learning rate, weight decay constant, batch size, etc., are usually not learned by the network, instead, they are determined and fixed by a human. Selecting the best hyperparameters is usually a bottleneck as this would require series of trial and error over several options, and even without a guarantee of the best choice. Most times, these hyperparameters are by extensive grid search leading to training of hundreds of models.

Apart from hyperparameter tuning, model selection can be done through another time-consuming process known as cross-validation. In this case, a dataset is split into k-folds and the model is trained k-times on different k-1 folds for model selection. Cross-validation which usually indicates the mean ± std of the model performance is also used for model selection as well as hyperparameter tuning.

In the course of training different models for cross-validation, hyperparameter tuning, etc., several model weights are stored which is also memory-consuming.

1.3.2.4 Multiple Instance Bottleneck

Deep learning is traditionally designed to learn and approximate a function that directly maps input to output. In most cases, the output is designed as a vector of log probabilities which can be thresholded at 50% for binary classification. In the case of multiple classes, an arg-softmax is used to determine the model prediction. A more difficult case is the multi-label classification where zero, one, or more class labels can be predicted. Even in the multi-label situation, there can still be multiple instances of the same label. Deep learning models, in most cases, are not designed to handle this kind of situation. This is largely due to the dataset not being annotated to account for the number of occurrences for the class labels. One backdrop of this effect is that it makes it impossible for deep learning models that are weakly supervised for localization on binary data labels to be able to infer the number of instances for each localized object class.

Also, there could be class overlap, especially in fine-grained action labels making a direct input-output mapping insufficient to correctly differentiate multiple instance cases as can be seen in action triplet recognition.

1.4 Terminology in Surgical Tool-Activity Recognition

Several terms related to surgical workflow analysis are not well-defined. Since they will be used throughout this thesis, it becomes imperative to explicitly clarify their definitions.

The terms *instrument* and *tool* are used interchangeably to mean the devices for carrying out desired effects during surgery which usually involves the manipulation of the anatomies. While *tool* may be loosely defined to include computer systems and AI solutions used during a surgical procedure, it is, however, in this thesis, limited to the hardware devices that have direct contacts with the tissue, and performs a specific action on them. They perform such functions as cutting, dissecting, grasping, holding, retracting, etc. The term *tool* is more widely used in the community, however, without changing the meaning, we prefer *instrument* when discussing triplet to have a better acronym ⟨*i,v,t*⟩ for ⟨*instrument, verb, target*⟩ rather than ⟨*t,v,t*⟩ for ⟨*tool, verb, target*⟩ which would introduce ambiguities in the text. Still on tools, the term *detection* means the localization and classification/recognition of surgical instruments. Sometimes, we simply use the term *localization* in this thesis to also mean *detection.* Also,

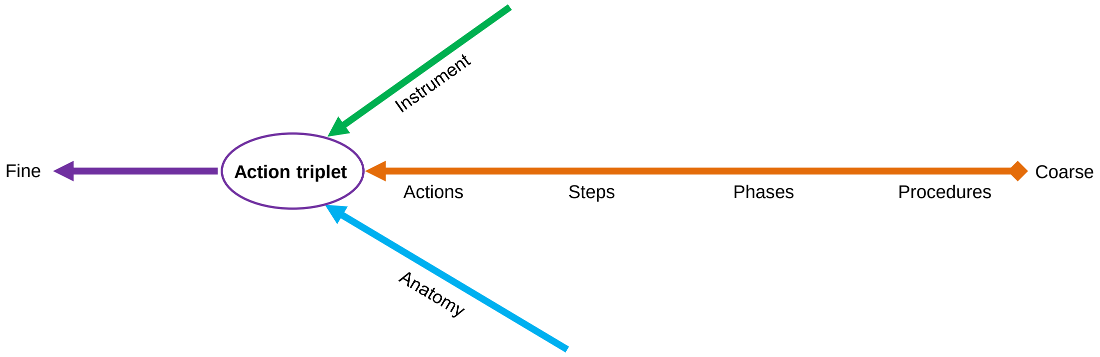


Figure 1.6 – Axis showing the granularity of activity in the OR, from the coarsest level (right) to the finest level (left). The granularity axis is modified from [Lalys 2014] to show the intersection with instrument and anatomy recognition to form a triplet.

detection can be used to simply mean the binary presence detection.

The definition of the term *activity* is subjective and overlapping, depending on the context and level of abstraction at which they are described. We adopt the notion of activity granularity presented in [Lalys 2014] also represented by the axis in Figure 1.6. The axis represents activity from the coarsest to the finest level. Following the axis (right to left), a *procedure* describes the full central activity of a full surgery such as cholecystectomy, cataracts, etc. Within the procedure, there are *phases*, which are the meaningful sequence of tasks carried out to achieve a procedure. A phase usually describes a series of actions on anatomy (e.g.: gallbladder retraction) or a group of anatomies (e.g.: calot triangle dissection) intended for a unified purpose. When a phase is further split into smaller units while still retaining the "action on an anatomy" description (e.g.: cutting cystic-artery, pushing needle, etc.), it is called a *step*. Going more finer, the term *action* ignores the anatomy and identifies the verb, such as cutting, pushing, etc. as the fundamental element in the semantic interpretation of a surgical scene. We modify the axis in Figure 1.6 to rightfully highlight the type of activity tackled in this thesis which is at the intersection of the fine-grained action with the instrument and the anatomy which are being left out as the activity becomes finer. This intersection is referred to as *action triplet* or simply *triplet*. Hence, surgical action triplet is not only *fine-grained* but also a *detailed* and *comprehensive* modeling of surgical activities.

Surgical videos capture workflow activities that can be recognized at different levels of granularity depending on the focus of the research. If the recognition of the activities at any time step utilizes all observations made from the beginning of the procedure up to that time step, it is referred to as *online*. This type of recognition is usable intra-operatively or in real-time. Whereas it is called *offline* if the recognition utilizes all observations in the entire procedure including the ones ahead of the given prediction time step. This type of recognition can only be used post-operatively. All our proposed methods in this thesis are designed for real-time benefits.

1.5 Summary and Thesis Overview

We have presented in this chapter the general overview of the thesis. We equally discussed the medical background including the clinical motivation, challenges, and highlighted as well the research gaps which the work is designed to fill. To conclude this chapter, we summarize our contributions and thesis outline in this section.

1.5.1 Contributions

The fundamental aim of this thesis is to address the problem of surgical activity recognition by developing deep learning methods that can detect and track surgical instruments and also recognize their fine-grained detailed activities in laparoscopic videos. The contributions of this thesis mainly revolve around two main points: (1) the recognition, detection, and tracking of surgical instruments using deep learning methods that are weakly supervised on binary presence labels; (2) the recognition of surgical actions at a fine-grained level described in the form of a triplet ⟨*instrument, verb, target*⟩. The contributions are discussed in detail as follows:

The first contribution is the study and development of novel deep learning models that can exploit weakly annotated data for the detection, localization, and tracking of surgical instruments. Existing works [Sznitman 2012a, Rieke 2016, Sznitman 2014] on surgical instrument recognition rely on full supervision: a situation whereby the detection and tracking models are trained on data in which the spatial positions of the instruments are manually annotated. However, creating spatial annotations such as region boundaries and pixel-wise masks is expensive, tedious, and time-consuming [Jia 2017]. This bottleneck has limited the exploitation of deep learning methods on only a very tiny fraction of the dataset that could be annotated spatially [Vardazaryan 2018]. Since generating binary annotations just indicating the presence of the instruments requires less effort, we propose a new deep learning object detection and tracking method that circumvents the lack of spatially annotated surgical data with weak supervision on binary presence labels [Nwoye 2019]. While existing work [Vardazaryan 2018] localizes a point on the instruments using weak supervision, it is not trivial to model their temporal consistency or track surgical instruments across frames without requiring spatial annotations. Hence, we propose an RNN that could leverage the temporal information in video data in a manner that still allows for weakly supervised learning, resulting in an elegant end-to-end tracking method that models the spatio-temporal motion of the surgical instruments. In the first instance, we propose a deep learning model that can model localization features in its inner convolution layer without requiring spatial annotations. We show that the activation at this inner layer, also known as heatmaps, can sufficiently describe the position of the surgical instruments when trained on binary labels. Then, we learn smooth trajectories of the instruments by modeling the temporal consistency of the localization heatmaps. This we achieve by employing a ConvLSTM layer, which is known for its spatiotemporal capability to infuse temporal smoothing while retaining a 3D spatial dimension of the input features. The ConvLSTM leverage the temporal information inherent in video data to model the motion tracking of the detected instruments [Nwoye 2019] without spatial training labels. Combining the convolution's spatial localization and ConvLSTM temporal modeling, we built a weakly-supervised tracking model with SOTA performance on the three tasks of presence detection, localization, and tracking in the Cholec80 dataset [Twinanda 2016b].

For the second contribution [Nwoye 2020], we build upon the foremost research to recognize the activities of the surgical instruments. Previous research has mostly focused on phase recognition [Blum 2010, Dergachyova 2016, Twinanda 2015], gesture recognition from robotic data [DiPietro 2016, Malpani 2016] and event recognition [Loukas 2015]. The coarse nature of these tasks' output leaves out substantial semantics for helpful AI assistance. Hence, we propose a more detailed analysis of recognizing fine-grained activities representing the instrument-tissue interactions in endoscopic videos. We model these activities as *surgical action triplets* of ⟨*instrument, verb, target*⟩ and develop deep learning models to recognize these triplets. As an ablation experiment, we build a naive baseline for simple classification of the triplet IDs which unfortunately proves to be sub-optimal. Then, we extended the approach to a multi-task learning method to capture the interacting components of the triplets, namely: instrument, verb, and target, and learn their association using a Multiple Layer Perceptron (MLP). We observed that we can better inform the verb and target detection by leveraging the appearance cue of the instrument, of course, as the triplet is instrument-centric. To this end, we have the proposed a model which utilizes the instrument class activation to guide the other components' detection. Since MLP could not conserve the semantic structure of the triplet association in its dense connectivity, we also modeled a learnable higher dimensional space for the tripartite association of the triplet components. The proposed approach was evaluated on a new dataset, *CholecT40*, which has been generated in collaboration with our clinical partners from 40 videos of the *Cholec80* dataset and annotated with 128 action triplet classes. Albeit action triplets information is used in [Katić 2014, Katić 2015] to improve surgical phase recognition, this is the first work to recognize action triplets directly from surgical videos.

Even though the proposed triplet recognition model outperformed the baseline models, there are potential areas of improvement in the recognition pipeline, one of which concerns the low performance recorded for the verbs and targets. Thus, as a third contribution, we introduced a new form of spatial attention mechanism [Nwoye 2021] to capture the individual action triplet components in a scene. This technique focuses on the recognition of the verbs and targets using the activations resulting from the instruments. In performance, the proposed attention method outperforms the previous proof of concept model in triplet component detection.

Motivated by the performance improvement of the attention-guided modeling, as the fourth contribution, we extended the model to capture even longer range attention for the triplet component association. In this, we proposed a transformer-inspired model [Nwoye 2021] that semantically models the association of the detected components of the triplets. In practice, it leverages both self and cross attentions with interacting components of the triplets to learn their association. Different from self-attention in Natural Language Processing (NLP), we propose to also utilize several cross attentions to benefit from the learned representative features of components. The transformer-inspired model sets a new SOTA performance in the triplet recognition task.

The last contribution is the generation of large datasets for tool and activity recognition.

For the tool detection task, we generated spatial labels on 5 videos for the evaluation of weakly supervised models on tool localization and tracking. For the tool-activity recognition task, we generated CholecT40 [Nwoye 2020] for surgical action triplet recognition, thanks to our clinical collaborators at Strasbourg (IRCAD and IHU) in the CONDOR project. To standardize the data and label, we extended the dataset to CholecT50 with additional 10 videos and standardized classes. CholecT50 is now a mixture of annotations from different surgeons that capture more variability of surgical expertise, and the label mediation thereafter. To further encourage research in this domain, the CholecT50 is used to organize an EndoVis sub-challenge under the name *Surgical Action Triplet Recognition 2021 (CholecTriplet2021)*[5] held jointly with MICCAI 2021 in Strasbourg, France. This challenge will help to navigate the activity recognition in surgical workflow analysis to a new level and to match the pace of similar research, such as HOI in the Computer Vision community. We plan to release the CholecT50 (train set) as the largest fine-grained dataset for surgical action triplet recognition to date containing videos of cholecystectomy recordings annotated with 100 action triplet classes.

1.5.2 Outline

This thesis is organized into three parts as follows:

- The first part introduces the clinical context and motivation in chapter 1. Chapter 2 presents a review of related works available in the literature and a comparative analysis of their task formulations, methodologies, levels of supervision, and significance.

- The second part contains the main body of the thesis. It spans from chapters 3 - 6 representing different tasks and methods that are assembled to achieve the thesis objectives. Chapter 3 presents a weakly supervised method for surgical instrument detection, localization, and tracking in laparoscopic videos. The methods presented in this chapter have been published in [Nwoye 2019]. Chapter 4 presents a method for fine-grained action recognition as well as an action triplet dataset. The method presented in this work has been published in [Nwoye 2020]. Chapter 5 presents a method based on a spatial attention mechanism for improving the triplet components detection. This chapter additionally includes the improved dataset for action triplet recognition. Some of the results presented in this chapter have been submitted for publication [Nwoye 2021]. Chapter 6 presents a transformer-inspired architecture for enhanced action triplet recognition, specifically improving the triplet association. Some of the results presented in this chapter have been submitted for peer-review [Nwoye 2021].

- Finally, the third part of this thesis concludes the work by first discussing the existing and potential clinical applications of the proposed methods in chapter 7. Afterward, a summary of the thesis is presented in chapter 8, along with discussions about the possible future directions to improve the methods.

[5] https://cholectriplet2021.grand-challenge.org/

2 A Review of Related Work

Learn as much as you can from those who know more than you do,
who do better than you, who see more clearly than you.
– Dwight Eisenhower

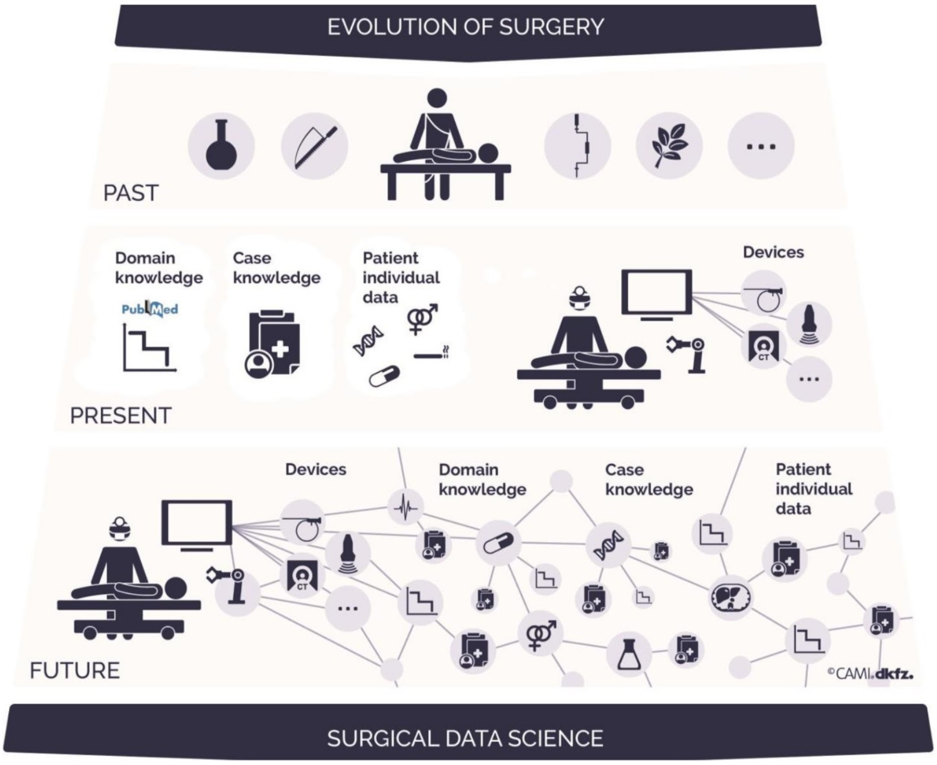


Figure 2.1 – Surgical Data Science (SDS) in the evolution of surgery [Maier-Hein 2017]

Chapter Summary

Motivated by the need for the real-time information about surgical instruments and their activities in CAI systems, lots of methods have been investigated to detect instruments on varieties of surgical data including robot kinematics [Reiter 2012a], electromagnetic signals [Lahanas 2016, Fried 1997], ultrasound [Hu 2009], fluoroscopy [Weese 1997] and images in laparoscopic videos. Among all these modalities, the image-based approaches have become increasingly attractive because they do not require a modification of the instrument design nor the OR [Lalys 2014].

In this thesis, we focus on using laparoscopic video recordings to perform surgical tool tracking and activity recognition. Thus, our review of related works in this chapter will focus mostly on works that employ vision-based approaches. These works are found by keyword search on Google search engine, Google scholars, ResearchGate, PubMed, ArXiv.org, Semantic scholars, Refseek, Microsoft academic search, Scopus, Web of Science, etc., also by connected papers, cited or reference papers, and referencing papers. We will discuss in the following sections the different levels of complexity at which the tasks have been exploited highlighting their investigated methodologies, benefits, limitations, and inter-dependencies across tasks and methods. Where necessary, we will take a tour of the task in the wider computer vision community for a broader overview and emphasis. We then finalize the review by explaining how our work is related to the existing literature.

2.1 Tool Detection

In SDS, detection of surgical instruments has been tackle at different levels of complexity. While some detections are only concerned with identifying the instruments in surgical images, others model the properties of the identified instruments such as their location, pose, shape, motion, etc. In the literature, some work independently tackle one aspect of the detection, whereas others jointly modeled inter-dependent tasks. Each aspect of tool detection provides some information that can be useful in the development of CAI and other medical applications. In this section, we review the related works on surgical instrument detection as follows.

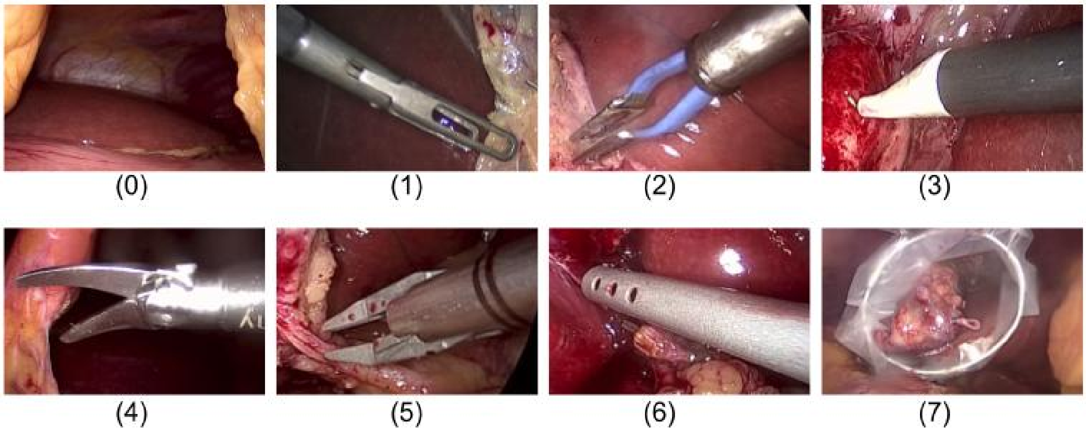


Figure 2.2 – List of the seven surgical instruments used in the Cholec80 dataset [Twinanda 2016b]. Labels: (0) no instrument, (1) grasper, (2) bipolar, (3) hook, (4) scissors, (5) clipper, (6) irrigator, (7) specimen bag. No instrument label (0) is not a distinct label in the dataset. It is added here to demonstrate when all instruments are absent.

2.1.1 Presence Detection

Surgical tool presence detection is one of the key problems in automatic surgical video content analysis. It involves the detection of surgical instruments by providing binary information denoting which instruments are used at each time in surgery. This goes beyond the image-level classification [Krizhevsky 2012] in computer vision task as zero, one, or several types of instruments used in laparoscopic surgery can be detected in one image at the same time: one image can't be classified by a single instrument class. Hence, instrument presence detection is cast as a multi-label classification problem. The instrument presence labels are determined by the visual information from the laparoscopic videos. They are annotated solely by their visibility per frame and do not require localization information. Solving the instrument presence detection can benefit many applications such as the evaluation of surgical instrument usage, video database indexing based on the tools used in each video, and automatic surgical report generation. Also, the presence detection information can be combined with other signals to detect a potential upcoming complication such as the detection of instruments that should not appear in certain surgical phases.

With the success of deep learning in image classification tasks, earlier work [Twinanda 2016b] proposed Endonet, which is a CNN architecture with a multitask branch for phase and instrument recognition. The Endonet model, shown in Figure 2.3, predicts the binary presence probabilities of seven laparoscopic instruments namely grasper, bipolar, hook, scissors, clipper, irrigator, and specimen-bag as shown in Figure 2.2. The work also introduced the widely used Cholec80 [Twinanda 2016b] dataset which consists of 80 videos of cholecystectomy recording annotated with phase and instrument labels. At that early stage, an endoscopic vision challenge is launched and code-named M2CAI 2016 workshop [1] to establish this research in the community. The challenge introduced the m2cai-tool dataset,

[1] http://www.camma.u-strasbg.fr/m2cai2016/index.php/tool-presence-detection-challenge-results

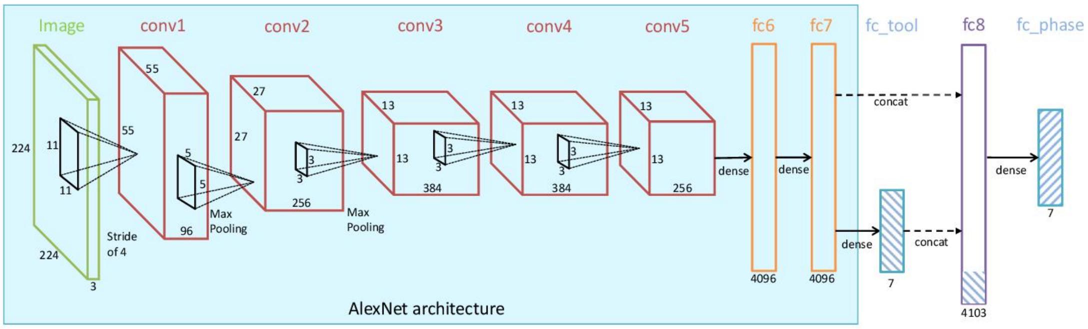


Figure 2.3 – Architecture of EndoNet proposed in [Twinanda 2016b] as a multi-task deep learning framework for the recognition of surgical tools and phases in laparoscopic videos.

among others, comprising 15 videos of cholecystectomy procedures from University Hospital of Strasbourg (Strasbourg, France) and annotated with binary presence labels of the same 7 surgical instruments as in Cholec80. In the challenge, [Luo 2016] utilized multiple CNN to model the recognition of each instrument class independently. The unsatisfactory performance suggests that the intrinsic association among the instruments is important. Others [Twinanda 2016a, Sahu 2016, Raju 2016, Zia 2016] explored well known deep learning models from the computer vision community using transfer learning from ImageNet dataset [Deng 2009], and finetuned them on the m2cai-tool challenge dataset. For Example, ToolNet [Twinanda 2016a] and [Sahu 2016] finetuned the popular AlexNet architecture [Krizhevsky 2012] while [Raju 2016, Wang 2017] finetuned GoogLeNet [Szegedy 2015] and VGG-16 [Simonyan 2014]. Also, [Zia 2016] finetuned AlexNet, VGG-16 and Inception-v3 [Szegedy 2016], all for instrument presence detection on m2cai-tool dataset. Beyond the challenge, the performance of the deep learning models on the instrument presence detection tasks has been remarkably improved. This comes from advanced features modeling including model ensemble [Wang 2017, Al Hajj 2018], class label balancing [Sahu 2017, Mondal 2019, Alshirbaji 2018], and multi-tasking with complementary phase recognition task [Twinanda 2016b, Mondal 2019, Jin 2020]. In an effort to hasten model training, residual CNN (ResNet) [He 2016] has also been used including the densely connected CNN (DenseNet) [Iandola 2014] in [LIN 2019] for instrument presence detection.

To consider the long-term relationships in the sequential video frames, [Sahu 2016] proposed to combined ImageNet pretrained and finetuned features that capture both phase and tool co-occurrence. The combined features are used to create contextual features for tool detection coupled with a label set sampling technique to reduce bias. In [Roychowdhury 2017], long-term relationship between images is exploited using a Markov Random Field (MRF) modeling. The drawback is that online video analysis is not possible with their proposed approach which requires approximately 20K temporal sequence. The idea of temporal modeling has since advanced to graphical modeling of continuous video frames [Wang 2019a] where a Graph Convolutional Networks (GCN) is used to analyze the video as a whole and find correlations useful for the instrument presence detection. These days, notwithstanding, exploring temporal information for instrument detection is mostly done using RNN. This is because RNN keeps a temporal memory to remember past information. On the task of presence detection, a simpler version of RNN know as Gated Recurrent Unit (GRU) [Cho 2014] has been utilized to extract spatiotemporal features in [Namazi 2019]. Its single-cell state makes it less memory consumption and faster than the popular LSTM [Hochreiter 1997]. However, LSTM is more accurate on datasets with longer sequences such as Laparoscopic videos. The LSTM model [Jin 2020, Mishra 2017, Al Hajj 2018] and its bidirectional counterparts [Mondal 2019] has been exploited on surgical videos for modeling the temporal dependencies, as well as smoothing the predictions for surgical tool detectors. Most times, these LSTM-based models are not trained end-to-end due to memory bottleneck as discussed earlier in Section 1.3.2.2.

Another way of exploring temporal information is by attention mechanism which allows a deep learning model to highlight only the important features in an input feature or across a sequence of inputs while suppressing the less relevant features. Of recent, a long-range attention modeling using an attention-guided network [Hu 2017] and a transformer-based method [Kondo 2020] has been explored for detecting the presence of surgical instruments in laparoscopic videos.

2.1.2 Spatial Localization

Instrument localization is the task of locating an instance of a particular instrument in an image. The location information is usually in form of coordinate points, or pixel masks indicating the spatial positions of instruments in real-world surgical video frames. In a demonstration, these coordinates can be plotted over the images in the form of tightly cropped bounding boxes, outline, overlay to precisely identify the detected instrument instance among several possible others. Hence, the data annotation involves manually specifying the region boundaries such as bounding box coordinates, center pixels, contours/outlines, etc., of the surgical instruments in the video frames. With the availability of spatial coordinate labels, instrument localization is mostly tackled as a regression problem where a learning network is trained to regress from either region proposals or fixed anchor boxes to nearby bounding boxes of pre-defined target instrument instances. In this case, the localization is cast as a distance optimization function such as L2 distance, Huber loss [Huber 1992], etc. Other methods which do not utilize spatial labels for their training mostly extract the box coordinates around some saliency or activation maps [Vardazaryan 2018].

Though sometimes used interchangeably, instrument detection is not the same as instrument localization. Surgical instrument detection is a more complex problem that combines the concepts of instrument localization and their classification. The classification in most cases is treated as an instrument presence detection task except that each positive presence label is attached to one localized instance of the instrument. In addition to the benefits of the presence detection mentioned in Section 2.1.1, the spatial position information from a surgical instrument detection model can be leveraged to understand the anatomy that the instrument is manipulating. The knowledge of tool location can be leveraged to build motorized camera systems that are adaptive to the surgeon's vision center of attention. It can also be useful in managing instruments that are off the screen thereby increasing patient's safety. Since the localization also contains information on the size of the detected instruments, some virtual measurement capabilities can be built around such information to obtain the accurate measurement of the sizes of various anatomical structures. In augmented reality, visual overlay on the tip of some instruments can benefit from their localization information.

In [Kranzfelder 2013a], radio frequency identification (RFID) tags is used to detect and categorize surgical instruments. At that period, the growing popularity of the vision-based approach was not unnoticed with work in [Ryu 2013] utilizing image processing techniques like K-means clustering and Kalman filtering to localize instruments in surgical videos. In those days, the traditional machine learning methods of feature engineering are widely explored. Its success usually depends on the wellness of the crafted feature representations which are mostly obtained from the image properties such as shape [Doignon 2005], color [Bodenstedt 2018b, Sznitman 2014, Allan 2012, Haase 2013], texture [Allan 2012, Reiter 2012a], gradients [Bouget 2015, Haase 2013], depth [Speidel 2008, Haase 2013], etc. The feature representation approach is not robust due to the diversity of surgical specializations, varying designs of surgical instruments, and visual ambiguity [Lalys 2014] affecting the image quality and visibility as discussed in Section 1.3.1.1. A combination of the image engineered features provides a potentially more discriminative feature space [Bouget 2017]. However, feature engineering is effort- and time-consuming.

There came deep neural networks to the rescue. As opposed to feature engineering, deep learning allows a CNN model to learn the most suitable features for the detection task without manual feature manipulations. In this regard, the region-based CNN has been widely used. Both the one-stage [Choi 2017] and two-stage Region-Based Convolutional Neural Network (RCNN) [Choi 2017, Jin 2018, Zhang 2020a] are been explored for tool localization. The simplest way of localizing surgical instruments is by point localization which entails the locating of a point coordinate that corresponds to a part of the detected instrument. This point could be the center coordinates or sometimes any coordinates which fall within the region boundaries of the instrument tips [Vardazaryan 2018]. Beyond the point coordinates, localizing the surgical instruments by their whole region boundaries using bounding boxes is the most common approach in the literature [Jin 2018, Choi 2017, Sarikaya 2017]. In [Jin 2018], a region-based CNN is employed to detect and localize surgical instruments in laparoscopic videos. Being the foremost work, [Jin 2018] extended the m2cai-tool dataset to m2cai-tool-localization by annotating a sample of 2532 frames with 3141 instances of bounding boxes. Beyond the instrument localization, the movements of the detected tools are also analyzed to automatically assess surgeon performance in the considered procedure. The dense anchoring scheme of this two-stage region proposal network is not cost- and time-efficient thereby affecting its real-time translation. Sequel to this, [Zhang 2020a] proposed to improve the inference speed of the two-stage region-based CNN using a modulated anchoring network. They also introduced another spatially annotated dataset AJU-Set of 3164 frames capturing 3952 instance boxes. Another way to improve the inference speed is by the use of lightweight models or single-shot detectors as done in [Choi 2017]. Besides the laparoscopic video, surgical instrument detection and localization can be seen in robotic and robot-assisted surgery [Sarikaya 2017, Choi 2017], and in eye surgeries [Al Hajj 2017]. In [Al Hajj 2017], optical flow information is used to analyze sequential features of consecutive images to exploit spatial

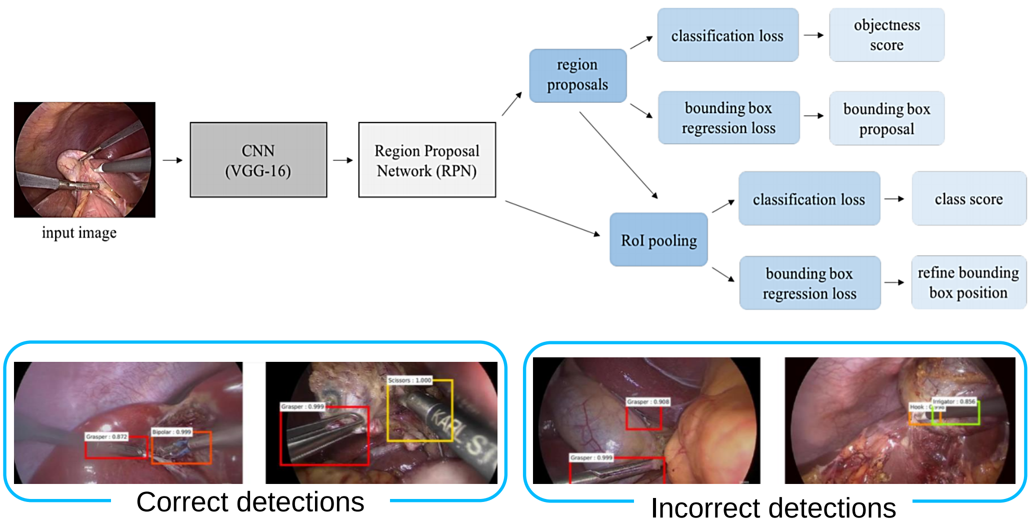


Figure 2.4 – The pipeline of the Faster R-CNN architecture for surgical tool detection proposed in [Jin 2018] including some qualitative results on the success and failure cases of the detection model.

redundancies between them in cataract surgery videos. However, optical flow information is not easy to generate and they introduce additional visual artifacts. Furthermore, optical flow algorithm performance is impaired when the spatial locations of a point change abruptly or when the spatial distance between objects in moving frames is inconsistent. Other works such as [Sarikaya 2017, Choi 2017] performed surgical tool localization in specific tasks of robot-assisted surgery videos. In [Sarikaya 2017], an architecture using multimodal CNN for fast detection and localization of tools in RAS videos is presented. The method applies a region proposal network (RPN) and a multimodal two-stream CNN for object detection to jointly predict objectness and localization on a fusion of image and temporal motion cues. In this vein, [da Costa Rocha 2019, Sestini 2021] presented a self-supervised approach that uses the kinematic model of the robot to generate the instrument segmentation masks for the training of a fully convolutional neural network for pixel-wise classification. Most of the methods explored on robot-assisted surgeries are specifically for surgical training tasks. Using robotic arms is limited in practice due to the relatively high cost. And there may be differences between specific training tasks and complete surgery.

The performance of frame-based detectors can be improved more if the deep learning models can learn more context from video data. Since an image frame is a part of a sequence, a prediction model can improve its confidence if it has access to temporal information around the frame. Some surgical contexts may not be correctly recognized by features extracted from a single image as they may be affected by some visual challenges such as blurring, noise, smoke, etc. Variations between consecutive frames may help to better detect and localize surgical instruments in some situations such as occlusion, deformation, and noise. Modeling spatio-temporal localization is non-trivial as most temporal modeling units such as LSTM, GRU, etc., do not preserve the spatial details of image features. A way of overcoming this was proposed in [Chen 2018] which used a 3D CNN to capture both the temporal and spatial features at

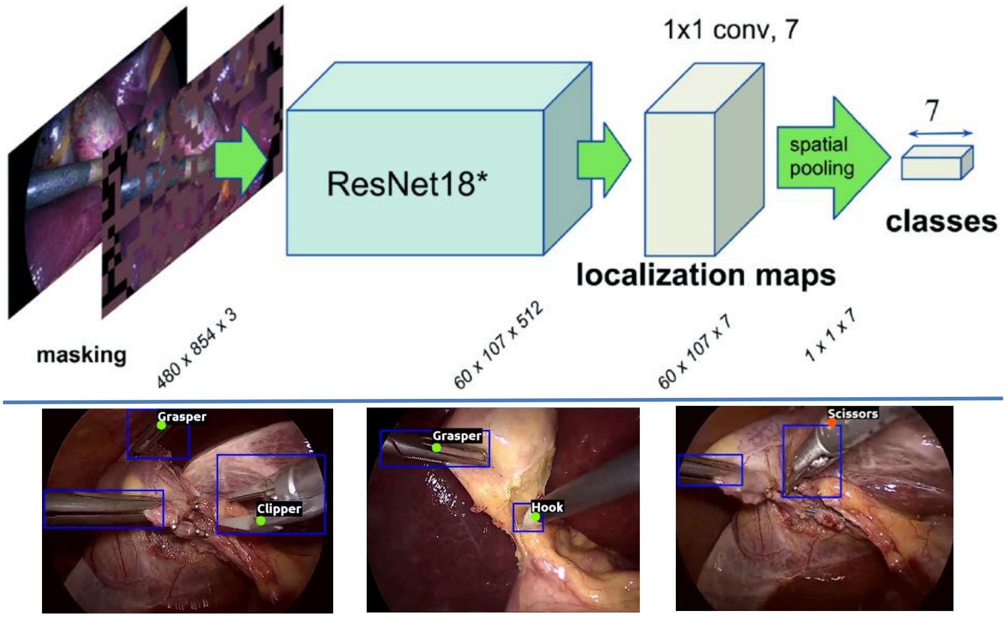


Figure 2.5 – The architecture of weakly-supervised model for instrument localization proposed in [Vardazaryan 2018] showing also some qualitative results of the model. Green dot represents the correct center point localization, red dot is incorrect, and bounding box is groundtruth label.

the same time. Another method in [Wang 2019a], used a GCN for temporal relationship modeling and a conventional CNN to preserve the spatial features. The GCN is used to analyze continuous video frames to find correlations. However, this approach has only be explored on a tool presence detection task. Nevertheless, the two approaches in [Chen 2018, Wang 2019a] only modeled a very short temporal sequence of few image frames instead of a full video. Another way of exploring temporal information is by attention mechanism. A long-range attention modeling using a transformer-based method has been explored in [Kondo 2020] for detecting the presence of surgical instruments in laparoscopic videos. However, the flattened nature of their temporal attention modeling does not preserve spatial features and hence makes the attention model formulated in this manner unsuitable for tool localization.

Despite the progress in surgical instrument detection, the research is limited by the lack of spatially annotated datasets as already discussed in Section 1.3.1.2. Generating such region boundaries such as bounding boxes, or outlines is indeed time-consuming, tedious, and expensive. As part of their contributions, work in [Jin 2018, Zhang 2020a] had to generate a tiny fraction of spatial labels from a large dataset which is further split into training and tiny validation sets for their experiments. However, it is common knowledge that deep learning models are data-hungry as discussed in Section 1.3.2.1. Also, the tiny dataset is insignificant for model evaluation especially on such delicate tasks as surgical procedures. An interesting bypass to the issue of limited spatially annotated datasets is to weakly supervise a detection model on the easier-to-generate dataset.

Weak supervision is a learning technique whereby deep learning models are trained on data with imperfect or weaker labels. For instance, learning to localize objects using the image-level labels, learning to segment objects using bounding boxes coordinate labels, etc. This is mostly due to the annotation efforts needed to generate the more complex labels. Weakly supervised models while learning a weaker function based on the available labels such as binary presence detection or recognition is expected to also capture some features usable for the recognition of a higher complex task such as localization.

In laparoscopic videos, weak supervision has been employed in [Vardazaryan 2018] to circumvent the lack of spatially annotated data using binary presence labels. In their work, a global pooling operation is applied on the output of an Fully Convolutional Network (FCN) to force the network activations on the most salient features needed to localize surgical instruments in laparoscopic images. However, their localization is limited to only a point on the instruments. A more recent work [Fuentes-Hurtado 2019][2]proposed to use easy labels, provided as stripes over different objects in an image, in combination with a partial cross-entropy loss function to obtain dense pixel-level prediction maps for scene segmentation in laparoscopic videos. Other interesting applications of weak supervision in medical imaging are seen in the segmentation of cancerous regions in histopathological images [Jia 2017] and in the detection of the Region of Interest (ROI) in chest X-rays and mammograms [Hwang 2016]. A closely related approach such as semi-supervised learning combines a chunk of labeled data with large unlabeled data for model training as done in [Yoon 2020] targeting surgical tool detection in gastrectomy videos. Semi-supervised learning has also been explored on surgical gesture recognition [van Amsterdam 2019] where a limited amount of demonstration labels are used to generate an appropriate initialization for a Gaussian Mixture Models (GMM) based algorithm.

Another way to detect instruments is by segmentation. This goes beyond the boundary localization, which classifies the patches of an image containing an instrument, to the classification of every pixel in an image. Instrument segmentation provides the exact outline of the instruments by grouping every image pixel belonging to the same instrument and assigning them their corresponding category label, while the rest of the non-instrument pixels are assigned a background label which is usually 0. The pixel classified label is known as a segmentation mask. Generating the groundtruth segmentation mask is very expensive and time-consuming. Depending on the interest, the instrument segmentation can focus on producing a binary mask, where every pixel belonging to an instrument is assigned a foreground label (usually 1) and the rest of the image pixels are classified as background with a label value of 0. By using precise segmentation, the pose of the surgical instrument of interest can be efficiently estimated in laparoscopic surgery. In most cases, the instruments are segmented according to their classes where all pixels belonging to the same class of instrument are assigned a unique label. This type of segmentation is called semantic segmentation. The benefit is that it would facilitate a better understanding of tool-tissue interaction since the different instruments are designed for a specific type of manipulation on the anatomies. Aside the semantic segmentation, another widely used type of segmentation in surgery is instance segmentation. In this case, the image pixel of every distinct instrument is assigned a distinct label. These instance labels can be very useful in tracking to maintain distinct instrument identities over

[2] Method published after our proposed method in this thesis.

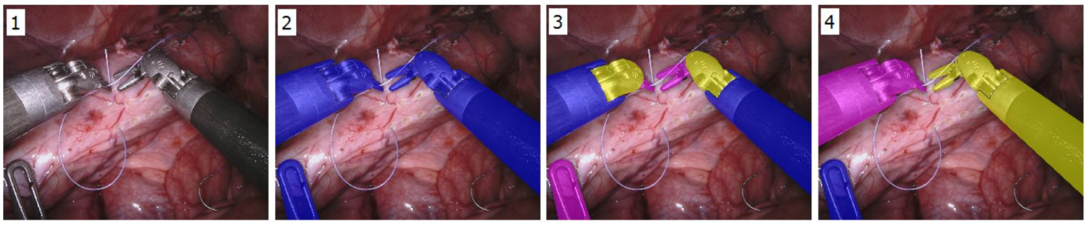


Figure 2.6 – A snapshot from a robotic surgical video from MICCAI 2017 Endoscopic Vision Robotic Instrument Segmentation SubChallenge [Allan 2019]: (1) original video frame; (2) binary segmentation; (3) part segmentation; (4) semantic segmentation. Images contained in [Shvets 2018].

time. Surgical instrument segmentation can go as far as parts segmentation, where different parts of one instrument such as the tips, wrist, shaft, etc., are segmented differently. overlaying part of the segmentation masks, such as the tip, on the anatomies can provide surgeons with valuable information that can improve decision making during complex procedures as it could be offered a better analysis of the relations between the patient anatomy and operating instrument. In general, instrument segmentation provides useful information for surgical report generation. It is also important in augmented reality visualization where precise pixel-based segmentation of the tools is necessary for handling occlusions and providing the user with the correct perception [Vercauteren 2019].

Earlier work [Speidel 2009] propose an automatic method for detection of instruments from endoscopic images by segmenting the tip of the instrument and then recognizing them based on 3D instrument models. Also, [Zisimopoulos 2017] utilize a commercially available surgical simulation to train tool detection and segmentation based on deep convolutional neural networks and generative adversarial networks. Their model which is trained on a simulated dataset was tested on a real cataract dataset. A three-term MICAI EndoVis challenges in 2015, 2017 [Allan 2019] and 2018 [Allan 2020] cascadingly built a dataset for robotic instrument segmentation. The challenge participation was characterized by the exploration of CNN [Pakhomov 2019], FCN and variants of U-Nets [Shvets 2018] architectures for surgical instrument segmentation. Later in 2019, a challenge focusing on the robustness of the segmentation introduced another dataset (ROBUST-MIS) [Ross 2020] for estimating the generalization ability of the algorithms across interventions and institutions. Due to the annotation efforts in generating segmentation masks, recent works are now focusing on weak [Lee 2019] and self [da Costa Rocha 2019, Sestini 2021] supervisions. [da Costa Rocha 2019] propose a self-supervised approach that uses the kinematic model of the robot to generate the instrument segmentation masks for the training of a fully convolutional neural network for pixel-wise classification.

2.2 Motion Tracking

The aforementioned tool detection approaches do not exploit the temporal coherence of a video sequence and do not perform tracking. Surgical instrument tracking goes beyond the classification and localization of the instrument position or region boundaries to tracing a consistent movement of the detected instrument over time. To have a better overview of instrument tracking as done in this thesis, we would start this section by reviewing object tracking as it is tackled in the computer vision community. This will be followed by instrument tracking in laparoscopic videos.

2.2.1 Object Tracking in Computer Vision

Object tracking involves the prediction of trajectories of targets in a video sequence. It is one of the most important tasks in computer vision that underpins many practical applications such as traffic monitoring, detection of traffic accidents, smart video analysis, medical diagnosis systems, human activity recognition, robotics, autonomous vehicle tracking, and so on. In object tracking, the target specifies in the first frame and must be detected and tracked in the next frame of the video [Soleimanitaleb 2019]. While this seems straightforward, tracking can be affected by illumination variation, background clutters, low resolution, scale variation, occlusion, deformation of the target shape and position, fast and infrequent motion, and so on.

The earlier methods leverage feature-based approaches which is one of the simplest ways of object tracking. It usually involves the extraction and computation of unique features of an object in an image and finding the object in the next frame by exploiting similarity criteria. Many machine learning approaches has exploited many hand-crafted features for object tracking such as color [Li 2004, Fotouhi 2011], texture [Zhao 2009, Wagenaar 2010], optical flow [Hariyono 2014, Kim 2016], etc.

Tracking has been widely tackled as an estimation problem in which an object is represented by a state vector describing its dynamic and constantly updating behaviors such as position, velocity, etc. Popular in this regards are the Kalman [Najafzadeh 2015, Nordsjo 2004] and particle [Pérez 2002, Jiang 2003] filters. Most deep learning trackers even used these filters on their detected objects to track their identity and motion in the next frame. Deep learning methods consider tracking as a decision-making process. It is usually jointly modeled with object detection where a deep learning model first detects objects in an image frame and attempt to associate them linearly with the objects detected in the previous frames to maintain a consistent trajectory of the objects over time. This is known as *tracking-by-detection*. One of the most widely used algorithms for the data association part is the Munkres algorithm [Munkres 1957] modified from Hungarian [Kuhn 1955] algorithm for transportation and assignment problems. Apart from direct use of the Hungarian algorithm for object tracking [Mills-Tettey 2007], more robust methods that re-adjust this algorithm to consider the bounding box parameters of the detection results, as well as the information about the tracked objects to associate the detections in a new frame with previously tracked objects, has been developed. Such new methods include the Simple Online and Realtime Tracking (SORT) algorithm [Bewley 2016] which uses the Hungarian method with an association metric that measures bounding box overlap. Also, the deep learning counterpart, DeepSORT [Wojke 2017], integrates the appearance information to better distinguish the objects and handle the linear assignment problem even under occlusion.

Some methods formulate the association using a graph [Brasó 2020, Weng 2020, Li 2020a] by modeling the detected objects as nodes and casting their edge linking possibilities as a cost function learnable by minimizing some fixed [Jiang 2007, Zhang 2008] or learned [Leal-Taixé 2014] cost variables. More complex optimization of this cost is seen in methods such as multi-cuts [Tang 2017, Keuper 2016, Tang 2016] and minimum cliques [Zamir 2012].

Since the data association is based on dynamic features, some methods leverages temporal information in video data to model tracking using RNN models such as LSTM [Poormehdi Ghaemmaghami 2017, Milan 2017, Milan 2017], GRU [Ma 2018, Li 2020c], ConvLSTM [Liu 2020, Liu 2019, Liu 2018], etc.

Since tracking involves both object identification and path association, it becomes a problem how to correctly judge the performance of a tracker. The mean AP metrics used in the object detection evaluation failed to capture the association capability of a tracker. This led to the introduction of the widely used CLEAR metrics [Bernardin 2008] that to allow for objective comparison of tracker characteristics, focusing on their precision in estimating object locations, their accuracy in recognizing the object configurations, and their ability to consistently label objects over time. CLEAR metrics consist of the Multiple Object Tracking Precision (MOTP) which measures the tracker's ability to estimate precise object positions only, and Multiple Object Tracking Accuracy (MOTA) which accounts for all object configuration errors, false alarms, misses, and mismatches, made by the tracker across all frames. Aside the two CLEAR metrics, the rate of objects identity switch, track fragmentation, average track length, Mostly Lost (mL), Mostly Tracked (mT), and Self Quality Evaluation (SQE) [Huang 2020] are other metrics being used in evaluating the performance of a tracking model.

Object tracking has been approached at different levels. Some methods [Bertinetto 2016, Dong 2019, Dong 2016, Henriques 2014, Liu 2016a] track only one object across video frames. This type of tracking is known as Single Object Tracking (SOT). The task is to detect and maintain the identity of one object over time. SOT has long been advanced to consider Multiple Object Tracking (MOT) [Bae 2014, Tang 2016, Bergmann 2019, Keuper 2016, Milan 2017]. In most cases, it simply entails tracking multiple instances of the same object such as multiple persons, or multiple cars, etc. Advancingly, some methods now consider tracking multiple instances of different objects [Lee 2016]. This is known as Multi-Class Multi-Object Tracking (MCMOT). In MCMOT, a tracker will maintain separate trajectories for every instance of all objects in a video: separate trajectories for all moving persons, animals, cars, etc. Quite uncommon, Single Class Multiple Object Tracking (SCMOT) would describe a situation where there is only one instance of different object classes in a scene. Since there are multiple objects to track, this situation is also described as MOT. Examples of this scenario could be found in medical applications where there could be only one instance of different surgical instruments for a particular procedure. For instance, only one instance of the hook, clipper, scissors, bipolar, irrigator, specimen-bag, etc. is possible in laparoscopic cholecystectomy.

The computer vision community has developed centralized benchmarks object tracking such as single-object short-term tracking [Kristan 2015], PETS dataset [Ellis 2010] for surveillance, KITTI benchmark [Geiger 2012] for autonomous driving, Pedestrian [Dollár 2009], and MOTchallenge [Leal-Taixé 2015] for multiple object tracking. The MOT challenge which introduces multiple tracking of pedestrians and vehicles increases the difficulty of the challenge over the years such as MOT15 [Leal-Taixé 2015], MOT16 [Milan 2016], MOT17, MOT20 [Den-

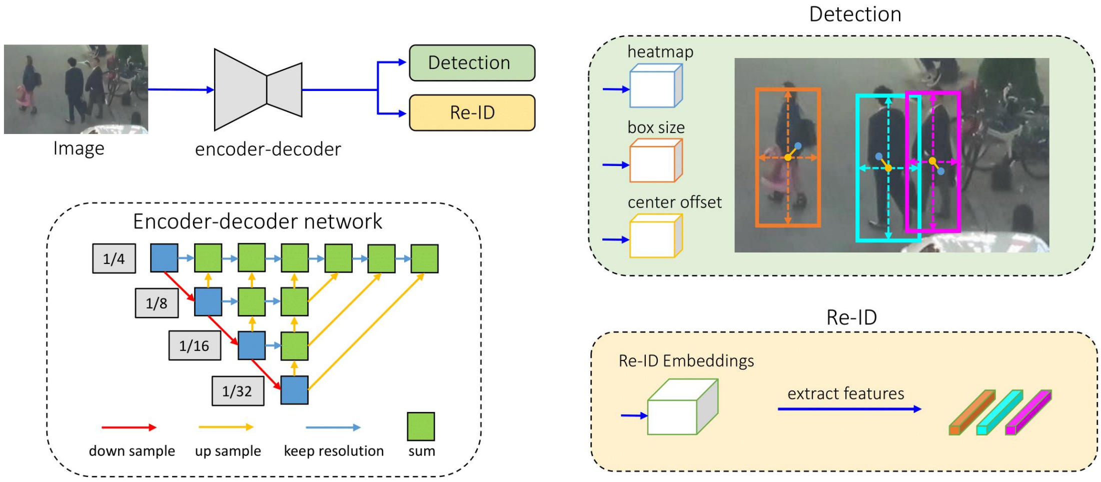


Figure 2.7 – Architecture of one-shot FairMOT tracking model proposed in [Zhang 2020b] for joint object detection and re-identification (re-ID) in a single network.

dorfer 2020], etc. This benchmark has encouraged the exploration and development of more deep learning models for MOT tracking.

Tracking-by-detection, which is the dominant strategy, breaks tracking into two steps of detection and association. Many works [Wojke 2017, Yu 2016, Yu 2016, Mahmoudi 2019, Fang 2018] develop Separate Detection and Embedding (SDE) deep learning models for the task. The embedding is for object re-identification. Some models handle tracking in an end-to-end manner. This means the model is trained to conduct both feature extraction and candidate evaluation in one pipeline. One of the commonest methods in this regard is the Siamese tracker [Tao 2016] which measures the similarity between two captured images to determine whether the same objects exist in both images. Others such as [Sun 2019] use deep learning to learn the data association by jointly modeling object appearance and their affinities between frames in one single pipeline. Also, [Zhu 2018] employ the integration of both the detection and data association networks to handle noisy detections and frequent interactions between targets. They leverage a dual attention network to selectively learn temporal and spatial features while helping the network suppress noisy observations.

Recently, a method [Wang 2020] proposed Joint Detection and Embedding (JDE) by integrating the detection and embedding models into a single network to facilitate the development of real-time systems for object tracking which is essential in practice. Such an approach in a single-short detector will encourage high-speed tracking. Unfortunately, the accuracy of these one-shot approaches [Kokkinos 2017, Voigtlaender 2019, Wang 2020] drop remarkably with the number of identity switches increasing by a large margin compared to two-step methods. An alternative study [Zhang 2020b] examines the causes of failures in the JDE-based one-shot trackers and pinned the failure to uncertainty caused by anchors in differentiating objects' identities. They proposed FairMOT [Zhang 2020b] which is an anchor-free single-shot deep network that relies on center features instead of the anchors for object identity embedding while maintaining a multiple layer feature aggregation.

2.2.2 Surgical Instrument Tracking in Laparoscopic Videos

Instrument tracking involves detecting an inserted surgical instrument and tracing its path inside a patient's body as it moves from one anatomy to the other until its extraction from the body. Motion tracking is often used in surgical skill assessment [Speidel 2006, Jin 2018] because the instrument's path length and movement range between varying surgical skill levels. Instrument tracking can be extended to motion estimation [Quellec 2014b] which can be useful in action anticipating especially when an instrument is approaching a no-zone area in the patient's body [Speidel 2008]. As demonstrated in other endoscopic videos such as cataract videos [Quellec 2014a, Quellec 2014b], the instrument's motion features can help to improve activity recognition particularly in the understanding of the tool-tissue interaction in surgical videos. Surgical instrument tracking is barely researched in laparoscopic procedures particularly for the vision-based approach. This is most likely due to the lack of spatially annotated dataset [Bodenstedt 2018a]. Most of the spatial annotated datasets do not encompass the entire video but only cover selected sample images due to the tedious efforts required to annotate long laparoscopic videos and hence, they are not suitable for modeling trajectories. Another factor limiting the tracking research is the visual challenges prevalent in laparoscopic videos as discussed in Section 1.3.1.1: notably the fast instrument motion, multiple simultaneous instruments, anatomy-instrument occlusions, and re-initialization due to out-of-view conditions [Robu 2020].

Earlier work on instrument tracking are based on the machine learning approach of feature engineering. This includes hand engineered features such as shape [Sznitman 2012b], color [Reiter 2010, Zhou 2014], texture [Kumar 2013, Rieke 2015, Reiter 2012b], depth [Speidel 2008], motion [Kumar 2013, Speidel 2014] etc. [Wolf 2011] combines color-segmentation with prior geometrical instrument models for a more robust feature engineering [Bouget 2017]. Other robust features such as SIFT have been used to build a 2D tracker based on a Generalized Hough Transform [Du 2016]. The performance of methods based on Radio Frequency Identification (RFID) [Kranzfelder 2013a] and Electromagnetic (EM) tracking [Liu 2016b, Sastry 2017] are limited by the manual re-initialization and magnetic interference respectively.

Aside from the feature engineering, surgical instrument tracking has been approached using template matching [Sznitman 2012a], such that [Reiter 2012a] proposed to generate virtual templates using robot kinematics. However, configuring a large number of templates for different instruments, their rendering, and pattern matching is not time and cost-efficient. Some of the earlier successful tracking approaches [Ye 2016, Du 2018, Colleoni 2019, Ryu 2013, Du 2016] were implemented on robotic and robot-assisted surgery, which provides robot kinematics as additional information. For instance, [Ye 2016] combined kinematic data with online part-based templates generated by tool CAD models for 3D tracking of articulated tools. However, the need for additional information from the CAD models might hinder clinical translation. These CAD-reliant models also failed to handle instrument occlusion and are unable to recover temporal trajectories. Another work by [García-Peraza-Herrera 2016] employed optical flow to propagate FCN segmentation in real-time. While their tracking-by-segmentation approach provides a definite region boundary coverage of the tracked instrument, it is greatly affected by the instrument's fast motion and deformation. Alternatively, a method that focuses on pose estimation of the surgical instrument to account for a more flexible representation of the detected instruments is proposed in [Du 2018]. This method allows for multiple instrument tracking but with an insufficient real-time capability. A major stumbling block to the clinical translation of these robotic-based methods is that they are not usable in non-kinematized surgery: suggesting the suitability of a vision-based approach.

Feasibility studies on vision-based approaches [Bouget 2017, Bodenstedt 2018a, Du 2019] to surgical instrument tracking recommended the use of modern deep learning approaches for instrument segmentation and tracking [Bodenstedt 2018a]. The analysis emphasized the lack of spatially annotated dataset as one of the major challenges affecting instrument tacking and segmentation research [Bodenstedt 2018a, Du 2019]. Though these studies are targeted at single-object tracking, which does not easily translate to the real surgical scenario where surgery is performed by simultaneous use of multiple instruments, they however proffer a comparative analysis on methods that can be extended to track multiple instruments. In the meantime, the vision-based approach to surgical instrument tracking are concentrated more in retinal microsurgery [Richa 2011, Sznitman 2011, Sznitman 2012a, Li 2014, Bouget 2017] and cataract surgery [Liu 2002, Baldas 2010, Banerjee 2019, Morita 2020] than in laparoscopic surgery, howbeit, these eye surgeries do not exhibit as fast instrument motion as in laparoscopic surgery. Recently, a new method[3]has been developed in [Robu 2020] to track multiple instances of surgical instruments in endoscopic videos. Their localization is not centered on the tool-tips but encompasses the whole tool's body. An alternative work in [Zhao 2017] considered the surgical instrument as two parts: end-effector and shaft. Albeit their model could detect the instrument's shaft, it, however, focuses its tracking only on the more meaningful end-effector. The works in both [Zhao 2017] and [Robu 2020] are fully supervised on bounding box labels.

2.3 Activity Recognition

In this section, we provide a review of works on surgical activity recognition at both coarse- and fine-grained levels of granularity. Since our interest in this thesis is more on the finer activities which are not very much explored in the field, we first review the SOTA methods for activity recognition in the computer vision community. Afterward, we present vision-based approaches that have been proposed to address surgical activity recognition.

2.3.1 Activity Recognition in Computer Vision

Activity recognition is one of the most active research fields in the computer vision community. It is concerned with the recognition of actions of one or more objects from a series of observations and environmental conditions. The prominent activity being studied in the literature is Human Activity Recognition (HAR). In this case, human is the central subject interacting with other objects and its environment. The detection of human-object interaction (HOI) is an important and relatively new class of visual relationship detection tasks, essential for deeper scene understanding. It is used for a wide range of applications, such as video surveil-

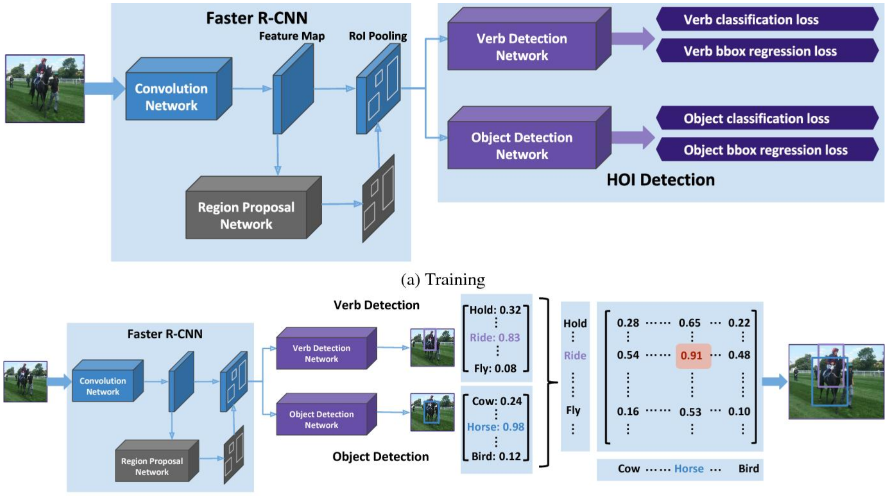
[3] Method published after our proposed method in this thesis.

lance [Lin 2008, Kushwaha 2012], video retrieval [Bendersky 2014], and human-computer interaction [Rautaray 2010]. Just like in medical computer vision, action recognition was once at a single verb recognition stage when [Gkioxari 2014] extended Pascal VOC [Everingham 2010] with action labels, and the task involves identifying the activity of an object in a given image. Beyond the images, there are video datasets such as UFC101 [Soomro 2012], Kinetics [Kay 2017], Thumos [Idrees 2017], etc, with a focus on detecting actions in short video clips that describe the central message of the videos. Nonetheless, their formulation does not represent real-life scenarios as they are more or less like summarizing all activities in a video clip as one. The goal would be to expand action recognition to every image frame that may exhibit varying characteristics within a video dataset. There has also been some works that tackle action recognition as generating captions for images [Blank 2005,Laptev 2008,Fang 2015,Karpathy 2015].

For more understanding of the human activities in images and videos, an activity is formulated as a triplet of ⟨subject, verb, object⟩. In the beginning, this action triplet recognition is mainly centered on still images [Delaitre 2011, Hu 2013, Chao 2015, Mallya 2016] where actions are defined at the intersection of the detected subject and objects within an image frame. Research at this level of granularity began to deepen with the introduction of a benchmark dataset, *"Human Interacting with Common Objects"* (HICO) [Chao 2015], for the recognition of HOI. The dataset demonstrates the key HOI features such as a diverse set of interactions with common object categories, a list of well-defined, sense-based HOI categories, and exhaustive labeling of co-occurring interactions with an object category in each image. It is important to note that in this dataset, human is the only subject of interaction. Ever since then, lots of deep learning methods have been explored for the recognition of HOI in images. In [Mallya 2016], CNN-based appearance features are extracted from human and object detections to obtain state-of-the-art results on recognition.

To provide more training data, a large-scale Human Activity Knowledge Engine (HAKE), that bridges the relationship between instance activity and body part states, is introduced in [Li 2019]. This knowledge-based dataset is continually being enlarged and enriched to make it more powerful towards promoting activity understanding.

Since the triplet recognition tasks do not localize the regions of the action, [Chao 2018] introduces HICO-DET which extends the HICO dataset with spatial annotations. They also introduce a region-based network that characterizes the spatial relations between the bounding boxes of the detected human and objects to identify their interactions. Ever since then, many deep learning works have exploited these spatial annotations for triplet detection. [Mallya 2016] modeled human and object detections using CNN-based appearance features, while [Chao 2018] extended the approach with spatial relationship modeling using a multi-stream architecture. Following a simple observation that actions are accompanied by strong contextual cues, a method in [Gkioxari 2015] developed an action recognition system that uses RCNN to localize multiple regions of actions and classify them. Analysis in [Gkioxari 2018] suggests that HOI is human-centric, which means that every action is human-driven. To prove this, a method that leverages human appearance cues to predict an action-specific density over the location of the correct target objects is proposed. The ap-



(a) Training

(b) Testing

Figure 2.8 – Architecture of a multi-task learning framework for detection of HOI proposed in [Shen 2018]. The verb-object pairing for zero-shot prediction in this architecture inspired our trainable 3D interaction space for data association.

pearance cues are generated by FasterRCNN fully supervised on the human bounding boxes. Another work in [Qi 2018] argued that the detection networks generally lack the structural knowledge of relationship in HOI, and thus, proposed to incorporate this using a differentiable Graph Parsing Neural Network (GPNN). The GPNN provides a generic HOI representation that is applicable in both spatial and spatial-temporal domains. From another perspective, [Gupta 2015] argued that complete understanding of action in a scene is by being able to associate every object in the scene with a different semantic role describing their actions. To this effect, they extend the famous MS-COCO dataset [Lin 2014] with visual semantic role labels in a 10K image dataset known as VCOCO. Other interesting datasets for HOI detection include HCVRD [Zhuang 2018], Bongard-HOI [Jiang 2021], Ambiguous-HOI [Li 2020b], etc.

Despite the enormous progress in providing both methods and data for the HOI detection, [Shen 2018] argued that the space for possible human-object interaction is inexhaustible and so, it is impractical to obtain labeled training data for all interactions of interest. Hence, they propose to scale HOI recognition to a long tail of categories through a zero-shot learning approach. Their proposed model disentangles reasoning on verbs and objects during training and entangles them at test-time to produce detections for unseen verb-object pairs.

The overall performance on HOI recognition has been low with 31.3% on HICO-DET, 47.1% on HICO, and 58.8% on V-COCO leaderboards. It easily comes to the mind to leverage temporal information to better the performance. Work in [Do 2017] annotated HOI taking into consideration their temporal dynamic, and also build an LSTM sequential model, with Conditional Random Field (CRF) for the refinement of the outputs to improve performance. More research [Do 2017, Qi 2018, Almushyti 2019] also explored temporal information in video data to better analyze the interactions between objects captured in image frames.

Alternatively, [Almushyti 2019] applied attention mechanism to the LSTMs forcing them to focus only on the essential parts of human and object for temporal HOI understanding. It appears like the whole idea is now shifting to attention modeling for its capability to suppress non-crucial context information when modeling subtle interactions between elements. Since the advent of the attention mechanism [Bahdanau 2014], lots of deep learning tasks are learnt by exploiting attention in all shades, from self [Vaswani 2017] to cross [Mohla 2020], and from spatial [Fu 2019] to temporal [Sankaran 2016] attentions. A turn of events in the HOI research saw massive exploitation of attention mechanisms for action triplet recognition and detection. Work in [Ulutan 2020] concluded that attention modeling, better than feature concatenation, is the appropriate way of enforcing spatial configurations. [Gao 2018] proposed an instance-centric attention module that learns to dynamically highlight regions in an image conditioned on the appearance of each instance. Such attention modeling allows the network to selectively aggregate features relevant for recognizing HOIs. In [Wang 2019b], an attention model that modulates the global features to highlight only the image regions with relevant context information to detect HOIs is proposed. [Kolesnikov 2019] decomposed their model to first capture box attention and augmenting the second module with that for a more directed detection. Attention has become so popular that even graphical models [Zhou 2019] are also be infused with attention to guide the parsing of human-object body parts relationships for correct HOI prediction. When it seems that attention is helping, another work [Zou 2021] directly modeled HOI instances using a Transformer without decoupling the task into separated stages of object detection and interaction classification. Meanwhile, another work in [Kim 2021] with superior performance on the V-COCO dataset showed that even with an end-to-end self-attention Transformer, it is still better to learn the interacting components of the triplets before associating them.

2.3.2 Vision-Based Surgical Activity Recognition

Just like in computer vision, surgical activity recognition was equally explored at a very coarse-grained level of granularity. This includes the recognition of the surgical procedure being performed which is akin to recognizing the central activity in a short video clip of UFC101 [Soomro 2012] or Kinetics [Kay 2017] dataset in the vision community. Surgical procedure recognition will help in the organization and indexing of surgical databases. In the literature, a kernel SVM model [Twinanda 2014] has been used to identify the type of surgery being performed. To advance its usefulness for the OR scheduling, [Kannan 2019] proposed an early recognition of laparoscopic surgery type from a video within the first few seconds. This is achieved using a future-state predicting LSTM model that learns to approximate its current state to the state of another LSTM ahead in time. The early surgery classification will be useful in real-time fully automated context-aware assistance and automatic acquisition of information in the OR without a workflow or proprietary system interruption.

Beyond the procedure recognition, workflow analysis also entails the automatic recognition of surgical phases in a given procedure. The phases are the different semantic sequences of a surgical procedure such as *calot triangle dissection*, *gallbladder packaging*, etc. The automatic recognition of these phases plays an important role in surgical process modeling (SPM) [Lalys 2014] and can be introduced in computer-assisted intervention (CAI) systems

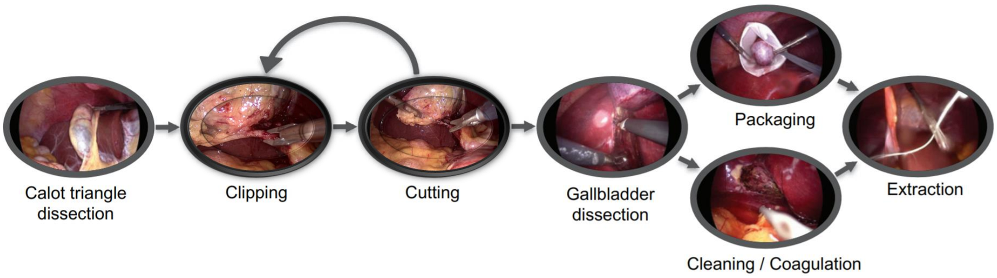


Figure 2.9 – Presentation of the surgical phases of the cholecystectomy procedures from the Cholec120 dataset according to [Padoy 2019].

to improve situation awareness [Maier-Hein 2017] in the OR. Surgical phases recognition has been approached in the past using different imaging modalities such as endoscopic videos [Lo 2003, Ahmadi 2006, Blum 2010, Dergachyova 2016, Twinanda 2016b, Funke 2018, Zisimopoulos 2018, Yu 2018], or from ceiling mounted cameras [Twinanda 2015, Chakraborty 2013]. The task of phase recognition is one of the most researched areas in surgical data science. For this reason, methods ranging from feature engineering [Padoy 2007, Blum 2008] to using features learned by convolution networks [Cadene 2016, Twinanda 2016b, Lea 2016a] have been exploited. Owing ti the temporal nature of the task, RNN-based methods [Al Hajj 2018] and their variants [Yu 2018] including the recent temporal convolution network [Czempiel 2020], attention mechanism and transformer-based models [Gao 2021, Czempiel 2021] has all been explored to improve the precision of surgical phase prediction models. Recently, some research [Sahu 2020] are also considering the modeling of the transition between the surgical phases. However, the coarse-grained level of surgical phase modeling leaves out some fine details that would provide better information in the development of CAI systems in the OR.

In robotic surgery, research has focused on gesture recognition from kinematic data [DiPietro 2016, DiPietro 2019], videos from robotized surgery [Zia 2018, Kitaguchi 2019, Sarikaya 2020, Park 2021], system events [Malpani 2016] and the recognition of other events, such as the presence of smoke or bleeding [Loukas 2015]. [van Amsterdam 2019] used at a minimum one expert demonstration and its ground truth annotations to generate an appropriate initialization for a GMM-based algorithm for gesture recognition. Hence, the potential of weak supervision could be to improve unsupervised learning while avoiding manual annotation of large datasets.

Notwithstanding the benefits of surgical phase recognition, each phase comprises finer actions that could present a more detailed understanding of the tool-tissue interactions in the procedure. And so more works [Charriere 2014, Lecuyer 2020, Ramesh 2021]are now examining the recognition of the several steps within the phases. Nonetheless, both surgical phases and steps are a composite of several fine-grained activities that could require different forms of intervention. The need for deeper analysis of workflow activities necessitates the introduction of surgical action recognition [Rupprecht 2016, Khatibi 2020] to recognize the key verbs of the activities, e.g.: *dissection, cutting, coagulation, clipping, suturing, etc.*. Similarly, the action recognition challenge [Wagner 2021b] is also contested within the Endoscopic

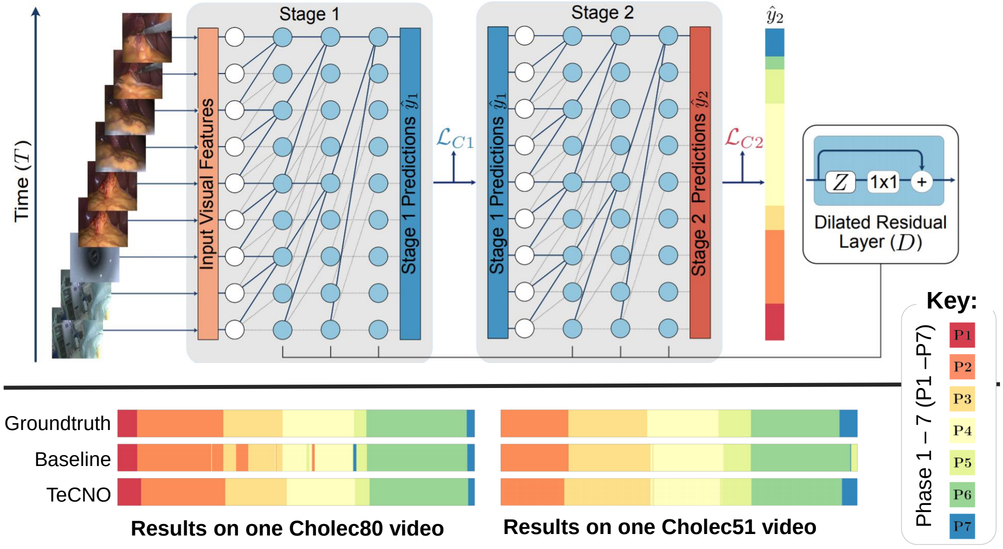

Time (T)
Stage 1
Stage 2
Input Visual Features
Stage 1 Predictions $\hat{y}_1$
$\mathcal{L}_{C1}$
Stage 1 Predictions $\hat{y}_1$
Stage 2 Predictions $\hat{y}_2$
$\mathcal{L}_{C2}$
$\hat{y}_2$
Z
1x1
+
Dilated Residual Layer (D)
Key:
Phase 1 – 7 (P1 –P7)
P1
P2
P3
P4
P5
P6
P7
Groundtruth
Baseline
TeCNO
Results on one Cholec80 video
Results on one Cholec51 video

Figure 2.10 – Architecture of TeCNO: a multi-stage TCN hierarchical refinement model for surgical phase recognition proposed in [Czempiel 2020]. The qualitative results of the method in comparison with ResNetLSTM baseline on two different surgical datasets is also presented in this figure.

Vision (EndoVis) challenge [4] at MICCAI 2019. These action recognition tasks are aimed at understanding tool-tissue interaction in surgical videos but their formulation ignores the information about the instruments performing the actions.

One of the earliest works on tool-tissue interaction recognition in videos is [Lo 2003] where a pipeline to segment laparoscopic videos is proposed. Using a naïve Bayesian network on top of several visual cues related to shape, deformation, change in light reflection, and other low-level visual features, the method yields promising results in segmenting five videos into four major events: idle, retraction, cauterization, and suturing. [Haro 2012] used tool-tissue interaction videos along with the kinematic data obtained from the robotic console to classify surgical gestures in the suturing action, such as insert needle and pull suture. [Charriere 2016] proposed a real-time method to jointly recognize two granularity levels of activities in cataract surgeries. By their results, they show that tool usage signals outperform visual information. This is expected since the tool usage signals contain more discriminative and semantic information compared to the low-level visual features extracted from the videos. However, tool signal information is not readily available. Despite the fine-grained nature of these action recognition tasks, they failed to capture the interacting elements (instruments and anatomies), and their relationships over time. Besides, detecting critical anatomy is crucial in automating safety warnings in CAI [Vercauteren 2019].

Based on this, [Katić 2014, Katić 2015, Neumuth 2006] formalized surgical activities as a triplet consisting of the used instrument, the performed action, and the organ acted upon. Also, [Speidel 2009] describe a surgical situation following a triplet formalism. This formulation introduced a deeper understanding of the image contents in laparoscopic videos thereby

[4] https://endovissub-workflowandskill.grand-challenge.org/

taking the workflow analysis to a new higher level. It also places surgical activity recognition at par with what is obtainable in the HOI in the vision community. The formulation of surgical action triplets offers great insight in modeling surgical situations and is a very expressive way of representing tool-tissue interaction. For this reason, [Katić 2014, Katić 2015] leverage the triplet formulation to better recognize surgical phases. Ever since then, there has not been continual research either directly modeling surgical actions as a triplet or using triplet information in surgical workflow analysis. Indeed, [Twinanda 2017] cites the difficulty in generating a triplet dataset as a major stumbling block to recognize surgical actions as triplets.

What comes very close to modeling surgical action triplets in videos is seen in the recent Medical Imaging and Deep Learning (MIDL) 2020 challenge [5][6]where a fine-grained action detection challenge is introduced. The challenge produced an ESAD dataset consisting of 4 prostatectomy videos annotated with 21 action classes that allow for the benchmark of several models during the challenge. One interesting feature of the ESAD dataset [Bawa 2021][6]is the provision of spatial labels for surgical actions which must have been motivated by the spatial labels for HOI detection [Zhuang 2018, Li 2020b, Jiang 2021, Chao 2018, Lin 2014] in the computer vision community. However, the dataset does not directly formulate surgical actions as a triplet. It instead describes the actions happening on the anatomies without taking note of the instrument performing the action. The instrument is especially important because what is considered a safe action on one tissue might be considered unsafe when using a different instrument. For example, actions of the coagulating instruments on the liver might be of risk if cutting instruments were used. Also, the anatomy classes in the dataset are not precise, as most organs are simply grouped as tissue. This would impair the needed information for fostering surgical safety using CAI. A report [Bawa 2021] following the SARAS-ESAD challenge summarized several methodologies used in the challenge. This includes the baseline model which is based on a feature pyramid network (FPN) that predicts the class scores and bounding box coordinates. Other methods in the challenge are based on single and double stage region-based CNN detectors, temporal modeling using LSTM and ConvLSTM, and attention mechanisms. Most of the models rely on deepening the baseline backbone and tweaking their data augmentation strategy to improve their performances.

A more recent work [Xu 2021][6]build upon two robotic minimally invasive surgery datasets that already have spatial labels and extend their annotations with action captions in the format of ⟨object1, predicate, object2⟩. However, their method does not offer a diverse set of interaction classes for each triplet category. Leveraging these triplet-like labels, their method generates captions for images in the surgical dataset which is synonyms to what video captioning in the computer vision community [Blank 2005, Laptev 2008, Fang 2015, Karpathy 2015]. Note that this work [Xu 2021] is published after the proposed method in this thesis.

[5] https://saras-esad.grand-challenge.org/

[6] Method published after our proposed method in this thesis.

2.4 Thesis positioning

In this thesis, we address the problem of surgical activity recognition in laparoscopic videos. We focus on fine-grained activities that are instrument-driven within a patient's body. Since the activities of interest are instrument-centric, we begin the thesis with methods detecting and tracking surgical instruments. Despite the efforts made by [Jin 2018, Zhang 2020a] to generate spatially annotations for the training of the detection model, we could observe two limitations as follows: (1) Due to the task difficulty, only a tiny fraction of the available dataset could be spatial annotated, (2) Using a fully-supervised approach, the generated dataset is split further leading to a more tiny test set which is not significant enough for model performance evaluation. An effort is made by [Vardazaryan 2018] to address some of these limitations by proposing to supervised a tool localization model on the easier to generate binary presence labels which are readily available in the research community. However, their localization is limited to only a coordinate point on the tip of the instruments. In this thesis, we improved on the weakly supervised method proposed in [Vardazaryan 2018] by first extending the localization from point coordinates to whole region boundaries of the tooltips enclosed with bounding boxes. Addressing the second limitation, we fully annotated 5 videos with spatial bounding boxes only for the evaluation of our weakly-supervised models. The entire 12K instances of the spatially annotated dataset are set aside only for model evaluation thereby increasing the reliance on the performance of the models evaluated on such a large dataset. This test set is even larger than the entire 2532 annotation instances in the existing m2cai-tool-localization datasets [Jin 2018] and 3164 instances in AJU-Set [Jin 2018, Zhang 2020a] which the authors even further split between training and tiny validation set.

In addition to localizing surgical instruments by a weakly supervised approach, we extend our proposed method to track the localized instruments across time. One can observe that the number of studies addressing instrument tracking in laparoscopic videos is still limited and the problem is still unsolved. Most of the existing methods in the literature rely on hand-crafted features [Sznitman 2012a, Speidel 2008, Du 2016]. Despite being widely used in the vision community, the deep learning method is hardly used for surgical instrument tracking. This is not unconnected to the lack of spatially annotated data. In most of the available datasets [Jin 2018, Zhang 2020a], their spatial labels do not encompass the entire video but only cover selected sample images due to the tedious efforts required to annotate long laparoscopic videos, and hence, they are not usable for modeling consistent trajectories. The models built for instrument tracking in robotic and robot-assisted surgery [Ye 2016, Du 2018, Colleoni 2019, Ryu 2013] rely on kinematic information which are not available for non-robotized surgeries. We tackle these problems in two steps: (1) by providing a method that does not require spatial annotations or kinematic information for their training, (2) by exploiting the inherent temporal information in the video data to model surgical instrument tracking in a weakly-supervised deep learning model.

Without spatial labels, it is not straightforward to learn the trajectory of a moving instrument. Using temporal modeling units like LSTM, GRU, etc., in this case, would flatten the features without preserving the pixel spatial relationships. One would expect that 3D-

CNN [Chen 2018] and GCN [Wang 2019a] could have been the interesting ways of capturing temporal features alongside the CNN spatial features. However, the lack of state management in both models limits their approach to be implemented only for short temporal videos. It would not be feasible to build a 3D-CNN or GCN model that can fit a localizable size of spatial features from every frame of a long laparoscopic video as a single input for its computation. Tailoring such a model for offline processing with pre-recorded videos where the CNN is first saved to a buffer before temporal refinement as is the case of many methods in the computer vision community does not support real-time and online inference which is expected in the OR. We tackle this limitation by providing a convolutional LSTM (ConvLSTM) method that can process a full laparoscopic video of any length by propagating a sequential temporal state one frame at a time till the entire duration of a video without any memory bottleneck.

A recent method [Robu 2020][7]in the literature approached tool tracking with a box localization that encompasses the entire instrument body including the shaft and end-effector. This would be more useful in segmentation tasks but considering tracking with bounding box localization, the boxes would cover the entire frame whenever an instrument is inserted vertically, horizontally, or diagonally across the opposite end of the frame. In such a case, a slight motion of the instrument would be untracked if the handle position could still result in a bounding box covering the entire frame. In laparoscopic surgery, it can be observed that the instrument's handles are mostly similar across instrument classes. And since the tool-tips are the clinically relevant part of the instruments when describing their interaction with the anatomy, we propose a method that precisely localizes the tool-tips as was also the case in [Zhao 2017].

The review in section 2.2.1 shows that tracking development in the computer vision is aided by several benchmark challenges [Leal-Taixé 2015, Ellis 2010, Kristan 2015], datasets [Geiger 2012, Dollár 2009, Milan 2016, Dendorfer 2020] and evaluation standards [Bernardin 2008, Huang 2020]. We positioned our work with several take home from these reviews. First, We align our tracking as a MOT following the same categorization of similar methods [Bae 2014, Tang 2016, Bergmann 2019, Keuper 2016, Milan 2017, Lee 2016] that tracks multiple objects per frame. However, our case would fall under a special case known as SCMOT owing to the single instance of the multiple instrument classes in the laparoscopic video except for grasper. Secondly, we developed an end-to-end detection and tracking that is suitable for real-time applications. Our proposed method follows the *tracking-by-detection* approach. Thirdly, just like in [Liu 2020, Liu 2019, Liu 2018] we employed a convolutional LSTM to natively learn the data association part of the tracking leveraging the temporal information in the surgical videos. Furthermore, we also follow the established CLEAR MOT metrics [Bernardin 2008] standard in judging our model performance.

Beyond the instrument detection and tracking, we tackle surgical activity recognition in laparoscopic videos. We observed that most of the existing works in the medial computer vision literature approached this by recognizing coarse-grained activities such as phases [Ahmadi 2006], steps [Ramesh 2021], events [Loukas 2015] or gestures [DiPietro 2016]. Our review on the task handling in the computer vision community shows that activity recognition

[7] Method published after our proposed method in this thesis.

has long been disentangled to recognizes the individual actors such as humans and objects in the activity formulation which could present more details for situation awareness modeling and safety checking in laparoscopic procedures. Despite the formulation of surgical actions as a triplet of ⟨instrument, verb, target⟩ this problem remains largely unsolved. In this thesis, we address this limitation in two steps: (1) by generating a large dataset for surgical action triplet recognition, (2) by proposing novel approaches to recognize surgical actions as triplets. To the best of our knowledge, this work is the first study that recognizes action triplets directly from surgical videos.

Other methods tackling fine-grained surgical action recognition or detection are either ignoring the information about the surgical instruments [Bawa 2021], the anatomy [Park 2021] or both [Khatibi 2020]. A more recent work [Xu 2021][8], considers both instrument and anatomy but their data labels do not offer a diverse set of interaction classes for each triplet category. Different anatomies are simply grouped as tissue. We provide specific class details of the anatomies involved in the tool-tissue interaction. Since their work also provides spatial labels for the instruments performing the actions, our method could stand as a bridge between recognizing more details about the surgical actions and localizing the regions of the action. Meanwhile, owing to the weak supervision utilized in our method, we also provide qualitatively the weak localization of the regions of the surgical action.

Since it is relatively new to recognize surgical actions as triplets, we examined the level of works done in triplet detection and recognition in the computer vision community. Most of the works [Qi 2018, Gkioxari 2018, Shen 2018, Xu 2019] on HOI decompose triplet detection into two stages of detecting the components and associating their interaction. We follow suit to propose deep learning models that first learn the instrument, and their target tissues before deciphering how these components interact in a given surgical video frame.

Detecting the individual components of the triplet is not trivial especially when the target is not explicitly determined by their visual presence but by their involvement in the interactions. [Gkioxari 2018] shows that a visual appearance cue learnable using a fully supervised object detector like FasterRCNN can be useful in detecting the correct triplet components. We follow a similar approach by proposing a method that can guide the anatomy detection by the instrument's localization information. A challenge exists in that our triplet dataset does not contain spatial labels to train an object detector for this purpose as is the case in [Gkioxari 2018]. We solve this problem by leveraging weak supervision on binary presence labels.

It is still not straightforward to associate the detected components to form a triplet especially in a multi-label situation like in laparoscopic instruments and anatomies. In [Shen 2018], an outer product of the detected object's logits and detected verb's logits is employed to form a 2D matrix of interaction between the verbs and the objects in HOI. The challenge is that while a single-class human is the only subject in the HOI's triplet, the instrument class in the surgical triplet is multi-labeled thereby increasing the complexity of the data association. We solve this by proposing a learnable 3D interaction for the triplet components association.

While surgical action triplet recognition is still largely unexplored, the computer-vision counterpart of HOI is widely researched in the literature with lots of works exploring varying

[8] Method published after our proposed method in this thesis.

methodologies including graphical modeling [Qi 2018], temporal modeling [Do 2017], attention [Zou 2021], etc. One would also observe a gradual shift towards utilizing attention features to better detect the action of interest in HOI which is supported by a study in [Ulutan 2020]. This inspired the exploration of the attention mechanism in this thesis for triplet components detections and their association leading to the development of a transformer-inspired neural network for this purpose.

Methods, Results, and Discussion Part II

3 Weakly-Supervised Method for Surgical Tool Detection and Tracking

Doing little things well is a step towards doing big things better.
– Vincent Van Gogh

It is the little details that are vital. Little things make big things happen...
– John Wooden

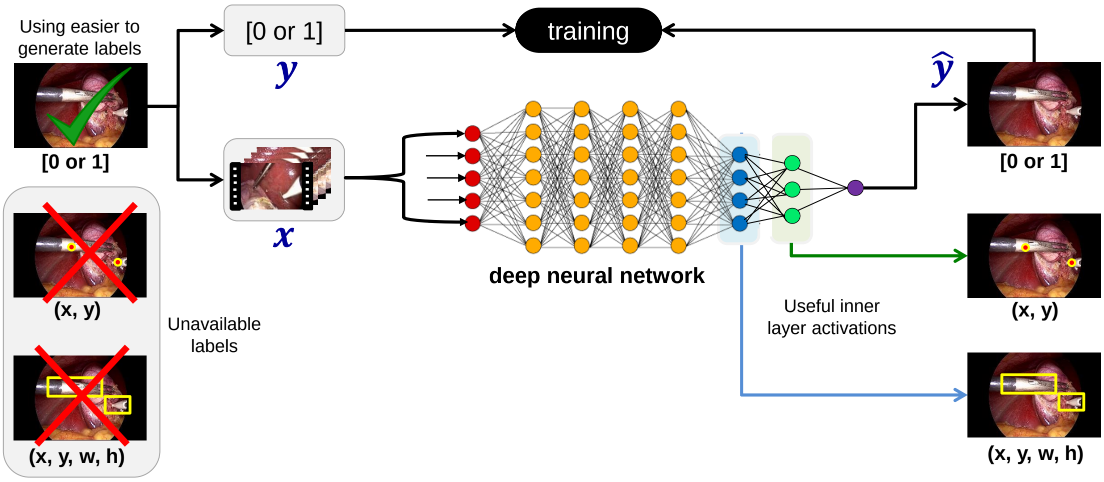


Figure 3.1 – An illustration of weakly-supervised spatial localization using image-level labels of surgical image data. When faced with a lack of spatial annotation, a model can be trained in a fully convolutional manner using binary presence labels while its inner layers can learn the spatial localization task.

Chapter Summary

In this chapter, we present the weakly supervised deep learning pipeline for surgical instrument detection and tracking in laparoscopic videos [Nwoye 2019]. Firstly, we formalize detection and tracking as a joint task learning in the same pipeline in Section 3.1. Even as we rely on binary presence labels to train our model, we manually generate a significant amount of spatial labels for the model evaluation as will be presented in Section 3.2. In the rest of the sections, we present our methods, the baseline, the experiments done in this work, and the results in that order.

3.1 Formalization

We formalize the joint detection and tracking tasks using some notations as follows. Given a laparoscopic video $V = [\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N]$ containing N sequential images. We learn a function f that can map input features $\mathbf{X}$ to strong labels $\mathbf{Y}$:

$$f : \mathbf{X}_t \longrightarrow \mathbf{Y}_t, \quad 1 \leq t \leq N, \tag{3.1}$$

where $\mathbf{X}_t$ are features extracted from $\mathbf{I}_t$ at time t, and $\mathbf{Y}_t$ are the instrument binary presence labels at the current time.

To employ weakly supervised learning, we introduce an additional feature space $\mathbf{H}$ for the weak labels. And nicely incorporate this in-between our learning algorithm such that we could still predict the strong label $\mathbf{Y}$ from the image features, while possibly benefiting from the related information captured in $\mathbf{H}$:

$$f : \mathbf{X}_t \times \mathbf{H}_t \longrightarrow \mathbf{Y}_t \tag{3.2}$$

We ensure that the related information in $\mathbf{H}$ is of localization benefits by constraining that without any localized features in $\mathbf{H}$, the presence detection task would fail.

Extending weak supervision for motion tracking is non-trivial especially while relying only on binary presence labels for the training. Tracking solves a data association problem across a temporal direction. Given a sequence of images from the first to the t^{th} frame $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_t\}$, for each i^{th} image $\mathbf{I}_i$, we obtain a set of detections $D = \{d_1, d_2, ..., d_n\}$, where n is the number of detected instruments. Each detection $d_i = (b_i, c_i)$ is a pair of instrument's box coordinates and identity. We define a trajectory as a set of time-ordered detections $T_i = \{d_{i1}, d_{i2}, ..., d_{im}\}$ where m is the number of detections that form trajectory i. A tracker finds the set of trajectories $T_* = \{T_1, T_2, ..., T_k\}$ that best explains the detections in a temporal order. In a fully supervised network where the groundtruth coordinates of the detections are given, distance-based solvers such as Deep Sort [Wojke 2017], are usually employed.

For our proposed weakly-supervised approach, the detection coordinates comes from the activations in $\mathbf{H}$. Using a temporal model, we learn a function f that can smooth the activations in $\mathbf{H}$ by looking at the k previous features.

$$f : \mathbf{X}_{t..t-k} \times \mathbf{H}_{t..t-k} \longrightarrow \mathbf{Y}_t \tag{3.3}$$

With this, we can obtain a set of time-ordered trajectories using weak supervision.

The linear dependency in the joint formalization of these three tasks namely, presence detection, spatial localization, and motion tracking, ensure that the model could be trained in an end-to-end manner.

3.2 Dataset Generation

Our training data is Cholec80 dataset [Twinanda 2016b]. This consists of laparoscopic surgery recordings obtained at the University Hospital of Strasbourg/IRCAD. It consists of 80 videos of cholecystectomy surgeries aimed at removing the gallbladder laparoscopically, monitored through an endoscope. The videos are recorded at the frame rate of 25fps and downsampled to 1fps at which the tool presence binary annotations are generated.

Table 3.1 – The statistics of the tool bounding box dataset for model evaluation.

Videos	Frames	Bounding boxes								Grasper instances		
		Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator	Spec.Bag	Total	1	2	3
5	7168	6033	379	4313	327	384	332	354	12122	3546	1182	41

For our spatial tasks evaluation, we annotate 5 videos with tool centers and bounding boxes around the tool-tips. The tool shafts are excluded, following common practice. This annotation is carried out using in-house software, *Endolabeller*, developed by research group CAMMA[1] which allows class tagging, coordinates marking, and box drawing on image frames. It also allows label propagation across consecutive frames. With this, we generate 12K box instances from 7168 video frames for our model testing. The statistics of the dataset is presented in Table 3.1

Figure 3.2 shows some samples of images and their labels in the Cholec80 dataset for model training, including the spatial annotations generated for the evaluation of the instrument's localization and tracking in this experiment.

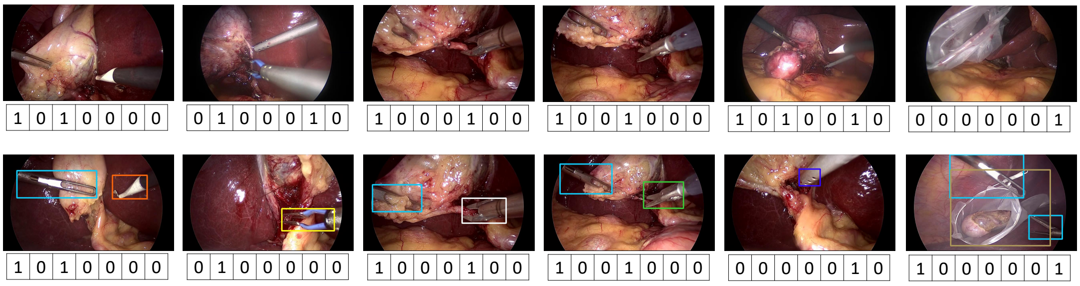


Figure 3.2 – Cholec80 data set showing: *First image row:* Sample images with binary presence labels used for training the model. The label vector captures the categories of the surgical instruments in the following order: grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag. *Second image row:* Sample images for testing the model, in addition to the binary presence labels, we also provide bounding box annotations.

3.3 Weak Supervision

Weak supervision in deep learning is a technique where imperfect labels are used to address a more challenging pattern recognition task. By imperfect labels, we mean that the annotations do not explicitly represent the groundtruth of the proposed task, they are either imprecise, inexact, inaccurate or represent something else. In other words, the training labels provide limited signals for the intended learning objectives.

Ideally, Deep Convolutional Neural Networks (DCNN) trained in a fully supervised manner, where the target labels describe the objective function, are superior in performance, however, these supervised learning approaches are "data-hungry", making them impractical in real-world industrial applications. Even with some amount of labeled data, a point of concern for supervised learning would be: how perfect is a perfect label? Can the "perfectness" of dataset labels be guaranteed? If not, what could be the performance of a model trained solely on the acclaimed perfect labels, when making inferences on noisy data. Would such models be able to make deductive reasoning beyond what they were trained on? Furthermore, if full supervision is the only way, would it be possible to annotate every real-world case in a dataset? Coupled with the fact that there is an insufficient quantity of labeled data and insufficient subject-matter expertise and time to label and prepare data, these many questions point to

[1] http://camma.u-strasbg.fr

the direction of weak supervision which is designed to innovate truly intelligent models that can learn beyond their pre-defined objective function.

This approach alleviates the burden of obtaining a hand-labeled dataset, which can be costly or impractical, however, it requires skillful formulation of the simple-complex task relationship in the training scheme. A weakly-supervised deep learning model while learning a weaker function such as recognition is expected to also capture some features usable for the recognition of a higher complex task such as localization.

Weak supervision can be *incomplete*- where only a subset of training data is given with labels, while the rest are left unlabeled. This can also be referred to as semi-supervised learning [Yoon 2020, van Amsterdam 2019]. Weak supervision can also be *inaccurate* - where the given labels are not always groundtruth. Most times, the model learns to remove the noise or is being regularized using the inaccuracies or label smoothing. Our interest is more on the third category known as *inexact* weak supervision. In this case, the training data are given with only coarse-grained labels, and a model is expected to learn more complex labels from the data. This approach is mostly used when there are no direct training labels for the objective tasks. Most weakly supervised models [Vardazaryan 2018, Jia 2017, Hwang 2016] relying on inexact labels are built on fully convolution networks (FCNs) where the inner convolution layers are designed to capture heatmaps that could inform the localization or segmentation. This type of weak supervision has been approached in the past in different ways. While some works configure new loss functions, others design novel training schemes to concurrently distillate knowledge to the weakly supervised layers. In all cases, the modeling should be intuitive enough to deduct or distillate knowledge from the weak training labels to a more complex and challenging pattern which is the hidden objective function.

Weak supervision is here motivated by the idea that when a CNN is trained in a fully convolutional manner for a classification task, some of the convolution layers before the dense layer learn a general notion about the detected object. The activations in these inner layers can therefore be exploited for other tasks than the ones they were originally trained for. Based on this, we employed weak supervision to learn surgical instrument localization and tracking while relying on binary presence labels. Our weakly supervised learning approach is formulated using an intuitive architectural design and end-to-end training scheme for the proposed model. As illustrated in Figure 3.1, the weakly-supervised models would be able to localize the present instruments in their last convolutional layers. They would also track their motion leveraging temporal information while relying on the easier to generate binary presence labels for their training.

3.4 End-to-End Architecture for Tool Detection and Tracking

The proposed approach is composed of a CNN + Convolutional LSTM (*ConvLSTM*) neural network trained end-to-end, but weakly supervised on tool binary presence labels only. We model our spatial localization using a convolution layer in such a manner that the presence probability is determined by the values on pixel activations (heat maps) in the convolution channels. We use the ConvLSTM unit to model the temporal dependencies in the motion of

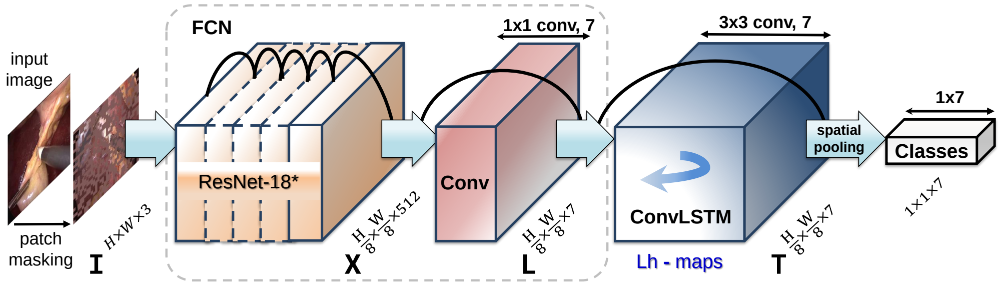


Figure 3.3 – The architecture of the proposed ConvLSTM tracker showing the *XLT* configuration.

the surgical tools and leverage its spatio-temporal ability to smooth the class peak activations in the localization heat maps.

3.4.1 Spatial Localization Modeling

3.4.1.1 Feature Encoding

Our models are built on the ResNet-18 architecture [He 2016], which is popular for its excellent performance on object detection. Before the era of residual networks, a simple neural network learns a function f that maps a given input x to an output y as follows:

$$y = f(x) \tag{3.4}$$

Stacking multiples of such functions increases the number of the learning parameters and computations, which helps the network to learn a deeper and more abstract representation of the input at every layer. Research has shown that deeper networks perform better than shallow counterparts since their increasingly complex layers favor the learning of non-linear functions. However, these deeper networks are prone to vanishing gradients due to the distance of some of the functions from the final layer, causing the exponential decay of their gradients to result in insignificant values too small to effectively change the model weights.

Residual neural networks overcome this bottleneck using an identity skip connection making it easier to optimize the residual mapping than to optimize the original, unreferenced mapping. It achieves this using a skip-connection (or shortcut) to the main function with reference to the layer inputs as follows:

$$y = f(x) + x \tag{3.5}$$

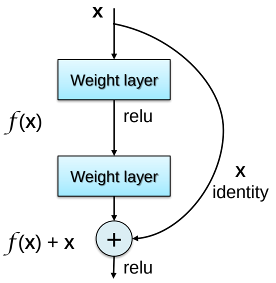


Figure 3.4 – Residual skip connection.

Residual networks also mitigate the degradation (accuracy saturation) problem: a situation where the training convergence becomes more difficult when adding more layers to a deep learning model.

training loss of a deep learning model increases with increasing layers.

Leveraging these properties, we employed a residual neural network known as ResNet [He 2016] for the feature extraction in our experiment. ResNet has several versions, but we choose a very lightweight version 18 for two reasons:

a. It produces higher resolution output feature maps: Since ResNet-18 has lesser layers, this minimizes the downsampling rate of the output features. Of course, higher resolution features maps are better for more precise localization. We even adjusted the strides of the last two blocks of the ResNet-18 from 2 to 1 pixel for the same purpose.

b. It is easier to train. Since we are working on medical images which are usually recorded at higher image resolutions, using a very deep feature extractor would increase their training time. And downsampling the original input image would lead to the loss of some tiny but salient landmarks of both the instruments and the anatomies that could help in the feature discrimination.

Finally, since full convolution architecture is the key for our spatial localization modeling, we remove the FC-layer of the ResNet.

Hence, we fed an (or a batch of) RGB input image(s) $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ to the ResNet-18 to extract high-level features $\mathbf{X} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 512}$ which are passed to the localization layer discussed in the next section.

3.4.1.2 Spatial Localization Layer

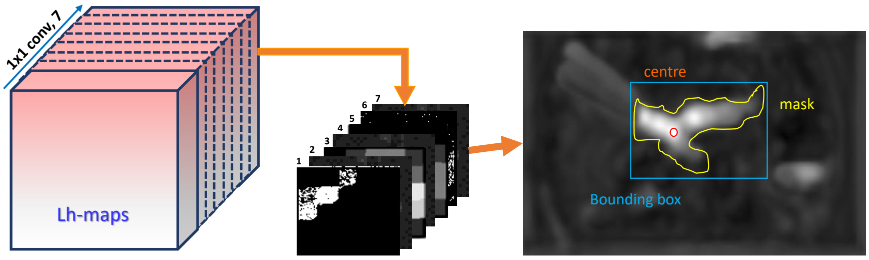


Figure 3.5 – A illustration of spatial localization modeling using a 7-channel convolution layer.

We model our spatial localization using a convolution layer with 7 filters to convolve the extracted features X into a 7-channel Localization heat maps (Lh-maps) $\mathbf{L} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 7}$ as shown in Figure 3.3. In our design, each channel is constrained to learn and localize a distinct tool type out of the 7 tools present in the considered laparoscopic procedure. Our intuition is also that the detected tool identities (IDs) would correspond to the IDs of the $\mathbf{L}$ channels whose heat map activation signals a positive localization in a given image.

The localization is cast on the activated regions of the $\mathbf{L}$ that surpass a threshold value. We noticed the peak of this activation mostly likely corresponds to the tools center point or the most discriminating part of the tool-tips. As shown in Figure 3.5, we added a post-processing step during evaluation where we automatically inspect each channel of the $\mathbf{L}$ so that a weak segmentation mask can be extracted from the connected component around the peak activated pixel using Otsu automatic thresholding [Otsu 1979]. And so a bounding box is fit over the mask to gather the tool location coordinates.

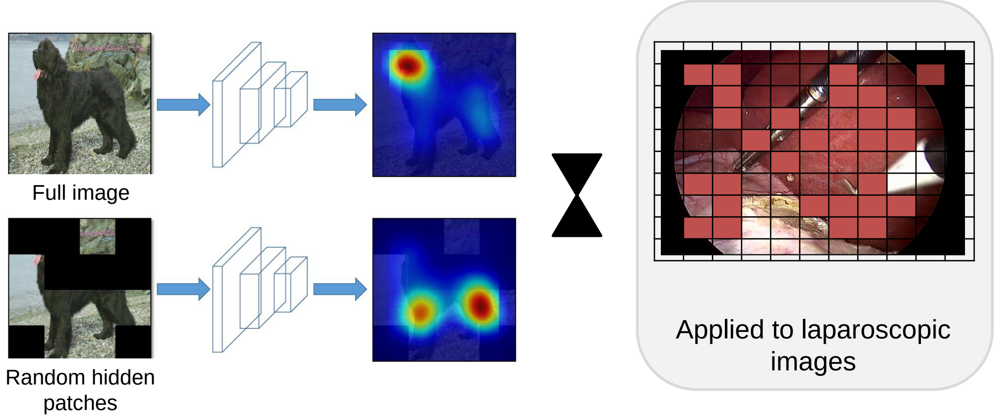


Figure 3.6 – A random patch masking on images as proposed in [Singh 2018] translated in this work for medical images.

Initially, the model tends to localize only this most discriminating region as this is a prevalent issue in weakly supervised localization. We employ random patch masking [Singh 2017] to counter this effect. This is achieved by creating 16 × 16 patches over the entire original images as shown in Figure 3.6. These patches are selected randomly at a probability of 0.5 on every forward pass. The pixel values of the selected patches are replaced with the mean pixel value of the entire training dataset as done in [Vardazaryan 2018]. According to [Singh 2017], this process enables the network to learn meticulously the necessary details of the object of interest by trying to hide randomly some already learned discriminate region, thereby forcing the network to discover other regions for the localization of the objects.

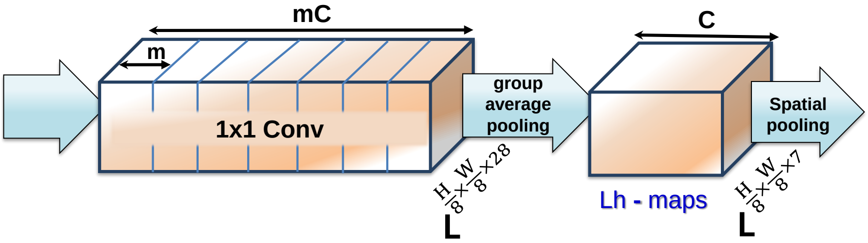

mC
m
1x1 Conv
group average pooling
$\frac{H}{8}\times\frac{W}{8}\times 28$
L
C
Lh - maps
Spatial pooling
$\frac{H}{8}\times\frac{W}{8}\times 7$
L

Figure 3.7 – A illustration of multi-maps of localization layer proposed in [Durand 2017] translated in this work for medical images.

For a special case, we also implemented a multiple mapping (multi-map) [Durand 2017]) of **L** to capture more localization details across multiple filters as done in [Vardazaryan 2018]. By multi-map, the localization layer is built with $m \times 7$ channels followed by an average pooling over each consecutive group of m channels to give the final 7 channels as shown in Figure 3.7. We retain $m = 4$ as used in [Vardazaryan 2018]. However, this multi-map did not bring any additional improvement to our localization model and was not continued in the tracking model.

3.4.2 Temporal Modeling for Motion Tracking

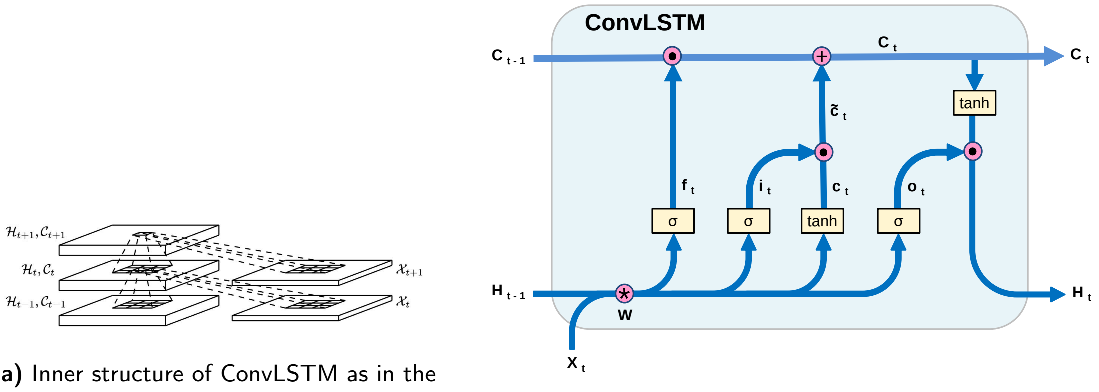
As shown in the Figure 3.3, we fed the extracted high-level features $\mathbf{X} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 512}$ to the localization layer to obtain the instrument's localized features $\mathbf{L} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 7}$ which is further passed to the temporal model for tracking. For the temporal modeling, we proposed we propose to use a RNN, to determine the current position of each tool from the input feature map along with information from prior images captured in RNN's state. This goes beyond the usual temporal refinement of an already extracted feature vector to instilling the full model pipeline with temporal awareness in their feature extraction and modeling. In designing this architecture, it is necessary to ensure that the overall network can still retain spatio-temporal information for each instrument when being trained in a weakly-supervised manner on binary presence data, namely that the localization information per tool is not lost but remains the key information used for predicting the binary presence. On this requirement, LSTM which is usable with flattened feature vectors could not be utilized for our proposed modeling as it does not preserve the spatial relationship of the localized pixels and cannot guarantee the preservation of the weakly learned localization heat maps captured in $\mathbf{L}$.

3.4.2.1 Convolutional Long Short Term Memory (ConvLSTM)

Since using a fully convolutional architecture is key in this regard, we, therefore, employ a ConvLSTM unit for its ability to learn the *spatio-temporal* dependencies of the localization heat maps while preserving the spatial dimension of the input features. The ConvLSTM achieves this by using a convolution kernel whose receptive field considers temporal information.

ConvLSTM is a RNN, just like the LSTM, but its internal matrix multiplications are exchanges with convolution operations (see Figure 3.8b). As a result, the data that flows through the ConvLSTM cells keeps the input dimension of the 3D input features instead of being just a 1D feature vector. Initially, a fully connected LSTM (FC-LSTM) is used for a similar purpose where images pass through the convolution layers and the outputs are flattened to 1D vector, collected over all images in the timestep to serve as the LSTM input. This too does not preserve the feature spatial dimension for the localization task. ConvLSTM is introduced in [Shi 2015] as an extension of FC-LSTM which has convolutional structures in both the input-to-state and state-to-state transitions. The ConvLSTM determines the future state of a certain cell in the grid by the inputs and past states of its local neighbors. This can easily be achieved by using a convolution operator in the state-to-state and input-to-state transitions (see Figure 3.8a).

The ConvLSTM takes an input feature $\mathbf{X}$, in our case, the input is the localized features L and produce a spatio-temporal localized features $\mathbf{T} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 7}$. Like the vanilla LSTM, the ConvLSTM unit uses two memory states, cell and hidden, to remember values over arbitrary time intervals. It also uses three gates to regulate the flow of information in and out of the cells. These gates take as inputs, the current inputs $\mathbf{X}_t$ and previous hidden states $\mathbf{H}_{t-1}$ and produce activations values that regulate usage and update of the information in the cell states C_{t-1} as shown in Figure 3.8b and describe as follows:


$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_t, \mathcal{C}_t$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_t$
ConvLSTM
C t-1
C t
C t
tanh
c̃ t
f t
i t
c t
o t
σ
σ
tanh
σ
H t-1
W
X t
H t

(a) Inner structure of ConvLSTM as in the original paper [Xingjian 2015].

(b) A ConvLSTM cell used in this work.

Figure 3.8 – Architecture of ConvLSTM Pipeline.

a. forget gate (f_t): A sigmoid layer for deciding on what (and amount of) information to forget/remember from the cell state.

b. input gate (i_t): A sigmoid layer that determines which current input values will be updated in the cell state. Before the update, a new candidate values c_t is obtained by a tanh function over the convolution of the previous hidden states and the current inputs. This helps to scale the update gate suggestions.

c. output gate (o_t): A sigmoid layer that decides the part of the cell states that should contribute to the final output. It first passes the current inputs x_t and hidden states h_{t-1} to a sigmoid layer to determine the output, then it passes cell states (with some parts forgotten and updated in previous stages) through a tanh layer to decide what part of the cell states should contribute to the output which will afterward serve as the current hidden state.

The gates are generated using convolutions on a concatenation of the current inputs $\mathbf{X}$ and previous hidden states $\mathbf{H}_{t-1}$. In practice, only a convolution operation using 4 times the original number of filters is performed on $\mathbf{X}$ and $\mathbf{H}$. The output is split into the 3 gates and the candidate values as shown in Figure 3.8b. In our implementation, we normalize the outputs of the convolution layer and the current cell states.

We summarize the key operations of ConvLSTM in the Equation 3.6, where (*) denotes the convolution operator and $\odot$ the Hadamard product. We ignore the normalization and bias terms for simplicity and ease of reading.

$$
\begin{aligned}
f_t &= \sigma\left(W_f * [H_{t-1}, X_t] + b_f\right)\\
i_t &= \sigma\left(W_i * [H_{t-1}, X_t] + b_i\right)\\
\tilde{c}_t &= \tanh\left(W_c * [H_{t-1}, X_t] + b_c\right)\\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{c}_t\\
o_t &= \sigma\left(W_h * [H_{t-1}, X_t] + b_o\right)\\
H_t &= o_t \odot \tanh\left(C_t\right)
\end{aligned}
\tag{3.6}
$$

The ConvLSTM units nicely integrates the functions on CNN in an LSTM pipeline. Compared to stacking a regular LSTM, the spatial relationships are maintained. And unlike using a simple convolution layer, the ConvLSTM takes into account the features from the previous frames, thereby enforcing consistency across time. With this, we have an RNN that suits the proposed weakly supervised localization as well as tracking task. The final *ConvLSTM Tracker* is constructed by adding a ConvLSTM unit to the localization model, as illustrated in Figure 3.3. The class peak activations in **L** are smoothed out in **T** using ConvLSTM leveraging the temporal information in the video data and resulting in consistent trajectories. With this smoothing effect, the ConvLSTM also replaces the IoU-based selection from the baseline tracker and naturally handles the birth and death of tracks for each tool.

3.4.2.2 History State Management

The nature of an RNN means that its prediction improves as it travels through time. This is because the RNN model is unrolled on a sequence of data. At the beginning of the sequence, it has little or no history. As it progresses over the sequence, it gains access to an increasing amount of previous data. This particular formulation is associated with a shortfall noticeable by a decline in model performance at the early (initialization) time. This is because the RNN cell are initialized from zero or random states which are not meaningful values.

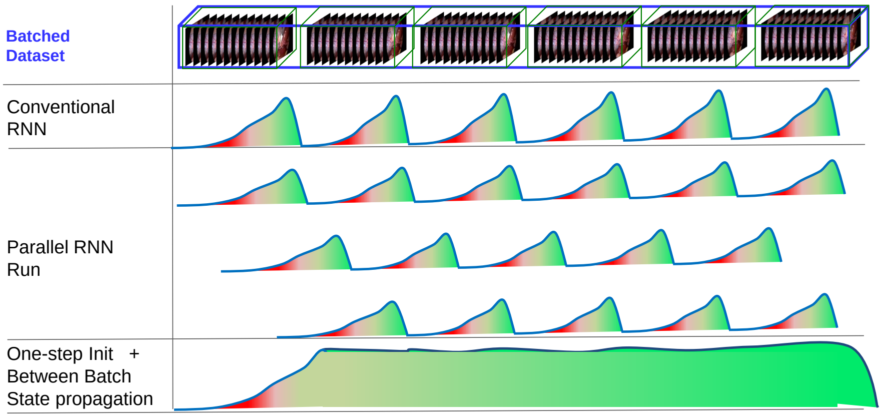


Figure 3.9 – An illustration of model usage of available temporal information with different RNN states management protocol.

Furthermore, a conventional RNN is known to propagate its internal states from one image to the next within an image batch by unrolling the model on the sequential data. Since the model is not also unrolled across batches of images, the RNN states are re-initialized at the beginning of every image batch as illustrated in Figure 3.9. This results in another performance drop at the beginning of every new sequential chunk (batch) of the same data due to the state re-initialization within the same video. Some works [Yu 2018, Czempiel 2020] attempt to overcome this by accumulating all extracted frame features so that full video features are given to an RNN network as a single batch. Aside from that the batch gradient descent associated with this modeling is sub-optimal compare to a stochastic one, longer videos with higher resolution features would require huge memory to fit the entire video as a single batch. In some cases, a parallel RNN is run with a shift batching protocol to account for the first few incorrect predictions within the batches as illustrated in Figure 3.9. Indeed, this would be time-consuming. Another work on surgical phase recognition [Yengera 2018] divides long videos into shorter sequences. In order to forward propagate the RNN states across the boundaries between consecutive sub-sequences, the model is trained with truncated back-propagation over accumulated gradients of the video sub-sequences. However, managing the batch normalization with this approach is complicated especially when training across multiple devices (or multi-GPUs).

We tackle these bottlenecks with two approaches to RNN memory state management.

a. **One-step initialization:** We ensure that the model is only initialized once at the beginning of a video. We use a *seek* counter, set to 0 at the first frame of a video and incremented afterward, to manage the state initialization and re-initialization.

b. **Between batch state propagation:** As illustrated in Figure 3.9, we propagate the ConvLSTM states *between* batches to maintain continuity in a video and capture temporal information over a longer sequence. This is achieved by re-initializing the ConvLSTM input states of every batch with the output states of the immediate previous batch without batch shuffling.

These setups are maintained at both the training and testing times.

3.4.3 Classification Layer

To perform weakly-supervised training on image-level labels y, we transform the *Lh-maps* into class-wise probabilities $\hat{y}$ using wildcat pooling [Durand 2017] as shown in Figure 3.3. We then learn a weighted cross-entropy loss function $\mathcal{L}$ for multi-label classification:

$$\mathcal{L} \longleftarrow \sum_{c=1}^{C} \frac{-1}{N} \left[\mathscr{W}_c y_c \log(\sigma(\hat{y}_c)) + (1 - y_c) \log(1 - \sigma(\hat{y}_c)) \right], \tag{3.7}$$

where y_c and $\hat{y}_c$ are respectively the ground truth and predicted tool presence for class c, σ is the sigmoid function, and $\mathscr{W}_c$ the weight for class c. The effect of the class weights $\mathscr{W}_c$ in this loss function is that $\mathscr{W}_c > 1$ decreases false negatives (FN) while $\mathscr{W}_c < 1$ decreases false positives (FP). With this, we counteract the polarizing effect of class imbalance by reducing FN for less frequent tools and reducing FP for dominant tools. The $\mathscr{W}_c$ is calculated as in Equation 3.8, where m is the median frequency of all tools in the train set and F_c is the frequency of the tools in class c:

$$\mathscr{W}_c \longleftarrow \frac{m}{F_c}. \tag{3.8}$$

The positive classes are selected by a threshold of 0.5 on the sigmoid values of the model predicted logits $\hat{y}$. The proposed model without the ConvLSTM unit is referred to as the FCN localization model proposed in [Vardazaryan 2018].

3.4.4 Variant of the Proposed Model

We explore three variants of the ConvLSTM tracker with similar architectures. Let **X** represent the feature e**X**traction unit, **L** the convolutional layer for the **L**ocalization modeling, and **T** the ConvLSTM for the **T**emporal modeling, we present the configurations of the model variants as follows:

a. **XLT Configuration:** In this configuration, illustrated in Figure 3.3, the ConvLSTM receives spatial input features from the **L** layer, refines them with temporal information and outputs *spatio-temporal Lh-maps*. The motivation for adding the ConvLSTM unit immediately after the FCN localization model (XL) is to refine the spatial *Lh-maps* with spatio-temporal information. This helps to smooth the class peak activations as well as the shape and size of the tools segmentation masks. It is important to note that the localization process is performed on the *spatio-temporal Lh-maps*, **T**.

b. **XTL Configuration:** With the ConvLSTM unit added before the last Convolution layer of the FCN localization model, it refines the extracted spatial features **X** with spatio-temporal information before localization by **L**. This guides the model in choosing relevant features based on temporal information across the video frames. By doing so, the receptive fields of **L** become aware of the temporal information. It is also important to note that the localization is on the **L** layer, which receives a spatio-temporal feature map, **T**, and outputs a *spatial Lh-map*, **L**. This model is expected to be more robust to occlusion and noise.

c. **XT Configuration:** The last variant replaces the **L** layer of the FCN localization model with a ConvLSTM (**T**) layer. Owing to its internal convolution process, the ConvLSTM layer takes over the task of localization from the **L** layer as well as the refinement of the feature map with temporal information. This results in a less complex architecture with the localization process on the **T** layer that produces *spatio-temporal Lh-maps*.

Note that we rename the configuration IDs, which are different from the published work, to avoid conflict with other standard abbreviations in the whole thesis.

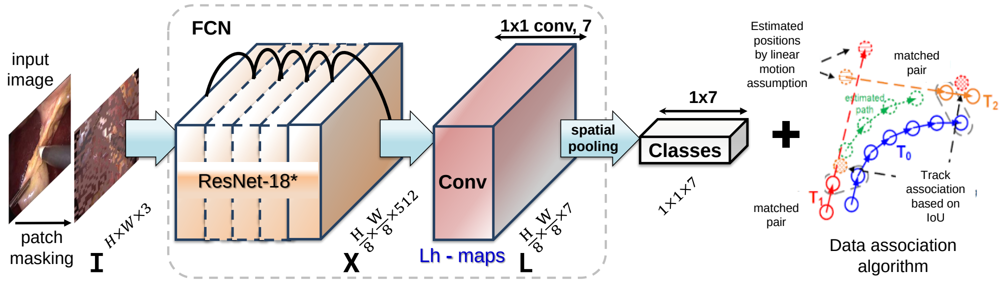


Figure 3.10 – The architecture of the FCN localization baseline (XL) + a tracking algorithm.

3.5 FCN Baseline with Tracking Algorithm

Our baseline is the FCN Tracking model which consists of the FCN localization model proposed in [Vardazaryan 2018] which we have re-implemented and extended with a handcrafted data association algorithm on the bounding boxes extracted from the Lh-maps as shown in Figure 3.10. Recall that the FCN localization model is the same as our proposed tracking model with the ConvLSTM part. We used a modified Hungarian [Kuhn 1955] track assignment algorithm to justify the contribution of the ConvLSTM part.

The FCN baseline has four variants models as proposed in [Vardazaryan 2018]. The difference in their configuration is accounted for by the use of random patch masking (rpm) and multi-map (m4) of the localization layer. These variants include the base configuration (FCN) using a single-map of the Lh-map (M1), a multi-map configuration (M4), a simple configuration trained on random patch masked images (M1+Msk), and a multi-map configuration trained on random patch masked images (M4+Msk). Again, we renamed the configuration IDs different from what it was in the published paper to avoid conflict with other standard abbreviations in the whole thesis.

We leverage the separation of the tool type in the 7-channel *Lh-map* from the FCN localization model to build a baseline for tool tracking. For localization, the raw *Lh-map* is resized to the original input image size by bilinear interpolation. Then, with a disc structuring element of size 12, we perform a morphological closing on the resized map to fill small holes in the image. On each channel of the *Lh-map*, a segmentation mask is extracted from the connected component around the pixel with maximum value using Otsu automatic thresholding [Otsu 1979]. A bounding box is then drawn over the mask to extract the tool location coordinates.

For tracking, the Intersection over Union (IoU) of the bounding boxes between the current frame $\mathbf{I}_t$ and the previous frame $\mathbf{I}_{t-1}$ is computed for each detected tool. The data association algorithm decides the inclusion and exclusion of detections in a trajectory. Ideally, instruments detected at time t are included in the previous trajectories if the IoU with previous detections at time $t-1$ is at least 0.5. In the case of multiple instances of the same tool, the closest tool instance compared to the detections in $\mathbf{I}_{t-1}$ is selected. The tracking algorithm also decides the death of a trajectory if the previous trajectory is untracked or has no matched pairs in the current time. Additionally, it initiates the creation of a new trajectory using unmatched instruments at the current time.

3.6 Experimental Setup

In this section, we discuss the experimental setup for training and evaluating the proposed instrument detector and tracker.

3.6.1 Data Setup and Pipeline

The dataset used in this experiment is *Cholec80* [Twinanda 2016b]. The resolution of images recorded in this video dataset varies from (854 × 480) to (1920 × 1080) maintaining the same aspect ratio. We extracted the frames at 1fps and unified their spatial dimensions by resizing them to (854 × 480) pixels during our experiment. We generate *seek* labels which are the index numbers for the consecutive frames. Only a seek value of 0, which signifies the first frame of a video, would trigger the initialization/re-initialization of the ConvLSTM model from an initial state.

The models are trained on 40 videos and validated on 10 videos. The validation set is used for hyperparameter tuning. We use the remaining 30 videos for the evaluation of the instrument presence detection. For evaluating the localization and tracking tasks, we use 5 videos from the test set which have been annotated with tool-tips centers and bounding boxes coordinates. The tool shafts are excluded, following common practice. The training data are augmented using slight rotation, horizontal flipping, and random patch masking of images. When finetuning the ConvLSTM layer, the dataset augmentation is limited to patch masking to reduce the training time, since the video dataset already contains lots of variability in the images, and the baseline is already trained with two other augmentation styles. For a high-performance data loading pipeline, our training data are stored as serialized TFRecords binaries.

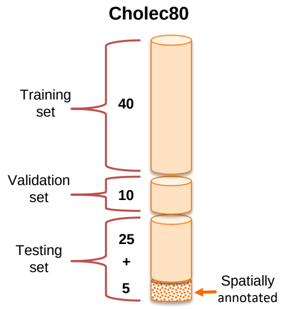


Figure 3.11 – Dataset splits.

3.6.2 Training and Loss Function

All the models presented in this chapter are trained by transfer learning. The feature extraction backbone is pretrained on ImageNet [Deng 2009] and so, during training, we reduce the learning rate for its optimization by $1e^{-2}$ to avoid swift override of universal features captured from the larger pretrained domain. All the models are trained using Stochastic Gradient Descent (SGD) with Momentum as optimizer (initial momentum $\mu = 0.9$). We maintained a step-wise learning rate ($\eta = 0.001$) policy, decayed ($\delta = 0.1$) after every 40 epochs.

All the models are trained for multi-label classification on the instrument binary presence labels. The optimized loss function $\mathscr{L}$ is the weighted cross-entropy with logits presented in Equation 3.7. An L_2 norm, with a weight decay constant of $1e^{-4}$ for the baseline and $1e^{-5}$ for the proposed models, is applied to regularize the optimization.

Owing to our GPU memory constraints and large input dimension, the network is trained with a batch size of 16 and the ConvLSTM models are unrolled to a timestep of 16. The ConvLSTM and the baseline models have the same backbone feature extractor for fair comparison which converges after 160 epochs of training. After which, every other layer is further trained for an additional 120 epochs maintaining a frozen backbone. Our model network is implemented in TensorFlow and trained for $14-21$ days on GeForce GTX 1080 Ti GPUs.

3.6.3 Inference and Evaluation Protocol

To quantify the instrument presence detection results, we use Average Precision (AP), which is defined as the area under the precision-recall curve. During inference, the predicted binary presence label is squashed into a sigmoid probability at which its AP with the groundtruth is calculated. At deployment, these probabilities are thresholded ($\Theta = 0.5$) to discrete binary values for instrument presence detection.

To quantify the network's ability to localize the distinct instruments in various frames, we compute the bounding box Intersection over Union (IoU) between the detected instruments and the groundtruths. This performance measure does not take into account the temporal consistency of the instruments across the frames. However, localization is only considered to be correct if and only if the $IoU \geq 0.5$. Because the binary presence annotation of the training dataset does not capture the number of instances per tool, at test time we compute the IoU of the detected tools to their closest groundtruth in the case of multiple instances [2].

For the tracking performance evaluation, we adopted the widely used CLEAR MOT metrics [Bernardin 2008]: MOTP and MOTA. MOTP is a measure of the localization precision which measures the average overlap between all the correctly matched hypotheses and their corresponding targets for a given IoU threshold (Θ).

$$MOTP = \frac{\sum_{t,i} D_{t,i}}{\sum_t C_t}, \tag{3.9}$$

where $D_{t,i}$ is the bounding box IoU of the tracked target i with the groundtruth, C_t is the number of matches in frame t. The value typically ranges between [Θ%, 100]. On the other hand, MOTA shows the tracker's ability at keeping consistent trajectories. It evaluates the effectiveness of the tracker from three errors, namely False Positive (FP), False Negative (FN) and Identity Switch (IDSW) in respect to the number of groundtruth objects (GT) as in equation 3.10:

$$MOTA = 1 - \frac{\sum_t FP_t + FN_t + IDSW_t}{\sum_t GT_t}. \tag{3.10}$$

The score, which usually ranges between (-∞, 100], can be negative in cases where the number of errors made by the tracker exceeds the number of all objects in the scene. Refer to [Bernardin 2008] for more details on the MOT metrics.

[2] This can only arise for the grasper in this dataset

3.7 Experimental Results

3.7.1 Quantitative Results

3.7.1.1 Presence Detection Results

Table 3.2 – Tool presence detection average precision (AP) for the evaluated models.

Method	Configuration	Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator	Spec.Bag	Mean
FCN Tracking Baseline	M1	96.7	91.9	99.4	50.6	80.3	85.2	88.3	84.6
	M1+Msk	**99.8**	92.6	99.8	85.1	96.9	60.9	78.6	87.7
	M4	95.9	89.4	99.5	69.3	85.4	**89.5**	87.1	87.9
	M4+Msk	99.6	90.9	99.8	48.5	88.5	66.2	91.0	83.6
ConvLSTM Proposed Tracker	XLT	99.7	**95.6**	99.8	86.9	**97.5**	74.7	**96.1**	**92.9**
	XTL	**99.8**	**95.6**	**99.9**	76.1	97.1	77.4	93.9	91.4
	XT	99.5	93.8	**99.9**	**90.3**	**97.5**	65.1	74.0	88.5

Comparing the AP of our model with the baseline (as presented in Table 3.2) shows that temporal information helps improve the instruments presence detection by over 5.0%. The performance improvement can also be seen across the instrument classes. This suggests that the temporal information helps the detection of instruments under occlusion and noise.

3.7.1.2 Spatial Localization Results

The localization results compared with our baseline model are presented in Table 3.3.

Table 3.3 – Localization accuracy of tools detected at IoU ≥ 0.5 for the evaluated models.

Method	Configuration	Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator	Spec.Bag	Mean
FCN Tracking Baseline	M1	05.9	20.5	34.7	03.5	06.4	**55.1**	44.4	24.3
	M1+Msk	15.5	10.1	27.8	20.0	13.3	53.7	06.4	21.0
	M4	05.0	11.5	15.5	25.1	8.7	42.5	14.8	17.6
	M4+Msk	08.7	0.01	25.6	20.0	**20.0**	49.0	02.2	17.9
ConvLSTM Proposed Tracker	XLT	33.8	**20.8**	41.9	21.1	12.6	52.1	23.8	29.3
	XTL	**54.5**	14.6	**50.0**	23.2	11.8	53.6	**60.1**	**38.2**
	XT	42.5	08.0	44.4	**25.3**	14.0	53.5	41.7	32.8

From the result in Table 3.3, our model improved the spatial localization of five out of the seven surgical instruments: grasper, bipolar, hook, scissors, and specimen bag. The proposed ConvLSTM trackers maintain comparable performance with the baseline in the localization of irrigator and clipper instruments. It is observed that the localizable tip of the irrigator is similar to its shaft, and hence there is no clear boundary in both the prediction and groundtruth spatial labels. This may account for the failure of the proposed model to outperform the baseline in both detection and localization. In Table 3.2, we observed that more than 90% of the tools occurrence can be successfully detected. Among these detections, four instruments that occur most frequently in the dataset (grasper, hook, irrigator, and specimen bag) can be correctly localized over 50% of the time using the strict $IoU \geq 0.5$ metric as shown in Table

3.3, which is a very promising result considering that no spatial information is used during training. Generally, the ConvLSTM shows a good performance on this metric by improving the mean accuracy by 13.9%. This justifies the benefits of using temporal information during training. We conclude that temporal data modeling helps in understanding the full spatial boundaries of moving objects. This is evident in the obtained results as all the ConvLSTM models outperform all the baseline models on mean spatial localization accuracy.

3.7.1.3 Motion Tracking Results

Table 3.4 – Tracking performance of the evaluated models.

Method	Configuration	Θ = 0.3		Θ = 0.5		Θ = 0.7		Mean	
		MOTP	MOTA	MOTP	MOTA	MOTP	MOTA	MOTP	MOTA
FCN Tracking Baseline	M1	58.1	29.8	**66.6**	19.3	77.3	05.3	67.3	18.1
	M1+Msk	49.9	47.9	61.2	21.2	75.3	02.7	62.1	23.9
	M4	46.6	29.6	60.4	09.6	75.4	-0.3	60.8	13.1
	M4+Msk	48.3	40.4	61.0	15.3	75.8	01.9	61.7	19.2
ConvLSTM Proposed Tracker	XLT	58.0	46.4	65.9	29.4	**77.4**	03.2	67.1	26.3
	XTL	**59.0**	**59.6**	65.9	**41.0**	77.3	**09.0**	**67.4**	**36.5**
	XT	54.4	47.7	63.3	26.1	76.7	00.3	64.8	24.7

We perform Multi-Class Tracking (MCT) of surgical instruments in laparoscopic videos owing that the dataset provides labels for different classes of the surgical instruments. The tracking performance is assessed across varying thresholds Θ in comparison with our baseline models as presented in Table 3.4. Our approach improved the baseline performance significantly. The results show that with comparable MOTP, ConvLSTM tracker can improve the MOTA baseline by 11.7% at Θ = 0.3, 19.8% at Θ = 0.5 and 3.7% at a strict Θ = 0.7. Generally, the ConvLSTM tracker shows its ability to learn a smoother trajectory by outperforming all the baseline in both mean MOTP and mean MOTA significantly.

Actually, we observed qualitatively that the proposed model can localize multiple instances of the same instrument as shown in Figure 3.12, but since the binary labels used for weak supervision does not differentiate these instances, we could not evaluate the model on MCMOT.

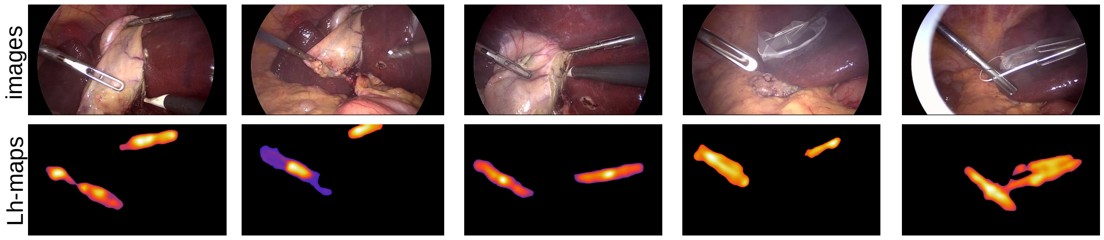


Figure 3.12 – Qualitative results showing the model's ability to localize multiple instance of the same instrument (grasper as the case is in Cholec80). However, the nature of the dataset labels does not allow for distinction of these instances for their evaluation. (best seen in colour).

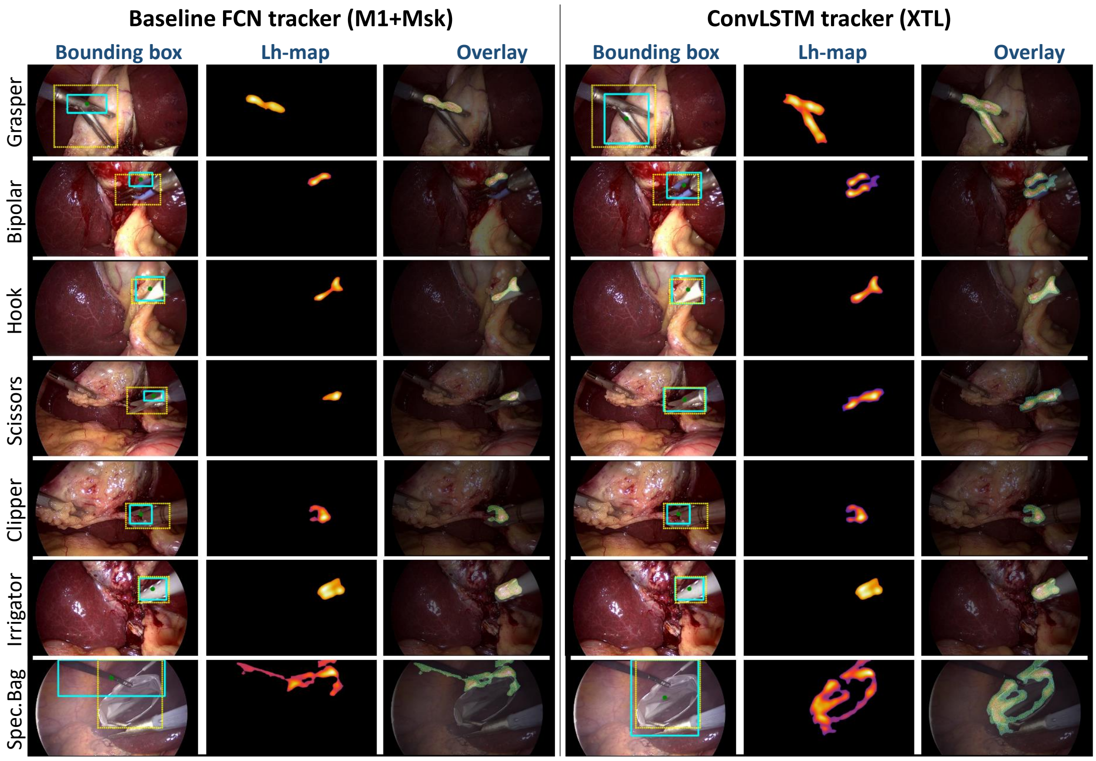


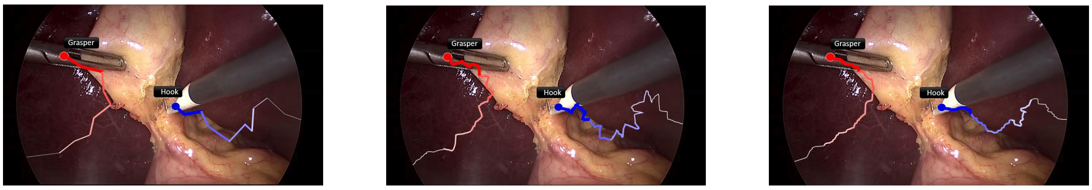
Figure 3.13 – Qualitative results showing the localization and tracking of the performance of the baseline and ConvLSTM models for the 7 tools. For each tool, we present a comparison of the detected bounding box (cyan in color) with the ground truth (dotted yellow box), the *Lh-map*, and the overlay of the segmented mask with the original image (best seen in color).

3.7.2 Qualitative Results

We present some qualitative results in Figures 3.12,3.13, and 3.14, to show what is learnt by the model in a weakly-supervised setting. We observe that using imperfect data for model training allows the model to figure out its one features that are needed to learn the complex task. This is evident in Figure 3.12 where a model shows the capability to localize multiple instances of the same instrument class. Recall that the models are trained on binary presence annotations, a label formulation that does not distinguish instrument instances, and yet the model inner layers could reason beyond these provided weak labels.

The qualitative results in Figure 3.13 show visually how the ConvLSTM is able to leverage the temporal coherence for tracking and localization for the 7 tools. From the positioning of the bounding boxes around the tools, it can be seen that the ConvLSTM model learns the region boundaries better than the baseline. The *Lh-maps* show that the ConvLSTM helps to smooth the localization and approximates the shape and size of the tools in each image. The overlay shows that it satisfactorily learns a trajectory close to the ground truth. A supplementary video that further demonstrates the qualitative performance of our approach can be found here: https://youtu.be/vnMwlS5tvHE. From the qualitative results, we notice that the *Lh-maps* produce a weak segmentation of the tool-tips, suggesting that this approach could be extended to segmentation.

Finally, our experiments also show that the ConvLSTM model trained on videos at 1fps can generalize to unlabeled videos at 25fps as shown in Figure 3.14, making it unconstrained by the fps. The likely indicator for this prowess is the one-state initialization and between-batch states propagation scheme introduced in this work which is maintained at both train and test time. With this, the pace of the instrument captured in the RNN state is propagated throughout the video which improves with more temporal histories. We also observed that the tracking trajectory is smoother at 25fps due to the very close motion of the instruments between the frames, unlike in 1fps with huge jumps. A video showing more results on variable fps tracking is provided here: https://youtu.be/SNhd1yzOe50.



(a) Trained on 1fps video data, tracking on 1fps test video.

(b) Trained on 1fps video data, tracking on 5fps test video.

(c) Trained on 1fps video data, tracking on 25fps test video.

Figure 3.14 – A qualitative results showing the ConvLSTM tracking ability at different frame rates.

3.7.3 Discussion

The evaluation presented in this work shows the positive contribution of the ConvLSTMs in modeling temporal data during weakly supervised training for surgical instrument tracking in laparoscopic videos. The most notable improvement is seen in the ConvLSTM XTL variant, which has the best results both in localization and in tracking. We believe that this is due to the fact that in this configuration, **T** refines the feature map from **X** with temporal considerations before they are localized separately by **L**. This is more robust than in XLT and XT, where the temporal refinement at the end of the pipeline may dilute the localization information and output a map with a slightly different semantic. In the XTL variant, the temporal information across the video frames guides the model in choosing relevant features for the *Lh-maps*, **L**.

In the qualitative results, we observe failure cases in different situations. First, due to the nature of the model, instruments might be missed when multiple instances of the same class are present. In the qualitative video, we could see that even when there are activations for multiple instances, the label formulation does not allow us to capture these on different localization channels. Sometimes, multiple instance activations are not usually high, making it difficult to ascertain the correct number of instances per instrument, and there is no formulation to decide the one that should be low or high among others. It would be interesting to see if the low activations in the *Lh-maps* could be exploited to estimate the number of instances for each class. The qualitative results also show that the models fail to detect an instrument when less than $\frac{1}{5}th$ of its tip is visible. We also observe that our models only localize the instrument's tip, not its shaft, likely because shafts are similar for all instruments and cannot be easily captured by a weakly supervised approach relying on binary presence.

3.8 Conclusion

This work aims at tracking tools in laparoscopic surgical videos without using any spatial annotation during training. A weakly supervised Convolutional LSTM approach that relies solely on binary tool presence information is proposed. First, we build a baseline tracker by performing a one-to-one data association on the localization results generated by the FCN proposed in [Vardazaryan 2018]. Then, we propose a fully convolutional spatio-temporal model for end-to-end tracking that is suitable for weakly supervised training. It relies on a ConvLSTM that leverages the temporal information present in the video to smooth the class peak activations and better detect the presence of tools, optimize their spatial localization and smooth their trajectory over time. This approach is evaluated on the Cholec80 dataset and yields 12.6% overall improvement on MOTA, 13.9% improvement on localization mean accuracy and 5% improvement on tool presence detection mAP. The results justify that the ConvLSTM can leverage the spatio-temporal coherence of consecutive image frames across a surgical video to improve tool presence detection, spatial localization, and motion tracking. The quantitative and qualitative results also suggest that the proposed approach could be integrated into a surgical video labeling software to initialize the tool annotations, such as their bounding boxes and segmentation masks.

4 A Multitask Learning Method for the Recognition of Surgical Action Triplets

Details make perfection, and perfection is not a detail
– Leonardo da Vinci

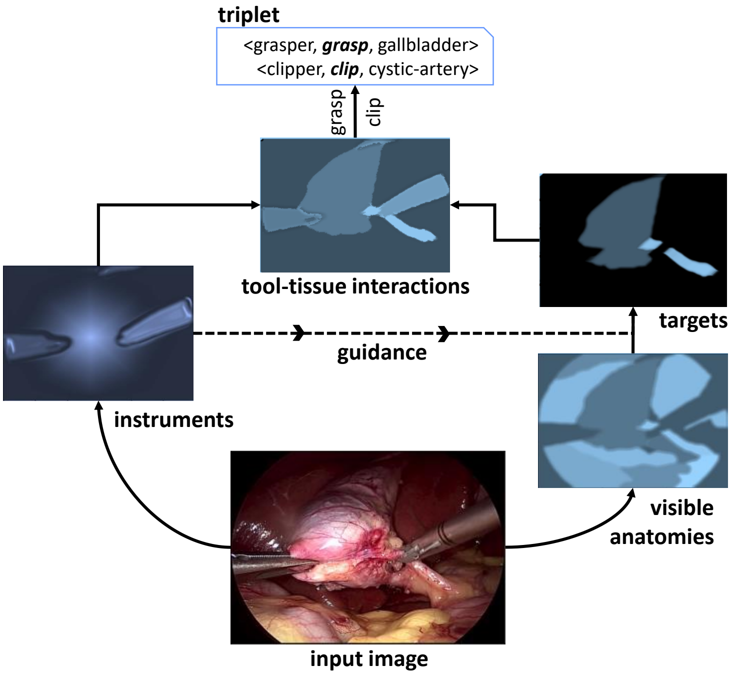


Figure 4.1 – Demonstrating the instrument-centric property of surgical action triplet.

Chapter Summary

In the previous chapter, we presented an end-to-end architecture for joint detection and tracking of surgical instruments in laparoscopic videos, which met the objectives of the first part of the thesis. In this chapter, we present a study [Nwoye 2020] that examine the activities of the surgical instruments at a fine-grained level. We first present an overview of surgical action triplet in Section 4.1, followed by a description of the first dataset generated to support research at this level of granularity in Section 4.2. We then present, in Section 4.3, a recognition pipeline for surgical action triplets that follows the characteristics of surgical triplets to provide a proof-of-concept to the defined task. This will be accompanied by some empirical baselines and the breadth of experiments carried out in Section 4.4. Finally, we discuss the experimental results in Section 4.5, highlighting their significance in achieving the AI needed for safety in the OR.

4.1 Surgical Action Triplet

In general, action triplet recognition can be defined as the automatic identification of an action instance as a triplet of ⟨*subject, verb, object*⟩ (SVO). The subject is the performer of an action. For instance, in the computer vision community, where this is mostly tackled as HOI recognition, *human* is the subject, resulting in ⟨*human, verb, object*⟩ triplet formalism. In surgical data science, the subject is the *instrument*, whereas the object is the *target* acted upon by an instrument. In laparoscopic surgery, these targets are mostly anatomies, although they can be other foreign bodies such as clip, suture, specimen-bag, water, etc. The target is the commonly ignored part of the triplet in other fine-grained action recognition framework.

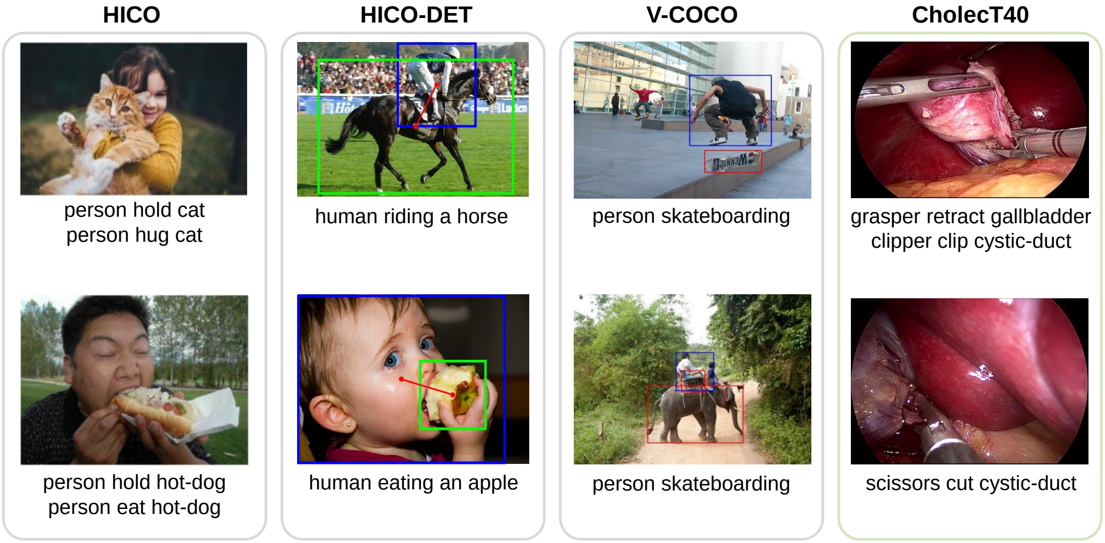


Figure 4.2 – Cross section of triplet datasets from HICO [Chao 2015], HICO-DET [Chao 2018], V-COCO [Gupta 2015], and CholecT40 [Nwoye 2020] datasets.

Despite being more challenging to annotate and detect, the target adds substantial semantics to the recognized action/instrument. The verb is a term that describes the action performed. It essentially describes the relationship between the instrument and target in a surgical action instant. Hence, it can also be regarded as the *interaction term*.

A description of surgical action to includes the used instrument, the action performed, and the treated anatomical structures is first given in [Neumuth 2006] as an aid to analyze surgical interventions in detail. [Speidel 2009] also describes surgical situations following the same analogy. [Neumuth 2010] used the terminology to unambiguously describe surgical processes when modeling complex behaviors. Specifically, [Katić 2014] presented an ontological formalism of surgical actions as a series of triplet $\langle$*instrument, verb. target*$\rangle$ (IVT) which was also re-established in [Katić 2015]. Both [Katić 2015] and [Katić 2014] leveraged triplet formulation provided by manual annotation to better recognize surgical phases.

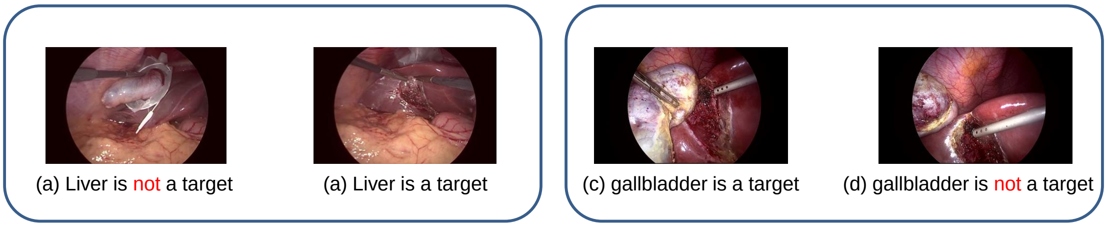

(a) Liver is not a target
(a) Liver is a target
(c) gallbladder is a target
(d) gallbladder is not a target

Figure 4.3 – Illustrating instrument-centric property of triplets.

Surgical action triplets are very unique and possess some characteristics that make them different from the conventional single verb actions. One of which is that action triplets are **instrument-centric**: meaning that an action is only performed if an instrument is present. Indeed, clinically an action can only occur if a hand is manipulating the instrument. As an example, the *liver*, which is visible most of the time in laparoscopic cholecystectomy, is labeled a target only when being acted upon by an instrument as illustrated in Figure 4.3(a-b). The same goes to the gallbladder illustrated in Figure 4.3(c-d). Building a recognition system that can selectively predict the target anatomies among other visible ones is non-trivial. The instrument-centric property is also strengthened by the fact that a verb that describes the action of an instrument, cannot be possible without the instrument itself.

Another property of the triplet is that each of the components, namely: instrument, verb, and target, is **multi-label**; meaning that multiple instances are possible for all the three when there are multiple triplets in one image. Solving the triplet association, in this case, is a tripartite graph matching problem, which is an NP-hard optimization problem.

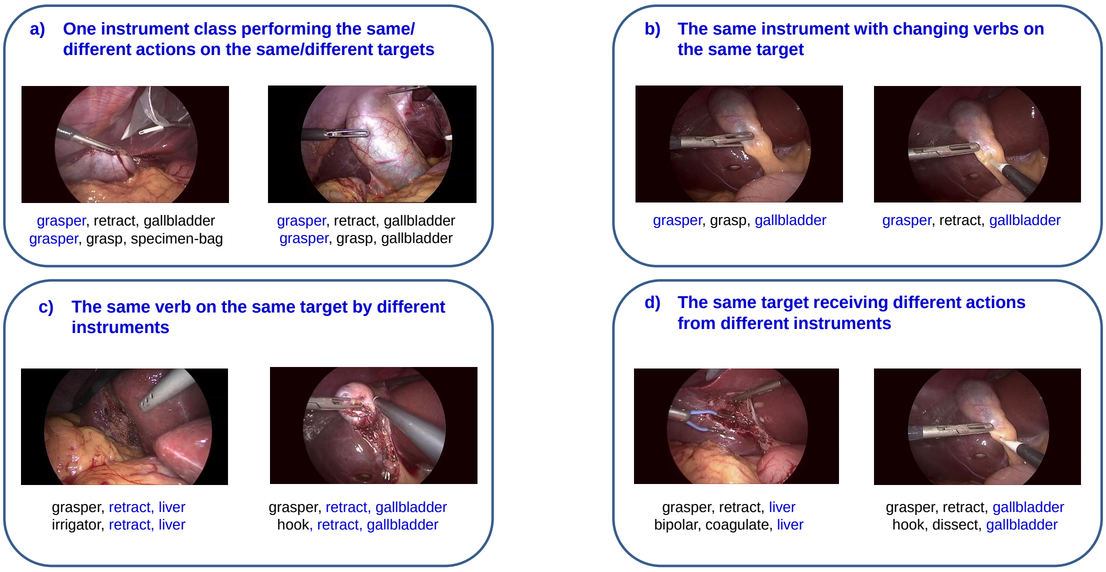

a) One instrument class performing the same/ different actions on the same/different targets
grasper, retract, gallbladder
grasper, grasp, specimen-bag
grasper, retract, gallbladder
grasper, grasp, gallbladder
b) The same instrument with changing verbs on the same target
grasper, grasp, gallbladder
grasper, retract, gallbladder
c) The same verb on the same target by different instruments
grasper, retract, liver
irrigator, retract, liver
grasper, retract, gallbladder
hook, retract, gallbladder
d) The same target receiving different actions from different instruments
grasper, retract, liver
bipolar, coagulate, liver
grasper, retract, gallbladder
hook, dissect, gallbladder

Figure 4.4 – Illustrating multiplicity and overlap property of triplets.

Furthermore, **multiplicity** and **overlap** exist in surgical action triplets. This can occur in all three components of the triplet. On the aspect of the instruments: one instrument class can be involved in multiple actions. As an example, one *grasper* can be grasping a *specimen-bag* while another is packing a *gallbladder* (see Figure 4.4(a)). A more complex scenario is when the same instrument is interacting with multiple targets such as a *grasper* grasping the collection of *blood-vessels* including the *cystic-artery* at the same time. An overlap can be found when different instruments are used for the same action (or verb), e.g.: *dissection* performed by *bipolar, grasper, hook, irrigator*, and *scissors* (see Figure 4.4(c)). Even the role (verb) of an instrument on a target can imperceptibly change within a short interval. So to say, the applications of the surgical instruments vary according to the surgeon's intention for use. For instance ⟨*grasper, retract, gallbladder*⟩, ⟨*grasper, grasp, gallbladder*⟩, ⟨*grasper, dissect, gallbladder*⟩ are visually similar but different action which are tough to distinguish even for experienced surgeons and require careful observation of the area surrounding the tool-tip (see Figure 4.4(b)). Similarly, one target can be simultaneously involved in multiple distinct actions. When operating on an organ or structure, multiple instruments can interact with the same target. e.g.: ⟨*grasper, retract, cystic-duct*⟩ and ⟨*hook, dissect, cystic-duct*⟩ happening at the same time. A more familiar case is where a *grasper* is retracting *liver* permitting the *bipolar* to coagulate it as shown in Figure 4.4(d). With all these multiplicity and overlap of actions possible in a single frame, action triplet in itself is also a multi-label.

An interesting property of the triplet is that the formalism is very **expressive** and in human-readable form. This makes it easier to use for fine descriptive feedback in safety monitoring, surgical report generation, and documentation, as well as the generation of video caption and subtitles for surgical education.

4.2 Dataset Generation

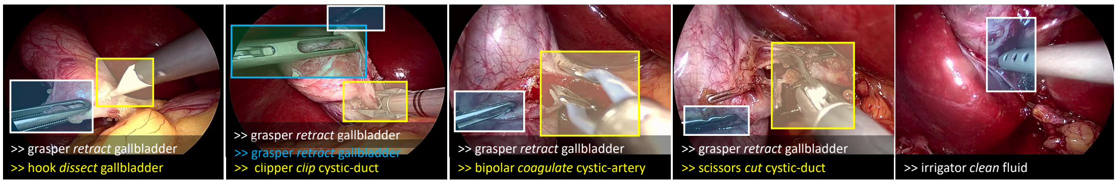

>> grasper *retract* gallbladder
>> hook *dissect* gallbladder
>> grasper *retract* gallbladder
>> grasper *retract* gallbladder
>> clipper *clip* cystic-duct
>> grasper *retract* gallbladder
>> bipolar *coagulate* cystic-artery
>> grasper *retract* gallbladder
>> scissors *cut* cystic-duct
>> irrigator *clean* fluid

Figure 4.5 – A sample of surgical images showing some action triplet instance labels. The localization in the images is not part of the dataset, but a representation of the weakly-supervised output of our recognition model.

To encourage progress towards the machine recognition of instrument-tissue interactions, we introduce *CholecT40* [Nwoye 2020]: an endoscopic video dataset consisting of 40 videos from the public Cholec80 [Twinanda 2016b] dataset, annotated with surgical action triplet information. These are videos of laparoscopic cholecystectomy surgery recorded at the University Hospital of Strasbourg, France, and collected by the CAMMA research group.

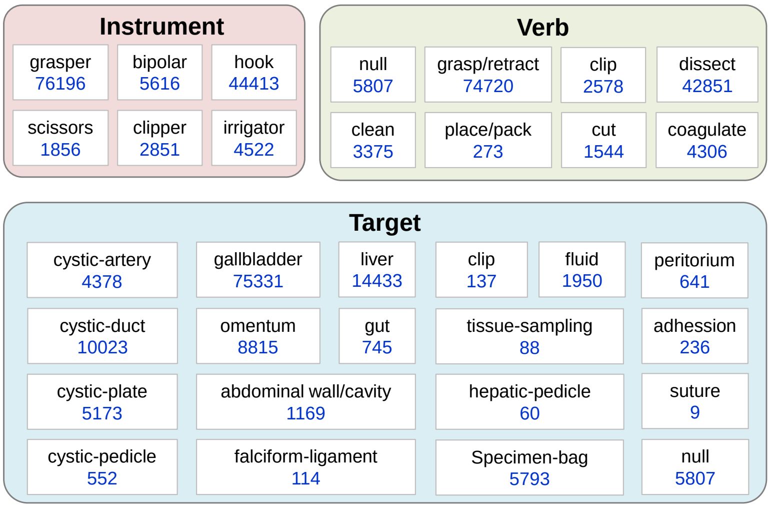
The annotations were carried out by a surgeon using the software *Surgery Workflow Toolbox-Annotate* from the B-com institute[1]. A team of surgeons, who are involved in both clinical practice and research, first developed an ontological dictionary containing a list of items to annotate, their definitions, identification protocol, and a naming convention to guide the annotation process. Training on the use of the annotation software is also provided.

The annotation is video-based. A video is a single laparoscopic intervention on one patient. It starts from the first insertion of the laparoscopic camera into the patient's body and stops with the last removal of the camera. An endoscopic video captures several coarse-grained surgical activities which can be broken down into finer division formulated as *action triplets*. The action triplets are annotated by marking the beginning and end for every temporal triplet instance. A temporal triplet label comprises of a continuous combination of an instance of the same *instrument, verb* and *target* configuration describing the tool-activity within a defined timeline. A change in the triplet configuration marks the end of the current action and the beginning of a different one. This occurs when the corresponding instrument exits the frame, or if the verb or target changes. An action is annotated as a triplet if an instrument is visibly interacting with the tissue/target in a given timeline, upholding the instrument-centric

[1] https://b-com.com/

property of the triplet. A *null* verb or target is annotated when a visible instrument is idle or when there is no instrument. Invisible surgical actions, whereby the performing instruments are out of the field of view, are not considered in this dataset. Out-of-frame actions are not reported, and video frames that are recorded outside the patient's body are zeroed out for privacy preservation.

The annotation process is followed by label mediation which is carried out by another clinician. We then define classes for the triplet. Theoretically, there is a large number of the observed instruments, verbs, and targets in the recorded videos. Their combinatorial possibilities are totally high. We sub-sample the labels based on their number of occurrences. Also, since the annotations are generated by marking the action timelines, more images with labels can be generated at higher frame rates. But for our experiments, we downsample the videos to 1 fps yielding a total of 83.2K frames annotated with 135K action-triplet instances. The resulting annotations span 128 triplet classes composed from 6 instruments, 8 verbs, and 19 target classes. We present these three components and their instance counts in Figure 4.6.



Instrument

grasper 76196
bipolar 5616
hook 44413
scissors 1856
clipper 2851
irrigator 4522

Verb

null 5807
grasp/retract 74720
clip 2578
dissect 42851
clean 3375
place/pack 273
cut 1544
coagulate 4306

Target

cystic-artery 4378
gallbladder 75331
liver 14433
clip 137
fluid 1950
peritorium 641
cystic-duct 10023
omentum 8815
gut 745
tissue-sampling 88
adhession 236
cystic-plate 5173
abdominal wall/cavity 1169
hepatic-pedicle 60
suture 9
cystic-pedicle 552
falciform-ligament 114
Specimen-bag 5793
null 5807

Figure 4.6 – Dataset statistics showing the frequencies of the instruments, verbs and targets in the triplet dataset.

The video dataset is randomly split into training (25 videos, 50.6K frames, 82.4K triplets), validation (5 videos, 10.2K frames, 15.9K triplets), and testing (10 videos, 22.5K frames, 37.1K triplets) sets as shown in Table 4.1.

Table 4.1 – Statistics of the dataset split.

Data split	No. of Videos	No. of Frames	No. of Label instances
Training	25	50.6K	82.4K
Validation	5	10.2K	15.9K
Testing	10	22.5K	37.1K
Total	40	83.2K	135K

Examples of such action triplets include: ⟨*grasper, retract, gallbladder*⟩, ⟨*hook, dissect, omentum*⟩, ⟨*bipolar, coagulate, liver*⟩ ⟨*clipper, clip, cystic-artery*⟩, ⟨*scissors, cut, cystic-duct*⟩, ⟨*irrigator, aspirate, fluid*⟩, etc, as also presented also Fig. 4.5. These annotations are in the form of binary labels which are positives for the classes of the action triplet occurring at each time in the videos.

The full dataset is presented in Table 5.2. Additional statistics on the co-occurrence distribution of the triplets are presented in terms of the ⟨instrument, verb⟩ and ⟨instrument, target⟩ in Tables 4.2 - 4.3 respectively.

Table 4.2 – Dataset statistics showing the instrument-verb occurrence frequency.

Verb	Instrument					
	Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator
clean	40	7	-	-	-	3328
clip	-	-	-	-	2578	-
coagulation	-	3756	534	16	-	-
cut	-	-	8	1536	-	-
dissect	767	892	40772	151	-	269
grasp/retract	72394	589	1006	45	59	627
null	2722	372	2093	108	214	298
place/pack	273	-	-	-	-	-

4.3 Tripnet: Proposed Approach for Action Triplet Recognition

To recognize the instrument-tissue interactions in the CholecT40 dataset, we build a new deep learning model, called *Tripnet,* by following a Multi-Task Learning (MTL) strategy. The MTL network models the detection of various components of the triplet following a baseline study which shows that naively classifying the triplet IDs without considering the individual components is insufficient for the recognition of action triplets from videos. Notwithstanding, the conventional MTL setup does not favor the instrument-centric property of surgical action triplet. This is because their individual branches do not provide any form of interaction with each other, and hence are context-free from the instrument cue. Meanwhile, our second baseline study shows that the intrinsic dependency of other components on the instrument's appearance cue is important for their correct detections. Hence, we propose a special kind of MTL strategy where the other branches leverage the activations from the instrument branch to better their detections. This we called the *Class Activation Guide* (CAG). Another novelty of the proposed model is seen in the use of *3D Interaction Space (3Dis),* proposed in this work, to learn the relationships between the components of the triplets. This is a 3D feature space where the relationship between the triplet components is resolved, providing a solution to the complex tripartite matching of the components. In the following sections, we describe in detail the proposed method and the supporting baselines.

Table 4.3 – Dataset statistics showing the instrument-target occurrence frequency.

Verb	Instrument					
	Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator
abdominal wall/cavity	36	361	-	-	-	772
adhesion	1	73	9	154	-	-
clip	137	-	-	-	-	-
cystic artery	38	190	2639	558	953	-
cystic duct	786	215	6710	670	1572	70
cystic pedicle	112	90	48	4	58	240
cystic plate	1451	478	2959	32	54	199
fallciform ligament	81	33	-	-	-	-
fluid	7	-	-	-	-	1943
gallbladder	48720	731	25750	57	-	73
gut	709	19	6	-	-	11
hepatic pedicle	10	46	4	-	-	-
liver	10919	2399	356	90	-	669
null	2722	372	2093	108	214	298
omentum	4413	521	3553	110	-	218
peritoneum	298	-	286	57	-	-
specimen bag	5685	79	-	-	-	29
suture	1	-	-	9	-	-
tissue sampling	72	9	-	7	-	-

Table 4.4 – CholecT40 dataset statistics showing the frequency of occurrence of the triplets.

Name	Count	Name	Count	Name	Count
bipolar, clean, gallbladder	7	grasper, grasp/retract, cystic-artery	38	hook, grasp/retract, liver	189
bipolar, coagulate, abdomenal-wall/cavity	361	grasper, grasp/retract, cystic-duct	786	hook, grasp/retract, omentum	11
bipolar, coagulate, cystic-artery	84	grasper, grasp/retract, cystic-pedicle	112	hook, null-verb, null-target	2093
bipolar, coagulate, cystic-duct	56	grasper, grasp/retract, cystic-plate	1373	irrigator, clean, abdomenal-wall/cavity	768
bipolar, coagulate, cystic-pedicle	75	grasper, grasp/retract, falciform-ligament	81	irrigator, clean, cystic-duct	29
bipolar, coagulate, cystic-plate	412	grasper, grasp/retract, fluid	7	irrigator, clean, cystic-pedicle	104
bipolar, coagulate, falciform-ligament	33	grasper, grasp/retract, gallbladder	47894	irrigator, clean, cystic-plate	152
bipolar, coagulate, gallbladder	341	grasper, grasp/retract, gut	709	irrigator, clean, fluid	1943
bipolar, coagulate, liver	2132	grasper, grasp/retract, hepatic-pedicle	10	irrigator, clean, gallbladder	14
bipolar, coagulate, omentum	262	grasper, grasp/retract, liver	10919	irrigator, clean, liver	291
bipolar, dissect, adhesion	73	grasper, grasp/retract, omentum	4381	irrigator, clean, specimen-bag	27
bipolar, dissect, cystic-artery	106	grasper, grasp/retract, peritoneum	286	irrigator, dissect, cystic-duct	41
bipolar, dissect, cystic-duct	135	grasper, grasp/retract, specimen-bag	5680	irrigator, dissect, cystic-pedicle	89
bipolar, dissect, cystic-plate	54	grasper, grasp/retract, suture	1	irrigator, dissect, cystic-plate	10
bipolar, dissect, gallbladder	348	grasper, grasp/retract, tissue-sampling	57	irrigator, dissect, gallbladder	29
bipolar, dissect, omentum	176	grasper, null-verb, null-target	2722	irrigator, dissect, omentum	100
bipolar, grasp/retract, cystic-duct	24	grasper, place/pack, abdomenal-wall/cavity	18	irrigator, grasp/retract, abdomenal-wall/cavity	4
bipolar, grasp/retract, cystic-pedicle	15	grasper, place/pack, clip	94	irrigator, grasp/retract, cystic-pedicle	47
bipolar, grasp/retract, cystic-plate	12	grasper, place/pack, gallbladder	141	irrigator, grasp/retract, cystic-plate	37
bipolar, grasp/retract, gallbladder	35	grasper, place/pack, specimen-bag	5	irrigator, grasp/retract, gallbladder	30
bipolar, grasp/retract, gut	19	grasper, place/pack, tissue-sampling	15	irrigator, grasp/retract, gut	11
bipolar, grasp/retract, hepatic-pedicle	46	hook, coagulate, cystic-artery	20	irrigator, grasp/retract, liver	378
bipolar, grasp/retract, liver	267	hook, coagulate, cystic-duct	41	irrigator, grasp/retract, omentum	118
bipolar, grasp/retract, omentum	83	hook, coagulate, cystic-pedicle	15	irrigator, grasp/retract, specimen-bag	2
bipolar, grasp/retract, specimen-bag	79	hook, coagulate, cystic-plate	9	irrigator, null-verb, null-target	298
bipolar, grasp/retract, tissue-sampling	9	hook, coagulate, gallbladder	213	scissors, coagulate, omentum	16
bipolar, null-verb, null-target	372	hook, coagulate, liver	159	scissors, cut, adhesion	154
clipper, clip, cystic-artery	952	hook, coagulate, omentum	77	scissors, cut, cystic-artery	551
clipper, clip, cystic-duct	1558	hook, cut, liver	8	scissors, cut, cystic-duct	655
clipper, clip, cystic-pedicle	14	hook, dissect, adhesion	9	scissors, cut, cystic-plate	20
clipper, clip, cystic-plate	54	hook, dissect, cystic-artery	2582	scissors, cut, liver	90
clipper, grasp/retract, cystic-artery	1	hook, dissect, cystic-duct	6509	scissors, cut, peritoneum	57
clipper, grasp/retract, cystic-duct	14	hook, dissect, cystic-plate	2899	scissors, cut, suture	9
clipper, grasp/retract, cystic-pedicle	44	hook, dissect, gallbladder	25022	scissors, dissect, cystic-plate	12
clipper, null-verb, null-target	214	hook, dissect, omentum	3465	scissors, dissect, gallbladder	45
grasper, clean, gallbladder	40	hook, dissect, peritoneum	286	scissors, dissect, omentum	94
grasper, dissect, cystic-plate	78	hook, grasp/retract, cystic-artery	37	scissors, grasp/retract, cystic-artery	7
grasper, dissect, gallbladder	645	hook, grasp/retract, cystic-duct	160	scissors, grasp/retract, cystic-duct	15
grasper, dissect, omentum	32	hook, grasp/retract, cystic-pedicle	33	scissors, grasp/retract, cystic-pedicle	4
grasper, dissect, peritoneum	12	hook, grasp/retract, cystic-plate	51	scissors, grasp/retract, gallbladder	12
grasper, grasp/retract, abdomenal-wall/cavity	18	hook, grasp/retract, gallbladder	515	scissors, grasp/retract, tissue-sampling	7
grasper, grasp/retract, adhesion	1	hook, grasp/retract, gut	6	scissors, null-verb, null-target	108
grasper, grasp/retract, clip	43	hook, grasp/retract, hepatic-pedicle	4	Total	135456

4.3.1 Naive Approach

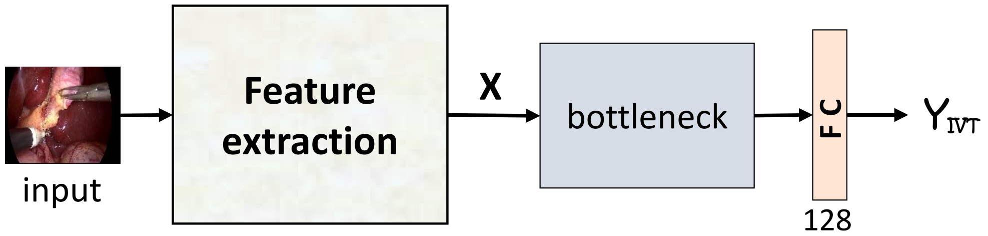


Figure 4.7 – The architecture of a naive CNN model for action triplet recognition.

The study starts with building and training a simple deep learning algorithm to recognize surgical action triplets from a given laparoscopic video V^N containing N sequential image frames. The model learns a function f that can naively map features $\mathbf{X}$ extracted from image data to their corresponding triplet IDs $\mathbf{Y}_{ivt}$, [∀ ivt ∈ IVT] without any consideration of the interacting components that constitute the triplets:

$$f : \mathbf{X} \longrightarrow \mathbf{Y}_{ivt} \tag{4.1}$$

Architecturally, the Naive CNN baseline is composed of a feature extraction layer, a bottleneck layer, and a classification layer as shown in Figure 4.7. The feature extraction backbone is based on ResNet-18. Generally, residual networks are popular for their excellent performance in image recognition at scale. This is largely for their ability to leverage residual skip connections to boycott the issues associated with vanishing gradients while maintaining a very deep layered network for a better approximation of non-linearity functions. We choose a more shallow version of the residual network (ResNet-18), to allow for better quantification of the actual contribution of the proposed method. The bottleneck layer consists of two additional 3x3 convolutional layers with $(256, 64)$ filters which helps to reduce the dimensionality of extracted features from the backbone. The refined features obtained at this layer are context-free and do not consider the interacting components of the triplets. The final layer is a fully connected (FC) layer with N units for the classification of the triplets, where $N = 128$ corresponds to the number of triplet classes. We use this naive CNN model as the first baseline in this study.

The naive CNN model is faced with lots of challenges as described below:

1.) There are too many triplet classes due to a large amount of triplet combinatorial possibilities. Without a special configuration of a highly parameterized function to learn such a huge class size, a model would overfit the most frequently occurring classes.

2.) The correlation between the triplet composition and their assigned IDs is not captured in the naive modeling. This is a huge problem in this type of data where both similar and dissimilar triplets are assigned different IDs without any special consideration. As depicted in the analogical example in Figure 4.8, while some triplets may differ from each other by just one component, others may differ in virtually all the components, and it is not possible to deduct these from the nature of the IDs. Hence, the convergence of a deep learning function that does not understand these label distances becomes highly unlikely, and the model is difficult to train.

We then approach the non-semantic representation of the triplet composition by their label IDs using a multiple task learning of the triplet components in the following section.

4.3.2 Multi-Task Learning Approach

Multiple task learning (MTL) is a deep learning approach for joint and parallel learning of multiple and different but related tasks simultaneously. In MTL modeling, all the tasks share some representations up to a certain level thus allowing each task, while leveraging inductive transfer, to exploit commonalities and differences across tasks and by so doing, improve their efficiency and accuracy. It also serves as a form of regularization based on inductive bias. MTL network strategy has been exploited in [Jin 2020, Mondal 2019, Twinanda 2016a] for the parallel modeling of surgical instruments and phase recognition in surgical videos. These works have

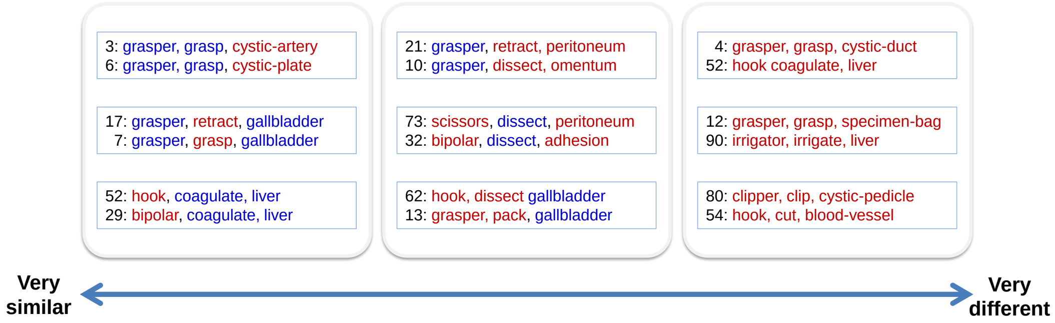

3: grasper, grasp, cystic-artery
6: grasper, grasp, cystic-plate
17: grasper, retract, gallbladder
7: grasper, grasp, gallbladder
52: hook, coagulate, liver
29: bipolar, coagulate, liver
21: grasper, retract, peritoneum
10: grasper, dissect, omentum
73: scissors, dissect, peritoneum
32: bipolar, dissect, adhesion
62: hook, dissect gallbladder
13: grasper, pack, gallbladder
4: grasper, grasp, cystic-duct
52: hook coagulate, liver
12: grasper, grasp, specimen-bag
90: irrigator, irrigate, liver
80: clipper, clip, cystic-pedicle
54: hook, cut, blood-vessel
Very similar
Very different

Figure 4.8 – Axis of action triplet similarity from very similar (right) to very different (left), showing that the triplet IDs does not portray the similarity in triplet composition.

shown that with MTL, correlated tasks can share deep learning layers and features to improve performance. Following this observation, we build a MTL network with three branches for the instrument (I), verb (V), and target (T) recognition sub-tasks. And so, instead of learning a simple function that directly, but naively, maps the image features $\mathbf{X}$ to the associated triplet labels $\mathbf{Y}_{IVT}$ as illustrated in Equation 4.1, we decompose this to multiple functions to learn the triplet components in parallel as follows:

$$\begin{aligned} f_I &: \mathbf{X} \longrightarrow \mathbf{Y}_I, \\ f_V &: \mathbf{X} \longrightarrow \mathbf{Y}_V, \\ f_T &: \mathbf{X} \longrightarrow \mathbf{Y}_T, \end{aligned} \tag{4.2}$$

and afterwards, learn an association function that maps the outputs of the multi-tasks functions to the triplet labels describing their interactions:

$$f : (\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T) \longrightarrow \mathbf{Y}_{IVT} \tag{4.3}$$

In implementation, the MTL uses the same feature extraction backbone as in the naive baseline model for fair comparison. This is followed by an MTL layer as shown in Figure 4.9. Each of the three branches of the MTL layer is modeled using two layers of convolution followed by a fully connected (FC)-layer for the individual task classification. All the three branches share the same feature extraction backbone.

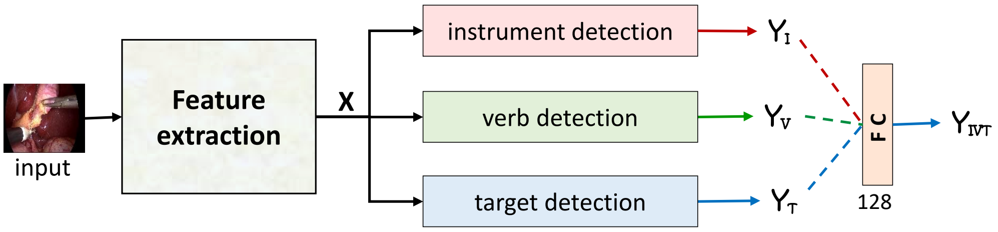


Figure 4.9 – The architecture of a MTL baseline model for action triplet recognition.

Having obtained three outputs logits from the MTL branches representing the components of the triplet, the next task is to associate them to form the final triplets. For this MTL network, we concatenate the three feature vectors, and using an FC-layer, for high-level reasoning, we learn to classify the action triplets. This ensures that the triplets are learned on the prior features that consider the constituting components of the triplets. We use this MTL method as our second baseline in this study. While the MTL model improves the naive modeling with triplet components consideration, it falls short in two respects:

1.) The MTL branches fail to model the instrument-centric property of surgical action triplet. Following prior knowledge, while there may be many visible anatomies, only the ones interacting with an instrument are labeled targets. To learn the correct target, a model would have to leverage the instrument appearance cue. In HOI as shown in Figure 4.10, the search space for the interacting objects is constrained to the predicted locations of humans in an image obtained using a region proposed model as done in [Gkioxari 2018, Xu 2019, Qi 2018, Shen 2018]. This allows their model to ignore other objects that have no form of interaction with humans. However, this method needs to be trained on data in which the human and object bounding boxes have been manually annotated.

2.) The triplet structure is lost in the classification layer. This is because the dense connectivity in an FC-layer does not preserve the structuring of the concatenated triplet components in the meaningful format of ⟨I,V,T⟩.

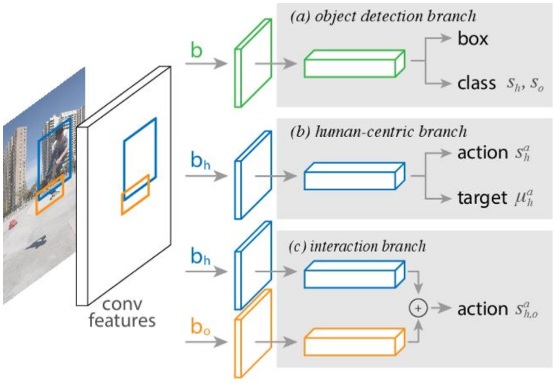


Figure 4.10 – The architecture of the action triplet detection model presented in [Gkioxari 2018]. In this method, a precise target object localization is constrained on person's appearance.

We tackle these two observed problems in the triplet modeling using a CAG module integrated within the MTL framework and a 3Dis module leading to the proposed model called *Tripnet* as shown in Fig. 4.11. The architecture of Tripnet is conceptually divided into three: the **base** - for feature extraction, the **neck** - for the components detection and the **head** for the triplet association. We retain the same ResNet-18 for feature extraction as in the baseline models for a fair comparison. The neck consists of two modules: the instrument subnet and verb-target subnet which we regrouped to form a new module: CAG. In it, lies the first novelty of this work.

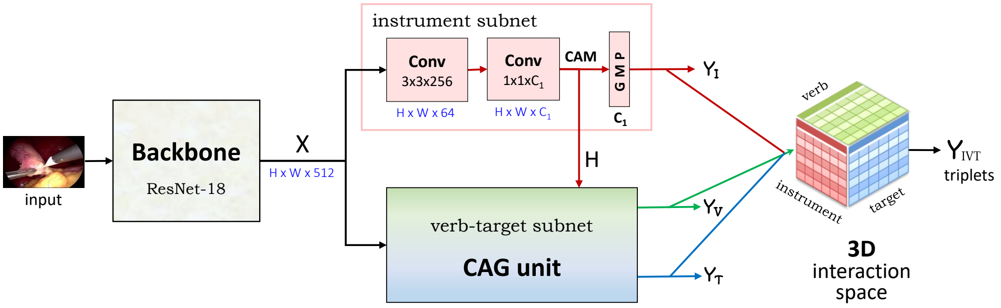


Figure 4.11 – Tripnet: the architecture of the proposed model for action triplet recognition. *Feature dimension values* ($H = 32,\ W = 56,\ C_I = 6$)

4.3.3 Class Activation Guide (CAG) for Triplet Component Detection

The CAG is a special kind of MTL pipeline introduced to moderate the model search space of one MTL's branch on the features from another branch. In this case, the verb and target detection branches are conditioned on the instrument detection branch since the pose of the instruments is indicative of their interactions with the tissues. In the CAG, we modify the MTL function responsible for learning the instrument recognition to additionally captures the spatial position of the instruments. But since we have no spatial annotation to crop the action locations as done in [Gkioxari 2018, Xu 2019, Qi 2018, Shen 2018], our best bait is to employ weak supervision. And so, we introduce another feature space **H** for the weak localization

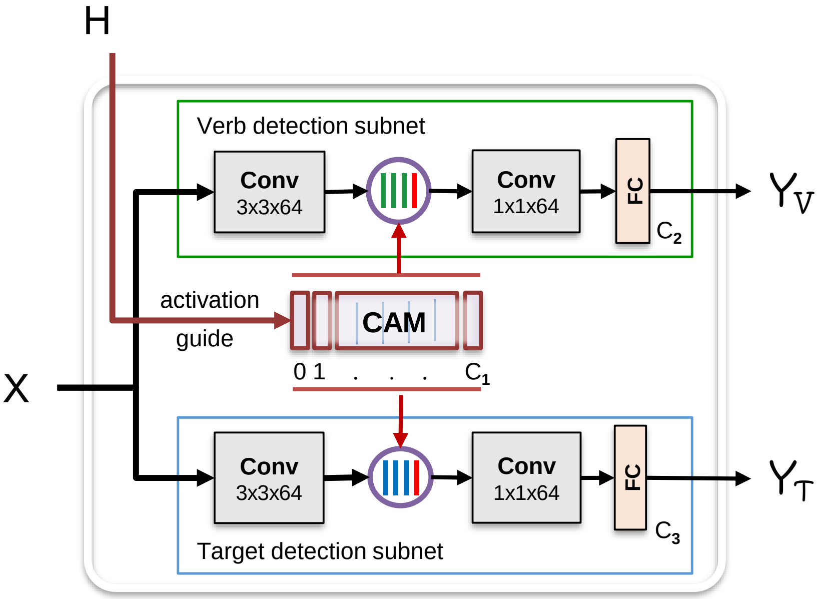


Figure 4.12 – The diagram of the class activation guide (CAG) for a directed detection of the verb and target components of the triplets.

which can be co-learned in the same function without the need for additional labels:

$$f_I : \mathbf{X} \times \mathbf{H} \longrightarrow \mathbf{Y}_I \tag{4.4}$$

We hypothesize that the instrument's Class Activation Map (CAM) **H** from the instrument branch has sufficient information to direct the verb and target detection branches towards the likely regions of interest for the actions. And so going by the instrument-centric property, we can condition the learning space of the remaining MTL functions on the instrument's appearance cue captured in the **H** as follows:

$$\begin{aligned} f_V &: \mathbf{X}|\mathbf{H} \longrightarrow \mathbf{Y}_V, \\ f_T &: \mathbf{X}|\mathbf{H} \longrightarrow \mathbf{Y}_T, \end{aligned} \tag{4.5}$$

In implementation, we regroup the three branches of the MTL into two branches: the *instrument* subnet and the *verb-target* subnet for convenience as illustrated in Figure 4.11. The instrument subnet which consists of two layers of convolutions is now terminated with a global max-pooling (GMP) layer (instead of an FC layer) to learn the CAM of the instruments for their weak localization, as suggested in [Nwoye 2019]. This helps the model to localize the instruments performing the actions while relying only on binary presence labels for their training. The intuition here is that to a large extent, the weakly localized boundaries of the instruments also represent the regions of the actions and enable weakly supervised action detection.

The verb-target subnet is then transformed to a *class activation guide (CAG)* unit as shown in Figure 4.12. Each of the branches in the CAG module consists of two convolution layers and an FC-layer for the task recognition. Both branches receive the instrument's CAM as additional input. This CAM input is then concatenated with the verb and target intermediary features, concurrently, to guide and condition the model search space of the verb and target on the instrument appearance cue.

The instrument subnet and the CAG respectively provides raw output vectors ($\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$), also called logits, of the instrument (I), verb (V) and target (T) branches. These logits are passed to the model head for the final triplet association.

4.3.4 3D Interaction Space (3Dis) for Triplet Association

Recognizing the correct action triplets involves associating the right ($\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$) components depicting the tool-tissue interaction in the image.

In the existing work [Shen 2018], where the data association problem involves only the *object-verb* pair, the outer product of their logits is used to form a 2D matrix of component interaction at test time as shown in Figure 4.13. In a similar manner, we innovate a *3D interaction space* 3Dis for associating the triplets. This sits on the head of the proposed Tripnet architecture in Figure 4.14. Unlike in [Shen 2018], where the data association is not learned by the trained model, we model a trainable interaction space. Given the m-logits, n-logits and p-logits for the $\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$ respectively, we learn the triplets $\mathbf{Y}_{IVT}$ using a 3D projection

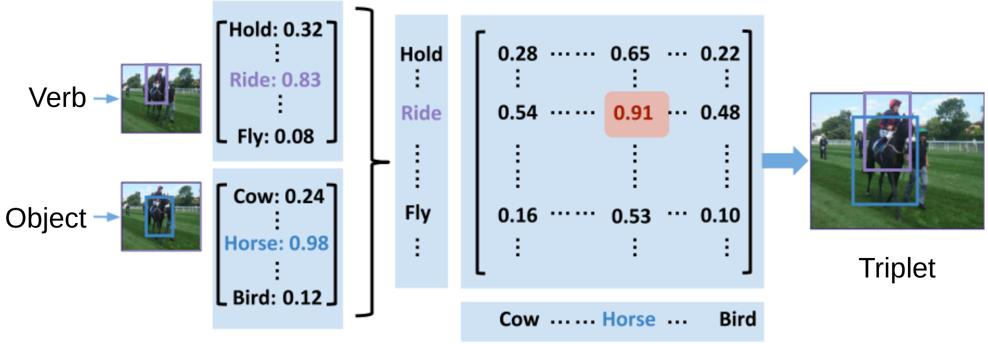


Figure 4.13 – A portion of the architecture of the action triplet detection model presented in [Shen 2018] showing the (verb,object) data association. A verb-object matrix is generated at test time by outer product of the direct prediction probabilities of each of the classes. The full architecture of this method has been presented in Figure 2.8.

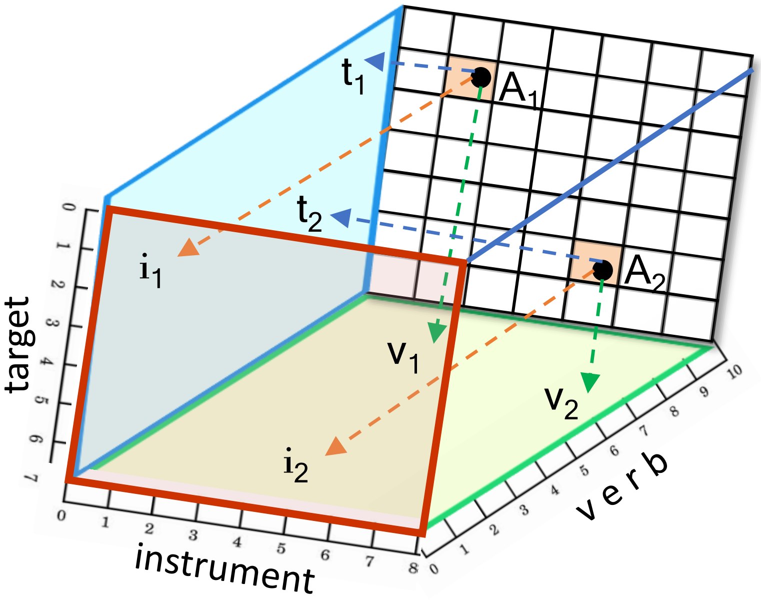


Figure 4.14 – A trainable 3D interaction space (3Dis) for complex tripartite data association of the triplet components.

function Ψ as follows:

$$\mathbf{Y}_{IVT} \longleftarrow \Psi(\alpha\mathbf{Y}_I, \beta\mathbf{Y}_V, \gamma\mathbf{Y}_T), \tag{4.6}$$

where α, β, γ, are the learnable weight vectors for projecting $\mathbf{Y}_I$, $\mathbf{Y}_V$ and $\mathbf{Y}_T$ to the 3D space and Ψ is an outer product operation. This gives an $m \times n \times p$ grid of logits with the three axes representing the three components of the triplets. For all $y_i \in \mathbf{Y}_I, y_v \in \mathbf{Y}_V, y_t \in \mathbf{Y}_T$ the 3D point $y_{(i,v,t)} \in \mathbf{Y}_{IVT}$ represents a possible triplet. A 3D point with a probability above a threshold is considered a valid triplet. The trainable 3Dis module handles the tripartite multi-label data association task. It also allows the recognition of multiple triplets in the same frame.

In practice, there are more 3D points in the space than valid triplets in the CholecT40 dataset. Therefore, we mask out the invalid points, obtained using the training set, at both train and test times. Meanwhile, the actual modeling of the 3Dis will likely support zero-shot learning of unseen triplets if invalid label points are not masked. This is because, instead of learning $\mathbf{Y}_{IVT}^{n=128}$ labels, the 3Dis is modeling $\mathbf{Y}_{IVT}^{N>128}$ where N comprises all possible combinations of the triplet components (including invalid ones), and n comprises of only the triplet classes provided in CholecT40. In other words, 3Dis is re-purposed to detect up to $|V| \times |I| \times |T|$ triplet classes despite requiring training data with lesser triplet classes. However, at this stage of research, it is important to concentrate on learning the available label set, and hence the 3Dis masking.

4.4 Experimental Setup

In this section, we present the details of our experiments. This includes the input data pipeline, the model training, and the evaluation protocol followed in this research.

4.4.1 Data Setup and Pipeline

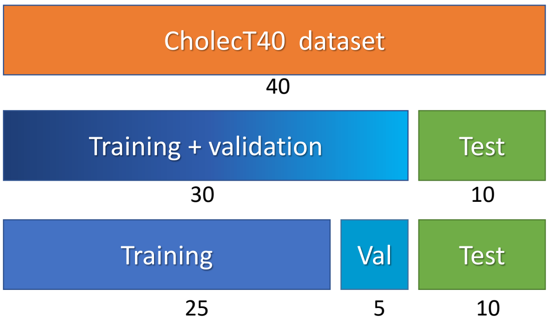


Figure 4.15 – Dataset splits.

We perform our experiments on the newly introduced CholecT40 dataset. The images extracted from the videos are resized to $256 \times 448 \times 3$ for model training and inference. The models are trained on 25 videos. The model hyperparameters are tuned on 5 videos of the validation set. Inference for their test performance evaluation is conducted on 10 videos following the split schedule shown in Figure 4.15. During training, we employ three types of data augmentation techniques which are slight rotation, horizontal flipping, and patch masking [Singh 2017]. There was no image preprocessing during the training and validation as we want to model to autonomously determine the relevant features from the video data which already contains lots of variability in the images and artifacts that may be wrongly removed during image preprocessing. For a high-performance data loading pipeline, our training data are stored as serialized TFRecords binaries.

4.4.2 Training and Loss Function

We leverage transfer learning from the ImageNet dataset [Deng 2009] to train our models. All the individual tasks are trained for multi-label classification using the weighted sigmoid cross-entropy with logits as loss function, regularized by an L_2 norm with $1e^{-5}$ weight decay. The class weights are calculated as in [Nwoye 2019]. All the models are trained using SGD with Momentum as optimizer (initial momentum $\mu = 0.9$). All the experimented models are trained using exponentially decaying learning rates with initial values of $1e^{-3}, 1e^{-4}, 1e^{-5}$ for the component detection modules, the pretrained feature extraction backbone, and 3D interaction space, respectively. The learning rates and other hyperparameters are tuned from the validation set using the grid search method. The network is trained with a batch size of 32 with flexible size for the last set of frames in the video. The model networks are implemented in TensorFlow and trained for 6 days for 200 epochs on GeForce GTX 1080 Ti GPUs. The number of training parameters for the MTL baseline and Tripnet models is 14.94M and 14.95M respectively.

4.4.3 Inference and Evaluation Protocol

Predicted outputs are probability scores that can be threshed ($\Theta = 0.5$) to indicate class presence or absence. To evaluate the capacity of a model at recognizing correctly a triplet and its components, we use two types of metrics:

a. **Component average precision:** This measures the AP of detecting the correct components of the triplet, computed as the area under the precision-recall curve per class. Using this, we measure the AP for instrument (AP_I), verb (AP_V), and target (AP_T) detections. With the task being instrument-centric, we also show the class-wise performance of the instrument's presence detection. To use these metrics for the naive models or for any model that predict only the triplet labels $\mathbf{Y}_{IVT}$, we decompose their predictions into the constituting components ($\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$) following Equation 4.7:

$$\begin{aligned}
\mathbf{Y}_I &= [\ max(\mathbf{Y}_{IVT}|I = i) \quad \forall i \in \{0,1,..,C_1\}\ ],\\
\mathbf{Y}_V &= [\ max(\mathbf{Y}_{IVT}|V = v) \quad \forall v \in \{0,1,..,C_2\}\ ],\\
\mathbf{Y}_T &= [\ max(\mathbf{Y}_{IVT}|T = t) \quad \forall t \in \{0,1,..,C_3\}\ ],
\end{aligned} \tag{4.7}$$

where C_1, C_2 and C_3 are the class sizes for the instrument, verb, and target components respectively. This directly translates to obtaining the probability of a given component class as the maximum probability value among all triplet labels having the same component class label in a given frame. For instance, the predicted probability of a *grasper* instrument in a frame is the maximum probability of all triplet labels having *grasper* as their instrument component label. The ground truth for these components is also derived in the same manner.

b. **Triplet average precision:** This measures the AP of recognizing the tool-tissue interactions by observing elements of the triplet in conjunction, i.e.: looking at different sets of triplet components. Thus, we measure the APs for the instrument-verb (AP_{IV}), instrument-target (AP_{IT}), and instrument-verb-target (AP_{IVT}). During the AP computation, a prediction is registered as correct if all of the components of interest are correctly identified (e.g. instrument and verb for AP_{IV}). Meanwhile, the AP_{IVT} evaluates the recognition of the complete triplets making it the main metric in this study.

A test case is a single procedure on one patient represented by a full laparoscopic video. Thus, all the AP scores are video-specific computed as follows:

a. Per-category AP is computed across all frames in a given video.

b. Category AP is obtained by averaging per-category APs across all videos.

c. Mean Average Precision (mAP) is obtained by averaging category AP, serving as the main metric.

4.5 Experimental Results

The approaches are evaluated on the CholecT40 test set. We present the experimental results and their discussion as follows:

4.5.1 Ablation Studies

Table 4.5 – Ablation study for the CAG unit and 3D interaction space.

Study				Performance			
FC	3Dis (untrained)	3Dis (trained)	CAG	AP_I	AP_{IV}	AP_{IT}	AP_{IVT}
✓				74.6	14.02	7.15	6.43
	✓			89.3	14.28	6.99	6.03
	✓		✓	**89.7**	16.72	7.62	6.32
		✓		89.5	20.63	12.08	12.06
		✓	✓	**89.7**	**35.45**	**19.94**	**18.95**

We conducted an ablation study to quantify the contribution of the two novel modules in our proposed Tripnet method. The results as tabulated in Table 4.5 shows that the positive contributions of the two modules. At each introduction of the CAG module, we witness a minimum of 1 in the verb detection and 1 in the correct target detection, justifying the need for using instrument cues in the verb and target recognition. This unit is also indirectly affecting the correct recognition of the (instrument-verb) and (instrument-target) pairs as these combinations rely on the correct detections of the individual elements.

We also observe that learning the instrument-tissue interactions is better with a trainable 3D projection than with either the untrained 3D space or with an FC-layer. The preservation of the triplet ordering in the 3Dis helped the model to better understand the relationship between the components. Making this unit trainable helps the model to learn some weights needed to navigate the complex selection and tripartite matching problem posed by the nature of the triplet formulation. This results in a large 6.0% improvement of the AP_{IVT}. The two units complement each other and improve the results across all metrics. Thus, we record the best performance in all the metrics by combining the CAG unit and the trained 3D interaction space.

4.5.2 Quantitative Results

4.5.2.1 Component Detection Precision

Table 4.6 – Model performance on triplet components detection.

Method	Instrument						Mean (AP_I)	Verb (AP_V)	Target (AP_T)
	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator			
Naive CNN	75.3	04.3	64.6	02.1	05.5	06.0	27.5	–	–
MTL Baseline	96.1	**91.9**	**97.2**	55.7	30.3	76.8	74.6	43.5	22.3
Tripnet	**96.3**	91.6	**97.2**	**79.9**	**90.5**	**77.9**	**89.7**	**53.6**	**24.8**

Table 4.6 presents the recognition AP of the components across all triplets for all the experimental models. The results show that the naive model does not understand these individual triplet components. This comes from the fact that it is designed to learn the triplets using their IDs. Intuitively, this setup becomes a problem where two different triplets sharing the same instrument or verb still have different IDs. The confusion posed by this can complicate the recognition of the final triplet as can be justified from the results. On the other hand, the MTL and Tripnet networks overcome this issue by inculcating the explicit modeling of the triplet components in their recognition pipeline. Both models show competing performance on instrument detection. Moreover, Tripnet outperforms the MTL baseline by 15.1% mean AP. This can be attributed to its use of the CAG module and 3D interaction space to learn better semantic information about the instrument behaviors. The Tripnet improves over the MTL baseline by leveraging the instrument's weakly supervised heat maps. The Tripnet performance shows that the CAG unit helps the model to better detect the verbs and targets of interests, and outperform the MTL baseline on the duo by 10.1% and 2.5% respectively. Compare to the conventional use MTL, we observed that the CAG helps the verb detection with a higher margin than it did for the target detection. The likely rationale for this is that without an instrument's cue, it is almost impossible to infer the performed verb, whereas the correct target detection may also be affected by their visibility and deformation.

4.5.2.2 Triplet Association Precision

Table 4.7 – Action triplet recognition performance for instrument-verb (AP_{IV}), instrument-target (AP_{IT}) and instrument-verb-target (AP_{IVT}) association.

Method	AP_{IV}	AP_{IT}	AP_{IVT}	**Mean**
Naive CNN	7.54	6.89	5.88	6.77
MTL Baseline	14.02	7.15	6.43	9.20
Tripnet	**35.45**	**19.94**	**18.95**	**24.78**

Table 4.7 presents the triplet recognition performance at various levels of component's association. The naive CNN model has again the worst performance for the AP_{IV}, AP_{IT} and AP_{IVT} metrics, as expected from the previous results. This justifies that recognizing the

components of the triplets from a surgical scene is essential for the understanding of their interactions. The MTL baseline model, on the other hand, performs only slightly above the naive model despite its high instrument detection performance in Table 4.7. This is because the MTL baseline model, after learning the components of the triplets, dilutes this semantic information by concatenating and feeding the output to an FC-layer. This is not the same with the proposed Tripnet which models a structured triplet component's relationship in its 3Dis unit to learn better triplet association. With this, the proposed model improves over the MTL baseline by increasing the AP_{IVT} by 12.5% on average. The results show that Tripnet understands the articulation of the instruments better with a margin of 21.43% over the baseline. In the midst of all the visible anatomies, the Triplet is able to associate assign the instrument to the right target better than the closest baseline with a margin of 12.79%. Overall, the Tripnet outperformed all the baselines in instrument-tissue interaction recognition by a minimum of 15.6%.

In general, it can be observed that it is easier to learn the instrument-verb components than the instrument-target components. This is likely due to the fact that:

a.) a verb has a more direct association to the instrument creating the action.

b.) the dataset contains many more target classes than verb classes.

c.) many anatomical structures in the abdomen are usually discriminated with difficulty by non-medical experts.

While the action recognition performance appears to be low, it follows the same pattern as other models in the computer vision literature on action datasets of even lesser complexity. For instance, on the HICO-DET dataset [Chao 2018], [Gkioxari 2018] achieves 10.8%, [Qi 2018] achieves 14.2% and [Xu 2019] achieves 15.1% action recognition AP, also known as AP_{role}. In fact, the current state-of-the-art performance on the HICO-DET dataset is 21.2% as reported on the leaderboard server at the time of this work. Similarly, the winner of the MICCAI 2019 sub-challenge on action recognition, involving only four verb classes, scores 23.3% F1-score [2]. This shows the challenging nature of fine-grained action recognition.

4.5.3 Qualitative Results

To reveal and better appreciate the performance of the proposed model in understanding instrument-tissue interactions, we overlay the predictions on several surgical images in Figure 4.16. The qualitative results show that Tripnet does not only improves the performance of the baseline models but also localizes accurately the regions of interest of the actions. This is another benefit of the weakly supervised learning of the instrument localization. It is also observed that the majority of incorrect predictions are due to one incorrect triplet component. Instruments are usually correctly predicted and localized. This shows how closely the model approximates even the incorrect triplet predictions. The prior understanding of the underlying

[2] https://www.dropbox.com/s/n4fdbsc4zhdyug0/Presentation_EndoVis_SurgicalWorkflowandSkill2019.pdf?dl=0

components of the triplets in the MTL methodology introduces this bonus. Nonetheless, it is not straightforward to predict the verb/target directly from the instrument due to the multiple possible associations as can be seen from the complete statistics provided in the Tables 4.2 and 4.3.

4.6 Conclusion

In this work, we tackle the task of recognizing action triplets directly from surgical videos. Our overarching goal is to detect the instruments and learn their interactions with the tissues during laparoscopic procedures. To this aim, we present a new dataset, which consists of 135k action triplets over 40 videos. We study the characteristics of triplets and how the instrument determines the annotation of the other components of the triplet. We also explain the semantic overlap and multiplicity in the action triplet composition.

For recognition, we propose a novel model that relies on instrument class activation maps to learn the verbs and targets. We also introduce a trainable 3D interaction space for learning the ⟨*instrument, verb, target*⟩ associations within the triplets. Experiments show that our model outperforms the baselines by a substantial margin in all the metrics, thereby demonstrating the effectiveness of the proposed approach. While triplet formulation captures surgical activities at a more useful fine-grained level of granularity, their recognition is still challenging.

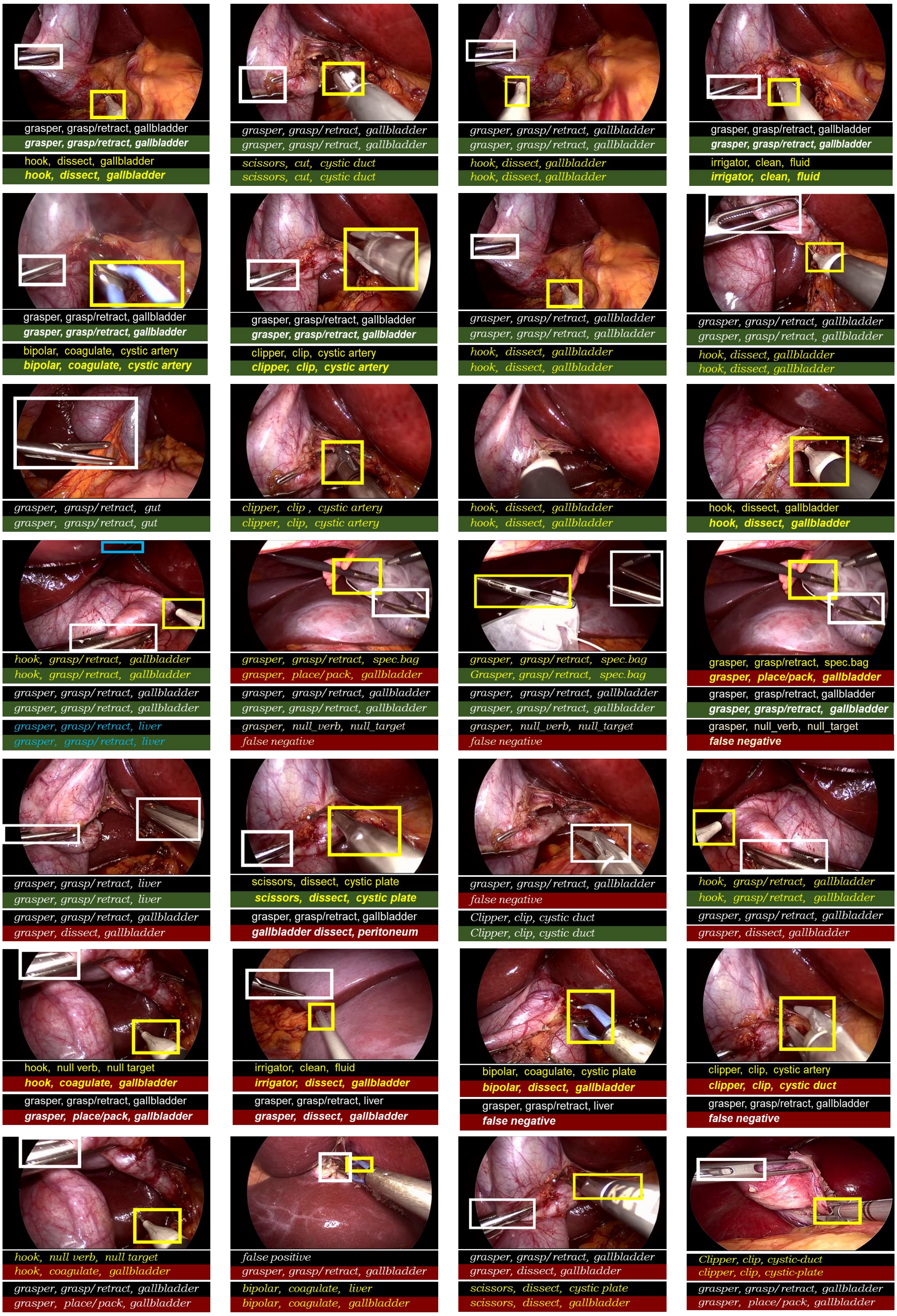

grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
scissors, cut, cystic duct
scissors, cut, cystic duct
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
irrigator, clean, fluid
irrigator, clean, fluid
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
bipolar, coagulate, cystic artery
bipolar, coagulate, cystic artery
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
clipper, clip, cystic artery
clipper, clip, cystic artery
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
grasper, grasp/retract, gut
grasper, grasp/retract, gut
clipper, clip , cystic artery
clipper, clip, cystic artery
hook, dissect, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
hook, grasp/retract, gallbladder
hook, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, liver
grasper, grasp/retract, liver
grasper, grasp/retract, spec.bag
grasper, place/pack, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, null_verb, null_target
false negative
grasper, grasp/retract, spec.bag
Grasper, grasp/retract, spec.bag
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, null_verb, null_target
false negative
grasper, grasp/retract, spec.bag
grasper, place/pack, gallbladder
grasper, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, null_verb, null_target
false negative
grasper, grasp/retract, liver
grasper, grasp/retract, liver
grasper, grasp/retract, gallbladder
grasper, dissect, gallbladder
scissors, dissect, cystic plate
scissors, dissect, cystic plate
grasper, grasp/retract, gallbladder
gallbladder dissect, peritoneum
grasper, grasp/retract, gallbladder
false negative
Clipper, clip, cystic duct
Clipper, clip, cystic duct
hook, grasp/retract, gallbladder
hook, grasp/retract, gallbladder
grasper, grasp/retract, gallbladder
grasper, dissect, gallbladder
hook, null verb, null target
hook, coagulate, gallbladder
grasper, grasp/retract, gallbladder
grasper, place/pack, gallbladder
irrigator, clean, fluid
irrigator, dissect, gallbladder
grasper, grasp/retract, liver
grasper, dissect, gallbladder
bipolar, coagulate, cystic plate
bipolar, dissect, gallbladder
grasper, grasp/retract, liver
false negative
clipper, clip, cystic artery
clipper, clip, cystic duct
grasper, grasp/retract, gallbladder
false negative
hook, null verb, null target
hook, coagulate, gallbladder
grasper, grasp/retract, gallbladder
grasper, place/pack, gallbladder
false positive
grasper, grasp/retract, gallbladder
bipolar, coagulate, liver
bipolar, coagulate, gallbladder
grasper, grasp/retract, gallbladder
grasper, dissect, gallbladder
scissors, dissect, cystic plate
scissors, dissect, gallbladder
Clipper, clip, cystic-duct
clipper, clip, cystic-plate
grasper, grasp/retract, gallbladder
grasper, place/pack, gallbladder

Figure 4.16 – Qualitative results: triplet prediction and weak localization of the regions of action (*best seen in color*). Predicted and ground-truth triplets are displayed below each image: black = ground-truth, green = correct prediction, red = incorrect prediction. A missed triplet is marked as false negative and a false detection is marked as false positive. The color of the text corresponds to the color of the associated bounding box.

5 Attention Mechanisms for Enhanced Component Detection

It's attention to detail that makes the difference between average and stunning
– Francis Atterbury

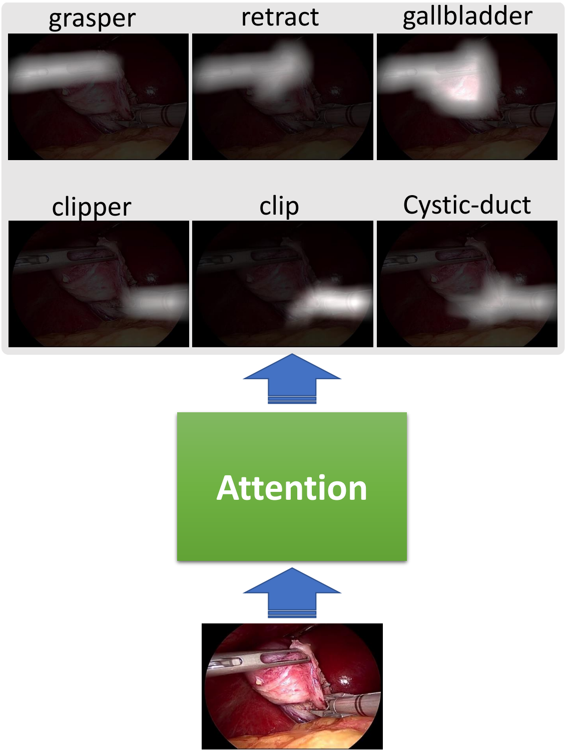


Figure 5.1 – The Concept of Attention Mechanism for highlighting focal feature representation.

Chapter Summary

In this chapter, we extend the research on surgical action triplet recognition presented in the previous chapter. Our extension is both in data (Section 5.2) and in method (Section 5.3) as presented further.

5.1 Objectives

While the initial results on action triplet recognition are encouraging, the experiments reveal potential areas of improvement. Specifically, we observe that the correct detection of the individual components positively influences the final triplet recognition mAP. Thus, this chapter focuses on improving the triplet components detection, more specifically the verbs and the targets.

In our preliminary experiments, we explore several attention mechanisms and decisively, design a novel spatial attention mechanism that is guided by the instrument activations to improve the detection performance of the verb and target components of the triplet. We observe that, correspondingly, this also improves the overall triplet recognition performance.

Since deep learning models are data-hungry especially for complex tasks like triplet recognition, we are motivated to increase the quantity of the training data for better generalization of the model. Beyond adding more data samples and annotations to the foremost triplet dataset [Nwoye 2020], CholecT40, we also improve the standard of the annotations via expert knowledge aggregation and label mediation.

In the following sections, we discuss the updated triplet dataset and the proposed attention methods for the recognition of surgical action triplets in laparoscopic videos, and an insightful discussion on the experimental results.

5.2 Dataset Generation

We present an improved action triplet dataset known as *CholecT50* [Nwoye 2021], which is an extension of CholecT40 dataset [Nwoye 2020] with additional 10 videos and standardized classes. The 50 video dataset is a collection of 45 videos from the famous Cholec80 dataset [Twinanda 2016b] and 5 videos from an in-house dataset of the same surgical procedure.

Different from CholecT40, two surgeons annotated the dataset, capturing a wider spread of surgical expertise in the annotation. The first surgeon annotated 40 videos and the second, 10 videos. The dataset follows the same annotation process as in CholecT40 as discussed in Section 4.2.

Just like in CholecT40, there is a large number of observed instruments, verbs, and targets, the theoretical number of all possible triplet configurations (900) is prohibitively high. Even limiting those configurations to the approximately 300 observed in the dataset has little clinical relevance due to the presence of several overlapping surgical semantics (i.e., two slightly different triplets denoting the same surgical action). More still, defining and selecting classes for the triplet dataset, annotated by different surgeons, is more challenging. However, we leverage the variability and diversity in the annotations to standardize the annotations and sub-sample the labels to a reasonable number of classes with maximum clinical utility. To achieve this, a team of clinical experts selected the top relevant labels for the triplet dataset which are determined by a two-step process. First, class grouping ($\cup$) is carried out to super-class triplets that are semantically the same. Some examples of overlapping triplets grouped include:

a. $\langle$*grasper, grasp, gallbladder-fundus*$\rangle$ $\cup$ $\langle$*grasper, grasp, gallbladder-neck*$\rangle$ $\cup$ $\langle$*grasper, grasp, gallbladder*$\rangle$ $\cup$ $\langle$*grasper, grasp, gallbladder-body*$\rangle$ $\longrightarrow$ $\langle$*grasper, grasp, gallbladder*$\rangle$

b. $\langle$*irrigator, aspirate, bile*$\rangle$ $\cup$ $\langle$*irrigator, aspirate, fluid*$\rangle$ $\cup$ $\langle$*irrigator, aspirate, blood*$\rangle$ $\longrightarrow$ $\langle$*irrigator, aspirate, fluid*$\rangle$

c. $\langle$*grasper, pack, gallbladder*$\rangle$ $\cup$ $\langle$*grasper, store, gallbladder*$\rangle$ $\longrightarrow$ $\langle$*grasper, pack, gallbladder*$\rangle$

d. $\langle$*grasper, retract, gut*$\rangle$ $\cup$ $\langle$*grasper, retract, duodenum*$\rangle$ $\cup$ $\langle$*grasper, retract, colon*$\rangle$ $\longrightarrow$ $\langle$*grasper, retract, gut*$\rangle$

e. $\langle$*bipolar-grasper, coagulate, liver*$\rangle$ $\cup$ $\langle$*bipolar, coagulate, liver-bed*$\rangle$ $\cup$ $\langle$*bipolar, coagulate, liver*$\rangle$ $\longrightarrow$ $\langle$*bipolar, coagulate, liver*$\rangle$

In addition to class grouping, surgical relevance ratings of the labels are carried out by three clinicians. For the rating, the clinicians assigned a score from a range of [1-5] to each triplet composition based on their possibility and usefulness in the considered procedure. Their average scores, as well as the triplet's number of occurrence, is used in ordering the triplet classes, after which the top relevant classes are selected. Where there is ambiguity or label disagreement, label mediation is done by the third clinician.

Table 5.1 – Statistics of the triplet's component labels in the dataset.

instrument		Verb		Target	
Label	Count	Label	Count	Label	Count
bipolar	6697	aspirate	3122	abd-wall/cavity	847
clipper	3379	clip	3070	adhesion	228
grasper	90969	coagulate	5202	blood-vessel	416
hook	52820	cut	1897	cystic-artery	5035
irrigator	5005	dissect	49247	cystic-duct	11883
scissors	2135	grasp	15931	cystic-pedicle	299
		irrigate	572	cystic-plate	4920
		null-verb	10841	fluid	3122
		pack	328	gallbladder	87808
		retract	70795	gut	719
				liver	17521
				null-target	10841
				omentum	9220
				peritoneum	1227
				specimen-bag	6919

The resulting dataset comprises 100 triplet classes that follow the format of ⟨*instrument, verb, target*⟩. The triplets are composed of 6 instruments, 10 verbs, and 15 target classes. We provide instance counts for these selected triplet components in Table 5.1. We present the CholecT50 dataset triplet labels including their number of occurrences in Table 5.2. We also present the co-occurrence distribution of ⟨*instrument, target*⟩ and ⟨*instrument, verb*⟩ pairs in Tables 5.3 and 5.4 respectively.

For our experiment, we down-sampled the videos to 1fps yielding 100.86K frames annotated with 161K action triplet instances. On average, a video contains 2.08K frames. The video dataset is split into training, validation, and testing sets as presented in Figure 5.2. The videos in the dataset splits are distributed in the same ratio to include annotations from each surgeon.

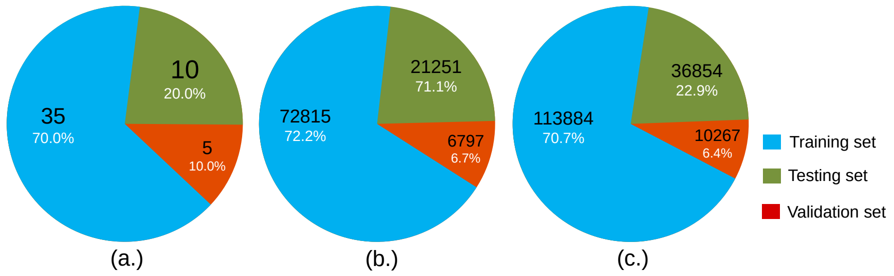


Figure 5.2 – Statistics of the dataset split showing: (a) number of videos, (b.) number of frames, and (c.) number of instance labels.

Table 5.2 – Dataset statistics showing the number of occurrences of the triplets.

Name	Count	Name	Count	Name	Count
bipolar, coagulate, abdominal-wall/cavity	434	grasper, grasp, cystic-artery	76	hook, dissect, gallbladder	29292
bipolar, coagulate, blood-vessel	251	grasper, grasp, cystic-duct	560	hook, dissect, omentum	3649
bipolar, coagulate, cystic-artery	68	grasper, grasp, cystic-pedicle	26	hook, dissect, peritoneum	337
bipolar, coagulate, cystic-duct	56	grasper, grasp, cystic-plate	163	hook, null-verb, null-target	4397
bipolar, coagulate, cystic-pedicle	77	grasper, grasp, gallbladder	7381	hook, retract, gallbladder	479
bipolar, coagulate, cystic-plate	410	grasper, grasp, gut	33	hook, retract, liver	179
bipolar, coagulate, gallbladder	343	grasper, grasp, liver	83	irrigator, aspirate, fluid	3122
bipolar, coagulate, liver	2595	grasper, grasp, omentum	207	irrigator, dissect, cystic-duct	41
bipolar, coagulate, omentum	262	grasper, grasp, peritoneum	380	irrigator, dissect, cystic-pedicle	89
bipolar, coagulate, peritoneum	73	grasper, grasp, specimen-bag	6834	irrigator, dissect, cystic-plate	10
bipolar, dissect, adhesion	73	grasper, null-verb, null-target	4759	irrigator, dissect, gallbladder	29
bipolar, dissect, cystic-artery	187	grasper, pack, gallbladder	328	irrigator, dissect, omentum	100
bipolar, dissect, cystic-duct	183	grasper, retract, cystic-duct	469	irrigator, irrigate, abdominal-wall/cavity	413
bipolar, dissect, cystic-plate	54	grasper, retract, cystic-pedicle	41	irrigator, irrigate, cystic-pedicle	29
bipolar, dissect, gallbladder	353	grasper, retract, cystic-plate	1205	irrigator, irrigate, liver	130
bipolar, dissect, omentum	176	grasper, retract, gallbladder	48628	irrigator, null-verb, null-target	573
bipolar, grasp, cystic-plate	8	grasper, retract, gut	686	irrigator, retract, gallbladder	30
bipolar, grasp, liver	95	grasper, retract, liver	13646	irrigator, retract, liver	350
bipolar, grasp, specimen-bag	85	grasper, retract, omentum	4422	irrigator, retract, omentum	89
bipolar, null-verb, null-target	632	grasper, retract, peritoneum	289	scissors, coagulate, omentum	17
bipolar, retract, cystic-duct	8	hook, coagulate, blood-vessel	57	scissors, cut, adhesion	155
bipolar, retract, cystic-pedicle	9	hook, coagulate, cystic-artery	10	scissors, cut, blood-vessel	21
bipolar, retract, gallbladder	32	hook, coagulate, cystic-duct	41	scissors, cut, cystic-artery	613
bipolar, retract, liver	164	hook, coagulate, cystic-pedicle	15	scissors, cut, cystic-duct	808
bipolar, retract, omentum	69	hook, coagulate, cystic-plate	9	scissors, cut, cystic-plate	20
clipper, clip, blood-vessel	51	hook, coagulate, gallbladder	217	scissors, cut, liver	90
clipper, clip, cystic-artery	1097	hook, coagulate, liver	189	scissors, cut, omentum	27
clipper, clip, cystic-duct	1856	hook, coagulate, omentum	78	scissors, cut, peritoneum	56
clipper, clip, cystic-pedicle	13	hook, cut, blood-vessel	15	scissors, dissect, cystic-plate	12
clipper, clip, cystic-plate	53	hook, cut, peritoneum	92	scissors, dissect, gallbladder	52
clipper, null-verb, null-target	309	hook, dissect, blood-vessel	21	scissors, dissect, omentum	93
grasper, dissect, cystic-plate	78	hook, dissect, cystic-artery	2984	scissors, null-verb, null-target	171
grasper, dissect, gallbladder	644	hook, dissect, cystic-duct	7861		
grasper, dissect, omentum	31	hook, dissect, cystic-plate	2898	Total	161005

Table 5.3 – Dataset statistics showing the instrument-verb co-occurrence frequency.

Instrument	Verb									
	aspirate	clip	coagulate	cut	dissect	grasp	irrigate	null	pack	retract
bipolar	-	-	4569	-	1026	188	-	632	-	282
clipper	-	3070	-	-	-	-	-	309	-	-
grasper	-	-	-	-	753	15743	-	4759	328	69386
hook	-	-	616	107	47042	-	-	4397	-	658
irrigator	3122	-	-	-	269	-	572	573	-	469
scissors	-	-	17	1790	157	-	-	171	-	-

5.3 Attention Tripnet for Enhanced Component Detection

In this section, we describe the proposed method. Recall that action triplet recognition requires two steps process: (a) simultaneous solving of three multi-label classification problems. and (b) performing their association while accounting for multiple instances. The proposed method in this chapter is focused on the first point, which is for enhanced component detection. Surgical action triplets are instrument-centric. Detecting the correct verbs and target anatomies is very challenging, because the visibility as well as the subtly involvement of a tool and anatomy in an action have to be taken into consideration. A limited effort is made in our

Table 5.4 – Dataset statistics showing the instrument-target co-occurrence frequency. Target ids 0 ... 14 correspond to *gallbladder, cystic-plate, cystic-duct, cystic-artery, cystic-pedicle, blood-vessel, fluid, abdominal-wallcavity, liver, adhesion, omentum, peritoneum, gut, specimen-bag, and null respectively.*

Instrument	Target														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
bipolar	728	472	247	255	86	251	-	434	2854	73	507	73	-	85	632
clipper	-	53	1856	1097	13	51	-	-	-	-	-	-	-	-	309
grasper	56981	1446	1029	76	67	-	-	-	13729	-	4660	669	719	6834	4759
hook	29988	2907	7902	2994	15	93	-	-	368	-	3727	429	-	-	4397
irrigator	59	10	41	-	118	-	3122	413	480	-	189	-	-	-	573
scissors	52	32	808	613	-	21	-	-	90	155	137	56	-	-	171

previous method, Tripnet [Nwoye 2020], to handle this using a CAG module that conditioned the detection of verbs and targets on the instruments activations, via concatenated features.

Inspired by the findings in [Ulutan 2020] that attention modeling is better than feature concatenation for enforcing spatial configurations, we explore several types of attention mechanisms and propose a new form of spatial attention, known as *Class Activation Guided Attention Mechanism (CAGAM)*. This spatial attention explicitly uses tool types and location features captured in the instrument activation features to highlight the discriminative features for the verb and targets respectively.

Since the inception of the attention mechanism [Bahdanau 2014], many deep learning models have exploited attention in various forms, from self [Vaswani 2017] to cross [Mohla 2020], and from spatial [Fu 2019] to temporal [Sankaran 2016]. Recently, [Ji 2019] showed that attention can be informed from saliency features. While [Ji 2019] uses a combination of spatial and textual attention modules to capture the fine-grained cross-modal correlation between an image and a sentence, another work by [Yao 2020] utilizes image saliency to guide an attention network for weakly-supervised object segmentation. In medical imaging, attention u-net [Oktay 2018] is used to learn the focus on target structures for pancreas segmentation.

Our proposed CAGAM focuses its spatial attention on the relevant features for verbs and targets informed by the instrument's appearance cue. The CAGAM is achieved by redesigning the saliency-guided attention mechanism in [Ji 2019, Yao 2020] to utilize a more adequate and easier to learn class activation map (CAM). While our approach is similar in the attention guiding principle, it contrasts in three regards: (a) our attention network is guided by the instrument's activations which are learnable in the same network, using a global pooling layer without relying on a *third-party* saliency generation network, (b) our attention guide implements a combination of position and channel attention for the target and verb detection tasks respectively, (c) we employ cross-attention from the instrument domain to the other task domains (i.e.: *verb and target*) contrary to self-attention in [Yao 2020].

Meanwhile, the CAGAM is an improvement on the class activation guide (CAG) module introduced in [Nwoye 2020] which is a concatenation of the model's intermediary features with the instrument's activations. We show the improved performance of our previous Tripnet model, only upgraded with the CAGAM. This upgraded model is called *Attention Tripnet*.

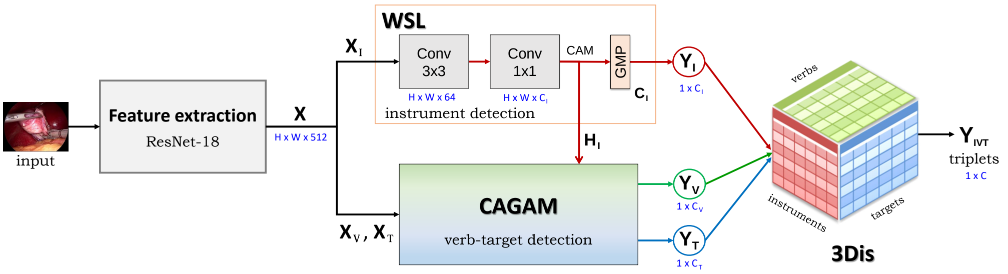


Figure 5.3 – Architecture of the Attention Tripnet showing the base (feature extraction backbone), neck (instrument detection branch and CAGAM module), head (3D interaction space). *Feature dimension values ($H = 32$, $W = 56$, $C_I = 6$, $C_V = 10$, $C_T = 15$, $C = 100$)*

The architecture of the *Attention Tripnet* is shown in Fig. 5.3. The model is conceptually divided into three parts: the **base** is for feature extraction. The **neck** performs triplet component detection - this is where Attention Tripnet's innovation is situated. The **head** here performs triplet association using the 3Dis proposed in [Nwoye 2020].

5.3.1 Feature Extraction

The base, a Resnet-18 model, takes an RGB frame (of $\mathbb{R}^{H\times W\times 3}$ dimension) from a video and extracts convolutional features $\mathbf{X} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 512}$. In the neck, the extracted feature is triplicated into ($\mathbf{X}_I$,$\mathbf{X}_V$,$\mathbf{X}_T$) for multitask learning of the instrument, verb, and target components of the triplets respectively.

5.3.2 Weakly Supervised Instrument Detection

The feature $\mathbf{X}_I$ is fed to the instrument detection branch otherwise known as the (Weakly-Supervised Learning (WSL)) module, where it is refined by a 3×3 convolution layer (Conv) of 64 channels, then followed by a 1×1 Conv of $C_1 = 6$ channels for instrument localization in form of class activation maps (CAM). The outputs of the instruments' CAM, marked as ($\mathbf{H}_I$), are trained for localization via their Global Maximum Pooled (Global Maximum Pooling (GMP)) values $\mathbf{Y}_I$ representing instrument presence class-wise probabilities as done in [Vardazaryan 2018]. This CAM feature ($\mathbf{H}_I$) alongside the remaining extracted features ($\mathbf{X}_V$, $\mathbf{X}_T$) is passed to the CAGAM module for verb and target detections.

5.3.3 Class Activation Guided Attention Mechanism (CAGAM)

According to [Vaswani 2017], an attention function can be described as matching a query ($\mathbf{Q}$) and a set of key-value ($\mathbf{K}$, $\mathbf{V}$) pairs to form an output. The output is computed as a weighted sum of the values ($w\mathbf{V}$), where the weight ($w = \mathbf{Q}\mathbf{K}^T$) is computed by an affinity score function of the query with the corresponding key.

CAGAM is a new form of spatial attention mechanism that propagates attention from known to unknown context features thereby enhancing the unknown context for relevant pattern discovery. It is used, in this case, to discover the verbs and targets that are involved in tool-tissue interactions leveraging instrument CAM features. The CAM serves as the known

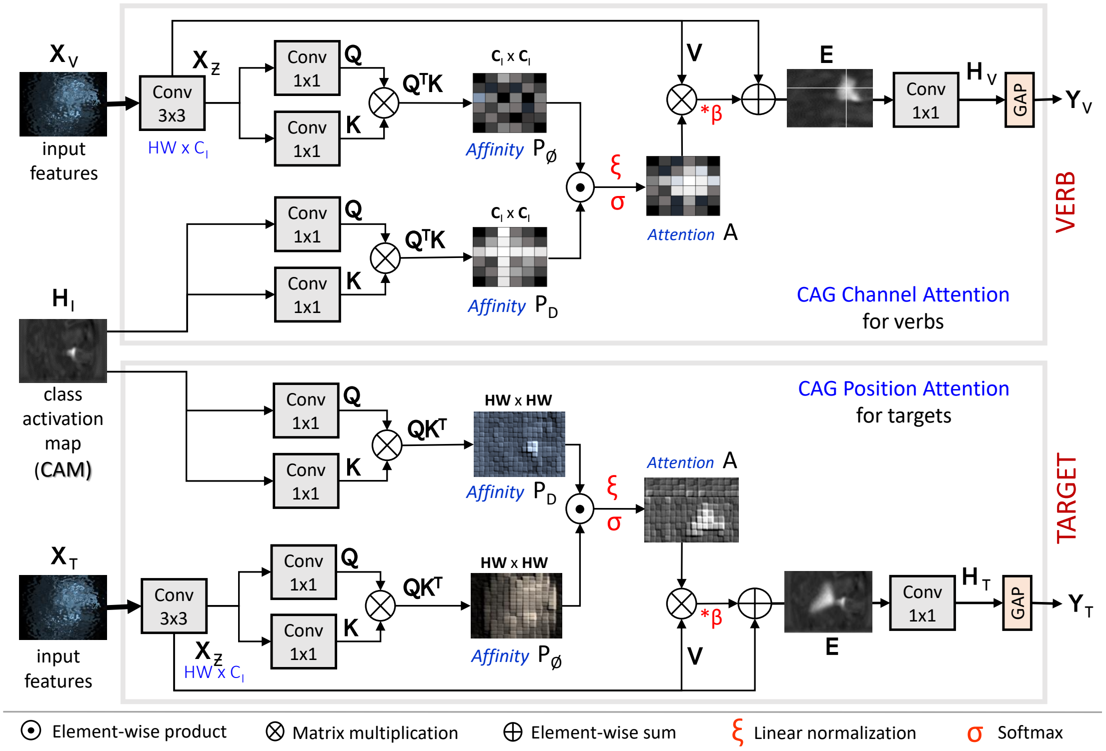


Figure 5.4 – Class Activation Guided Attention Mechanism (CAGAM): uses the attention learned from the instrument's CAM to highlight the verb class (top) and target class (down). *Feature dimension (height $H = 32$, width $W = 56$, depth $D = 64$, instrument's class size $C_I = 6$, verb's class size $C_V = 10$, target's class size $C_T = 15$).*

context features in this regard, since they are already class-wisely discriminated and localized for the instruments. We model the CAGAM to enhance the verb's and target's unfiltered features by element-wise addition of an *enhancement*: this enhancement is a computed spatial attention ($\mathbf{A}$) from the instrument affinity maps ($\mathbf{P}_D$) as well as the component affinity maps ($\mathbf{P}_\varnothing$) themselves. The $\mathbf{P}_D$s are termed discriminative because they originate from the instrument CAM features whereas $\mathbf{P}_\varnothing$s are termed non-discriminative because they are formed from the unfiltered component features.

We observe that verbs and targets behave differently with regards to their instrument; that is, verbs are mostly affected by the instrument's type, while targets tend to be determined by its position. This distinction is a key factor in the choices of attention mechanism in the CAGAM which indeed combines **channel attention** for verb detection (Fig. 5.4: top) and **position attention** for target detection (Fig. 5.4: bottom). This choice is well validated in ablation studies shown further (Table 5.5).

Both types of spatial attention mechanisms are similar, except for the dimensions used, and therefore the nature of the information attended to. The channel attention is captured in the $C_I \times C_I$ channel dimensions; this is possible when the $\mathbf{Q}$ are transposed before being multiplied by $\mathbf{K}$ resulting in affinity maps $(\mathbf{P}_D, \mathbf{P}_\varnothing) \in \mathbb{R}^{C_I \times C_1}$ and a subsequent attention map $\mathbf{A} \in \mathbb{R}^{C_I \times C_1}$ of the desired size $C_I \times C_I$, informed by instrument type. On the other hand,

Algorithm 1: CAG channel attention for verbs.

```
input  : a set $\langle \mathbf{X_V}, \mathbf{H_I} \rangle$ for unfiltered feature and instrument CAM feature
output : a set $\langle \mathbf{H_V}, \mathbf{Y_V} \rangle$ for per-verb activation maps and verb logits
begin
    $\mathbf{P_D} \leftarrow$ AFFINITY($\mathbf{H_I}$) ;                ▷ get discriminative affinity map
    $\mathbf{X}_\mathcal{Z} \leftarrow$ CONVOLUTION($\mathbf{X_V}$) ;   ▷ get context feature from unfiltered feature
    $\mathbf{P}_\emptyset \leftarrow$ AFFINITY($\mathbf{X}_\mathcal{Z}$) ;   ▷ get non-discriminative affinity map
    $\mathbf{A} \leftarrow \sigma\left(\frac{\mathbf{P_D}\mathbf{P}_\emptyset}{\xi}\right)$ ;   ▷ get attention map: $\sigma$ is softmax function, $\xi$ is scale factor
    $\mathbf{V} \leftarrow$ CONVOLUTION($\mathbf{X_V}$) ;               ▷ get Value feature
    $\mathbf{E} \leftarrow \beta(\mathbf{VA}) + \mathbf{X}_\mathcal{Z}$ ;   ▷ get enhanced feature: $\beta$ is learnable temperature
    $\mathbf{H_V} \leftarrow$ Convolution($\mathbf{E}$) ;               ▷ get per-verb activation maps
    $\mathbf{Y_V} \leftarrow$ GAP($\mathbf{H_V}$) ;                      ▷ get verb logits using global average pooling
    return $\langle \mathbf{H_V}, \mathbf{Y_V} \rangle$

function AFFINITY($\mathbf{X}$)                    ▷ compute the affinity map of a given context feature
    $\mathbf{Q} \leftarrow$ CONVOLUTION ($\mathbf{X}$) ;                ▷ get query feature
    $\mathbf{K} \leftarrow$ CONVOLUTION ($\mathbf{X}$) ;                ▷ get query feature
    $\mathbf{P} \leftarrow \mathbf{Q^T K}$ ;                            ▷ compute affinity map
    return $\mathbf{P}$
```

multiplying $\mathbf{Q}$ by $\mathbf{K}^T$ (instead of $\mathbf{Q}^T$ by $\mathbf{K}$) produces affinity maps $(\mathbf{P}_D, \mathbf{P}_\emptyset) \in \mathbb{R}^{HW \times HW}$ and a subsequent position attention map $\mathbf{A} \in \mathbb{R}^{HW \times HW}$ that is informed by instrument position.

Algorithm 2: CAG position attention for target.

```
input  : a set $\langle \mathbf{X_T}, \mathbf{H_I} \rangle$ for unfiltered feature and instrument CAM feature
output : a set $\langle \mathbf{H_T}, \mathbf{Y_T} \rangle$ for per-target activation maps and target logits
begin
    $\mathbf{P_D} \leftarrow$ AFFINITY($\mathbf{H_I}$) ;                ▷ get discriminative affinity map
    $\mathbf{X}_\mathcal{Z} \leftarrow$ CONVOLUTION($\mathbf{X_T}$) ;   ▷ get context feature from unfiltered feature
    $\mathbf{P}_\emptyset \leftarrow$ AFFINITY($\mathbf{X}_\mathcal{Z}$) ;   ▷ get non-discriminative affinity map
    $\mathbf{A} \leftarrow \sigma\left(\frac{\mathbf{P_D}\mathbf{P}_\emptyset}{\xi}\right)$ ;   ▷ get attention map: $\sigma$ is softmax function, $\xi$ is scale factor
    $\mathbf{V} \leftarrow$ CONVOLUTION($\mathbf{X_T}$) ;               ▷ get Value feature
    $\mathbf{E} \leftarrow \beta(\mathbf{VA}) + \mathbf{X}_\mathcal{Z}$ ;   ▷ get enhanced feature: $\beta$ is learnable temperature
    $\mathbf{H_T} \leftarrow$ CONVOLUTION($\mathbf{E}$) ;               ▷ get per-target activation maps
    $\mathbf{Y_T} \leftarrow$ GAP($\mathbf{H_T}$) ;                      ▷ get target logits using global average pooling
    return $\langle \mathbf{H_T}, \mathbf{Y_T} \rangle$

function AFFINITY($\mathbf{X}$)                    ▷ compute the affinity map of a given context feature
    $\mathbf{Q} \leftarrow$ CONVOLUTION ($\mathbf{X}$) ;                ▷ get query feature
    $\mathbf{K} \leftarrow$ CONVOLUTION ($\mathbf{X}$) ;                ▷ get query feature
    $\mathbf{P} \leftarrow \mathbf{Q^T K}$ ;                            ▷ compute affinity map
    return $\mathbf{P}$
```

In both cases, their respective enhanced features (**E**) are obtained when the resulting attention **A** is multiplied to its originating unfiltered context ($\mathbf{X}_{\mathcal{Z}}$) and added to the same context feature. Processing these enhanced features with separate convolutions of the desired number of filters produces per-class activation maps ($\mathbf{H}_V, \mathbf{H}_T$) for each component task, which is, ultimately, transformed to their respective class-wise logits ($\mathbf{Y}_V, \mathbf{Y}_T$) using global average pooling.

Specific implementation details of the CAG channel attention and CAG position attention are presented in Algorithm 1 and 2 respectively.

5.3.4 The 3D Interaction Space (3Dis) for Triplet Association

As in previous chapter, having obtained the instrument $\mathbf{Y}_I$, verb $\mathbf{Y}_V$, and target $\mathbf{Y}_T$ logits from the WSL and CAGAM respectively, the three logits are fed to the 3Dis proposed in [Nwoye 2020] which learns their association using a projection function Ψ as follows:

$$Y_{IVT}^{N} = \Psi(\alpha Y_I, \beta Y_V, \gamma Y_T), \tag{5.1}$$

where N is the number of all possible triplet combinations, α, β, γ, are the respective learnable weight vectors for projecting Y_I, Y_V and Y_T to the 3D space and Ψ is an outer product operation. The 3Dis preserves the triplet structure: an interaction is formed by a feature point from each of the three components. We streamline all possible triplet combinations in the CholecT50 dataset by mapping only the valid triplet points in the 3Dis to an embedding space $Y_{IVT}^{N} \Rightarrow Y_{IVT}^{C}$, containing vectors of probability scores for each of the $C = 100$ valid triplet class.

5.4 Experimental Setup

5.4.1 Data Setup and Pipeline

All our experiments are performed on CholecT50. Due to variability in the video dataset, frame resolution varies from 480 × 854 to 1080 × 1920. We unified their spatial dimensions by resizing to 256 × 448. We also employed random scaling $[0.5, 1.5]$ and brightness/contrast shift ($delta = 0.2$) data augmentation, during training. For a high-performance data loading pipeline, our training data is stored as serialized TFRecords binaries. To obtain specific labels for the component tasks, we design a mapping function, which extracts per-component labels from the triplet labels; those are three vectors of binary presence labels with length $N = [6, 10, 15]$ per frame, where $n \in N$ is the class size for each triplet's component trained as an auxiliary task.

The models are trained on 35 videos, validated on 5 videos, and tested on 10 videos according to the data split in Figure 5.2. For the cross-validation experiment, the 50-video dataset is partitioned into $k = 5$ equal-sized samples. We conducted K repeated experiments with a different j^{th} subsample hold out for evaluation while the rest $K - j^{th}$ is split into Train/Val sets in a 7:1 ratio. The Val set is for hyperparameter tuning. The final estimation is the average performance of the K held-out subsamples from the K experiments.

5.4.2 Training and Objective Loss

Since classifying each triplet component, namely instrument, verb and target, is a multi-label classification problem, we employ weighted sigmoid cross-entropy to learn their various losses: L_I, L_V and L_T respectively. The weighted cross-entropy with logits is as follows:

$$L = \sum_{c=1}^{C} \frac{-1}{N}\Big(W_c y_c log\big(\sigma\left(\hat{y}_c\right)\big) + \left(1 - y_c\right) log\big(1 - \sigma\left(\hat{y}_c\right)\big)\Big), \tag{5.2}$$

where y_c and $\hat{y}_c$ are respectively the ground truth and predicted labels for class c, σ is the sigmoid function, and W_c is a weight for class balancing. The three components detection tasks are jointly learned in a multi-task manner following the uncertainty loss procedure given in [Kendall 2018] that uses learnable parameters w_I, w_V, w_T to automatically balance the tasks training as follows:

$$L_{comp} = \frac{1}{3}\left(\frac{1}{e^{w_I}}L_I + \frac{1}{e^{w_V}}L_V + \frac{1}{e^{w_T}}L_T + w_I + w_V + w_T\right). \tag{5.3}$$

This is only used for the auxiliary tasks captured by multi-task learning.

The triplet association loss L_{assoc} is also modeled as a sigmoid cross-entropy. To jointly learn the complete tasks end-to-end, we define the total loss using the equation:

$$L_{total} = L_{comp} + \rho L_{assoc} + \lambda L_2, \tag{5.4}$$

where ρ is a warm-up parameter that allows the network to focus solely on learning the individual components' information within the first 18 epochs. $\lambda = 1e^{-5}$ is a regularization weight decay for the L_2 normalization loss.

5.4.3 Hyper-parameters

The feature extraction backbone is pretrained on ImageNet. All the models are trained using Stochastic Gradient Descent with Momentum ($\mu = 0.95$) as optimizer. We maintain a step-wise learning rate ($\eta = 0.001$) policy, decayed by $\delta = 0.1$ after every 50 epochs. The models are trained in batches of size 8 for 200 epochs. The final model weights are selected based on their validation loss saturation. All the hyper-parameters are tuned on the validation set (5 videos) using the grid search approach and the best selected by validation loss.

5.4.4 Hardware and Schedule

Our networks are implemented using TensorFlow and trained on GeForce GTX 1080 Ti, Tesla P40, RTX6000, and V100 GPUs provided by CAMMA-ICube, the Unistra Mesocentre, and GENCI-IDRIS (Grant 2021-AD011011638R1). The training time is approximately 118-180 hours on a GTX 1080 Ti. Total storage space consumption for the model, input data, output weights, and summaries is under 10GB. The parameter count for the Attention Tripnet is 11.81M.

5.4.5 Inference and Evaluation Protocols

The model is tested in online mode. The output is a multi-label vector of probabilities per frame for each task. We follow the same video-based evaluation protocol in Section 4.4.3 to compute the *component average precision* for (AP_I, AP_V, AP_T) and *triplet average precision* for $(AP_{IV}, AP_{IT}, AP_{IVT})$.

Due to high similarities between triplets, we additionally measure the ability of a model to predict the exact triplets within its top N confidence scores. We call this metric, the **Top-N recognition performance** of the triplets. For every given test sample x_i, a model made an error if the correct label y_i does not appear in its top N confident predictions $\hat{y}_i$ for that sample. Using this setup, we measure the top-5, top-10, and top-20 accuracies for the triplet prediction.

5.5 Experimental Results

In this section, we present the results of Attention Tripnet in comparison with the baseline and state-of-the-art (SOTA) methods.

5.5.1 Ablation Studies

5.5.1.1 Ablation Study on the Encoder's Attention Type

Table 5.5 – Ablation study on the task-attention suitability.

Guided detection	AP_V	AP_T
None (*as in* MTL baseline)	48.4	28.2
CAM (*as in* Tripnet's CAG)	51.3	32.1
CAM + Channel attention	59.0	31.5
CAM + Position attention	51.2	35.1
CAM + Dual[1] attention	**61.1**	**40.2**

[1] Dual = (channel + position) attentions

For the choice of the attention type in the CAGAM module, we present an ablation study on the use of different spatial attention: channel and position. As shown in Table 5.5, we compare with a baseline model (MTL) [Nwoye 2020] without attention (None), and show that attention guidance helps better detect the components in general. We also justify the distinct attention types for verbs and targets. Firstly, the channel and position attentions are each used for both verb and target detections (as reported in row 3 & 4 of Table 5.5), before they are combined (Dual) in the last row. Channel attention is better suited for verbs than targets, with +10.6% vs +3.5% improvement respectively. The likely indication for this is that the verbs are more sparse, with each more tied to unique instrument classes, which are captured channel-wisely in the instrument's CAM, suggesting that channel attention guidance would be more beneficial for verb search. Position attention behaves the opposite: +2.8% vs +6.9%. We explain that the anatomies are approximately in the same locality across frames. Position attention captures the anatomical spatial location relative to the activated pixels in

Table 5.6 – Performance summary of the proposed models compared to state-of-the-art and baseline models.

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Naive CNN	57.7	39.2	28.3	21.7	18.0	13.6
Naive TCN	48.9	29.4	21.4	17.7	15.5	12.4
MTL baseline	84.5	48.4	28.2	26.6	21.2	17.6
Tripnet [Nwoye 2020]	**92.1**	54.5	33.2	29.7	26.4	20.0
Attention Tripnet	92.0	**60.2**	**38.5**	**31.1**	**29.8**	**23.4**

the instrument's CAM. Matching and pairing verbs, with channel attention, and targets, with position attention, give the most balanced and highest improvement: +12.4% verbs, +12.0% targets. This choice is, therefore, retained in the proposed model.

5.5.2 Quantitative Results

5.5.2.1 Component Detection and Association mAP

For ease of reference, we present a summary of the component detection precision and triplet association precision for the proposed model in comparison with the baselines in Table 5.6. The baseline is a simple CNN model known as naive CNN in the previous chapter. Our second baseline applies temporal refinement on the outputs of the naive CNN using a TCN [Lea 2016b]. The performance of these baselines shows that it is not sufficient to naively classify the triplet IDs without considering the triplet components. Multi-task learning (MTL) of the triplet components helps the model gain in performance, but still scores low on triplet association. The previous SOTA, Tripnet [Nwoye 2020], leverages the CAG to improve the MTL in the triplet components detection. It also improves the interaction recognition AP_{IVT} by 2.4% using the 3Dis.

The proposed Attention Tripnet uses the CAGAM to further improve Tripnet's verb detection by 5.7% and target detection by 5.3%. The Attention Tripnet is on par for instrument detection AP; this is likely due to the instrument detection being already saturated. The overall performance does increase, with indeed a 3.4% improvement for triplet recognition.

Next, we present a performance breakdown of the triplet components per-class for all the experimented models.

5.5.2.2 Per-Class Instrument Detection Performance (AP_I)

In Table 5.7, we analyze the model performance on the instrument detection per category. Here, we observed that Tripnet and Attention Tripnet networks all detect the various instruments category at a performance higher than 80.0%. This is not the case with the naive networks which do not consider the triplet components in their modeling. The grasper and hook are the most correctly detected while the scissors and irrigator are the least detected.

Table 5.7 – Result breakdown for instrument (AP_I) per class detections.

Method	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	mAP
Naive CNN	91.4	47.9	89.1	24.0	50.2	43.2	57.7
Naive TCN	90.5	37.6	86.2	15.9	33.3	29.6	48.9
MTL baseline	95.5	85.8	96.6	74.8	85.8	68.2	84.5
Tripnet	**97.8**	91.2	**98.1**	**90.7**	92.1	**82.7**	**92.1**
Attention Tripnet	**97.8**	**91.5**	**98.1**	89.7	**92.8**	82.1	92.0

This is likened to their usage frequency and inter/intra-class variance.

5.5.2.3 Per-Class Verb Detection Performance (AP_V)

Table 5.8 – Result breakdown for verb (AP_V) per class detections.

Method	Grasp	Retract	Dissect	Coagulate	Clip	Cut	Aspirate	Irrigate	Pack	Null	mAP
Naive CNN	48.6	82.1	80.5	30.5	49.5	23.8	32.4	16.0	09.2	15.9	39.2
Naive TCN	24.9	80.2	66.4	27.4	31.9	14.7	14.8	13.9	2.0	15.4	29.4
MTL baseline	47.9	85.0	84.8	55.0	79.1	44.1	35.4	13.4	18.0	17.0	48.4
Tripnet	45.8	88.1	86.7	66.3	85.1	68.3	44.9	12.2	**22.5**	20.1	54.5
Attention Tripnet	**62.4**	**89.4**	**89.4**	**69.7**	**88.5**	**84.3**	**48.5**	**20.8**	21.4	**22.7**	**60.2**

In Table 5.8, it is observed that the most dominantly used verbs such as *retract, dissect, clip,* and *cut* are correctly recognized over 80.0% of the time by the proposed model; this is likely because these have the strongest affinity with a particular instrument. Otherwise in cases where an instrument has more than one frequent verb, detection performance tends to spread out over those verbs according to their prevalence: *grasp (*62%), *retract (*89%*)*, & *pack (*21%) for *grasper*, or *aspirate (*49%*)* & *irrigate (*21%) for *irrigator*, etc. as can be seen in the Attention Tripnet model. *Null, Pack* and *irrigate* verbs are the least recognized verbs. *Null*, as the default verb class, carries complex semantics, conveying not only inaction but also any verb uncategorized in our dataset. *Pack* slightly overlap with *retract* and *grasp* which more frequent in the dataset. *Irrigate* is often mistaken for *aspirate*, since distinguishing those is mostly based on the fluid's dynamics (expulsion or suction) over time. A temporal model may be a way to improve on this. In summary, the proposed model, while maintaining a comparable performance with the SOTA on *pack*, outperforms the SOTA Tripnet and the baseline models in verb detection across all categories.

5.5.2.4 Per-Class Target Detection Performance (AP_T)

The target appears to be the most challenging component to correctly detect. Certain targets are easier to detect than others. As can be seen in Table 5.9, the *gallbladder* and *specimen-bag* are the most recognized targets, with the proposed models exceeding 80.0% AP. Other targets such as *liver, fluid,* and *omentum*, are moderately detected at an AP above 40%. This is likely due that their obvious nature and clearer boundaries compared to the less detected ones.

Table 5.9 – Results Breakdown for Target (AP_T) Per-Class Detection. The target ids 1..14 correspond to *gallbladder, cystic-plate, cystic-duct, cystic-artery, cystic-pedicle, blood-vessel, fluid, abdominal-wall-cavity, liver, omentum, peritoneum, gut, specimen-bag, null* respectively.

Method	Target														mAP
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Naive CNN	84.2	14.8	26.3	18.7	14.3	03.6	32.4	10.1	49.8	35.2	**08.4**	08.4	69.3	15.9	28.3
Naive TCN	79.9	10.0	21.4	19.6	07.0	01.3	14.8	06.9	43.1	27.9	01.9	09.0	37.4	15.4	21.4
MTL baseline	85.1	12.2	29.3	18.6	06.5	06.4	30.6	09.8	55.7	35.8	02.1	08.4	71.1	17.5	28.2
Tripnet	87.0	**22.5**	29.7	21.9	04.7	15.0	42.9	32.3	57.5	36.7	02.0	**11.9**	74.1	20.9	33.2
Attention Tripnet	**87.8**	15.6	**37.1**	**30.1**	**16.6**	**26.5**	**53.2**	**37.5**	**59.8**	**48.7**	**03.5**	08.3	**85.6**	**23.5**	**38.5**

Interactions with those targets are easier to ascertain than interactions with much smaller structures such as *peritoneum, cystic-artery*, and other *blood-vessels*. Within the *cystic-pedicle*, the *cystic-duct* is the most detected tubular structure. The *cystic pedicle* in itself is not well recognized. This is likely because its definition as *cystic pedicle* instead of *cystic-artery* and *cystic-duct* depends on whether the pedicle is dissected or not. This shows how deceptively complicated the task of anatomical target detection is.

5.5.2.5 Top-N Triplet Recognition Performance

Table 5.10 – Top N Accuracy of the triplet (I,V,T) predictions.

Method	Top-5	Top-10	Top-20
Naive CNN	67.0	80.0	90.2
Naive TCN	54.5	69.4	84.3
MTL baseline	70.2	80.2	89.5
Tripnet	70.5	81.9	91.4
Attention Tripnet	**75.3**	**86.0**	**93.8**

Since the large number of triplet classes makes it complex to precisely assess the model performance. Here we focus on the top N predictions of the final triplets as shown in Table 5.10. This reveals the model's confidence in approximating its prediction to the most likely classes. When considering the model's top 20 predictions, the proposed model records an AP of ≈ 94%. The model's confidence however decreases when considering more top predictions, suggesting how closely related most of the triplet classes could be. Interestingly, the performance remains above 75% with the top 5 predictions, better than all compared baselines.

5.5.2.6 k-Fold Cross Validation

To ascertain the confidence interval with a less biased and less optimistic estimate of the model, we compute their cross-validation performance averaged over the 5 hold-out test splits. The results as presented in Table 5.11, shows that the baseline model has the least performance in both the component detections and triplet association. Even with the standard deviation

Table 5.11 – K-fold Cross-Validation Performance.

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
MTL Baseline	84.4±1.4	46.6±3.5	26.0±3.0	25.2±3.0	19.2±4.1	16.7±1.9
Tripnet	91.6±1.8	51.3±3.2	32.1±2.9	30.9±2.3	24.4±2.0	19.7±1.3
Attention Tripnet	90.3±1.9	57.6±3.9	34.5±2.4	32.4±3.0	27.0±3.6	22.4±2.1

(std), it is still not sufficiently comparable to the SOTA or the proposed models in all the six metrics.

The Attention Tripnet model has a lower performance on instrument detection than the SOTA model, however, its std upper-bound shows that it has a performance spread that is comparable to the SOTA model. On the verb detection, The attention model maintains a score large enough to outperform the SOTA in its std spread justifying the use of attention for verb detection. This is similar in target detection, however, the SOTA maintains a high std that can approximate the lower-bound performance of the Attention Tripnet model. This means that the worst-performing Attention Tripnet is as good as the best performing Tripnet in this regard.

Concerning the triplet association recognition, the Attention Tripnet again maintain mean APs with deviation spreads that consistently place it above the SOTA model in all the three metrics (AP_{IV}, AP_{IT}, AP_{IVT}) used in judging the recognition of tool-activities.

The average results over the 5 hold-out test splits in comparison with the baselines help to ascertain the confidence interval with a less biased and less optimistic estimate of the proposed model.

5.5.2.7 Statistical Significance Analysis

Table 5.12 – The p-values obtained in Wilcoxon signed rank test of the proposed methods using the SOTA model (Tripnet) as the alternative method. (*Lower p-value is preferred*).

	Tasks	p-value
Component Detection	AP_I	$p \approx 0.327$
	AP_V	$p \ll 0.001$
	AP_T	$p \ll 0.001$
Triplet Association	AP_{IV}	$p \approx 0.018$
	AP_{IT}	$p \approx 0.010$
	AP_{IVT}	$p < 0.005$

We also measure the statistical significance of the proposed model performance using the SOTA model as the alternative method. Using the Wilcoxon signed-rank test, we sample $N = 30$ random batches of 100 consecutive frames instead of 30 random frames to simulate the evaluation on video clips. The null hypothesis (H_0) states that the difference between the proposed method and the alternative method follows a symmetric distribution around zero. The level of statistical significance is expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence to reject the null hypothesis. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. Similarly, a p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. We perform the statistical significance analysis for each task, and based on the obtained p-values as presented in Table 5.12, we draw the following conclusions:

a. The proposed model does not significantly improve the instrument detection sub-task. The p-value (0.327) falls short of the standard 0.05 significant level. This is mainly because the instrument detection performance is already saturated in the alternative method; there is no new modeling in the proposed methods targeting the improvement of the instrument detection sub-task. Being a two-tailed test, the p-value also shows that the SOTA does not outperform the proposed model on instrument detection.

b. The guided attention mechanism is very useful in improving the verb and target detections in the Attention Tripnet model. Its contribution is significant enough to even beat a more narrow 0.01 significant level.

c. While improving the component detection, we observe a corresponding improvement in the triplet association which are also significant at a 0.005 significance level on AP_{IVT} association.

In summary, we reject the null hypothesis H_0 at a confidence level of 5%, concluding that there is a significant difference between the proposed models and the alternative method.

5.5.3 Qualitative Results

5.5.3.1 Triplet Recognition with Weak Localization

Given a model prediction as a vector of class-wise probability score per frame. The predicted class labels are obtained by applying a 0.5 threshold on these outputs. Localization heatmaps are obtained from the weakly supervised learning (WSL) module of the instrument detection. Bounding boxes are obtained by connected components to the maximum pixel in the thresholded heatmaps. The bounding boxes are overlaid on the original images to depict the localization of the regions of actions. These information are presented as qualitative results in Figure 5.5. This localization shows the focus of the model when it makes a prediction, thereby providing insight into its rationale. These results are solid arguments in favor of the model's ability for spatial reasoning when recognizing surgical actions.

5.5.3.2 Attention Map Visualization

To understand the benefit of the CAGAM's attention mechanism, we visualize its attention maps in Figure 5.6. For each input image, we randomly selected a few points (marked $i \in [1,2,3,4]$) in the images and reveal the corresponding attention maps for the tool-tissue interaction captured in the CAGAM's position attention map. We observe that the attention

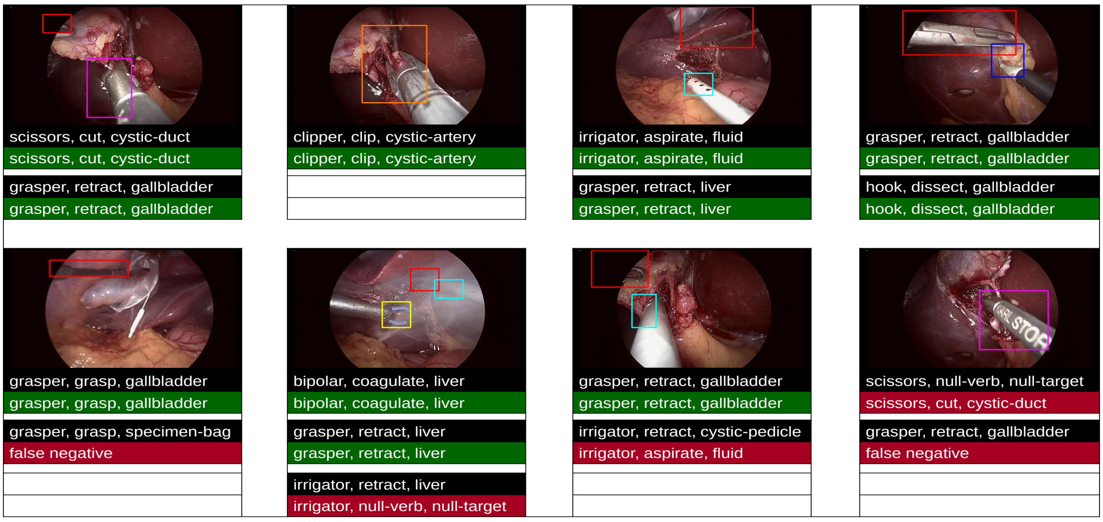

scissors, cut, cystic-duct
scissors, cut, cystic-duct
grasper, retract, gallbladder
grasper, retract, gallbladder
clipper, clip, cystic-artery
clipper, clip, cystic-artery
irrigator, aspirate, fluid
irrigator, aspirate, fluid
grasper, retract, liver
grasper, retract, liver
grasper, retract, gallbladder
grasper, retract, gallbladder
hook, dissect, gallbladder
hook, dissect, gallbladder
grasper, grasp, gallbladder
grasper, grasp, gallbladder
grasper, grasp, specimen-bag
false negative
bipolar, coagulate, liver
bipolar, coagulate, liver
grasper, retract, liver
grasper, retract, liver
irrigator, retract, liver
irrigator, null-verb, null-target
grasper, retract, gallbladder
grasper, retract, gallbladder
irrigator, retract, cystic-pedicle
irrigator, aspirate, fluid
scissors, null-verb, null-target
scissors, cut, cystic-duct
grasper, retract, gallbladder
false negative

Figure 5.5 – Qualitative results of the Attention Tripnet showing the triplet predictions and the localization of the regions of the actions. Localization bounding boxes are obtained from the instrument detection branch of the model. Predicted and ground-truth triplets are displayed below each image: black = ground-truth, green = correct prediction, red = incorrect prediction. A missed triplet is marked as false negative and a false detection is marked as false positive (Best viewed in colour).

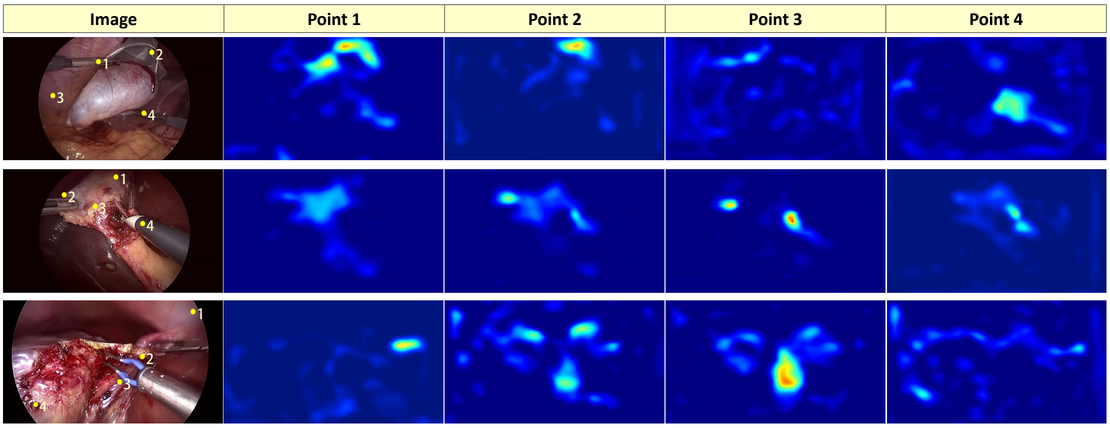


Figure 5.6 – Attention maps in the CAGAM module on the CholecT50 test set. The left column is the input image, the subsequent columns are the attention maps captured by the different points as marked in the input image. The attention map shows the focus on the target (best seen in color).

module could capture semantic similarity and full-image dependencies, which change based on the contribution of the selected pixel to the action understanding. This shows that the model learns attention maps that contextualize every pixel in the image feature with respect to the action performed. For instance in the top image: point 2, a pixel location on the instrument - *grasper*, creates an attention map that highlights both the instrument and its target - *gallbladder*. Indeed, the attention guidance introduced in this model helps to highlight the triplet's interest regions while suppressing the rest. This effect is shown further in the supplementary video.

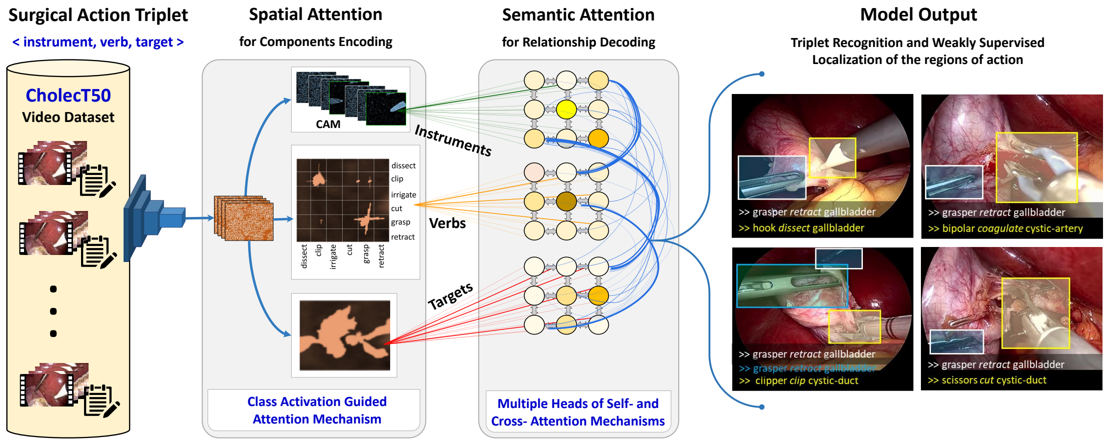
5.6 Conclusion

We have presented a new state-of-the-art method for surgical action triplet recognition. The presented method features a new form of spatial attention for enhanced tripnet components detection. The spatial attention leverages the activation features of surgical instruments to detect the verbs and targets of interest more efficiently. We also introduce, in this work, a new large-scale endoscopic video dataset, CholecT50, for action triplet recognition.

We validated our proposed method on the newly introduced CholecT50 dataset surpassing the previous methods in both the components detection and their association precision. Levering the introduced CAGAM, it is observed that the proposed model outperformed the baselines and SOTA methods in verb and target detection by AP > 5.0, and AP > 3.0 on the triplet association. Hence, future work will focus on improving the association part. The qualitative results show that while performing recognition, the model learned spatial reasoning about the triplets suggesting a possibility of segmenting regions of triplet interaction.

6 Transformer-inspired Method for Enhanced Triplet Association

An idea does not come out fully mature,
they become clearer as you work on them.
– Mark Zuckerberg, Facebook



Surgical Action Triplet
< instrument, verb, target >
CholecT50
Video Dataset
Spatial Attention
for Components Encoding
CAM
dissect
clip
irrigate
cut
grasp
retract
Class Activation Guided Attention Mechanism
Instruments
Verbs
Targets
Semantic Attention
for Relationship Decoding
Multiple Heads of Self- and Cross- Attention Mechanisms
Model Output
Triplet Recognition and Weakly Supervised Localization of the regions of action
>> grasper retract gallbladder
>> hook dissect gallbladder
>> grasper retract gallbladder
>> bipolar coagulate cystic-artery
>> grasper retract gallbladder
>> grasper retract gallbladder
>> clipper clip cystic-duct
>> grasper retract gallbladder
>> scissors cut cystic-duct

Figure 6.1 – An illustration of attention mechanisms for the modeling of triplet components detection and their association.

Chapter Summary

So far, we have tackled action triplet recognition with two novel deep learning methods, with the latter improving the former. The Attention Tripnet is designed to enhance the triplet component detection using an attention mechanism. The improvements on this task also result in a corresponding improvement on the final triplet association without additional modeling.

In this chapter, we propose a new and more advanced unit to handle the triplet association [Nwoye 2021]. Our interest here is to effectively capture interactions between the detected triplet components with optimal precision. This is achieved using a longer range attention mechanism as will be discussed further. The proposed model here is trained end-to-end for both triplet components detection and their association recognition.

6.1 Objectives

Previously, in Section 5.1, we disentangle the task of surgical action triplet recognition into: (1) triplet components detection, and (2) the triplet components association, for easy modeling. The Attention Tripnet presented in the previous chapter uses spatial attention to recognize the triplet components at a precision better than the previous SOTA method. Though the Attention Tripnet marginally improves also on the final triplet recognition, the association would improve if specifically modeled to fit the new attention performance.

Hence, the objective of this chapter is to complete the disentangled tasks by developing a new deep learning module that can resolve the association of the detected components. This involves a module that would replace the less advanced 3Dis used in the previous chapters for the following reasons:

a. Surgical action triplets exhibit a high level of semantic overlap which has not been effectively tackled with a primitive matrix multiplication in the 3Dis.

b. The 3Dis constructs interaction points for all possible triplet combinations, given the three-component classes. Apart from the large size of these interaction points, they also include the invalid ones (i.e.: impracticable triplets and those not in the used dataset). Implementing a 3Dis layer, the knowledge of valid class distribution in the training data is needed for masking out the invalid points at both training and testing times. Added to this demand, the positions of the invalid triplets in the 3D space need to be correctly estimated for error-free masking.

c. By modeling all triplet possibilities resulting in large triplet classes, the 3Dis is hard to train.

d. Attention Tripnet reveals that attention modeling is useful in highlighting the triplet components of interests in each surgical image frame. Hence, it becomes more interesting to investigate how attention modeling can help also in the components relationship understanding.

Taking everything into account, the work presented in this chapter is targeted at the development of a semantic attention mechanism that can help to resolve the triplet components relationship for more efficient recognition of tool-tissue interaction in laparoscopic videos.

6.2 Rendezvous for Enhanced Triplet Component Association

Encouraged by the utility of the attention mechanism in detecting the triplet components, we extend the attention modeling to also capture the triplet components' association without using the 3Dis. For this, we innovate a hybrid attention method: Multi-Head of Mixed Attention (MHMA) for surgical action triplet recognition. This models the required interactions efficiently using a Transformer-like architecture [Vaswani 2017, Dosovitskiy 2020, Chen 2021] that relies on long-range attention. The semantic properties of each component, represented in respective class maps, are used to consider the constituent components of the triplets. In MHMA, we do not divide class maps into patches like the Vision Transformer [Dosovitskiy 2020] does. We observed in a preliminary study, presented as ablation result, that the patch sequence approach degrades representations, especially information on instruments that is crucial for locating actions. The MHMA incorporates a new form of semantic attention: one that leverages the spatial and class-wise representations from different components of the triplets to decode the interactions between the detected instruments and tissues in a laparoscopic procedure. We then propose *Rendezvous* (RDV) - a transformer-inspired neural network that uses MHMA for action triplet recognition. The RDV network is conceptually divided into four segments: feature extraction backbone, encoder, decoder, and classifier as shown in Figure 6.2. The novel MHMA is situated in the RDV's decoder.

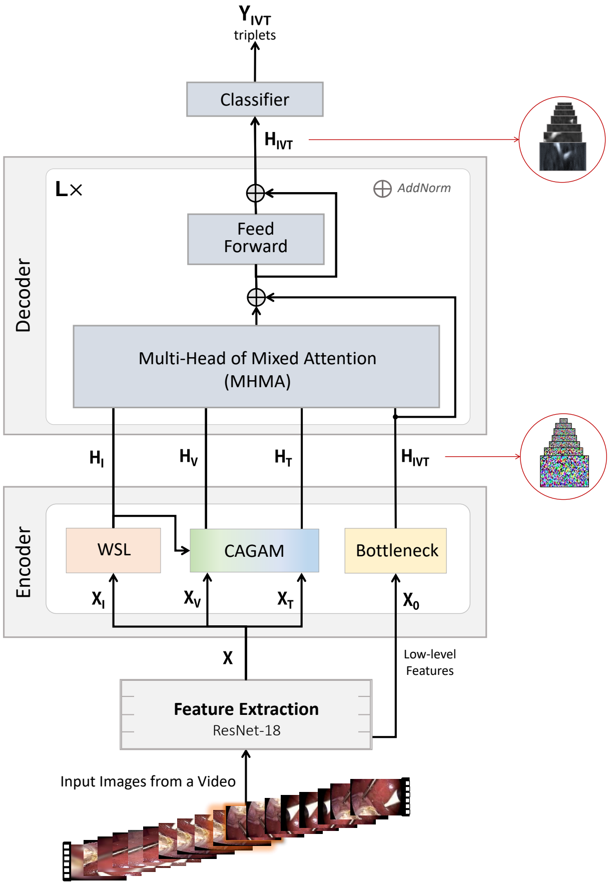


Figure 6.2 – Architecture of Rendezvous: a Transformer-inspired neural network model with multi-head of mixed attention mechanism for surgical action triplet recognition.

6.2.1 Feature Extraction

Still using the same feature extraction backbone, ResNet-18, as in the previous models, we extract visual feature $\mathbf{X}$ from RGB input images. This feature is replicated into a trio of $\mathbf{X}_I$, $\mathbf{X}_V$, $\mathbf{X}_T$ for the multitask learning of the instrument, verb, and target respectively.

6.2.2 Components Encoding

The encoder is responsible for detecting the various components of the triplets, while the decoder resolves the relationships between them. The encoder is composed of the WSL module for instrument detection, CAGAM module for verb and target recognition, and a bottleneck layer collecting unfiltered low-level features from Resnet-18's lower layer.

Specifically, $\mathbf{X}_I$ is fed to the WSL instrument detection branch returning the CAM ($\mathbf{H}_I$) and instrument logits $\mathbf{Y}_I$ as done in the Attention Tripnet model. The $\mathbf{H}_I$ and ($\mathbf{X}_V$, $\mathbf{X}_T$) are fed to the CAGAM module returning a pair of ($\mathbf{H}_V$,$\mathbf{Y}_V$) for the verb detection and a pair of ($\mathbf{H}_T$,$\mathbf{Y}_T$) for the target detection. To ensure that the $\mathbf{H}_I$,$\mathbf{H}_V$ and $\mathbf{H}_T$ class maps properly capture their corresponding components, we train their global pooled logits ($\mathbf{Y}_I$,$\mathbf{Y}_V$,$\mathbf{Y}_T$) as auxiliary classification tasks. In addition to those already refined component-specific features, a global context feature is also necessary since those are lost in component-specific features; which is why we also draw a low-level feature $\mathbf{X}_0$ from the first block of ResNet and feed it to the bottleneck layer that consists of $3 \times 3 \times 256$ and $1 \times 1 \times C$ convolution layers, where $C = 100$ is the number of triplet classes. This gives us the global context feature for triplets $\mathbf{H}_{IVT}$, with channels matched to the triplet classes. These features ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$, $\mathbf{H}_{IVT}$) are fed to the decoder layer.

6.2.3 Triplet Decoding

Having obtained the component features $\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$ and the global triplet feature $\mathbf{H}_{IVT}$ from the encoder, the correct triplets are recognized by resolving the relationship between these components. Hence, the RDV decodes all the intra- and cross-interactions between the triplet's global context feature and the three features corresponding to individual components, using scaled dot-product attention. In addition to self-attention, cross-attention adds the capability to better model the relationships with components participating in the action triplet. This is important when resolving interactions: for instance, an anatomical part can appear in the frame without being a target, often making the interaction with the instrument ambiguous.

To understand the attention decoder used in this work, we explain the **decoding-by-attention** [Nwoye 2021] concept below:

a. Firstly, attention decoding is described as a search process whereby a query ($\mathbf{Q}$), that is issued by a user (*sink* or *receiver*), is used to retrieve data from a repository (*source*). Normally, $\mathbf{Q}$ is a user's abridged description of the requested data also known as *search terms*.

b. The source context consists of a key-value ($\mathbf{K}, \mathbf{V}$) pair where $\mathbf{V}$ is a collection of several data points or *records* and $\mathbf{K}$ is the mean descriptor for each record also known as *keywords*.

c. To retrieve the requested data, the issued $\mathbf{Q}$ is matched with the available $\mathbf{K}$s to create an *affinity* ($\mathbf{P}$), also known as the *attention weight*.

d. The $\mathbf{P}$, when matched with $\mathbf{V}$, creates an **attention map** ($\mathbf{A}$) which helps retrieve the most appropriate data to the sink.

We implement a transformer-inspired decoder that is composed of a stack of $L = 8$ identical layers as shown in Figure 6.2. Each layer receives the triplet features $\mathbf{H}_{IVT}$ and the encoded class maps ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) as inputs which are processed successively by its two internal modules: MHMA and feed-forward, to produce refined triplet features, $\mathbf{H}_{IVT}$. The output of each module is followed by a residual connection and a layer normalization (AddNorm) as it is done in other multi-head attention networks. The entire cycle is repeated, with a more refined $\mathbf{H}_{IVT}$, until the L^{th} layer.

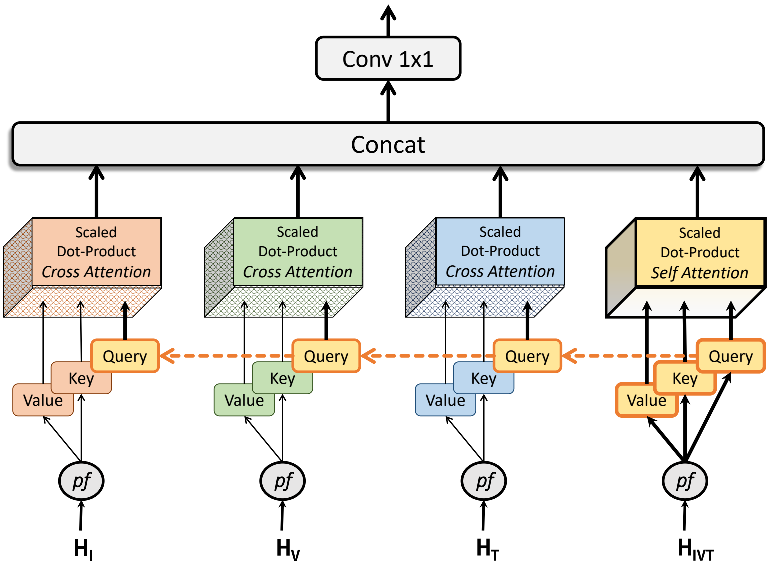


Figure 6.3 – Architecture of the multi-head of mixed attention (MHMA) mechanism in Rendezvous for triplet decoding :showing the feature projection into Q, K and V, and subsequent multiple heads of self and cross attentions using scale-dot product attention mechanism.

6.2.3.1 Multi-Head of Mixed Attention (MHMA) Mechanism

The multi-head attention combines both self and cross attentions, encouraging high-level learning of triplets from the interacting components as shown in Figure 6.3. It starts with a projection function, **pf**, which generates a set of value **V**, key **K**, and/or query **Q** for each context feature ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T, \mathbf{H}_{IVT}$). In the implementation as shown in Equation 6.1, the **pf** function generates vectors of $\mathbf{Q} \in \mathbb{R}^{1\times C}$ and $\mathbf{K} \in \mathbb{R}^{1\times C_z}$ that represent the abridged mean descriptors of the contexts by leveraging the global average pooling (GAP) operation. We map each descriptor to a feature embedding layer where we mask (dropout $\lambda = 0.3$) parts of **Q** to avoid repeating the same query in the L alternating layers. Using the **pf** function, we also obtain the $\mathbf{V} \in \mathbb{R}^{H\times W\times C_z}$ by a convolution operation on the feature context and reshape to $\mathbb{R}^{HW\times C_z}$. Hence, the extracted **Q**, **K** and **V** features follow the aforementioned concept on attention decoding (items 1 & 2). The **pf** function generates each **K** and **Q** using FC layers as done in [Vaswani 2017, Dosovitskiy 2020], and generates the **V** using convolution layers as done in [Fu 2019, Wang 2018, Huang 2019].

$$pf\,(H) = \begin{cases} Q: & FC\Big(DROPOUT\Big(GAP\,(\mathbf{H})\Big)\Big), \\ K: & FC\Big(GAP\,(\mathbf{H})\Big), \\ V: & CONV\,(\mathbf{H}). \end{cases} \tag{6.1}$$

Next, we build 4 attention heads for the instrument, verb, target, and triplet attention features. In the existing Transformer [Vaswani 2017] and Transformer-based models, each of the heads learns a self-attention. Self-attention helps a model understand the underlying meaning and patterns within its own feature representation. This is needed for the initial scene understanding. However, when each feature representation (such as a class-map) has been discriminated to attend to only one component in an image scene, understanding the underlying relationship will require a cross-attention with the other component features. In a cross-attention mechanism, the attention built from one context (the *source*) is used to highlight features in another context (the *sink*) as done in [Mohla 2020]. While the self-attention mechanism computes the focal representation on the same triplet features, cross attentions learn the triplet representations by drawing attention from the individual components: namely instrument, verb, and target. This mechanism models how the features of each component affect the triplet composition, by propagating the affinities from their respective context features to the required triplet features.

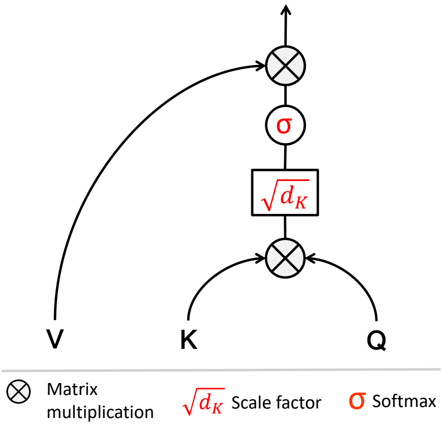


Figure 6.4 – Architecture of the Attention mechanisms used for self and cross attentions. In self-attention, the (K,V,Q) triple comes from one feature context. In cross attention the (K,V) pair comes from the source feature context while Q comes from the sink feature context.

To utilize both self and cross attention, we model the source context from the encoded class-map features ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) representing the triplet components and the sink context from the triplet features ($\mathbf{H}_{IVT}$). Of course, the source context remains the same as the sink in the self-attention mechanism. This means we generate the corresponding **K**s and **V**s from both the source and sink contexts, but generate the **Q** only from the sink context using the projection function, **pf**, as shown in Figure 6.3. With **Q** coming from the triplet features, we actually focus the image understanding on the actions of interest by pointing the cross-attention heads at the component's discriminative features ($\mathbf{H}_I$, $\mathbf{H}_V$, $\mathbf{H}_T$) in a manner that helps the attention network benefit from the learned class representations. This also respects the aforementioned decoding-by-attention concept. We then learn a scaled dot product attention of the **Q** on the (**K**,**V**) pair for each attention head as shown in Figure 6.4. Specifically, the scaled dot product attention is derived using the widely used attention formula [Vaswani 2017] in Equation 6.2:

$$\mathbf{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{V}.\sigma\left(\frac{\mathbf{K}\mathbf{Q}^T}{\sqrt{d_\mathbf{K}}}\right), \tag{6.2}$$

where σ is a softmax activation function, $\sqrt{d_K}$ is a scaling factor, and d_K is the dimension of K after linear transformation. The cross attention is implemented on the instrument, verb, and target attention heads, whereas self-attention is implemented on the triplet attention head. While each attention head simultaneously concentrates on its own features of interest, the multi-head module combines heads $A_{1..\mathcal{N}}$ to jointly capture the triplet features as in Equation 6.3:

$$\mathbf{A}_{1..\mathcal{N}} = \mathbf{W}\Big(\|_{i=1}^{\mathcal{N}} \mathbf{A}_i \Big), \tag{6.3}$$

where $\|$ is a concatenation operator for $\mathcal{N} = 4$ attention heads. A_1 is the triplet self-attention, $A_{2...\mathcal{N}}$ are the triplet cross attentions with the interacting components, and W is the matrix of convolution weights. This packed convolution scheme merges the information from all attention heads while preserving its spatial structure.

6.2.3.2 Feed-Forward

The output of the multi-head attention is further refined by a feed-forward layer which is a stack of 2 convolutions with an AddNorm. The output is a refined $\mathbf{H}_{IVT}$ with each channel attending to each triplet class.

6.2.4 Triplet Classification

The RDV model terminates with a linear classifier for the final classification of the triplets. The classifier is composed of a global average pooling (GAP) layer and an FC-layer (with $C = 100$ neurons) for the triplet classification. It receives as input triplet features $\mathbf{H}_{IVT}$ from the L^{th} layer of RDV decoder and output vector of logits ($\mathbf{Y}_{IVT}$) representing the class-wise probability of the action triplets. The triplet logits are trained jointly end-to-end with the auxiliary logits from the encoder.

6.3 Experimental Setup

6.3.1 Data setup and Pipeline

The dataset used in this experiment is *CholecT50* [Nwoye 2021]. Since this is a continuation experiment on the attention modeling, we reuse the same input pipeline, data loader, preprocessing, dataset split, and augmentation styles presented in the previous chapter (Section 5.4.1).

6.3.2 Training, Loss Function and Hyper-parameters

We train all the components tasks as well the main triplet recognition using weighted sigmoid cross-entropy as presented in the previous chapter (Section 5.4.2). The component tasks are trained by auxiliary loss minimization with a warmup parameter that forces the model to concentrate only on these sub-tasks for the first 18 epochs as done in Attention Triplet. The only difference being that we decay the components' learning rate harder ($1e-2$), forcing the network to switch and pay more attention to the association tasks after the warmup period elapse. All the hyper-parameters are tuned on the validation set (5 videos) with up to 74 grid search experiments. Other training configurations such as compute infrastructure, optimizer, epoch, batch size, learning rate policy, and decay schedules all follow the same setup as in the previous work (Section 5.4.3). The parameter count for an 8-layer RDV model is 16.61M.

6.3.3 Inference and Evaluation Protocols

All the evaluated models are tested in online mode producing a vector of multi-label probabilities per frame for each task. We follow the same video-based evaluation protocol in Section 4.4.3 to compute the Component Average Precision for (AP_I, AP_V, AP_T) and Triplet Average Precision for $(AP_{IV}, AP_{IT}, AP_{IVT})$, and the evaluation protocol in Section 5.4.5 to compute the *top-N recognition performance* of the triplets.

Additionally, we show the top 10 predicted triplet class labels and their AP scores for a more insightful analysis of the model's performance.

6.4 Results and Discussion

In this section, we rigorously validate new components of the Rendezvous (RDV) through careful ablation studies. We then provide a comparative analysis with baseline and state-of-the-art (SOTA) methods to show our methods' superiority.

6.4.1 Ablation Studies

6.4.1.1 Ablation Study on Decoder's Attention Type

Table 6.1 – Ablation study on the attention type in the multi-head decoder.

Model	Layer size	AP$_{IV}$	AP$_{IT}$	AP$_{IVT}$
Single Self	6	29.8	23.3	18.8
Multiple Self	6	35.7	32.8	26.1
Self + Cross (RDV)	6	**39.4**	**36.9**	**29.9**

One of the novel contributions of this work is its hybrid multi-head attention mechanism for resolving tool-tissue interactions, combining self and cross attention. This is a significant advancement over traditional sequence modeling transformers, which rely solely on multi-heads of self-attention. Our choice of multi-head attention is justified in the following ablation study presented in Table 6.1.

Our first ablation model in this regards (*Single Self*) uses a multi-head attention with the input feature coming from the high-level features (**X**) of ResNet-18 to compute a successive scale dot-product attention over 8 decoder layers as in RDV. It can be observed that using a multi-head of self-attention coming from a single source (triplet features) yields insufficient results for action triplet recognition.

The *Multiple Self* ablation model, as a "self-attention only" version of the RDV, uses self-attention in all four contexts: instrument, verb, target, and triplet. The RDV clearly performs the best in terms of association, justifying our use of cross-attention.

6.4.1.2 Ablation Study on Attention Sequence Modeling

Table 6.2 – Ablation Study on Attention Sequence Modeling: This study shows the usefulness of our class-wise mapping over the contemporally patch-base sequence in Vision Transformer [Dosovitskiy 2020].

Model	Layer Size	AP_{IV}	AP_{IT}	AP_{IVT}
Patch-base sequence	6	33.4	29.3	24.1
Class-wise mapping	2	36.0	34.1	27.0
Class-wise mapping	8	**39.4**	**36.9**	**29.9**

We conducted a further ablation study to assess our choice of features modeling in the attention heads, which has been chiefly sequence-based, especially on patches of images, in the literature. Here, we compare our approach, which models attention on class-wise representative features, to the patch-based sequence modeling in the Vision Transformer as shown in Table 6.2. The compared model is built by replacing the projection function in RDV with a linear projection of patch sequences on the CNN features as done in Vision Transformer. Our proposed approach is approximately 3.0% better than using the patch sequence. It appears that the breaking of the CNN features into patches may have diluted the spatial semantics of the encoded features suggesting that the division of features into patches for sequence modeling may be better done on the raw images than on the CNN features.

6.4.1.3 Scalability Study on Multi-Head Layer Size

Table 6.3 – A scalability study on the multi-head layer size: showing the mean average precision (mAP) for varying triplet associations, number of learning parameters (Params) in millions (M), and inference time (i-Time) in frame per seconds (FPS) on GTX 1080 Ti GPU. ↑ indicates that higher value is preferred whereas ↓ indicates that lower value is better.

Layer size	mAP_{IV} (%)↑	mAP_{IT} (%)↑	mAP_{IVT} (%)↑	Params (M)↓	i-Time (FPS)↑
1	35.8	30.7	24.6	**12.6**	**54.2**
2	36.0	**41.1**	27.0	13.1	47.9
4	38.5	32.9	27.3	14.3	39.2
8	**39.4**	36.9	**29.9**	16.6	28.1

We carried out a scale study to observe the performance of the RDV when increasing the number of multi-head layers while keeping track of the number of parameters and GPU requirements. As shown in Table 6.3, it is observed that the proposed model improves when scaled up, at the cost of increased computational requirements. This proves that the proposed model can still improve in performance with an increase in computing power. However, to balance performance and resource usage, we choose $L = 8$ as default settings in all our experiments. An 8-layer RDV with > 25 FPS processing speed can be used in real-time for OR assistance.

6.4.1.4 Ablation Study on Use of Auxiliary Loss

Table 6.4 – Ablation Study on Use of Auxiliary Loss.

Model	AP_{IV}	AP_{IT}	AP_{IVT}
Without aux-loss	33.6	27.0	21.2
With aux-loss	**36.0**	**34.1**	**29.9**

We also conducted an ablation study on the use of auxiliary loss to train the triplet components of the RDV model to quantify its contribution to the proposed model. As presented in Table 6.4, it is observed that learning the individual components of the triplets in the same network pipeline (as *Aux-Loss*) helps the model to better understand the triplets.

6.4.2 Quantitative Results

6.4.2.1 Component Detection and Association mAP

We present the model performance on both the triplet components detection and triplet association recognition in comparison with the baseline and SOTA models as shown in Table 6.5. On the detection of the components, the proposed model maintains a comparable performance with Tripnet and its Attention counterpart on instrument presence detection. This is so since the three models share the same module (i.e. WSL) in this regard. The proposed model outperforms the Tripnet by +6.2% on verb detection, and by +5.1% on target detection. Compare to the Attention Tripnet, which shares the same CAGAM module with the proposed model, we still record a +0.5% improvement on verbs with comparable performance on targets.

On triplet association recognition, which is the main objective of the RDV, the proposed method outperforms both the Tripnet and Attention Tripnet in all three metrics for judging the association task. Specifically, the RDV uses a multi-head attention decoder to improve the instrument-verb-target association performance mAP_{IVT} by +9.9% over the Tripnet and +6.5% better than the Attention Tripnet. Breaking down the recognition further, RDV is respectively +9.7% and +8.3% better than the Tripnet and Attention Tripnet on instrument-verb AP_{IV}. Similarly, RDV is +10.5% and +7.1 better than the Tripnet and Attention Tripnet respectively on instrument-target AP_{IT}. In all cases, RDV outperforms all the naive models and the MTL baseline.

A breakdown of per-class detection of the triplet components and their association performance is presented in the supplementary material.

Table 6.5 – Performance summary of the proposed models compared to state-of-the-art and baseline models.

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Naive CNN	57.7	39.2	28.3	21.7	18.0	13.6
Naive TCN	48.9	29.4	21.4	17.7	15.5	12.4
MTL baseline	84.5	48.4	28.2	26.6	21.2	17.6
Tripnet [Nwoye 2020]	**92.1**	54.5	33.2	29.7	26.4	20.0
Attention Tripnet [Nwoye 2021]	92.0	60.2	**38.5**	31.1	29.8	23.4
Rendezvous	92.0	**60.7**	38.3	**39.4**	**36.9**	**29.9**

6.4.2.2 Per-Class Instrument Detection Performance (AP_I)

Table 6.6 – Result breakdown for instrument (AP_I) per class detections.

Method	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	mAP
Naive CNN	91.4	47.9	89.1	24.0	50.2	43.2	57.7
Naive TCN	90.5	37.6	86.2	15.9	33.3	29.6	48.9
MTL baseline	95.5	85.8	96.6	74.8	85.8	68.2	84.5
Tripnet	**97.8**	91.2	**98.1**	**90.7**	92.1	**82.7**	**92.1**
Attention Tripnet	**97.8**	**91.5**	**98.1**	89.7	**92.8**	82.1	92.0
Rendezvous	97.7	89.4	**98.1**	**92.0**	92.2	**82.7**	92.0

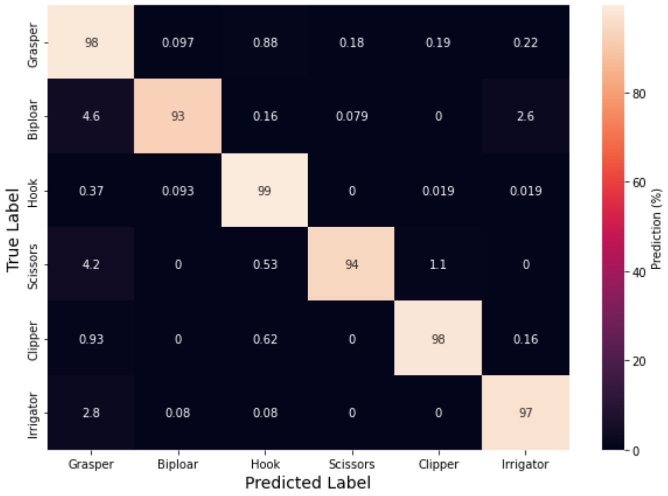


Figure 6.5 – A confusion matrix for tool recognition.

Results for instruments in Table 6.6 show that Tripnet, Attention Tripnet and RDV networks all detect the various instruments' category at a performance higher than 80.0%. The grasper and hook are the most correctly detected instruments. Apart from assessing the proposed model performance by AP metrics in Table 6.6, we also compute the model Accuracy at a threshold of 0.5. The accuracy scores, presented as confusion matrix in Figure 6.5, show the model's True Positive (TP), FP, and FN scores for each instrument's class. We convert these values to percentage (%) for ease of understanding. The confusion matrix in Figure 6.5 shows that the instrument recognition performance is saturated with no significant error rate in the matrix. This means that the proposed model records high true positives/negatives with almost no false positive/negative in all cases.

6.4.2.3 Per-Class Verb Detection Performance (AP_V)

Table 6.7 – Result breakdown for verb (AP_V) per class detections.

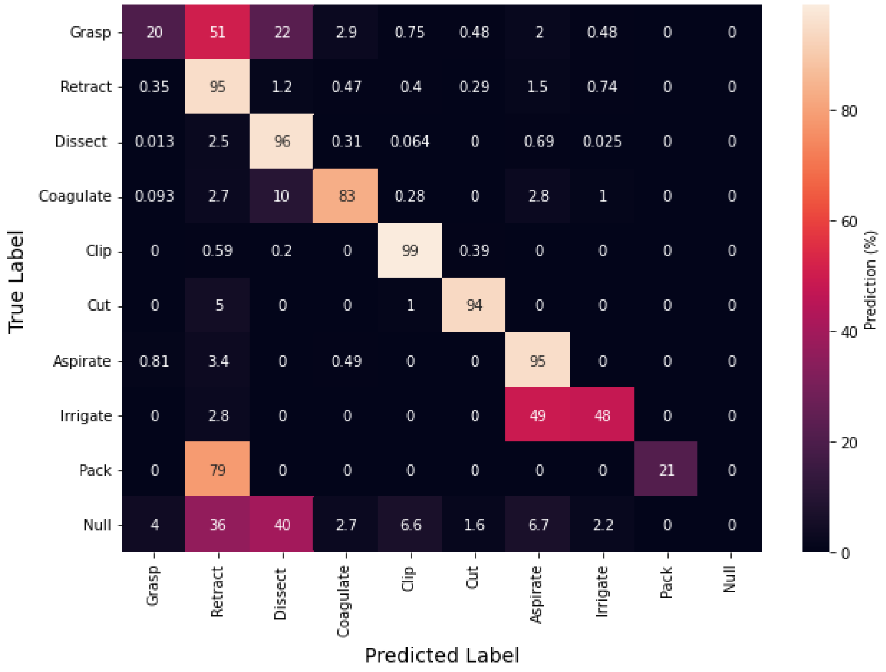
Method	Grasp	Retract	Dissect	Coagulate	Clip	Cut	Aspirate	Irrigate	Pack	Null	mAP
Naive CNN	48.6	82.1	80.5	30.5	49.5	23.8	32.4	16.0	09.2	15.9	39.2
Naive TCN	24.9	80.2	66.4	27.4	31.9	14.7	14.8	13.9	2.0	15.4	29.4
MTL baseline	47.9	85.0	84.8	55.0	79.1	44.1	35.4	13.4	18.0	17.0	48.4
Tripnet	45.8	88.1	86.7	66.3	85.1	68.3	44.9	12.2	**22.5**	20.1	54.5
Attention Tripnet	**62.4**	89.4	89.4	**69.7**	**88.5**	84.3	48.5	**20.8**	21.4	**22.7**	60.2
Rendezvous	60.4	**90.5**	**89.5**	68.7	86.7	**87.8**	**50.4**	17.4	**30.5**	21.0	**60.7**

We also analyze the verb recognition per category. As shown in Table 6.7, it is observed that the proposed model correctly recognizes the most dominantly used verbs such as *retract, dissect, clip*, and *cut* over 70.0% of the time; this is likely because these have the strongest affinity with a particular instrument. The verb *coagulate* which is strongly correlated to bipolar and hook is recognized near 70.0%. The overall performance (60.7%) of the proposed model is higher than all the baselines.

For the moderately recognized verb, inspecting the confusion matrix in Figure 6.6 reveals the type of errors made by the model. For instance, *grasp* is 51% of the time mistaken as retract and 22% of the time mistaken as dissect. This strong overlap between the three verbs can only be differentiated by a careful observation of the tooltips. This may be difficult for even an experienced surgeon. Another confusion is observed with the verb *irrigate*. 49% and 48% score suggest tight overlap. Indeed, the distinguishing factor is mostly based on the fluid's dynamics (expulsion or suction) over time which may be better captured with temporal modeling. However, due to the class frequency, *aspirate* is predicted more often. The last confusion in the matrix is seen with the *null* verb as expected because it represents both idle and undefined verbs in the dataset.

6.4.2.4 Per-Class Target Detection Performance (AP_T)

The target is the most challenging to recognize component; this can be attributed to the instrument-centric nature of the triplet. Meanwhile, some targets are easier to recognize than others. in Table 6.8, *gallbladder* is recognized ≈ 90% correctly by the proposed model. Also, *specimen-bag* is correctly recognized 84% of the time. The *liver* is ≈ 60% correctly recognized.



True Label	Grasp	Retract	Dissect	Coagulate	Clip	Cut	Aspirate	Irrigate	Pack	Null
Grasp	20	51	22	2.9	0.75	0.48	2	0.48	0	0
Retract	0.35	95	1.2	0.47	0.4	0.29	1.5	0.74	0	0
Dissect	0.013	2.5	96	0.31	0.064	0	0.69	0.025	0	0
Coagulate	0.093	2.7	10	83	0.28	0	2.8	1	0	0
Clip	0	0.59	0.2	0	99	0.39	0	0	0	0
Cut	0	5	0	0	1	94	0	0	0	0
Aspirate	0.81	3.4	0	0.49	0	0	95	0	0	0
Irrigate	0	2.8	0	0	0	0	49	48	0	0
Pack	0	79	0	0	0	0	0	0	21	0
Null	4	36	40	2.7	6.6	1.6	6.7	2.2	0	0

Figure 6.6 – A confusion matrix for verb recognition.

Table 6.8 – Results Breakdown for Target (AP_T) Per-Class Detection. The target ids 1..14 correspond to *gallbladder, cystic-plate, cystic-duct, cystic-artery, cystic-pedicle, blood-vessel, fluid, abdominal-wall-cavity, liver, omentum, peritoneum, gut, specimen-bag, null* respectively.

Method	Target														mAP
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Naive CNN	84.2	14.8	26.3	18.7	14.3	03.6	32.4	10.1	49.8	35.2	**08.4**	08.4	69.3	15.9	28.3
Naive TCN	79.9	10.0	21.4	19.6	07.0	01.3	14.8	06.9	43.1	27.9	01.9	09.0	37.4	15.4	21.4
MTL baseline	85.1	12.2	29.3	18.6	06.5	06.4	30.6	09.8	55.7	35.8	02.1	08.4	71.1	17.5	28.2
Tripnet	87.0	**22.5**	29.7	21.9	04.7	15.0	42.9	32.3	57.5	36.7	02.0	**11.9**	74.1	20.9	33.2
Attention Tripnet	87.8	15.6	**37.1**	30.1	16.6	**26.5**	53.2	37.5	**59.8**	48.7	03.5	08.3	**85.6**	23.5	38.5
Rendezvous	**89.1**	15.3	35.2	**34.5**	**22.7**	11.4	**53.7**	**40.6**	59.3	46.6	04.3	**12.5**	84.0	**25.0**	38.3

The *fluid* is ≈ 54% correctly recognized.

We leverage a confusion matrix to throw more light on the moderately recognized targets. As shown in Figure 6.7, the *gallbladder* interferes in the recognition on many tubular structures and blood vessels. In laparoscopic surgery, the gallbladder is closely connected to most of these structures such that a dissection of the *cystic-artery/duct* can also be interpreted as dissecting the gallbladder-*neck/fundus*. Even confusing these with *blood-vessel* can be attributed to the fact the *cystic-artery* is also a blood-vessel. Its distinction in this task is due to its special place in laparoscopic cholecystectomy. Also, instruments interacting with *omentum* and *peritoneum* is likely touching some parts of the *gallbladder*.

Another observation in the confusion matrix is the *cystic-plate*. This structure is always in contact with the *gallbladder*. The *cystic-artery* and *cystic-duct* are very difficult to distinguish,

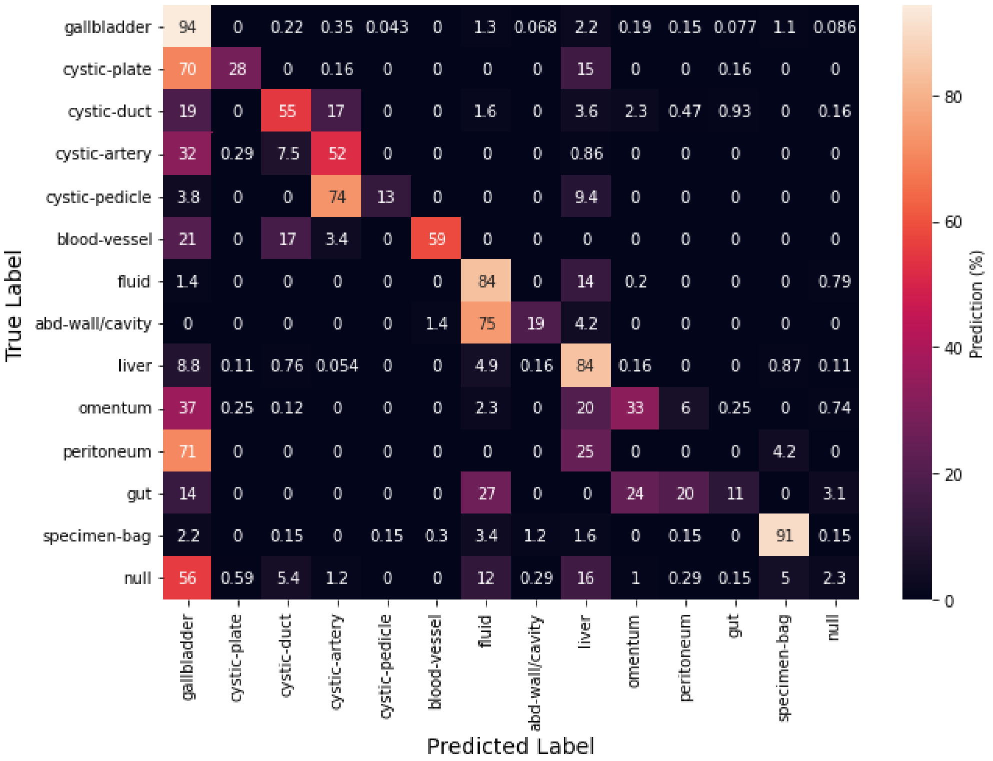


Figure 6.7 – A confusion matrix for target recognition.

with the major factor being their size and probably their location which is in proximity with each other. The *cystic-pedicle* is most times predicted as *cystic-artery*. This can be confusing since this *pedicle* is a yet-to-be dissected *cystic-artery* and *duct*. It is not surprising the overlap in the recognition of *abdominal wall/cavity* and *fluid*. While these two targets are clearly different, most times, *fluid/water* are actually aspirated from or irrigated on the *abdominal wall/cavity*. It is also surgically correct to annotated ⟨*irrigator, irrigate, fluid*⟩ as ⟨*irrigator, irrigate, abdominal-wall/cavity*⟩ whenever the anatomy being irrigated is not clearly defined. One of the most confusing targets is the *gut*; this target is coarsely defined to mean both *stomach, duodenum, small bowel*, etc., and so very difficult for a recognition model to focus. Finally, as expected, the *null* target is mostly incorrect since it encompasses no-target and other unconsidered targets.

6.4.2.5 Top-N Triplet Recognition Performance

Here, we present the top N predictions of the triplets (IVT combinations) to precisely assess the model performance which would have been more complex to assess individual classes due to a large number of the triplet categories. As shown in Table 6.9, the proposed model approximates its prediction to the most correct classes more than ≈ 95% of the time when top 20 predictions are considered. When the consideration window is reduced to 5, it still maintains a high 76% accuracy better than all the baseline methods. This suggests that this method would perform higher on a limited number of triplet classes.

Table 6.9 – Top N Accuracy of the triplet (I,V,T) predictions.

Method	Top-5	Top-10	Top-20
Naive CNN	67.0	80.0	90.2
Naive TCN	54.5	69.4	84.3
MTL baseline	70.2	80.2	89.5
Tripnet	70.5	81.9	91.4
Attention Tripnet	75.3	86.0	93.8
Rendezvous	**76.3**	**88.7**	**95.9**

6.4.2.6 Top-10 Triplets (AP$_{IVT}$) Association Performance

Here, we go into detail by presenting the result of the top 10 correctly detected triplets for the experimented models in Table 6.10. This helps to understand the individual strengths of the models in recognizing the tool-tissue interaction and as well provides room for analysis of surgical relevance of the model outputs.

We observed that all triplets predicted in the top results are clinically sensible, with none of the more unexpected instrument-verb or instrument-target pairings. Of importance, triplets with high surgical relevance in cholecystectomy procedure, i.e., ⟨*clipper, clips, cystic duct or artery*⟩ and ⟨*scissors, cut, cystic duct or artery*⟩, which are critical for safety monitoring, are better detected by the RDV than the SOTA. The triplet ⟨*grasper, grasp, specimen-bag*⟩ always appears in the top 2 even though its prevalence (6K) is not particularly high, compared to triplets such as ⟨*hook, dissect, gallbladder*⟩ (29K), ⟨*grasper, retract, liver*⟩ (13K), etc. This may be due to its consistent appearance in the workflow, usually towards the end; another factor could be the discernability of the *specimen-bag*. The proposed model recorded an average performance of 64.7% in its top 10 clearly above the Tripnet and its attention version. Remarkably, the entire top 10 for the RDV is recognized at an AP above 50%.

Interesting to note, the Attention Tripnet predicts in its top 10 rare but clinically important uses of surgical instrument, *irrigator*. This can be ambiguous, like the *irrigator* that is mostly used to aspirate or irrigate, but can as well be used to dissect in rare cases (⟨*irrigator, dissect, cystic-pedicle*⟩). Another detected rare case predicted by the same model includes ⟨*bipolar,*

Table 6.10 – Top-10 predicted Triplets for Instrument-Verb-Target Interaction).

Tripnet		Attention Tripnet		Rendezvous	
Triplet	AP	Triplet	AP	Triplet	AP
grasper,retract,gallbladder	77.30	grasper,grasp,specimen-bag	82.34	grasper,retract,gallbladder	85.34
grasper,grasp,specimen-bag	76.50	grasper,retract,gallbladder	78.41	grasper,grasp,specimen-bag	81.75
bipolar,coagulate,liver	67.39	bipolar,coagulate,liver	68.85	hook,dissect,gallbladder	75.90
hook,dissect,gallbladder	57.54	irrigator,dissect,cystic-pedicle	66.21	grasper,retract,liver	66.70
irrigator,aspirate,fluid	57.51	hook,dissect,gallbladder	63.22	bipolar,coagulate,liver	63.12
grasper,retract,liver	54.25	grasper,retract,liver	58.06	clipper,clip,cystic-duct	59.68
clipper,clip,cystic-artery	47.44	grasper,grasp,cystic-pedicle	55.35	bipolar,coagulate,blood-vessel	57.18
scissors,cut,cystic-duct	42.57	scissors,cut,cystic-artery	48.44	scissors,cut,cystic-artery	53.84
scissors,cut,cystic-artery	40.37	irrigator,aspirate,fluid	47.11	irrigator,aspirate,fluid	51.95
clipper,clip,cystic-duct	39.62	bipolar,coagulate,abdominal-wall-cavity	46.07	clipper,clip,cystic-artery	51.52
mean	56.05		61.41		64.70

Table 6.11 – Top-10 Predicted Instrument-Verb Association: The proposed models show a higher capability of detecting the top combinations that represent the most likely usage pattern of the individual instrument class, as well as the most clinical relevant instrument roles.

Tripnet		Attention Tripnet		Rendezvous	
Triplet	AP	Triplet	AP	Triplet	AP
bipolar,coagulate	88.71	grasper,retract	90.29	grasper,retract	90.51
grasper,retract	87.58	hook,dissect	87.18	hook,dissect	90.38
hook,dissect	86.88	bipolar,coagulate	78.17	bipolar,coagulate	88.05
scissors,cut	68.93	scissors,cut	77.99	scissors,cut	86.40
clipper,clip	67.10	clipper,clip	70.66	clipper,clip	82.65
irrigator,aspirate	57.51	irrigator,aspirate	57.10	irrigator,aspirate	51.95
grasper,grasp	23.54	grasper,grasp	37.10	grasper,grasp	48.97
irrigator,null-verb	16.28	irrigator,dissect	24.49	grasper,null-verb	28.91
clipper,null-verb	16.10	grasper,null-verb	20.81	scissors,null-verb	21.68
grasper,null-verb	12.47	irrigator,irrigate	16.47	irrigator,dissect	20.42
mean	52.51		56.03		60.99

coagulate, blood-vessel⟩. This suggests that the model effectively learned the surgical semantics of instrument usage even with small examples of peculiar classes.

We also analyze these top 10 predictions in lower division of considering the components association as presented in Section 6.4.2.7 and 6.4.2.8.

6.4.2.7 Top-10 Instrument-Verb (AP_{IV}) Association Performance

We tabulate the top 10 predicted instrument-verb classes and their recognition scores for the SOTA and the proposed models in Table 6.11. We observed that the top recognized instrument-verb combinations mostly represent the common usage pattern of the individual instrument class on multiple targets. In the three models, all the six instruments are represented with their associated frequent verbs within the top 7 predictions. It can be said that the models learn these actions by leveraging the instruments' activations which were class-wisely modeled. The CAGAM helped the Attention Tripnet and the RDV to outperform the Tripnet in all IV combinations detected across the three models. The attention mechanism helps the models to discover more important actions than idle instruments which are more clinically relevant. The RDV also outperforms the Attention Tripnet in their top 5 predictions. It also has the best performance in 8 out of the 9 common predicted labels by the two models. The RDV has the best upper and lower bound performances in the presented result. On average, the RDV is 8.48% and 4.96% better than Tripnet and Attention Tripnet respectively.

6.4.2.8 Top-10 Instrument-Target (AP_{IT}) Association Performance

Similarly, the top 10 predicted instrument-target classes are the most likely combination of the targets with the individual instrument class, as can be seen in Table 6.12 suggesting that the models capture the instruments usage pattern in laparoscopic cholecystectomy. Remarkably, we observe that the clinically most relevant situations which are the actions of the clipper and scissors on *cystic-artery* and *cystic-duct* are among the top detected instrument-target labels.

Table 6.12 – Top-10 Predicted Instrument-Target : The proposed models predict well the clinically most relevant situations, which are *clipping* and *cutting* of *cystic-artery* and *cystic-duct* among their top detected instrument-target labels.

Tripnet		Attention Tripnet		Rendezvous	
Triplet	AP	Triplet	AP	Triplet	AP
grasper,gallbladder	82.49	grasper,gallbladder	83.65	grasper,gallbladder	89.96
grasper,specimen-bag	76.50	grasper,specimen-bag	82.34	grasper,specimen-bag	81.75
bipolar,liver	67.09	bipolar,liver	68.37	hook,gallbladder	76.20
hook,gallbladder	57.96	hook,gallbladder	63.82	grasper,liver	66.22
irrigator,fluid	57.51	grasper,liver	54.26	bipolar,liver	62.45
clipper,cystic-artery	47.44	irrigator,cystic-pedicle	49.38	clipper,cystic-duct	59.68
grasper,liver	44.52	scissors,cystic-artery	48.44	bipolar,blood-vessel	57.18
scissors,cystic-duct	42.57	grasper,omentum	47.20	grasper,omentum	54.98
scissors,cystic-artery	40.37	irrigator,fluid	47.11	scissors,cystic-artery	53.84
clipper,cystic-duct	39.62	bipolar,abdominal-wall-cavity	46.07	irrigator,fluid	51.95
mean	55.61		59.06		65.42

This is interesting owing that these actions occur only once in a procedure over a very short interval. It is also interesting to observe that *null-target* which has a high-frequency count in the dataset but is less clinically relevant, is not detected in any of the model's top 10. Certain landmark actions in laparoscopic cholecystectomy workflow top the predictions. For example, usually the *grasper* would *grasp* the *gallbladder* during dissection and *grasp* the *specimen-bag* during gallbladder packaging. This could suggest that triplet information could be useful in surgical phase recognition as already exemplified by [Katić 2014]. On this set of interactions, the proposed model records higher performance compared to the baseline models.

Taking everything into account, the three compared models recorded a higher upper bound accuracy in detecting the instrument-verb than instrument-target and instrument-verb-target combinations but their performances drop much higher in the instrument-verb after the seven top classes. This is likely due to the first few verbs are directly associated with the individual instruments. Remarkably, the instrument-target performance stays high even after the top 10 classes (above 51% in RDV, 46% in Attention Tripnet, and 39% in Tripnet). This is likely because the interaction between the instruments and targets is more widespread as many instruments can interact with different targets using the same verbs, e.g: grasper can grasp any anatomy. *Irrigator* can *aspirate* any anatomy, etc. forcing the model to learn these individual anatomies. This explains why, even at low performance, the model attempt to pay similar attention to every target class more than verbs as seen in Tables 6.7 and 6.8.

Surprisingly, even with the AP_{IV} dropping below 20% after the top-10, the overall triplet AP_{IVT} remains higher (above 51% in RDV, 46% in Attention Tripnet and 39% in Tripnet) after the top 10 classes.

Finally, in addition to the top performance analysis, we present the full extent of the model's performance on all 100 classes using the AP box plots in Figure 6.8, showing upper and lower performance bounds for each model as well the spread around the mean score. The rectangular box represents the middle 50% of the score for each model also known as *interquartile range*. As can be seen from Figure 6.8, the proposed model maintains higher

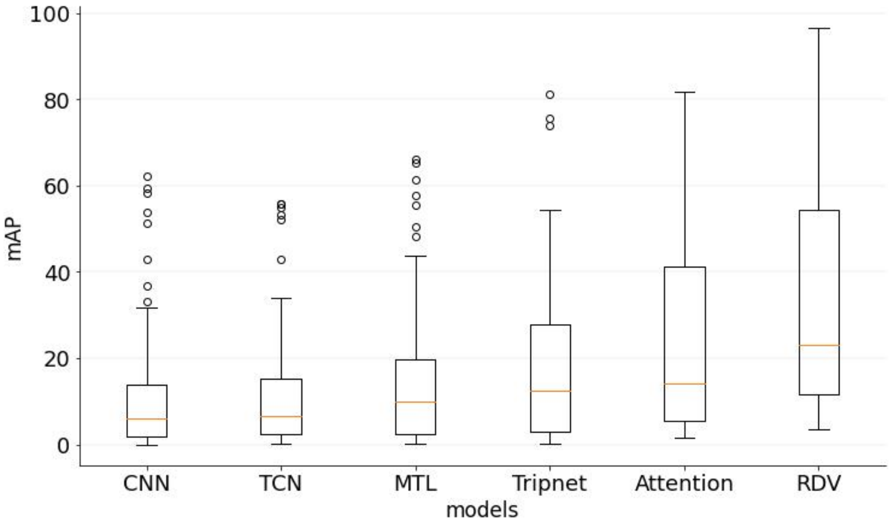


Figure 6.8 – Distribution of the model AP for the 100 triplet class predictions.

median and upper-quartile performance than the baselines. It also maintains higher upper-whiskers showing the extent of its performance distribution above the interquartile range.

6.4.2.9 Statistical Significance Analysis

Table 6.13 – The p-values obtained in Wilcoxon signed rank test of the proposed methods using the SOTA model (Attention Tripnet) as the alternative method. (*Lower p-value is preferred*).

	Tasks	p-value
Component Detection	AP_I	$p \approx 0.107$
	AP_V	$p \approx 0.130$
	AP_T	$p \approx 0.130$
Triplet Association	AP_{IV}	$p \ll 0.002$
	AP_{IT}	$p < 0.013$
	AP_{IVT}	$p \ll 0.002$

Using the Wilcoxon signed-rank test, we measure the statistical significance of the RDV model performance over the baseline. Here, we use the Attention Tripnet as the alternative method and define a null hypothesis (H_0) states that the difference between the proposed method and the alternative method follows a symmetric distribution around zero. If this hypothesis is true, then we discard the performance improvement of our proposed model as that which happens by chance. Following the same approach as in the previous chapter (Section 5.5.2.7), we sample $N = 30$ random batches of 100 consecutive frames from different test videos. The result of the statistical significance analysis for each task is tabulated in Table 6.13. Following the obtained p-values, we draw the following conclusions:

a. With p-values higher than the standard 0.05 significant level for AP_I, AP_V, and AP_T,

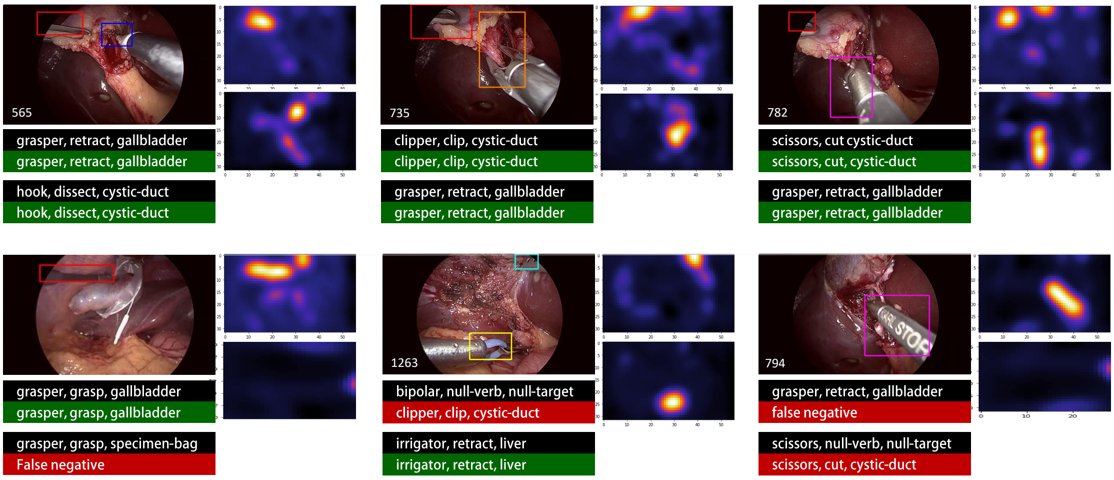

565
grasper, retract, gallbladder
grasper, retract, gallbladder
hook, dissect, cystic-duct
hook, dissect, cystic-duct
735
clipper, clip, cystic-duct
clipper, clip, cystic-duct
grasper, retract, gallbladder
grasper, retract, gallbladder
782
scissors, cut cystic-duct
scissors, cut, cystic-duct
grasper, retract, gallbladder
grasper, retract, gallbladder
grasper, grasp, gallbladder
grasper, grasp, gallbladder
grasper, grasp, specimen-bag
False negative
1263
bipolar, null-verb, null-target
clipper, clip, cystic-duct
irrigator, retract, liver
irrigator, retract, liver
794
grasper, retract, gallbladder
false negative
scissors, null-verb, null-target
scissors, cut, cystic-duct

Figure 6.9 – Qualitative results of Rendezvous model showing the triplet predictions, bounding boxes and heatmaps for the triplet detection. The localization bounding boxes are obtained from the WSL module of the proposed RDV model. Predicted and ground-truth triplets are displayed below each image: black = ground-truth, green = correct prediction, red = incorrect prediction. A missed triplet is marked as false negative and a false detection is marked as false positive (Best viewed in colour).

and being a two-tailed test, none of the models significantly outperformed each other in the three components detection. This is expected as both the proposed model and the alternative method share the same pipeline for triplet components detection.

b. On the triplet association performance, the proposed model's improvement is substantial to a 0.002 significant level for AP_{IV}, 0.01 significant level for AP_{IT}, and 0.002 significant level for AP_{IVT}. On the three metrics, the estimated p-values beat the standard 0.05 required to disprove the H_0.

Since the objective of the proposed model is to enhance the triplet association recognition, and having obtained meaningful p-values in the task metrics, we reject the null hypothesis H_0 at a confidence level of 5%, establishing an outstanding improvement over the alternative method.

6.4.3 Qualitative Results

6.4.3.1 Triplet Recognition with Weak Localization

The predicted class labels are obtained by applying a 0.5 threshold on the output probabilities of the proposed RDV model. These are presented in Figure 5.5, alongside the localization of the regions of action obtained from the weakly supervised learning (WSL) module of the network. This localization, depicted by bounding boxes overlaid on the image, shows the focus of the model when it makes a prediction, thereby providing insight into its rationale. Those results are solid arguments in favor of the model's ability for spatial reasoning when recognizing surgical actions. The semantic reasoning of the networks can be seen in the correct prediction of the triplet components association, and as well as predicting the correct number of triplet instances per frame. Looking at the class labels, the instruments are mostly correct. A closer look at these failed cases reveals that the incorrect prediction is very close to the groundtruth, for instance, ⟨*bipolar, null-verb, null-target*⟩ is actually aiming at *clipping* the *cystic-duct*, the ⟨*scissor, null-verb, null-target*⟩ is actually over the *cystic-duct* which only depth information could help the model understand whether a contact is made or not. For the false-negative cases, we observe that only a tiny portion of the instrument is visible. These qualitative results suggest that the model can be further exploited for action triplet detection and segmentation.

6.4.3.2 Qualitative Analysis of Top-5 Predicted Triplets

We also examine the top 5 prediction confidence of the proposed models compared to baselines on random frames (Figure 6.10). Fully correct predictions are signaled by the color green, while red indicates errors on all three components. Other colors indicate partially correct predictions. The performance here is judged by how much of the green labels are in the top level with all green at the first confident prediction being the best. Here, we can observe that the proposed model not only outperforms the baselines in most cases but also surrounds its predictions with the closely related triplets to the groundtruth. RDV model outperforms all the baselines each time, with the surgical actions correctly recognized each time within their top 5 predictions. Moreover, other actions in its top 5 have relevant components, showing the model's understanding of surgical actions by clustering triplets related to the performed actions. While Tripnet and Attention Tripnet also does very well in the top 5 predictions, it can be observed that RDV correct predictions are much closer to the top 2.

6.4.3.3 Supplementary Video

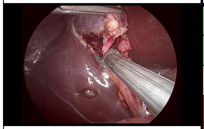
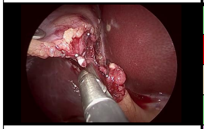
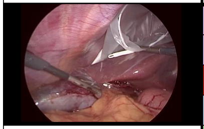
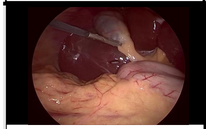
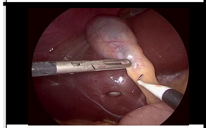
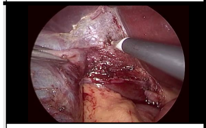
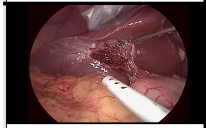
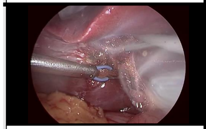
We all present a video showing some qualitative results of the RDV model for triplet prediction, action region localization, and visualization of the model's attention maps. The video is accessible online via https://youtu.be/d_yHdJtCa98

6.5 Conclusion

In this chapter, we presented a transformer-inspired method featuring a new hybrid attention mechanism that surpasses the state-of-the-art for the recognition of surgical actions as ⟨*instrument, verb, target*⟩ triplets. The proposed attention helps the model to semantically resolve triplet components relationships depicting the correct tool-tissue interaction in laparoscopic videos. This is achieved by leveraging multiple heads of both self and cross attentions on the component features.

We rigorously validated our performance claims on *CholecT50* which is a large-scale endoscopic video dataset for surgical action triplet recognition. We also discussed the benefits of the proposed methods in terms of clinical significance. Qualitative results suggest possible extensions to different tasks, including automated surgical report generation and spatial action segmentation.

While these initial results are encouraging, many challenges remain. One is the scalability on unseen triplets which may likely be tackled by zero-, one- or few-shot learning. Our results

Groundtruth	MTL Baseline	Tripnet	Attention Tripnet	Rendezvous
	clipper, clip, cystic-duct	grasper, retract, gallbladder	clipper, clip, cystic-duct	clipper, clip, cystic-duct
	grasper, retract, gallbladder	clipper, clip, cystic-duct	grasper, retract, gallbladder	grasper, retract, gallbladder
	clipper, clip, blood-vessel	clipper, clip, cystic-artery	clipper, clip, blood-vessel	clipper, clip, cystic-artery
	hook, retract, gallbladder	grasper, grasp, gallbladder	grasper, grasp, cystic-plate	clipper, clip, blood-vessel
clipper, clip, cystic-duct	bipolar, retract, liver	clipper, clip, blood-vessel	clipper, clip, cystic-plate	clipper, clip, cystic-plate
grasper, retract, gallbladder				
	grasper, retract, gallbladder	grasper, retract, gallbladder	scissors, cut, blood-vessels	grasper, retract, gallbladder
	hook, retract, gallbladder	grasper, grasp, liver	scissors, cut, cystic-artery	scissors, cut, cystic-duct
	grasper, retract, omentum	scissors, null-verb, null-target	grasper, retract, gallbladder	scissors, cut, cystic-artery
	scissors, cut, cystic-duct	scissors, cut, cystic-duct	scissors, cut, cystic-duct	scissors, null-verb, null-target
grasper, retract, gallbladder	grasper, null-verb, null-target	grasper, grasp, gallbladder	grasper, grasp, cystic-plate	grasper, grasp, gallbladder
Scissors, cut, cystic-duct				
	grasper, dissect, cystic-plate	grasper, grasp, specimen-bag	grasper, grasp, specimen-bag	grasper, grasp, specimen-bag
	grasper, null, null	grasper, null-verb, null-target	grasper, null-verb, null-target	grasper, grasp, gallbladder
	hook, dissect, gallbladder	grasper, grasp, gut	grasper, grasp, gallbladder	grasper, null-verb, null-target
	grasper, retract, gallbladder	grasper, retract, liver	grasper, retract, cystic-plate	grasper, retract, liver
grasper, grasp, specimen-bag	grasper, grasp, specimen-bag	grasper, grasp, liver	grasper, grasp, liver	grasper, pack, gallbladder
grasper, grasp, gallbladder				
	grasper, retract, gallbladder	grasper, retract, gallbladder	grasper, retract, gallbladder	grasper, retract, gallbladder
	grasper, null-verb, null-target	grasper, retract, peritoneum	grasper, retract, liver	grasper, retract, liver
	irrigator, retract, liver	grasper, grasp, gallbladder	grasper, retract, peritoneum	grasper, retract, cystic-plate
	grasper, retract, omentum	grasper, null-verb, null-target	grasper, retract, omentum	grasper, null-verb, null-target
grasper, grasp, gallbladder	hook, retract, gallbladder	irrigator, retract, omentum	grasper, grasp, gallbladder	grasper, grasp, gallbladder
grasper, retract, gallbladder				
	hook, dissect, gallbladder	grasper, retract, gallbladder	grasper, retract, peritoneum	grasper, retract, gallbladder
	grasper, grasp, gallbladder	hook, null-verb, null-target	grasper, retract, gallbladder	hook, null-verb, null-target
	hook, retract, gallbladder	hook, retract, liver	grasper, retract, cystic-pedicle	grasper, retract, peritoneum
	grasper, retract, peritoneum	hook, retract, gallbladder	hook, retract, liver	hook, dissect, gallbladder
grasper, retract, gallbladder	grasper, retract, omentum	grasper, grasp, gallbladder	hook, dissect, omentum	grasper, grasp, gallbladder
hook, dissect, gallbladder				
	hook, dissect, gallbladder	grasper, retract, gallbladder	hook, dissect, gallbladder	hook, dissect, gallbladder
	hook, coagulate, liver	hook, dissect, gallbladder	grasper, retract, gallbladder	grasper, retract, gallbladder
	hook, retract, gallbladder	grasper, retract, cystic-plate	hook, retract, liver	grasper, retract, cystic-plate
	hook, coagulate, gallbladder	hook, coagulate, liver	grasper, retract, liver	hook, dissect, cystic-plate
hook, coagulate, liver	bipolar, retract, liver	hook, coagulate, cystic-duct	grasper, dissect, cystic-plate	hook, coagulate, liver
grasper, retract, gallbladder				
	grasper, retract, liver	grasper, retract, liver	grasper, grasp, liver	grasper, retract, liver
	irrigator, retract, liver	irrigator, aspirate, fluid	grasper, retract, liver	irrigator, aspirate, fluid
	irrigator, null-verb, null-target	grasper, grasp, liver	irrigator, irrigate, liver	irrigator, retract, liver
	irrigator, irrigate, liver	irrigator, retract, liver	grasper, retract, cystic-pedicle	irrigator, irrigate, liver
irrigator, aspirate, fluid	irrigator, aspirate, fluid	irrigator, retract, omentum	irrigate, aspirate, fluid	irrigator, null-verb, null-target
grasper, retract, liver				
	bipolar, coagulate, liver	bipolar, coagulate, liver	grasper, grasp, liver	bipolar, coagulate, liver
	irrigator, retract, gallbladder	grasper, grasp, liver	bipolar, coagulate, liver	bipolar, coagulate, gallbladder
	irrigator, null-verb, null-target	grasper, retract, liver	irrigator, null-verb, null-target	grasper, retract, liver
	hook, coagulate, gallbladder	irrigator, retract, liver	grasper, retract, liver	irrigator, retract, liver
bipolar, coagulate, liver	bipolar, null-verb, null-target	irrigator, null-verb, null-target	irrigator, retract, liver	grasper, grasp, liver
irrigator, retract, liver				
grasper, retract, liver				

🟩	🟦	🟨	🟪	🟥
Correct complete triplet (IVT)	Correct pair (IT) instrument-target	Correct pair (IV) instrument-verb	Correct instrument (I)	Incorrect prediction

Figure 6.10 – Qualitative results showing the top-5 triplet predictions for the best performing baselines, and the proposed model (Best viewed in colour).

on rare triplets already hint at promising prospects for this approach. Inference speed is another challenge: increasing the number of layers generally drives up the performance, but is computationally very costly. Implementing a lightweight Rendezvous would help alleviate some of these costs. Further scaling a lighter version of Rendezvous with deeper decoding layers would likely lead to higher performance as already hinted by our promising results in this regard.

Applications, Conclusion, and Perspectives

Part III

7 Potential Clinical Applications in the OR

If you are not embarrassed by the product when you launch,
you've launched too late.
– Reid Hoffman

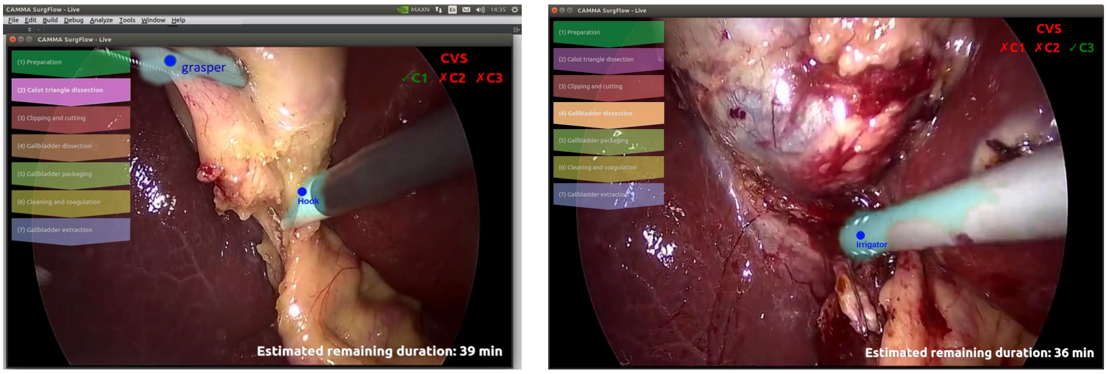

CAMMA SurgFlow - Live
File Edit Build Debug Analyze Tools Window Help
MAXN 14:35
CAMMA SurgFlow - Live
(1) Preparation
(2) Calot triangle dissection
(3) Clipping and cutting
(4) Gallbladder dissection
(5) Gallbladder packaging
(6) Cleaning and coagulation
(7) Gallbladder extraction
grasper
CVS
✓C1 ✗C2 ✗C3
Hook
Estimated remaining duration: 39 min
CAMMA SurgFlow - Live
(1) Preparation
(2) Calot triangle dissection
(3) Clipping and cutting
(4) Gallbladder dissection
(5) Gallbladder packaging
(6) Cleaning and coagulation
(7) Gallbladder extraction
CVS
✗C1 ✗C2 ✓C3
Irrigator
Estimated remaining duration: 36 min

Figure 7.1 – An online demonstration of surgical tool detection and overlay of the localization heatmap using SurgFlow System developed at CAMMA Research Lab.

Chapter Summary

The capability to automatically detect and track surgical instruments as well as recognize the instrument-tissue interaction in laparoscopic videos plays an important role in the development of CAI systems in the OR. Specifically, such a system would open up the possibility for many pre-operative, intra-operative, and post-operative applications, useful in surgical education, and foster further research in surgical data science. In this chapter, we first discuss the valid demonstration of the developed methods, followed by their potential applications in the OR, and conclude by highlighting how the work could facilitate more research in the field of surgical data science.

7.1 Demonstration of Methods

The first illustration of the benefits of the proposed model is the integration of the weakly supervised tool localization and tracking model in a demo software known as *SurgFlow*. This software is developed at CAMMA Lab for the demonstration of deep learning solutions in OR decision making. This system which is a C++ back-end software allows the integration of models developed in deep learning frameworks such as TensorFlow (TF) and Python. Developed models are frozen and imported using TF C++ libraries for their inferences. Currently, SurgFlow supports both online and offline inferencing.

Along with several other models, the weakly supervised tracking model has been successfully deployed in the SurgFlow and demonstrated in many seminars, lab tours, and live shows. In SurgFlow, the ConvLSTM tracking model is used to demonstrated tool presence detection, tool position localization by both center points and bounding coordinates, tool segmentation by the overlay of the Lh-maps, and tool tracking as shown in Figure 7.1. The demonstration by

SurgFlow shows that the output of the model can be integrated into a CAI system to provide assisted intervention in terms of early warning signals when a tool is approaching a no-go area in the patient's body [Madani 2021]. This would need to work with additional information marking out the go and no-go areas among the anatomies. By providing a consistent trajectory, the outputs are usable in generating post-surgery reports especially about the number of times the instruments such as hook have dissected in the safe and unsafe zones during a laparoscopic procedure.

7.2 Preoperative Application

7.2.1 Pre-operative Remaining Time Duration

To support routine surgical activities, it is imperative to display certain information at the most appropriate time. Information about the state of an ongoing surgery can help the OR staff prepare the next patients for surgery. Generating and sending a reliable notification to the OR staff about the remaining duration of an ongoing surgery is very crucial in OR since the next patient's anesthesia needs to be timely administered. A timely OR scheduling would improve patient outcome, reduce idle time and promote efficient utilization of the OR. While the tool and tool-activity recognition systems would be mostly used intra-operatively, the signals from the systems on one patient can be useful for the next patient's surgery preparation and anesthesia. This information can be formulated based on some detected signals such as the presence of specimen-bag which is usually used towards the end of the procedure. More specifically, the action triplet models can be more meticulous in detail here. A successive detection of packaging of the gallbladder in a specimen-bag, and their onward extraction would nicely present a constantly updating surgery remaining time. Aside from directly automating some of these signals for OR notifications, deep learning systems designed for the estimation of the remaining surgery duration (RSD) could benefit from the instrument and triplet information to improve their predictions.

7.3 Intraoperative Application

The bulk of the benefits of the developed systems in this thesis could be found in intraoperative applications. This is because they provide analysis of surgical activities to provide context-aware assistance and decision support in real-time surgery as would be discussed further.

7.3.1 Online Decision Support

Automatic recognition of surgical instruments and their activities provides information such as the used instruments, their locations, motion, actions, and the organs they are in contact with. Such information when combined with domain knowledge can be clinically formulated to guide the surgical decision-making process. AI systems based on the proposed methods would have a comprehensive understanding of the activities at each time step. It can as well replay performed activities for re-analysis and re-assessment of surgical situations and actionable feedback. With the expressive nature of action triplet recognition, it becomes easier to automate efficient browsing of related situations from a surgical database. This is useful for a real-time simulation of specific situations. Few-steps ahead simulation of critical situations could also be possible and helpful in early assistance intervention.

7.3.2 Safety Monitoring

Checklists of a safety monitoring system can be automated relying on feedback from an action triplet recognition system. During clipping and cutting phase in laparoscopic cholecystectomy, this could be the correct detection of the triplets that can ascertain the achievement of critical view of safety such *hook* dissecting *cystic-artery/duct/pedicle*, and *clipper* clipping *cystic-artery/duct*, without those, the use of *scissors* would be flagged.

An anomaly detection system can combine with action triplets recognition to predict and prevent prevailing complications. Such systems could also offer recommendations referencing past workflow.

Since the triplet model also recognizes the anatomical structures, the information could be utilized for validation of anatomies in the face of a cluster, occlusion, and visual illusion.

7.3.3 Interactive User Interface

One of the most fascinating potentials of the proposed methods is the provision of the right information for the creation of adaptive user-interface in the OR. The interface adaptation would be based on the surrounding context. This can be likened to a football video assistance referee (VAR), who tends to observe all actions on a football pitch from an adaptive monitor screen. Just like in football sports, the AI would easily and always detect the center of attraction and focus the camera for better coverage. Presently in the OR, the video screen is not yet equipped with such technologies. The endoscopic camera focus is manually controlled by an assistant surgeon. Since the camera field-of-view is limited, a recognition system, having the instrument position and motion information, could help to select a focal point, which will be the region of triplet interaction, at every point in a procedure. An automated system leveraging the triplet prediction and tool location details, could, at required intervals, selectively zoom the visualization of tiny structures especially when anatomy validation is necessary. It can also increase system brightness/contrast to bypass the effect of smokes from coagulation instruments, etc. This feature would enable CAI systems to selectively provide the most needed user interface in laparoscopy.

7.3.4 Automated Notification

The detected coordinates, time-based trajectories, and actions of surgical instruments are useful signals for automating timely notifications during surgery. Crucial warning signals may be formulated from monitoring the trajectories of certain instruments against some unsafe dissection areas in the body and analyzing how often an instrument, such as a hook, dissects unsafe regions. Triplet information will help automate warnings targeting the wrong instruments usage pattern for critical events. Several flagging of unconventional use of instruments, such as the using scissors for calot triangle dissection, would motivate the surgeon to switch to a more appropriate instrument such as a hook.

Additionally, notification systems can leverage detected tools and triplets information, as their events trigger, to remind surgeons of being mindful of safety checks at a certain phase of a procedure. The systems would give reliable feedback on the actions performed during this crucial period to validate the safety checks.

Also, a real-time notification system could leverage prolong detection of irrigator from the tool detection model, or more concisely irrigator, aspirate, blood from the triplet model to send signals to the senior surgeons for OR assistance.

7.4 Postoperative Application

7.4.1 Surgical Report Generation

The recordings of surgical procedures are usually stored in a video database. These videos are consulted post-operatively for many purposes. Firstly, they are used to generate surgical reports about a performed surgery. The operative report can be used for procedure evaluation, legal proceedings, recommendation, and documentation. Also, surgical reports contain valuable knowledge which can be utilized for various applications. Manually generating these reports can be tedious. Whereas automatic recognition of surgical activities could help to facilitate the reporting of surgical events and steps taken to avoid complications. The system can be designed to detect specific surgical report narratives that are essential for a particular procedure, such as CVS achievements, removal of a cyst, and other factors that could affect the normal workflow thereby leading to a longer procedure. The fine-grained nature of the triplets can support an adaptive system to generate reports tailored to a specific purpose, such as legal proceedings, skill analysis, action review, decision evaluation, and so on.

7.4.2 Offline Decision Evaluation

While it is most beneficial to evaluate surgical decisions intra-operatively, some decisions can also be on the stored surgical videos after the procedure. This evaluation is mostly done in comparison with the surgical outcome in patients. Since evaluating every surgical video by the senior surgeons and management could be time-consuming, there is a tendency to rely solely on the surgical report for this purpose. The automatic recognition of surgical activities could help to automatically evaluate the surgical decisions by comparing the taken surgical actions in videos with set standards and automatically ranking the series of surgical decisions taken at every crucial surgical step. The recognition systems developed in this thesis would help to provide all the surgical actions taken at every specified surgical phase or timeline for their evaluation.

7.4.3 Surgical Skill Evaluation

Just like decision evaluation, surgeons' skills can also be evaluated for certification and recommendation for further training. These skills are mostly assessed by the instrument usage pattern and sequence of actions taken to achieve a particular surgical task. With the developed models detecting instruments as well as tracking their trajectory, it would be easier to understand the professionalism of surgeons in instrument handling. The triplet recognition would help in judging the combination of sequence of actions taken to arrive at a given task. These detection and recognition models can be integrated into systems designed for surgical skill certification.

7.4.4 Surgical Skill Training

One of the many goals of documenting surgical videos is for training purposes. And with a large bank of video data, it would be difficult to browse through the videos to obtain ones with a specific surgical condition. Consequently, some surgical video databases have become redundant. Even on a single video, manually scrolling through a video timeline for a particular activity, such as gallbladder dissection, could be tedious and inexact with a possibility of a miss, leading to a repeated search. All these efforts could be ameliorated through automatic video indexing that can facilitate video retrieval. With the automation, even multiple indexes could be maintained for a particular video as well as multiple tagging on a particular surgical phase depicting several conditions such as bleeding, coagulation, and dissection of adhesion all in one surgical phase but may be consulted for different learning purposes. These multiple indexes could be tagged to the activities detected by the recognition systems such as triplet information, tool track duration, tool co-occurrence, etc. And so, the deep learning model can be used to index videos as well as retrieve videos based on key activities occurring in the videos for surgical training. It can also help in sorting and grouping videos based on their similarities or the presence of uncommon surgical situations in the videos.

The tool and activity recognition system can be integrated into surgical education software with automated instructions that could help a young surgeon practice some surgical skills and get automatic feedback.

7.5 Facilitating Research on Surgical Data Science

The first part of this thesis provides a bedrock for more research on weak supervision in this field. This will encourage the development of more models for complex tasks which are initially limited by the unavailability of spatial labels which are difficult to generate. The effect will be full utilization of large weakly-label medical data for the provision of CAI solutions in the OR. Another research impact of the proposed method is the facilitation of data annotation. The weakly supervised system can be used to integrated into a data labeling software to initialize the annotation such as bounding boxes and pixel segmentation mask. The model can also be used to initialize research on weakly-supervised segmentation.

Action triplet recognition would open up a new and novel approach to modeling and recognizing surgical activities in the community. Together with the associated dataset, this will create a new trend of research that will have more medical applications as already discussed. Such clinically relevant applications would attract more funding facilitating research in the community.

7.6 Conclusion

We have presented "SurgFlow", which is a feasibility demo of our proposed methods for clinical translation. We have also outlined the potential applications of our work in surgery, pre-operatively, intra-operatively, and post-operatively. Lastly, we highlighted how the work done in this thesis could help to facilitate more research in this domain.

8 Conclusion and Recommendation for Future Work

> *The future is completely open,*
> *and we are writing it moment to moment.*
> – Pema Ch odr on

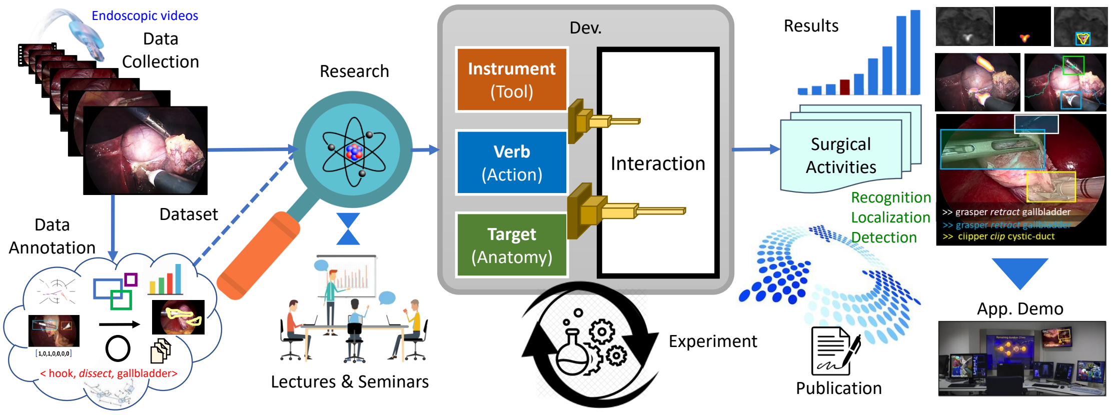


Figure 8.1 – Thesis at a glance.

Chapter Summary

We conclude the thesis in this chapter by presenting the summary of this work. We examine how the research tackles the highlighted aim of the thesis, how it achieves the stated objectives, and the solutions it offers to the stated problem and their limitations. We also highlight the significance and implications of the thesis findings. And conclude by suggesting avenues for future research.

8.1 Summary and Conclusion

The desire to have safe and efficient surgery, which is being approached with rapid development and the introduction of high-tech surgical systems, leads to increasing complexity in the OR. This motivates the need to optimize and support surgical workflow in many respects and particularly relying on the available and increasing amount of data captured by the information systems. The clinical utility for these data has not been sufficiently harnessed due to the coarse nature of the activities modeled by the previous and existing recognition systems, e.g. surgical phases. Sequel to these unresolved needs, this thesis addresses a novel research field; namely, the problem of recognizing surgical activities from videos to effectively understand the surgical semantics of every instance of tool-tissue interaction in the OR. Apart from capturing and modeling these activities at a fine-grained level of granularity, i.e., recognizing unit actions at every time step, the activities are also detailed. They provide comprehensive information about the recognized actions, including the instruments performing the actions and the underlying target anatomies receiving the actions.

This detailed activity recognition is formalized as triplet recognition ⟨*instrument, verb, target*⟩ representing the activities of the instruments. Being instrument-centric is one of the factors motivating the first part of the thesis, which tackles the detection and tracking of surgical instruments in laparoscopic videos. Given the unavailability of spatially annotated data and the cost of generating such spatial labels for the training of instrument localization models, we propose a new deep learning method that prevails over the lack of spatial annotations with weak supervision using only binary presence labels. Leveraging temporal information in the video data with a Convolutional LSTM, we also perform tracking, yet no spatial annotation is required. This answers the complementary research question seeking to utilize only the easier-to-generate binary presence labels for the training of a model for much higher and complex tasks such as spatial localization and motion tracking. One limitation of the proposed approach is that it is not designed to detect and track multiple instances of the same instrument. This is the case for only the instrument *grasper* in a laparoscopic procedure, which can have up to three instances of *grasper* depending on the phase of the procedure, especially at the *gallbladder packaging* phase. The binary presence label of this instrument is marked positive whenever at least one instance of the said instrument is visible. Thus, additional information or strategy would be needed in this situation to ascertain the number of instances per *grasper* in a surgical video frame. Notwithstanding, this study sets a bedrock for more deep learning research to utilize larger surgical data for their training while circumventing the lack of spatially labeled datasets and the tedious effort involved in generating spatial data annotation. Most importantly, the weakly supervised instrument detection builds a foundation for detailed and fine-grained activity recognition, which can benefit from the instrument position information, especially enabling such research on datasets without spatial labels.

The proceeding research tackles the surgical action triplet recognition. Action triplet is particularly interesting due to several characteristics, which make the recognition task non-trivial and unique. One of them is that visibility is not the only key, as in other classical vision tasks, to base a detection modeling. Also needed is an observed involvement in a tool-tissue interaction, manipulated using an instrument. For instrument-centric action triplets, visible anatomies are not always the targets until they are involved in a tool-tissue interaction. We tackle this problem in two ways: (1) with a class activation guide (CAG) module which learns to condition model feature encoding by the appearance of the instruments captured in their heatmap activation, (2) by advancing the CAG method with attention modeling to properly highlight interest features that can contribute to the correct recognition of the verb and target components of triplets while suppressing the irrelevant features. This approach is called the class activation guided attention mechanism (CAGAM) in this work. The two deep learning models, Tripnet and Attention Tripnet, developed using the two novel methods respectively, show significant contributions towards the detection of the correct verbs and targets involved in the tool-tissue interaction amid possible others. Results presented on these methods justify that our novel formulation of spatial reasoning is useful in guiding the recognition of the activities of the surgical instruments. They provide answers to understanding the details of surgical activities which are needed for the development of context-aware systems, especially one that would provide automated warnings and informative signals about the anatomy and action of instruments in a designated region of the body, most probably the unsafe dissection zone. One observed limitation of these approaches is that they are not able to localize the surgical targets. With the surgical instrument already weakly localized in the previous method, obtaining the spatial position of the target anatomies in the same model could provide supplementary information for a better feedback formulation in safety monitoring. An attempt to tackle target localization by weak supervision has not yielded a meaningful result. This is because the binary labels provided for the targets are not determined by only the presence of the anatomies considered targets in a given frame, as in the case of the instruments. Hence, weakly localizing the anatomical targets from such instrument-centric binary labels becomes very challenging. Notwithstanding, the recognition of the targets without their localization provides sufficient information, in this study, for the recognition of tool-tissue interaction in the considered procedure.

Surgical action triplet recognition does not end with recognizing the individual components of the triplets, it also involves the correct association of those components to form complete triplets, as there can be multiple instances per frame. This presents a whole new perspective to the task, as this association is non-linear and very challenging. In the first instance, the three components are each multi-label, making their association a tripartite matching problem, one described as NP-hard: a complex optimization task. Additionally, triplets exhibit a high level of semantic reasoning, which can be overlapping most of the time. For instance, the usage of an instrument (verb) depends on a surgeon's intention of use: *grasp, retract* or *dissect* using the same instrument on the same target. Distinguishing these actions would require careful observation of tooltips, and most probably its effect on the underlying target. Another characteristic is multiple instruments interacting with the same target, such as *hook, grasper, irrigator,* and *bipolar* dissecting the *gallbladder*. Last but not least is multiple targets interacting with the same instrument class: grasper grasping both *blood-vessels* and *cystic-artery*, two *graspers* grasping *specimen-bag* and *gallbladder*, or two *graspers* grasping the same anatomy. Besides the overlapping nature of these triplets, multiple triplets can occur at the same time. We first tackle the multiplicity and overlap by proposing to associate all the components in a 3D space while their interactions with each other would be learned. Improving on this, we leverage a long-range attention mechanism to efficiently capture these interactions in a novel method that is inspired by the transformers. This method, which we called Rendezvous (RDV), a meeting of attentions, leverages multiple heads of self-attention for initial action understanding, and cross-attentions for component entity-relationship modeling, to learn the tool-tissue interaction, solving the triplet association problem. This module is called multi-head of mixed attention (MHMA). It introduces a new way of modeling long-range attention by utilizing learned class-wise representative features without recurrence. The encouraging results show the potentials of implementing an activity recognition system based on our proposed method for real-time OR applications. With this, recognizing tool-tissue interactions captured as triplets provides, so far, the best clinical utility of surgical activity recognition, helpful in reducing surgeon's cognitive workload, and fostering safety and efficiency in the considered surgical procedure.

In this thesis, we also generated datasets to facilitate our research in the considered tasks. First, with the help of the CAMMA team, we generated a large dataset of spatial bounding box annotations from 5 laparoscopic videos, which are set aside for the evaluations of models trained in a weakly supervised manner for detection and tracking of surgical instruments. With special thanks to our clinical collaborators, we also generated two large action triplet datasets: CholecT40 and CholecT50, with the latter improving the former. This fine-grained dataset, which provides labels for the triplets, is the first and biggest in the domain and would be helpful to further research in the field. The CholecT50 has been used for the Endoscopic Vision grand challenge on action triplet recognition. While these triplet datasets do not contain spatial labels for action detection, they would encourage more intuitive exploration of deep learning methods, such as weak supervision, for the action localization. The binary labeled nature of the datasets would encourage the development of methods that would circumvent the tedious efforts of generating spatial annotations for images/videos. This will reduce research cost and time.

With high-profile potential applications such as safety monitoring, skill evaluation, and objective reporting, our methods, combined with the datasets; Cholec80 spatial labels, CholecT40, and CholecT50, bring considerable value to the field of surgical data science, particularly, on tool and activity understanding.

8.2 Recommendation and Future Work

Weakly Supervised Segmentation

We have tackled instrument recognition, localization, and tracking with state-of-the-art performance. Future studies may want to extend this to tracking by segmentation which would provide more precise localization and motion estimation of surgical instruments.

Few-shot Learning

For the surgical action triplet recognition, we developed deep learning methods that recognize and localizes the regions of the actions. While the initial results are encouraging, it is known that most triplet classes are largely unbalanced. Lots of these classes are super-classed to obtain a reasonable number of classes for model training. In a real-life application, there may be a need to handle some of these grouped triplets differently. Since 50 videos may not capture all triplet cases in the laparoscopic procedure, scalability on unseen classes is another interesting area of research. This can be tackled using zero-, one- or few-shot learning. Our effort at disentangling the triplets into their components and building a recognition model to first understand these components already lays a building block to few-shot learning. Furthermore, our results on rare triplets hint at promising prospects for this approach.

Active Learning

Currently, state-of-the-art methods are built on deep learning techniques. Research has shown that deep learning models improve their performance with more training data. Since fine-grained data annotation is non-trivial - expert knowledge is often needed - future works may want to exploit a large volume of unlabeled data using Active Learning techniques. The benefit is that a model, which can learn from the few annotated data already provided in CholecT40 and CholecT50, would also benefit from online expert annotations during training. This would also help to correct noisy annotations and improve the model prediction confidence.

Self Supervision

Using a fully supervised method for action triplet recognition at this stage is necessary to establish research in this field. In the future, it will be interesting to develop methods that explore unlabeled data for this purpose. Active learning method, on its own, would also require some element of expert involvement during training which could be tedious and unavailable. To better utilize unlabeled data that are usually left unused, it would be interesting to explore self-supervised methods for action triplet recognition. Self-supervision, which is concerned with learning semantically meaningful representations from unlabeled data, would allow a deep learning model to generate its training labels on the fly while solving some tasks that do not require human annotations. The representations learned in this way may be useful in understanding the surgical activities to both finer and detailed levels. Exploiting large unlabeled surgical data with self-supervision would improve the results for action triplet recognition. So far, the localization of regions of action in this task has been by weak supervision. Hence, my recommendation in this regard would be to fashion a new self-supervised approach that can learn both spatial and temporal useful representations.

Target Detection

The results of work on surgical action triplet show that recognizing the target is the most challenging. Part of the reason may be the instrument-centric property of the triplet. Looking at the triplet components, while the detection of surgical instruments is widely researched, recognition of surgical actions/verbs is fairly explored, thus providing insights for their modeling. However, existing research on anatomy detection tackles all visible anatomies. It becomes interesting to focus research on recognizing anatomies involved in a tool-tissue interaction (targets) from surgical videos. Localizing or segmenting these targets would be very useful to better understand the regions of interaction, as their areas of intersection with the tools bounding boxes or masks. This would in turn improve the triplet recognition performance.

Temporal Modeling

The results of some of our methods on action triplet recognition, especially the ones analyzed using confusion matrix, show that some verb classes cannot be effectively discriminated base on a single frame observation. As an example, differentiating *aspirate* from *irrigate* when using an irrigation and suction device would greatly depend on the temporal dynamics of the *fluid*. Also, the *grasp* and *retract* verbs of the *grasper* exhibits some temporal consideration. Generally, recognizing the verb component of the triplet can be improved by leveraging temporal information.

Triplet Tracking and Anticipation

Leveraging the temporal information in the video dataset, would not only improve the triplet performance but also may help in learning to track action triplet instances. This would be useful for action anticipation which may be beneficial in the context-aware assistance, particularly guiding against unintended actions.

List of Publications

International journals

Nwoye, Chinedu Innocent, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, *Weakly Supervised Convolutional LSTM Approach for Tool Tracking in Laparoscopic Videos*, International Journal of Computer Assisted Radiology and Surgery (IJCARS), 14(6), 1059-1067, 2019. https://youtu.be/vnMwlS5tvHE (Supplementary video 1),
https://youtu.be/SNhd1yzOe50 (Supplementary video 2)

Nwoye, Chinedu Innocent, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, *Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos*, Accepted at Medical Image Analysis (MedIA), 2021. https://youtu.be/d_yHdJtCa98 (Supplementary video)

International conferences with proceedings

Nwoye, Chinedu Innocent, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, *Weakly Supervised Convolutional LSTM Approach for Tool Tracking in Laparoscopic Videos*, International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), 2019. **Long oral presentation**. **Won the IPCAI Audience Choice Award for Best Paper Presentation**.

Nwoye, Chinedu Innocent, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, *Recognition of Instrument-Tissue Interactions in Endoscopic Videos via Action Triplets*, Medical Image Computing and Computer-Assisted Intervention (MICCAI), 364-374, 2020, **Long oral presentation**.

Others

Wagner, Martin, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky,.., Nwoye, Chinedu Innocent et al. *Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark.* arXiv preprint arXiv:2109.14956 (2021).

Luengo, Imanol, Maria Grammatikopoulou, Rahim Mohammadi, Chris Walsh, Chinedu Innocent Nwoye, Deepak Alapatt, Nicolas Padoy et al. *2020 CATARACTS Semantic Segmentation Challenge.* arXiv preprint arXiv:2110.10965 (2021).

Appendices Part IV

A Endoscopic Vision Challenges

Seek out strategic alliances,
they are essential to growth and provide resistance to bigger competition.
– Richard Branson

We are happy to announce

CholecTriplet 2021

An Endoscopic Vision challenge aimed at the recognition of surgical actions as a series of triplets of <*instrument, verb, target*>

Featuring

- ✓ CholecT50 action triplet dataset
- ✓ Colab code blog for quick start
- ✓ T50 slack community of ideas
- ✓ Grand prizes for the winning teams

March 15 – October 1, 2021

Visit and register @ http://cholectriplet2021.grand-challenge.org

MICCAI2021
CAMMA
>> grasper *retract* gallbladder
>> grasper *retract* gallbladder
>> clipper *clip* cystic-duct
>> grasper *retract* gallbladder
>> bipolar *coagulate* cystic-artery
ENDOSCOPIC
VISION
CHALLENGE

Figure A.1 – An ads inviting participants to the CholecTriplet 2021 challenge.

Chapter Summary

A.1 Challenge Organized

A.1.1 MICCAI EndoVis 2021 on Surgical Action Triplet Recognition

Excited about the future of surgical action triplet recognition, we organize an endoscopic vision challenge named: *Surgical Action Triplet Recognition Challenge 2021*, with an acronym **CholecTriplet2021**. Representing the CAMMA Research Lab, University of Strasbourg, France, a team of four including Chinedu Nwoye, Deepak Alapatt, Armine Vardazaryan, and Nicolas Padoy, form the organizing committee for the CholecTriplet challenge. We host the challenge on the EndoVis platform (accessible via https://cholectriplet2021.grand-challenge.org). The challenge is part of the Endoscopic Vision (EndoVis) 2021: a grand challenge that houses six different sub-challenges focusing on different aspects of surgical data and surgical workflow analysis. Among the other sub-challenges, CholecTriplet offers a new perspective to fine-grained surgical activities recognition. It formulates surgical activities as ⟨*instrument, verb, target*⟩ combinations.

Though, efforts have been made in previous challenges to model surgical activities in the OR, such as the m2cai workflow [1] which tackles phase recognition and surgical workflow and skill analysis [2], which recognizes fine-grained activities as single action verbs, CholecTriplet challenge offers a more detailed workflow analysis. It improves on the existing challenges by recognizing finer actions within the phases and enriching the recognized actions with information about the operating instruments and the operated anatomies. This provides a more comprehensive understanding of tool-tissue interaction in surgical videos for their optimal clinical utility.

A.1.1.1 Task

The challenge presents a single task that focuses on the recognition of surgical action triplets directly from the provided laparoscopic videos. This novel task investigates the state-of-the-art on surgical fine-grained activity recognition and establishes a new promising research direction in computer-assisted surgery.

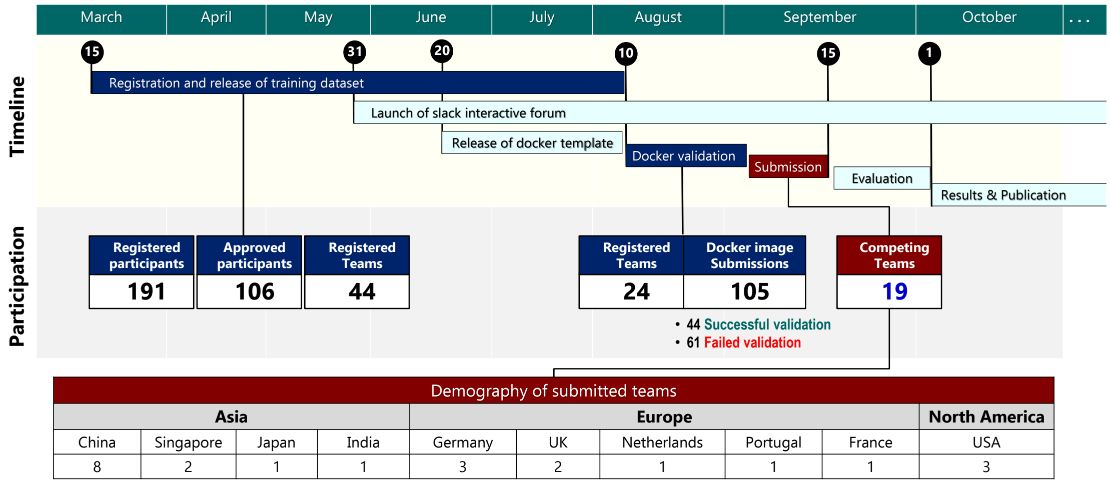
[1] http://camma.u-strasbg.fr/m2cai2016/index.php/workflow-challenge-results
[2] https://endovissub-workflowandskill.grand-challenge.org/

A.1.1.2 Dataset

The dataset for the challenge is CholecT50 that has been generated in this thesis. The dataset is hosted on a CAMMA private server. The download access is granted to only approved participants. For the challenge, we split the dataset into 45 training videos and 5 testing videos. While the training videos are part of public videos of the Cholec80 dataset [Twinanda 2016b], the test set is private. This allows the participants to pre-train their model on other public datasets. Apart from the triplet labels in the dataset, we also provide the groundtruth labels for the three components of the triplets.

A.1.1.3 Participation

The challenge, which runs for 7 months (14 March - 10 October), invites participants from several deep learning research laboratories and institutions across the world to compete with their proposed models. We record a total of 19 participating teams across 10 countries on 3 continents. Participation is by online registration and by signed consent to abide by all the terms and conditions regulating the challenge process and use of the released data. We provide several helpful resources for the smooth running of the challenge. This includes a collaborative blog for sample codes. The blog contains a guide for a quick start to the challenge. This includes snippets of code for an initial understanding of the dataset, loading of data, metrics for method evaluation, and docker template for the building of a submission docker image. The blog also contains samples of shallow models performing triplet recognition in different deep learning framework such as TensorFlow and PyTorch, and basic implementation of some building blocks of the recognition pipeline found in baseline methods.



Asia				Europe					North America
China	Singapore	Japan	India	Germany	UK	Netherlands	Portugal	France	USA
8	2	1	1	3	2	1	1	1	3

Figure A.2 – CholecTriplet 2021 Challenge timeline and participation statistics.

We set up a dedicated slack channel for efficient communication between the organizers and the participants. The slack helps to resolve issues with real-time feedback. Intuitively, we design a validation process, which runs few weeks before the challenge submission deadline. This process ensures that participants' docker has the correct identification, i/o pipeline, output format, and can run without run-time error. Method's quantitative performance is not evaluated at this stage. The participants submit their final method using a validated docker image. While the participants can update their method in the docker during final submission, they are to maintain the same template of the validated docker for error-free evaluation. Furthermore, we create several mailing lists to facilitate communication throughout the challenge.

A.1.1.4 Methods

The challenge focuses on exploiting machine learning methods for the online automatic recognition of surgical actions as a series of triplets. We provide a baseline method, Tripnet [Nwoye 2020], to the participants. We observe that most of the competing methods target the improvement of the baseline performance. We broadly classify the observed methods in the challenge into 5:

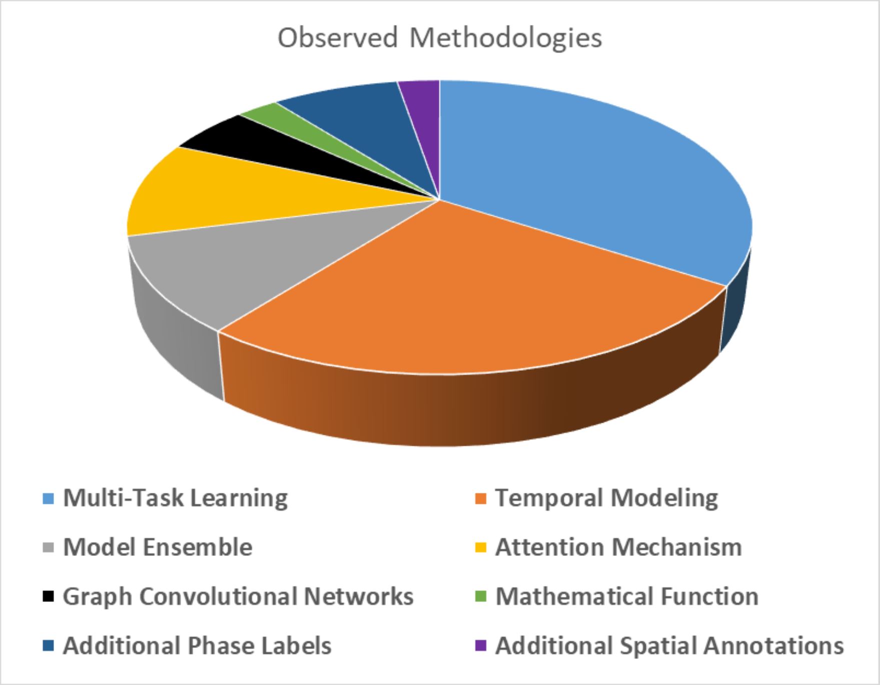


Figure A.3 – A pie chart summarizing the competing team's methods in the challenge.

a. Multi-task learning (MTL) method: many proposed recognition pipeline follows a similar approach in [Nwoye 2020] to first model the components of the triplets before learning their association. Apart from the multi-task learning strategy in [Nwoye 2020], where the multiple task branches share the same feature extraction backbone. Some of the methods employ different base models for the individual task. On the association part, the majority of the teams use a fully connected (FC) layer and an LSTM unit.

b. Temporal modeling method: Some methods leverage the video temporal information to recognize the triplets. It is reported that the temporal correlation between the triplets and the surgical phase information motivates this consideration. The most commonly used temporal modeling unit in the challenge is the LSTM. However, TCN, MCLnet, and ConvLSTM are also used.

c. Attention Mechanism: Some of the proposed methods rely on attention mechanisms, mostly spatial attention modeling, to detect the verb and target components of the triplets.

d. Model Ensemble: Some of the proposed methods ensemble multiple models' outputs to obtain their final predictions. The most commonly observed ensemble technique in the challenge is model averaging.

e. Graph Convolution Networks (GCN): this approach is becoming popular in the computer vision community for HOI recognition. But, it mostly relies on spatial bounding box annotations. Interestingly, the GCN is employed by two teams for surgical action triplet recognition.

f. Complementary Phase Modeling: It is also observed that some proposed model trained their model on additional phase information from the Cholec80 dataset. This is mostly designed as an additional branch in the MTL architecture.

g. Training with Spatial Labels: On a special case, a method presented in this challenge trained their deep learning method on a self-generated bounding box annotations for the tools.

In general, the presented methods cut across well-known deep learning methods using CNN, attention, and RNN frameworks.

A.1.1.5 Evaluation Protocol

Participants submit their challenge methods using a docker image containing the implemented models and their weights. We evaluate the submitted methods in an online mode. This means that a model makes inference on a given image frame leveraging only the current and, maybe, also previous frame information. At time t, a recognition model would have no access to future frame information at time $t+1, t+2, ...$. This would ensure the usability of the competing models in real-time intra-operative application in the OR.

The evaluation metrics is mean Average Precision (mAP), computed following the same approach in Section 4.4.3. During the evaluation, we exclude all triplet classes with either null-verb, null-target, or null-instrument components. This reason is to ensure a fair comparison of competing models since the null categories are not precisely defined.

Apart from the AP scores, we also statistically analyze the model performance in the top 5, 10, and 20 predictions. We extensively compute the AP for each triplet component as additional metrics for rank stability. The performance is also analyzed using Wilcoxon signed-rank test for performance significant level estimation.

A.1.1.6 Results

The challenge ends on 1st October with the results and awards presentation at the MICCAI EndoVis satellite event. The mAP on the triplet recognition task reported in the challenge ranges between 4.5 - 38.1%. A total of 6 teams score above 30%, whereas 5 teams score within the interval of 20-29%. Another 5 team scores fall within the interval of 10-19%. Only 3 teams score an mAP below 10%. The winning method scores 38.1% mAP on the triplet recognition task. The best performing baseline method, the Rendezvous [Nwoye 2021], which is re-trained on the challenge training split, obtains an mAP of 32.7% ranking 4th in the challenge leaderboard. Apart from the AP metrics, we also present further analysis of the results including the top-N accuracy, individual triplet components performances, and qualitative results. The winner receives a GPU award from NVIDIA, as well as cash prizes supported by Medtronic. The first and second runner-ups also receive cash prize awards. Additionally, they all receive award certificates acknowledging their state-of-the-art performance on the triplet recognition task.

A.1.1.7 Perspective

Surgical triplet recognition offers a new solution to fine-grained activity recognition in the OR. It is so far, the truly comprehensive framework of modeling surgical tool-tissue interaction in laparoscopic videos. The challenge helps to introduce and popularize this direction of research in the community. It recorded very good participation, a total of 19 teams, which is so far the highest number in the EndoVis sub-challenges. This is likely due to the novel nature of the task tackled, as well as the large-scale dataset of fine-grained surgical action triplet. It can also be attributed to the early start of the challenge including giving the participants a reasonable amount of time to develop methods and participate in the challenge. The organizers' fast and friendly communication with participants cannot be excluded from the participation huge turn out.

It is indeed a rewarding experience to the organizers and the participants, with a large number of very interesting approaches to solving the triplet recognition problem. It is also an interesting experience to host a challenge, manage a large group of people, and maintain the infrastructure to run the challenge.

Taking everything into consideration, it is ideal to conclude that triplet recognition is a solution for more fine-grained activity recognition, but triplet recognition, in itself, remains a challenge. Hence, this challenge set a bedrock for future research in the domain.

The challenge will be followed by a joint publication of the challenge methods including an ensemble of the best performing models for optimized results. For continuity, we are considering the possibility of setting up a benchmark server for continual evaluation of future methods on the dataset test set as it is being done in the computer vision community. This server could also provide the SOTA leader-board on the task for transparency and easy review of methods proposed in the field. We also plan to create a common GitHub repository for quick access of SOTA methods, which could help in further improvement of the methods. CholecTriplet 2021 challenge will indeed bring considerable value to the field of surgical data

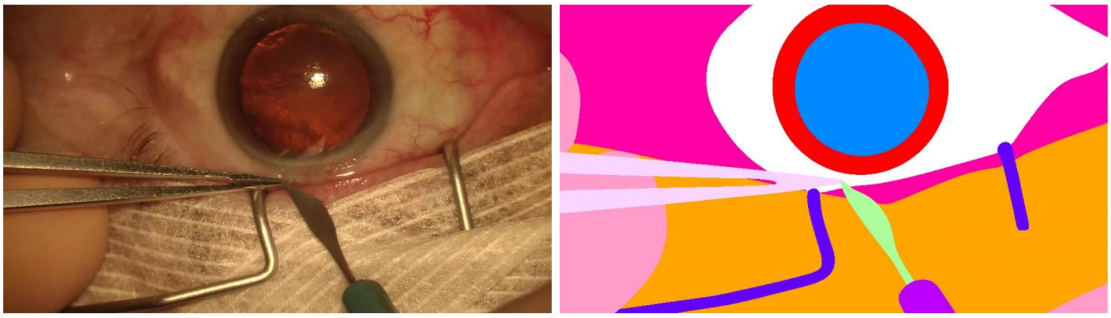


Figure A.4 – Example image frame (left) and semantic segmentation labels (right) for the Cataract dataset for image segmentation [Grammatikopoulou 2019].

science. The challenge will likely be repeated in 2022 following the widely declared interest from the 2021 EndoVis challenge attendants.

A.2 Challenge Participated

A.2.1 MICCAI EndoVis 2020 on Cataract Segmentation

Cataract segmentation challenge is an endoscopic vision sub-challenge focusing on the development of deep learning methods for the semantic segmentation of surgical instruments and anatomical structures in cataract eye surgical images. This challenge provides a dataset, Cataract Dataset for Image Segmentation (CaDIS) [Grammatikopoulou 2019], created by Digitial Surgery Ltd. The CaDIS dataset consists of 4670 images sampled from the 25 videos on public CATARACTS' [Al Hajj 2019] training set. The images are annotated with fine-grained semantic pixels labels. Each pixel in each image is labeled with its respective instrument or anatomical class from a set of 36 identified classes: 29 surgical instrument classes, 4 anatomy classes, and 3 miscellaneous classes. An example image and the corresponding segmentation mask is shown in Figure A.4. The 25 videos training set is provided to the challenge participants for method training. The organizers annotated additional 10 videos from their in-house dataset hold out as the hidden test set for the challenge.

The challenge consists of three tasks of segmenting cataract RGB images into body and instruments at three different levels of granularity: (1) instrument vs. background, (2) instrument category vs background, and (3) instrument category vs body organs. The first, second, and third tasks have 8, 17, 25 semantic class labels respectively.

A.2.1.1 Participation

[Figure: architecture] We formed a team, camma-cadis, of three members: Chinedu Nwoye, Deepak Alapatt, and Nicolas Padoy, all from the CAMMA research laboratory, University of Strasbourg, France, and participate in all the 3 sub-tasks. There are 11 participating teams in total.

A.2.1.2 Method

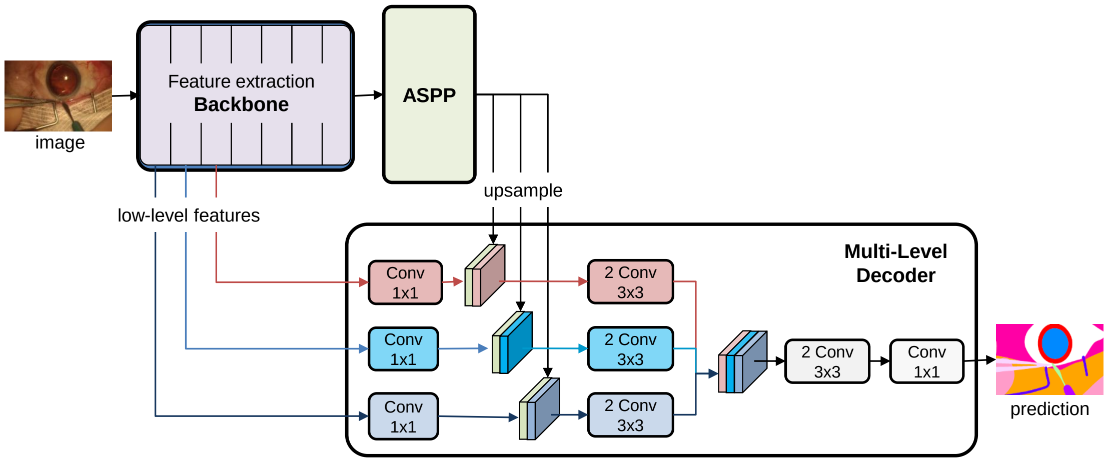


Figure A.5 – A Multi-Level Decoder Network for Cataract image segmentation.

We develop a multi-level decoder network for semantic segmentation: a deep learning method that can be used for all the 3 sub-tasks of semantic segmentation on cataract images. The proposed model follows an encoder-decoder architecture as shown in Figure A.5. The encoder is similar to DeepLab v3+. This consists of a Xception-65 [Chollet 2017] base model and an Atrous Spatial Pyramid Pooling (ASPP) [Chen 2017]. The Xception network extracts the visual feature from the RGB input image while the ASPP refines the extracted features using multiple convolutions at different dilation rates including a convolution on the global average feature.

To recover more spatial details for the segmentation, we build a multiple decoder of levels $L = \{l_1, l_2, ..., l_N\}$ that combines features from different consecutive blocks $B = \{b_1, b_2, ..., b_N\}$ of the feature extractors, such that each i^{th} block (b_i) is decoded at i^{th} decoding level (l_i), where N is the maximum decoding level. We maintain N=3 for our experiment.

For each decoding level, l_k, the ASPP output is refined using the low-level features obtained from block b_k of the encoder as shown in Figure A.5. In this way, the high-level features are decoded at different levels of encoding semantics. To generate the output, we concatenate the multilevel decoded features and apply two 3x3 convolutional layers followed by one 1x1 convolutional layer before generating the class-wise probabilities using the softmax activation.

A.2.1.3 Experiment

For our experiment, we downsample the input images to a resolution of $270 \times 480 \times 3$. We split the dataset into 23 videos for training and 3 for validation. We augment our training data by applying random scaling [0.5, 2] and varying brightness (delta=0.2). We train the proposed model for 200 epochs using categorical cross-entropy as a loss function and SGD with Momentum as the optimizer. We use a "poly" learning rate policy where the initial learning rate (7e-4) is multiplied by $\left(1-\frac{iteration}{maxiteration}\right)^{0}$.9. We fit a batch size of 4 on a Quadro P5000 GPU and train for approximately 36 hours. Initially, our proposed model is trained and

Table A.1 – Segmentation Results of the proposed model in comparison with the baseline. We

Task	Method	mIoU	PA	PAC
Task 1	DeepLab v3+	82.62	93.89	88.74
	Ours	**85.11**	**94.52**	**90.35**
Task 2	DeepLab v3+	72.26	93.49	80.80
	Ours	**77.50**	**94.40**	**84.67**
Task 3	DeepLab v3+	63.23	93.86	75.60
	Ours	**70.63**	**94.39**	**78.64**

tuned for the third sub-task which is the most challenging sub-task in the challenge. For the lack of time, we took our best model on this task and fine-tune only the last layer for the other two sub-tasks.

A.2.1.4 Performance and Discussion

We compare our proposed model to a baseline model which is a re-implementation of DeepLab v3+ on the CaDIS dataset. We evaluate the models on three metrics set by the challenge organizers: mean Intersection over Union (mIoU), Pixel Accuracy (PA), and Pixel Accuracy per-Class (PAC). The preliminary results on the validation set, presented in Table A.1, show that by multi-level decoding, our proposed segmentation pipeline outperforms the baseline by 3% in the first task, 5% in the second task, and 7% in the third task on mean IoU metrics. We also substantially outperformed the baselines in all other considered metrics in the 3 sub-tasks.

Our method is submitted by docker image to the challenge organizer for evaluation on the out-of-sample test set. In the challenge, our model is ranked $4th$ on the most difficult sub-task which has the finest level of granularity among the three sub-task. This is indeed the sub-task we focus our implementation and training on. We are ranked 6th and 10th for the 2nd and 1st tasks respectively.

A.2.1.5 Conclusion

Participating in the challenge for cataract semantic segmentation was beneficial for the organization of the succeeding action triplet challenge. We learnt from the challenge, the docker validation process, and the use of a dedicated slack channel to coordinate a challenge. The task in itself is useful, as it presents a different task on detecting surgical instruments on the different procedures. The detection, in this case, is at pixel-wise level, though fully supervised, presents a comparison on the use of deep learning frameworks for the detection of surgical instruments in different procedures. While the procedures are different, the data presents similar challenges especially with regards to visual ambiguities in surgical images. The challenge also presents the detection of the anatomical structures at a pixel-wise scale. While this is not directly comparable to target recognition in triplet, it gives insight in understanding the visual challenges facing the detection of anatomies especially a cases of unclear boundaries and tissue deformation. Furthermore, the challenge provides us the opportunity to compete and also learn from other participants the different considerations in modeling a deep learning task. In conclusion, while promising results are presented at the challenge, it is still challenging to distinguish anatomy vs surgical instruments with a limited-size dataset.

A.2.2 MICCAI EndoVis 2019 on Surgical Workflow and Skill Analysis

Understanding tool-tissue interaction in endoscopic surgery is essential for better surgical workflow analysis and skill assessment. For this reason, a new dataset, known as *Hei-chole*, is introduced by the National Center for Tumor Diseases (NCT) Heidelberg, Germany to aid research in this direction. The dataset is collected from 3 surgical centers and consists of 33 videos of laparoscopic cholecystectomies which has been annotated with binary labels of phase, action, tools and surgical skill information. In EndoVis 2019, the dataset is used to organize to a sub-challenge that focuses on online workflow analysis of laparoscopic surgeries. There are four sub-tasks within the challenge which include: (1) phase recognition, (2) tool detection, (3) action recognition, and (4) skill assessment. This novel kind of challenge investigates the current state-of-the-art results on surgical workflow analysis and skill assessment on one comprehensive dataset, *Hei-chole*.

A.2.2.1 Participation

The challenge recorded a total of 12 participated teams. Our team, camma, consists of four members: Tong Yu, Chinedu Nwoye, Armine Vardazaryan, and Nicolas Padoy, all from the CAMMA research group, University of Strasbourg, France. Originally, our team did not compete in the challenge, but we submit our methods after the challenge to augment the challenge participation for further result analysis and joint publication. We split our team into two sub-teams to focus on different tasks in the challenge. My sub-team tackles the action recognition task in the challenge. In the surgical action recognition task, there are 4 action classes, representing the most prevalent actions in laparoscopic cholecystectomy. These include grasp, cut, hold, and clip.

A.2.2.2 Method

Action is a process which is better modeled on a continuous sequential frames. For this reason, we considered temporal modeling as the key to learning the surgical actions in the laparoscopic videos. Fortunately, temporal information requires no additional annotation. Previously, some research exploits temporal information to model surgical dataset analysis. Work from [Al Hajj 2018] learns instrument detection using a long short term memory (LSTM) unit on an ensemble CNN architecture. Also, using a weakly supervised Convolution LSTM (ConvLSTM), [Nwoye 2019] models surgical instrument tracking in Cholec80 [Twinanda 2016b] dataset.

For our method, we build a temporal-aware deep neural network for action recognition. Our implementation is an adaptation of the ConvLSTM tracker in [Nwoye 2019] for action recognition. We observe that surgical actions are a direct expression of the instrument activities. We hypothesis that these activities can be derive from a temporal change in the pose, form, shape, and motion of the instruments. The ConvLSTM has been shown to learn the surgical instrument trajectory [Nwoye 2019] from only the binary presence labels, we adapt and train this spatio-temporal model for action recognition leveraging video temporal information and the ConvLSTM's long term dependency capacity.

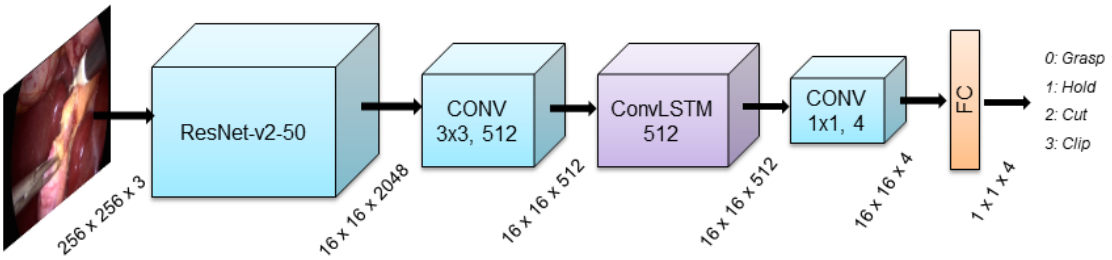


Figure A.6 – The CNN+ConvLSTM spatio-temporal model for action recognition.

The architecture of our proposed model is a CNN + Convolutional LSTM model trained end-to-end on the action binary labels for surgical action recognition. The base is a ResNet-50 [He 2016] model for feature extraction. We added an additional convolution unit as a bottleneck layer to reduce the feature space dimensionality. This is followed a Convolution LSTM (ConvLSTM) to take into account the temporal consistency of surgical actions. The output is further refined by a convolutional layer and a fully connected layer for higher-level reasoning on the spatio-temporal features for surgical action recognition.

A.2.2.3 Experiment and Results

Our network is trained on the provided training videos of Hei-Chole dataset which we further split into train/val in the ratio of 7:3. We adjusted the input dimension from 854x480 to 256x256 for easy fitting by the model. Following the recommendation for the challenge, we did not perform any data augmentation or data preprocessing. The ResNet-50 backbone is pretrained on ImageNet. The entire model is trained by truncated back-propagation with an initial learning rate of 1e-3 and 1e-5 for the pretrained and new layers respectively with a cosine decay policy. For class-balancing, we apply the same weighting scheme on sigmoid cross-entropy loss function as used in [Nwoye 2019].

The submitted methods are evaluated on three metrics: recall, precision, and F1-score. On the recall metrics, our methods scores 29.83% clinching the 1st position on this metric. On the precision metrics, we score 19.19% ranking 5th, and on F1-score our method obtains a score of 22.10% ranking 4th. The challenge winners are determined by the F1-score and on this we are placed 4th in the challenge leaderboard.

A.2.2.4 Conclusion

We present spatio-temporal modeling of surgical action in endoscopic videos using CNN and ConvLSTM. Leveraging video temporal information, we utilize the ConvLSTM to learn the instrument's actions. We observed that the model performs better for actions that occur in virtually all the videos than actions that occur in a few. This is likely due to its modeling of temporal consistency which may treat irregular actions as noise. We could improve on the less frequent actions with hard-negative mining but this would alter the temporal flow of a video and not suitable for an LSTM model.

This is the first challenge that I participated in the course of my Ph.D. program. The challenge provided me with the opportunity to learn how deep learning problems are approached by other participating teams.

B Résumé en français

Méthodes d'Apprentissage Profond pour la Détection et la Reconnaissance d'Outils et d'Activités Chirurgicaux dans les Vidéos Laparoscopiques

La science n'a pas de dimension morale. C'est comme un couteau...
Si vous le donnez à un chirurgien ou à un meurtrier, chacun l'utilisera différemment
– Werhner von Braun

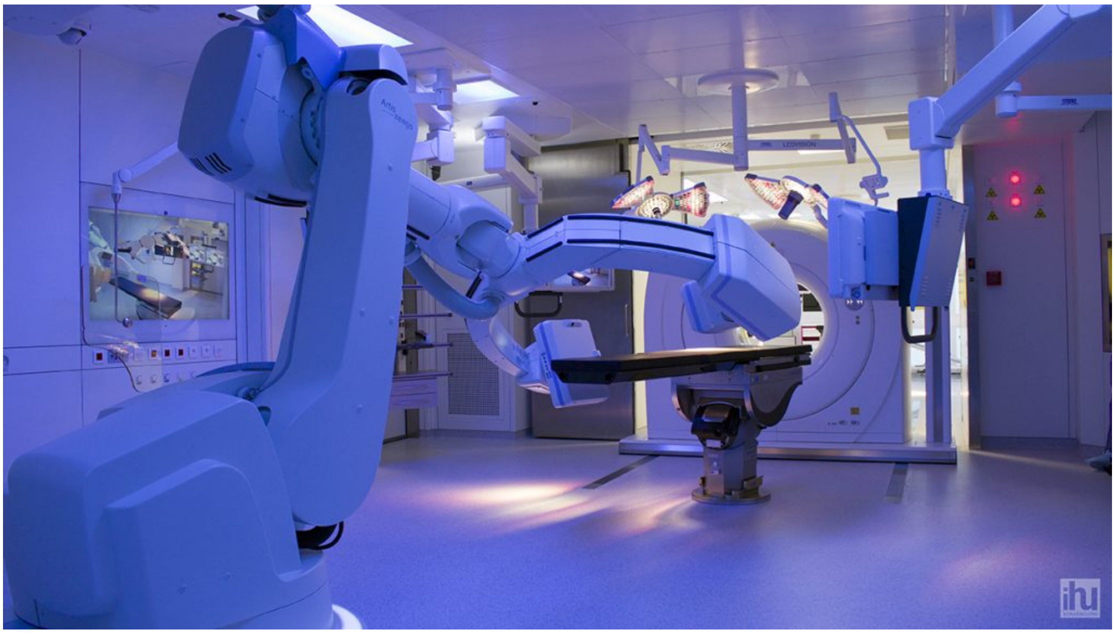


Figure B.1 – Un exemple de salle d'opération hybride qui combine une salle d'opération traditionnelle avec une salle d'intervention guidée par l'image. Suite interventionnelle de haute technologie à l'IHU Strasbourg, France.

Abstract

La chirurgie, une unité centrale du système de soins aux patients, s'améliore de plus en plus grâce aux innovations technologiques continues facilitant de meilleurs résultats pour les patients et fournissant de riches données peropératoires via des systèmes d'information. Ceci, cependant, augmente la complexité des flux de travail, ainsi que la charge de travail cognitive des chirurgiens. Par conséquent, il existe un besoin croissant d'optimiser le flux de travail chirurgical via des systèmes intelligents et analytiques qui peuvent fournir une aide à la décision et une assistance contextuelle aux chirurgiens. Malgré la vaste littérature sur la reconnaissance d'activité dans la vision médicale par ordinateur, la nature grossière des tâches principalement abordées, par exemple la reconnaissance des phases chirurgicales, ne fournit pas assez de détails pour une assistance IA plus utile en salle d'opération (OR). Les salles d'opération modernes de haute technologie nécessitent un système de reconnaissance d'activité plus détaillé: un système capable de capturer méticuleusement des actions plus fines, telles que les interactions entre l'instrument et les tissus, et de décrire de manière exhaustive les activités en cours.

Dans cette thèse, nous nous concentrons sur le développement de méthodes d'apprentissage en profondeur pour la détection et la reconnaissance d'instruments chirurgicaux et de leurs activités finement décrites dans des vidéos laparoscopiques. Ces activités sont formalisées sous forme de triplets de ⟨*instrument, verbe, cible*⟩ représentant l'activité-outil. Nous étudions, dans un premier temps, la détection et le suivi articulaires d'instruments chirurgicaux dans des vidéos laparoscopiques. Pour atténuer la difficulté de générer manuellement des annotations spatiales pour les instruments dans chaque image vidéo, nous développons une nouvelle méthode de localisation faiblement supervisée sur des étiquettes de présence binaires, qui sont plus faciles à générer. Pour tirer parti de la structure temporelle des vidéos chirurgicales, nous proposons l'utilisation d'un réseau de neurones récurrents pour suivre le mouvement des instruments, toujours sans nécessiter aucune forme d'étiquettes d'entraînement spatial. De plus, nous créons un grand ensemble de données vidéo avec des étiquettes spatiales, que nous utilisons pour valider la méthode proposée. En progressant vers la modélisation d'activité, nous générons un ensemble de données à grande échelle de triplets d'action chirurgicale et construisons plusieurs modèles d'apprentissage en profondeur pour leur reconnaissance. Tout d'abord, nous concevons un pipeline de reconnaissance qui apprend les composants individuels des triplets à l'aide de vecteurs caractéristiques générés par CNN et établit leur association dans un espace de vecteurs caractéristiques 3D, car une trame peut contenir plusieurs triplets. En améliorant la première méthode, nous proposons une nouvelle forme d'attention spatiale pour capturer plus efficacement les composants individuels du triplet en utilisant les activations résultant des instruments. De plus, nous introduisons une nouvelle forme d'attention sémantique, inspirée des réseaux Transformer, pour apprendre l'association des composants du triplet. Enfin, nous validons toutes les approches proposées sur les ensembles de données introduits dans ce travail, obtenant des performances de pointe sur chaque tâche.

Chapter Summary

B.1 Introduction

Les technologies émergentes en chirurgie, qui ont transformé la salle d'opération traditionnelle (OR) en un lieu de haute technologie, comme le montre la figure B.1, ont encouragé de nombreux algorithmes d'apprentissage en profondeur de pointe à être conçus pour l'analyse automatisée du flux de travail chirurgical afin de fournir une intervention assistée par ordinateur (CAI) dans la salle d'opération [Gibson 2018]. Un ingrédient clé pour développer des systèmes CAI qui peuvent fournir une aide à la décision contextuelle en chirurgie laparoscopique est d'avoir une connaissance en temps réel de la présence des instruments chirurgicaux, de leurs emplacements, de la pose par rapport à la caméra et de l'anatomie sous-jacente, leur mouvement dans le temps et comprendre leurs interactions avec les tissus environnants. Le CAI, en tant qu'un des domaines de recherche à l'intersection de la médecine et de l'informatique, pousse la recherche dans cette direction.

B.1.1 Contexte clinique et motivation

Cette thèse est menée dans le contexte de la cholécystectomie laparoscopique, qui est un type de chirurgie mini-invasive qui concerne l'ablation d'une vésicule biliaire défectueuse du corps [Olsen 1991]. Elle se caractérise par la dissection, la coupe des structures tubulaires entourant ou attachant la vésicule biliaire à d'autres organes du corps, et l'extraction de la vésicule biliaire détachée du corps [Massarweh 2007]. La laparoscopie est devenue une approche de référence pour la cholécystectomie [Pucher 2018] en raison de son faible risque attribué à l'ablation de la vésicule biliaire. Étant peu invasive, la procédure est moins traumatisante: le patient présente généralement une probabilité réduite d'infection nosocomiale, moins de douleur, moins de saignements et des temps de récupération plus rapides [Velanovich 2000] par rapport à la procédure ouverte, qui nécessiterait une coupe de la peau et des tissus de taille suffisamment pour qu'un chirurgien puisse avoir une vue complète des structures et des organes à opérer grande [Ballantyne 2002]. Cependant, ce succès a un prix pour le chirurgien, qui doit désormais faire face à une difficulté technique accrue provenant de la vision indirecte et de la manipulation non conventionnelle des équipements laparoscopiques avancés [Ballantyne 2002, Mascagni 2021b]. Cela augmente la charge de travail du chirurgien et rend la procédure plus complexe.

La complexité élevée de la laparoscopie est l'une des motivations de la recherche sur la CAI, qui concerne le développement de systèmes informatiques intelligents pour optimiser le flux de travail chirurgical et augmenter les capacités des cliniciens dans la salle d'opération [Lemke 2005, Stoyanov 2012]. La recherche dans cette direction porte sur le flux de travail chirurgical et l'analyse des compétences [Speidel 2009, Sznitman 2011, Jin 2018], la robotique médicale [Hager 1995, Speidel 2014, Vander Poorten 2020], l'imagerie médicale [Navab 1999, Fitzek 2021], la navigation interventionnelle [Navab 2002, Pfeiffer 2019b], réalité augmentée et visualisation [Navab 2007, Navab 2012, Rodas 2015], etc. Cette thèse se concentre davantage sur l'analyse du flux de travail chirurgical qui vise à la reconnaissance automatique d'un sous-ensemble prédéfini de tâches, d'activités d'intérêt ou d'opérateurs de telles activités en suivant le processus chirurgical avec une analyse en temps réel des données vidéo en direct acquises en peropératoire [Maier-Hein 2017]. Certaines des recherches dans l'analyse du flux de travail chirurgical comprennent: la détection d'outils [Bouget 2017], la classification des procédures [Kannan 2019], la reconnaissance de phase [Garrow 2021], l'estimation du temps de chirurgie restant [Aksamentov 2017], l'estimation de la pose du clinicien [Kadkhodamohammadi 2014], analyse des compétences chirurgicales [Reiley 2011], reconnaissance des gestes/événements chirurgicaux [DiPietro 2016], reconnaissance des étapes chirurgicales [Ramesh 2021], reconnaissance des activités/actions chirurgicales [Lalys 2014], etc.

B.1.2 Aperçu de la recherche

Des efforts ont été faits dans le passé pour modéliser les activités chirurgicales à partir de vidéos telles que la procédure, la phase, l'étape, la reconnaissance d'événements. Cependant, les activités modélisées dans la plupart de ces configurations sont de nature très grossière. Une telle modélisation granulaire ne fournit pas une image précise des activités en cours. Même les divisions à grain fin, telles que la reconnaissance d'action, omettent des détails substantiels sur l'anatomie. De telles informations sémantiques sont nécessaires pour une reconnaissance d'activité détaillée et complète qui est plus utile pour l'assistance d'IA nécessaire dans la salle d'opération.

Par conséquent, la principale question de recherche est **comment modéliser efficacement l'interaction outil-tissu pour déduire des actions précises à partir de vidéos pour la meilleure utilité clinique ?**. Pour tenter de répondre à cette question de recherche, nous sommes confrontés à un partitionnement des activités en entités constitutives impliquées dans l'interaction : l'instrument, son rôle et sa cible. Il semble maintenant que plusieurs tâches de reconnaissance soient impliquées, mais comme toutes les activités tournent autour des instruments, la localisation de ces instruments devient également impérative pour la reconnaissance des autres composants en interaction qui reposent sur les informations de position de l'instrument. Cependant, il y a un manque d'ensembles de données annotés spatialement pour former un modèle d'apprentissage en profondeur pour la détection d'instruments. Mais, comme il est plus facile de générer des étiquettes binaires indiquant la présence ou l'absence d'instruments chirurgicaux, **comment exploiter ces étiquettes de présence binaires plus faciles à générer pour la localisation et le suivi des outils ?** devient une question de recherche complémentaire. Le travail effectué dans cette thèse vise à fournir suffisamment de champ d'investigation, des réponses pratiques et une discussion perspicace à ces questions de recherche.

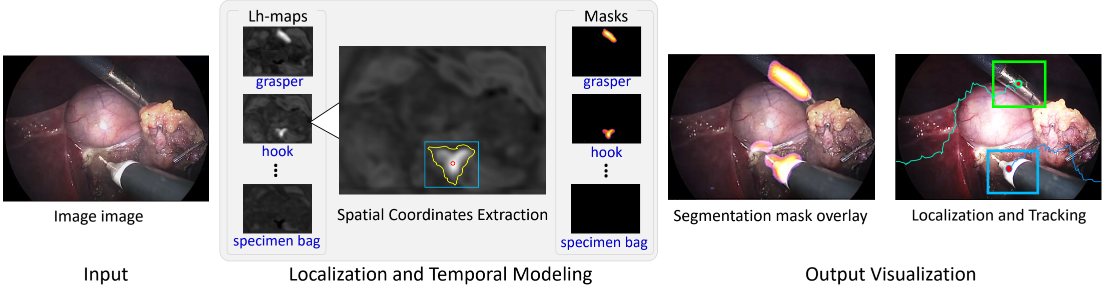

Lh-maps
grasper
hook
specimen bag
Masks
grasper
hook
specimen bag
Image image
Spatial Coordinates Extraction
Segmentation mask overlay
Localization and Tracking
Input
Localization and Temporal Modeling
Output Visualization

Figure B.2 – Flux séquentiel d'entrée - sortie du modèle ConvLSTM.

Dans la première partie de la thèse [Nwoye 2019], nous étudions et construisons des modèles capables d'exploiter des données faiblement annotées pour la détection, la localisation et le suivi des instruments chirurgicaux. La détection et le suivi des instruments faciliteront la compréhension des interactions outil-tissu, fourniront des signaux sur les situations chirurgicales, aideront à la décision chirurgicale et à l'évaluation des compétences, et seront utiles dans le suivi des instruments actionnés manuellement en chirurgie assistée par robot. Les travaux existants sur la détection d'instruments chirurgicaux reposent sur une supervision complète : une situation dans laquelle les modèles de détection et de suivi sont entraînés sur des données dans lesquelles les positions spatiales des instruments sont annotées manuellement. Cependant, la création d'annotations spatiales telles que des limites de région et des masques au niveau des pixels est coûteuse, fastidieuse et chronophage. En outre, la plupart des ensembles de données disponibles dans le domaine ne contiennent que des étiquettes de présence binaires qui sont générées par un simple marquage de 0 ou 1 pour indiquer la présence ou l'absence d'instruments chirurgicaux. Puisqu'il est plus facile de générer ces étiquettes binaires, nous avons proposé de les exploiter pour une tâche spatiale beaucoup plus complexe telle que la localisation d'instruments chirurgicaux. Notre modèle proposé exploite également les informations temporelles inhérentes aux données vidéo pour le suivi des instruments.

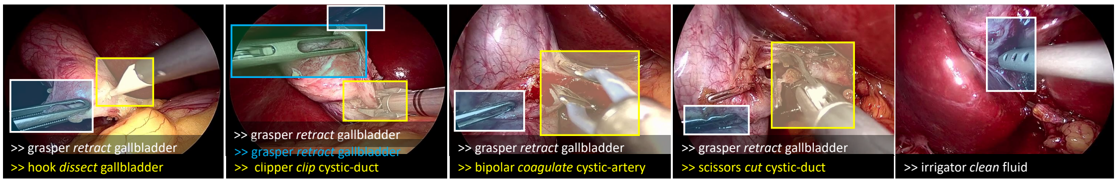

>> grasper *retract* gallbladder
>> hook *dissect* gallbladder
>> grasper *retract* gallbladder
>> grasper *retract* gallbladder
>> clipper *clip* cystic-duct
>> grasper *retract* gallbladder
>> bipolar *coagulate* cystic-artery
>> grasper *retract* gallbladder
>> scissors *cut* cystic-duct
>> irrigator *clean* fluid

Figure B.3 – Échantillons d'images chirurgicales montrant des étiquettes d'instance de triplet d'action.

Dans la deuxième partie de la thèse [Nwoye 2020, Nwoye 2021], nous nous appuyons sur les recherches les plus avancées pour reconnaître les activités des instruments chirurgicaux. Les travaux existants sur la reconnaissance d'activité se concentrent principalement sur la reconnaissance de phase, d'étape, d'événement, de geste ou même d'action à un seul verbe. [Twinanda 2016b, DiPietro 2016, Loukas 2015, Ramesh 2021, Khatibi 2020]. À ces niveaux de granularité, la reconnaissance laisse de côté une sémantique essentielle pour une assistance utile de l'IA. Par conséquent, nous proposons une reconnaissance plus détaillée des activités à grain fin représentant les interactions instrument-tissu dans les vidéos endoscopiques. Nous modélisons ces activités sous forme de triplets d'action chirurgicale de ⟨*instrument, verbe, cible*⟩ et développons des modèles d'apprentissage en profondeur pour reconnaître ces triplets. Le triplet représente l'instrument utilisé, l'action effectuée et l'anatomie sur laquelle on a agi comme défini dans l'ontologie existante [Neumuth 2006, Katić 2014]. Leur reconnaissance fournit des informations plus détaillées sur la situation chirurgicale qui permet de mieux comprendre l'interaction outil-tissu pendant la chirurgie. Il ajoute également des informations substantielles nécessaires à l'IA qui sont sûres, efficaces et approfondies. La reconnaissance de triplet peut également être utile en peropératoire pour surveiller les points de contrôle de sécurité critiques, l'aide à la sensibilisation au contexte, l'anticipation des actions pour une intervention précoce et en postopératoire pour le sous-titrage vidéo, la génération de rapports postopératoires, la validation des tissus et l'indexation et la récupération de vidéos spécifiques à l'action pour l'éducation chirurgicale. Pour soutenir la recherche dans cette direction, avec l'aide de nos collaborateurs cliniques, nous générons des ensembles de données à grain fin (*CholecT40* et *CholecT50*), les premiers du genre, pour la reconnaissance de triplets d'action chirurgicale. Voici quelques exemples de triplets d'action dans l'ensemble de données:⟨*grasper, retract, gallbladder*⟩, ⟨*hook, dissect, omentum*⟩, ⟨*scissors, cut,cystic-duct*⟩ comme on peut également le voir sur la figure B.3.

B.1.3 Littérature connexe

B.1.3.1 Travaux connexes sur la détection d'outils chirurgicaux

Dans cette section, les travaux réalisés dans cette thèse sont positionnés par rapport aux travaux connexes dans la communauté de recherche. Dans la littérature, de nombreux travaux ont été menés sur la détection de la présence d'outils chirurgicaux dans différents types de chirurgie : cholécystectomie laparoscopique [Twinanda 2016b, Zia 2016], chirurgie oculaire de la cataracte [Al Hajj 2018], etc, à l'aide de données vidéo. Alors que la plupart des méthodes utilisent l'apprentissage par transfert sur des architectures de réseau neuronal convolutif (CNN) de pointe [Twinanda 2016a, Sahu 2016, Raju 2016], d'autres utilisent différentes variantes de réseau neuronal récurrent (RNN) [Namazi 2019, Jin 2020, Mishra 2017]. Dans certains cas, des informations de phase supplémentaires sont utilisées dans une technique d'apprentissage multitâche pour capturer des caractéristiques de corrélation avec un biais inductif [Twinanda 2016b, Mondal 2019, Jin 2020].

Au-delà de la détection de présence, les instruments chirurgicaux sont localisés à l'aide de modèles supervisés sur des étiquettes de cadre de délimitation [Choi 2017, Jin 2018, Zhang 2020a]. Pour le manque et les difficultés associées à la génération d'étiquettes spatiales, une supervision faible est explorée pour apprendre des étiquettes plus faibles. [Vardazaryan 2018] apprend l'emplacement des outils chirurgicaux à partir d'étiquettes de présence binaires, mais la localisation est limitée aux coordonnées d'un seul point des outils dans une image. [Fuentes-Hurtado 2019] étend cela pour inclure les limites de régions entières des outils, cependant, leur modèle est supervisé sur des lignes longitudinales beaucoup plus difficiles à générer. Nous proposons une méthode faiblement supervisée [Nwoye 2019] pour la localisation et le suivi d'outils. Contrairement à [Vardazaryan 2018], nous localisons les coordonnées de la boîte englobante des outils, et contrairement à [Fuentes-Hurtado 2019], notre méthode est supervisée sur les étiquettes de présence binaires les plus faciles à générer.

Sur le suivi des outils chirurgicaux, les travaux antérieurs [Speidel 2008, Reiter 2010, Sznitman 2012b, Reiter 2012b] sont basés sur une approche d'apprentissage automatique de génération de caractéristiques. En plus d'être une approche manuelle, il est beaucoup plus difficile d'obtenir des caractéristiques robustes en utilisant cette approche. Les méthodes d'apprentissage en profondeur dans ce domaine sont principalement explorées sur la chirurgie robotique en utilisant des informations cinématiques supplémentaires, qui ne sont pas disponibles pour la chirurgie non robotique [Ye 2016, Du 2018, Colleoni 2019, Du 2016]. Les approches purement basées sur la vision [Zhao 2017, Robu 2020, Banerjee 2019] sur le suivi des outils sont principalement conçues pour une supervision complète avec des annotations de cadre global qui sont difficiles à générer. Par conséquent, nous proposons une méthode [Nwoye 2019] qui ne nécessite ni annotation spatiale ni information cinématique pour leur apprentissage. Au lieu de cela, nous exploitons les informations temporelles inhérentes aux données vidéo pour modéliser le suivi des outils. Le modèle proposé est seulement faiblement supervisé sur des étiquettes de présence binaires.

B.1.3.2 Travaux connexes sur la reconnaissance de l'activité chirurgicale

La définition de l'activité est subjective et dépend du niveau d'abstraction de l'activité réalisée. L'activité peut être décrite à un niveau de granularité différent d'une extrémité (à grain grossier, par exemple jouer au football, danser, etc.) à une autre (à grain fin, par exemple, donner un coup de pied, courir, faire signe de la main, etc.). Dans l'analyse du flux de travail chirurgical, la reconnaissance d'activité la plus grossière se concentre sur la reconnaissance des types d'interventions chirurgicales effectuées [Münzer 2013, Twinanda 2014, Petscharnig 2018a, Kannan 2019], par ex. cholécystectomie, cataracte, pontage gastrique, etc. Au sein de la procédure, les activités séquentielles sont reconnues comme des phases chirurgicales soit à partir de vidéos endoscopiques [Lo 2003, Ahmadi 2006, Blum 2010, Dergachyova 2016, Twinanda 2016b, Funke 2018, Zisimopoulos 2018, Yu 2018] ou de caméras montées au plafond twinanda2015data,chakraborty2013video. En approfondissant le niveau de granularité, d'autres travaux reconnaissent des événements [Malpani 2016, Loukas 2015], des gestes [DiPietro 2019, Kitaguchi 2019, Sarikaya 2020, Park 2021] ou des étapes [Charriere 2014, Lecuyer 2020, Ramesh 2021] se produisant dans les phases. Toutes ces activités étendues sont composées de plusieurs actions plus fines qu'il serait intéressant de reconnaître.

Pour fournir une image plus précise des activités en cours, la reconnaissance d'action, un niveau de reconnaissance d'activité plus fin, se concentre sur la reconnaissance des actions

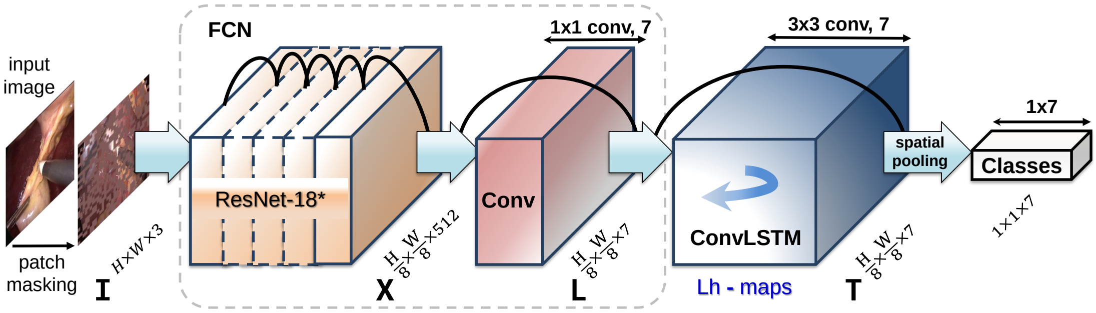


Figure B.4 – L'architecture du Traqueur ConvLSTM proposé.

effectuées par des verbes simples [Rupprecht 2016, Khatibi 2020, Wagner 2021b]. Bien que cette division soit de nature très fine, elle laisse de côté des détails sur l'anatomie. De telles informations sémantiques précieuses sont nécessaires pour une assistance plus utile de l'IA dans la salle d'opération, en particulier lorsqu'il s'agit de favoriser la sécurité et l'efficacité.

Et donc, le triplet d'action chirurgicale est proposé comme comprenant des informations sur les instruments chirurgicaux utilisés pour effectuer une action, des verbes - représentant l'action à grain fin effectuée et une cible - qui est l'anatomie sous-jacente sur laquelle agit [Katić 2014]. Avant le formalisme des triplets, plusieurs travaux [Neumuth 2006, Speidel 2009, Neumuth 2010] ont décrit la situation chirurgicale dans le cadre de triplets d'action. Les travaux de [Katić 2014] et [Katić 2015] ont utilisé les informations d'annotation pour améliorer la reconnaissance de la phase chirurgicale. Nous proposons la première méthode d'apprentissage en profondeur pour reconnaître des triplets d'action directement à partir de vidéos chirurgicales [Nwoye 2020].

B.2 Tâches et méthodes

Nous présentons les deux tâches principales de cette thèse ainsi que les différentes méthodes développées pour aborder ces tâches.

B.2.1 Détection d'outils chirurgicaux

B.2.1.1 Méthode faiblement supervisée pour la localisation et le suivi des outils

Le modèle proposé tel qu'illustré à la figure B.4 est une jointure de réseaux de neurones convolutifs (CNN) + mémoire convolutive à long court terme (ConvLSTM) qui est entraînée de bout en bout de manière entièrement convolutive . Le modèle est construit sur un modèle de base ResNet-18 sans les couches denses pour l'extraction des caractéristiques.

Notre première contribution principale ici est de proposer une modélisation faiblement supervisée de la localisation spatiale à l'aide d'une couche de convolution à 7 canaux qui sert de *localisation heat-maps* (Lh-maps), également appelées cartes d'activation de classe (CAM). Chaque canal est par conception contraint d'apprendre et de localiser un type d'outil distinct parmi les 7 outils présents dans la procédure laparoscopique considérée. Pour s'assurer que la localisation est apprise par une supervision faible sur l'étiquette de présence binaire, une aggregation spatiale "wildcat" [Durand 2017] est utilisé pour transformer la *Lh-map* en un vecteur 1 × 7 de valeurs de confiance par classe indiquant la probabilité de présence de l'outil. Les classes positives sont sélectionnées par un seuil de 0,5. Les activations faiblement supervisées se situent généralement sur la partie la plus discriminante des objets. Pour capturer méticuleusement plus de détails, un patch de masquage aléatoire [Singh 2017] des images d'entrée est appliqué pendant l'entraînement.

Notre deuxième contribution principale consiste à tirer parti de la cohérence spatio-temporelle à travers les images vidéo pour suivre les instruments chirurgicaux. Nous y parvenons en étendant notre modèle de localisation avec un ConvLSTM pour lisser le pic d'activation dans les Lh-maps et modéliser les trajectoires lisses des instruments. Nous utilisons l'unité ConvLSTM d'une manière qui permet toujours une formation faiblement supervisée. Il en résulte une élégante méthode de suivi de bout en bout, *ConvLSTM Tracker*, capable de modéliser le mouvement spatio-temporel des outils et également de s'adapter aux différents types de mouvement apparaissant dans une vidéo. Dans cette tâche, un ConvLSTM est préféré aux autres unités RNN en raison de sa capacité à maintenir la relation spatiale des pixels localisés. Pour maintenir la continuité dans une vidéo et capturer des informations temporelles sur une séquence plus longue, l'unité ConvLSTM est conçue pour maintenir l'initialisation des états d'un seul coup uniquement au début d'une vidéo, après quoi les états sont propagés sur (des lots d') images pendant toute la durée d'une vidéo. Le ConvLSTM gère nativement le problème d'association de données dans les trajectoires d'outils. Cela aide également le modèle de détection à devenir plus robuste à l'occlusion et au bruit.

Table B.1 – Résultats quantitatifs sur la détection, la localisation et le suivi des outils.

Model	Détection	Localisation	Suivi: $\Theta = \mu$ $(0.3-0.7)$	
	(% mAP)	(% $IoU \geq 0.5$)	MOTP	MOTA
Référence	87.7	21.0	62.1	23.9
Proposé	92.9	38.2	67.4	36.5

La méthode proposée est validée sur l'ensemble de données Cholec80 obtenant des performances supérieures par rapport aux lignes de base avec des améliorations de 12,6% sur la précision de suivi multi-objets (MOTA), 13,9% sur la précision de localisation ($IoU \geq 0.5$) et 5 Le résultat qualitatif démontré dans la vidéo (https://youtu.be/vnMwlS5tvHE) montre que le modèle proposé est capable de localiser les instruments et de les classer correctement. Il révèle également que les *Lh-maps* produisent une faible segmentation des instruments, suggérant que cette méthode pourrait être étendue à la segmentation. Une autre vidéo qualitative sur https://youtu.be/SNhd1yzOe50 montre que le modèle ConvLSTM formé sur des vidéos à 1 ips peut se généraliser aux vidéos non étiquetées à 25 ips, ce qui le rend non contraint par les ips.

B.2.2 Reconnaissance des triplets d'action chirurgicale

Dans la deuxième partie de la thèse, nous construisons progressivement 3 modèles d'apprentissage en profondeur pour reconnaître des triplets d'action chirurgicale directement à partir de vidéos chirurgicales.

B.2.2.1 Approche d'apprentissage multi-tâches centrée sur l'instrument

Le premier modèle à cet égard est *Tripnet*, qui est un modèle centré sur l'instrument avec des branches d'apprentissage multi-tâches (MTL) pour modéliser les différentes composantes du triplet, à savoir: instrument, verbe et cible. Comme le montre la figure B.5(a), l'architecture commence également par un backbone ResNet-18, suivi des branches MTL pour les trois composants. Chaque branche est une couche à deux convolutions et une couche de classification. La branche instrument suit la méthode de localisation faiblement supervisée dans notre première tâche (section B.2.1.1) pour apprendre la classe d'instruments ainsi que leurs emplacements. Cette couche est également connue sous le nom de module *Weakly Supervised Localization* (WSL).

Pour plus de commodité, les branches verbe-cible sont regroupées pour former un module appelé *class activation guide (CAG)* qui utilise les cartes d'activation de classe d'instrument (CAM) de la branche instrument (ou WSL) pour guider la reconnaissance du verbe et de la cible. comme illustré sur la figure B.5(b). Ceci est basé sur l'hypothèse qu'en l'absence d'annotation

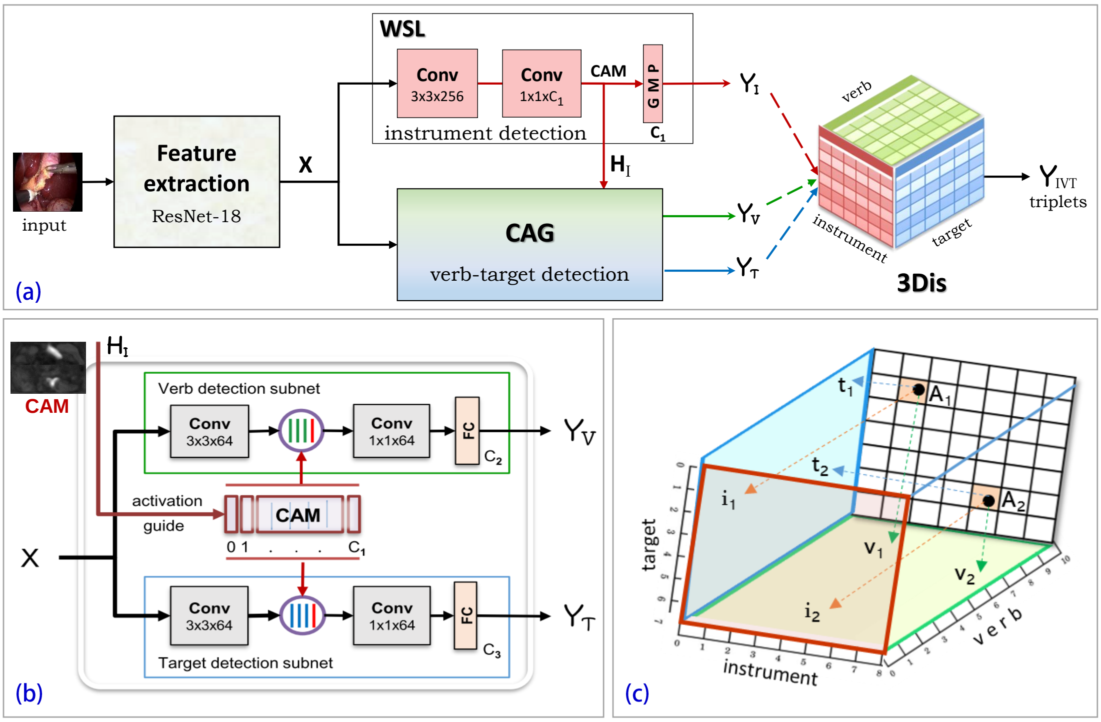


Figure B.5 – Tripnet : (a) l'architecture du modèle proposé pour la reconnaissance de triplet d'action. (b) module guide d'activation de classe (CAG), (c) espace d'interaction 3D (3Dis)

spatiale, la CAM de l'instrument dispose d'informations suffisantes pour conditionner l'espace de recherche du modèle et diriger les branches de détection de verbe et de cible vers la région d'intérêt probable des actions. Ceci est modélisé en concaténant l'entrée CAM avec les caractéristiques du verbe et de la cible respectivement pour leur fournir le repère d'apparence de l'instrument.

L'architecture se termine par un espace d'interaction 3D (3Dis) qui gère l'association tripartite complexe de la relation des composants du triplet en modélisant l'interaction outil-tissu correcte. Il est conçu à l'aide d'une fonction d'association qui est mise en œuvre par un produit externe des composants multi-étiquettes en interaction pondérés par certains vecteurs de projection. Il s'agit d'une amélioration par rapport à la matrice d'interaction 2D [Shen 2018] utilisée dans l'interaction homme-objet (HOI) qui est limitée aux seules relations verbe-objet. Le 3Dis permet également la reconnaissance de plusieurs triplets dans une même trame.

B.2.2.2 Mécanismes d'attention pour la détection améliorée des composants

Le deuxième modèle, connu sous le nom de *Attention Tripnet*, étend le Tripnet avec un mécanisme d'attention guidée par activation de classe (CAGAM), comme le montre la figure B.6, pour un conditionnement spatial plus précis des caractéristiques et une meilleure détection du verbe et des composants cibles du triplet.

Le CAGAM est basé sur un mécanisme d'attention. En règle générale, l'attention dans l'apprentissage en profondeur est la concentration d'un réseau de neurones sur les caractéristiques d'entrée les plus pertinentes pour produire la sortie souhaitée. Cette focalisation sélective est réalisable à l'aide d'un poids d'attention ou d'un score d'*affinité* résultant de la mise en correspondance d'une caractéristique de requête avec les caractéristiques clés correspondantes. Le CAGAM est un mécanisme d'attention spatiale où une affinité connue ou discriminante est utilisée pour améliorer une affinité inconnue pour la découverte de motifs pertinents. Nous implémentons CAGAM en utilisant des fonctionnalités d'affinité résultant de fonctionnalités CAM d'instruments connues pour guider le réseau pour la détection de verbes et de cibles, comme illustré dans la figure B.7.

Nous avons observé que les verbes et les cibles se comportent différemment vis-à-vis de

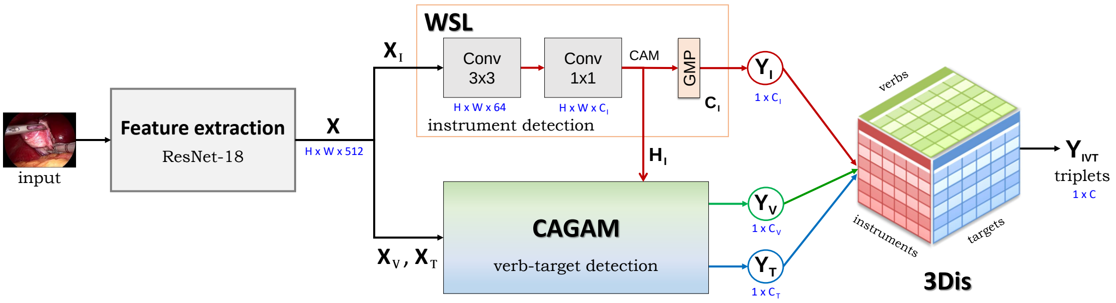


Figure B.6 – Attention Tripnet : l'architecture du modèle proposé pour la détection améliorée des composants triplet.

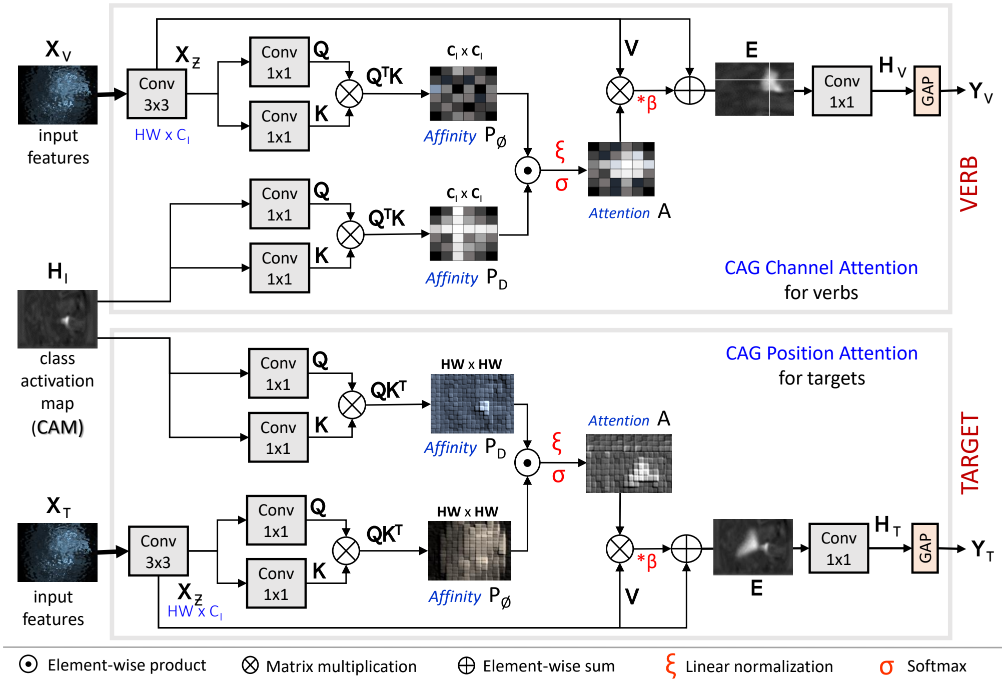


Figure B.7 – Mécanisme d'Attention Guidée par Activation de Classe (CAGAM) : utilise l'attention apprise du CAM de l'instrument pour mettre en évidence la classe verbale (en haut) et l'anatomie en contact avec l'instrument (en bas).

leur instrument. C'est-à-dire que le verbe est principalement affecté par le type d'instrument qui est discriminé par canal, tandis que la cible est principalement affectée par la position spatiale de cet instrument. Par conséquent, nous utilisons le mécanisme d'attention de canal pour la détection de verbe et le mécanisme d'attention de position pour la détection de cible. Les deux types d'attention sont similaires à l'exception de la dimension utilisée et donc de la nature de l'attention portée.

Le modèle Attention Tripnet est mis en œuvre en remplaçant le CAG de Tripnet par le nouveau CAGAM, enregistrant des performances améliorées à la fois dans la détection des composants triplet et la reconnaissance de leur association.

B.2.2.3 Méthode inspirée du Transformer pour une association de triplet améliorée

Le dernier modèle est un modèle inspiré du Transformer, connu sous le nom de *Rendezvous (RDV)*, qui tire parti de l'attention à plus long terme pour apprendre l'interaction outil-tissu comme illustré dans la figure B.8(une). La nouveauté réside dans son utilisation de Multi-Head of Mixed Attention (MHMA) qui combine à la fois des mécanismes d'auto-attention et d'attention croisée pour capturer l'interaction des trois composants d'un triplet. Le Rendezvous est mieux décrit comme la rencontre de plusieurs composants d'attention pour comprendre leur relation en tant que triplet. Il s'agit d'une amélioration par rapport au 3Dis [Nwoye 2020] qui est moins avancé, pour une meilleure association de composants de triplet.

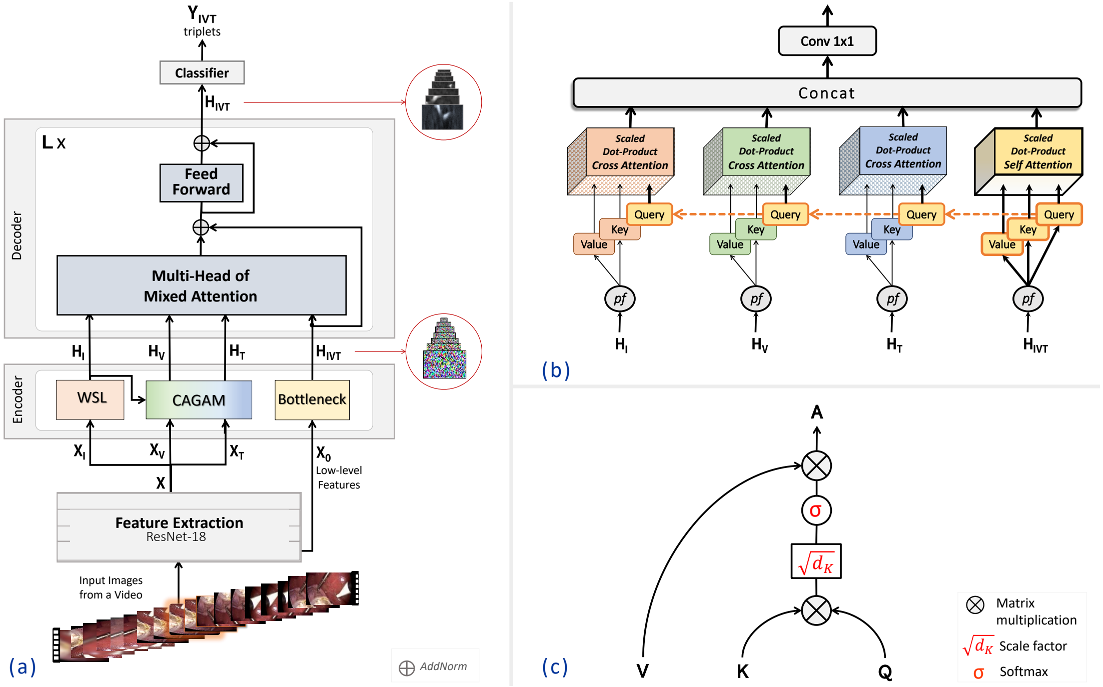


Figure B.8 – Vue d'ensemble du modèle: (a) architecture de Rendezvous: un réseau de neurones inspiré par Transformer pour la reconnaissance de triplet d'action, (b) une multi-tête de mécanismes d'auto-attention et d'attention croisée, (c) structure des mécanismes d'attention des produits scalaires à l'échelle: dans l'auto-attention, le triple (K,V,Q) vient d'un contexte de caractéristique, alors que dans l'attention croisée, la paire (K,V) vient du contexte de caractéristique source tandis que Q vient du contexte de caractéristique de puits.

Comme illustré dans la figure B.8(b), nous implémentons quatre têtes d'attention pour l'instrument, le verbe, la cible et le triplet. Nous implémentons une auto-attention sur la tête du triplet pour la compréhension initiale de la scène : dans ce cas, les caractéristiques de clé (K), de requête (Q) et de valeur (V) sont générées à partir des cartes de classe du triplet.

Avec chacune des caractéristiques des composants déjà discriminées (dans WSL et CAGAM) pour s'occuper d'un seul composant dans une scène d'image, la compréhension de leur relation sous-jacente nécessite une attention croisée entre les caractéristiques des composants. En plus de l'auto-attention, l'attention croisée est implémentée sur les trois têtes des composants, ajoutant la possibilité d'une meilleure modélisation des relations entre les composants participant à l'interaction outil-tissu. Ceci est important lors de la résolution des interactions : par exemple, une partie anatomique peut apparaître dans le cadre sans être une cible, rendant souvent l'interaction avec l'instrument ambiguë. MHMA modélise la façon dont les caractéristiques de chaque composant affectent la composition du triplet, en propageant les affinités de leurs caractéristiques de contexte respectives aux caractéristiques de triplet requises. Toutes les têtes d'attention sont implémentées en utilisant l'attention à produit scalaire normalisé, couramment employée, comme illustré dans la figure B.8(c). L'attention multi-têtes de sortie est encore affinée par une couche d'anticipation et le cycle de décodage continue jusqu'à $L = 8$ couches avant que les 8^{th} caractéristiques décodées soient classées à l'aide d'un perceptron multicouche (MLP).

Résultats et discussion

Table B.2 – Résumé des performances sur la reconnaissance des triplets d'action chirurgicale.

Méthode		Détection de composants			Association de triplés		
		AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Référence	Naive CNN	57.7	39.2	28.3	21.7	18.0	13.6
	Naive TCN	48.9	29.4	21.4	17.7	15.5	12.4
	MTL référence	84.5	48.4	28.2	26.6	21.2	17.6
Proposé	Tripnet	**92.1**	54.5	33.2	29.7	26.4	20.0
	Attention Tripnet	92.0	60.2	**38.5**	31.1	29.8	23.4
	Rendezvous	92.0	**60.7**	38.3	**39.4**	**36.9**	**29.9**

Les méthodes proposées sont évaluées sur le jeu de données CholectT50 pour la reconnaissance de triplet d'action. Les résultats des trois modèles proposés ainsi que les méthodes de référence sont présentés dans le tableau B.2. On observe que Tripnet, par sa stratégie MTL, obtient des performances de pointe sur la détection de présence d'instruments. Son amélioration des performances par rapport aux méthodes de base sur les détections de verbes et de cibles est encore amplifiée par l'Attention Tripnet exploitant le CAGAM. Le RDV améliore considérablement l'association des composants du triplet, établissant une nouvelle performance de pointe. En outre, la figure B.9 montre l'amélioration progressive du modèle proposé par rapport aux lignes de base pour la reconnaissance de triplet d'action chirurgicale, établissant une référence pour l'analyse future dans cette nouvelle tâche dans le domaine de la science des données chirurgicales.

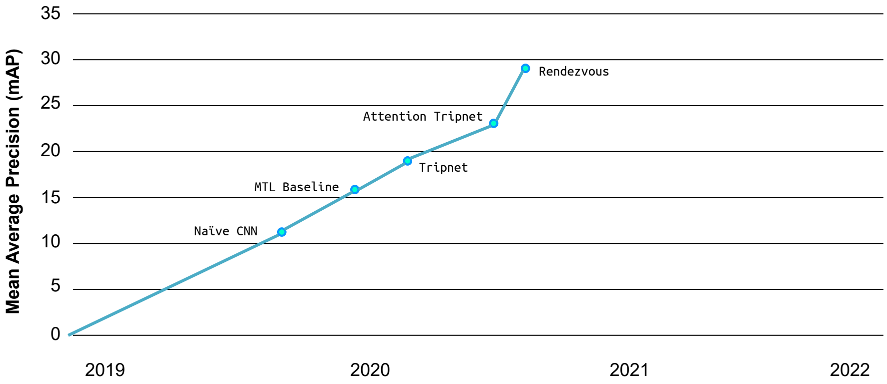


Figure B.9 – Résumé graphique des performances de tous les modèles évalués sur la reconnaissance de triplet d'action.

L'analyse des résultats qualitatifs de la figure B.10 montre la capacité du modèle RDV à reconnaître les triplets corrects ainsi qu'à localiser leurs régions d'interaction. Comme le montrent les images qualitatives, la majorité des prédictions incorrectes sont dues à un

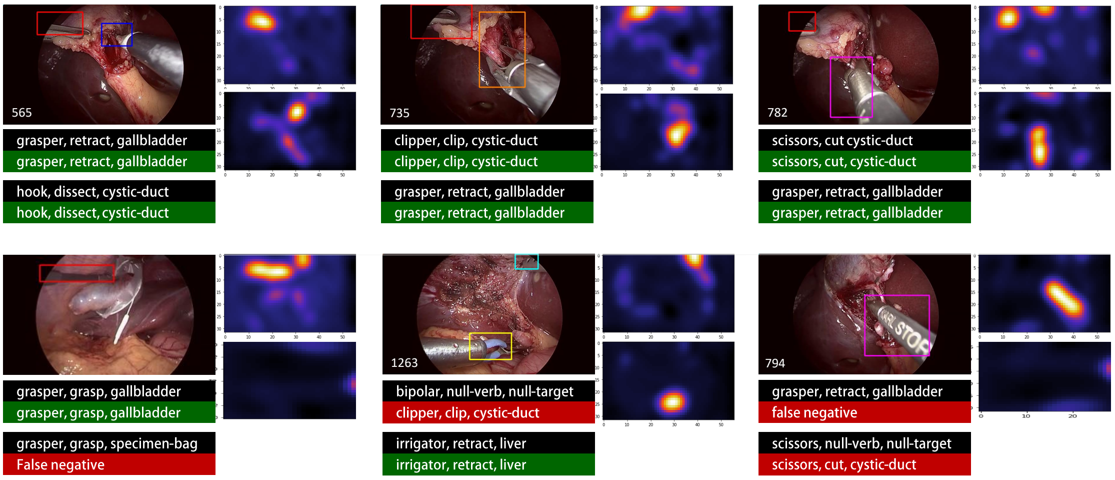

565
grasper, retract, gallbladder
grasper, retract, gallbladder
hook, dissect, cystic-duct
hook, dissect, cystic-duct
735
clipper, clip, cystic-duct
clipper, clip, cystic-duct
grasper, retract, gallbladder
grasper, retract, gallbladder
782
scissors, cut cystic-duct
scissors, cut, cystic-duct
grasper, retract, gallbladder
grasper, retract, gallbladder
grasper, grasp, gallbladder
grasper, grasp, gallbladder
grasper, grasp, specimen-bag
False negative
1263
bipolar, null-verb, null-target
clipper, clip, cystic-duct
irrigator, retract, liver
irrigator, retract, liver
794
grasper, retract, gallbladder
false negative
scissors, null-verb, null-target
scissors, cut, cystic-duct

Figure B.10 – Résultats qualitatifs du modèle RDV montrant les prédictions de triplet et les cartes thermiques pour la détection de triplet. Les cadres de délimitation de localisation sont obtenus à partir du module WSL du modèle RDV proposé. Les triplets prédits et vérité terrain sont affichés sous chaque image : noir = vérité terrain, vert = prédiction correcte, rouge = prédiction incorrecte. Un triplet manqué est marqué comme faux négatif et une fausse détection est marquée comme faux positif.

composant de triplet incorrect. Les instruments sont généralement correctement prédits et localisés. Cependant, il n'est pas simple de prédire le verbe/la cible directement à partir de l'instrument en raison des multiples associations possibles. Une démonstration vidéo intéressante des résultats qualitatifs est fournie au lien : https://youtu.be/d_yHdJtCa98. La vidéo montre également l'effet de l'attention CAGAM : pour cela, un pixel sur l'instrument crée une carte d'attention qui met en évidence la cible d'intérêt. Une telle carte d'attention est supprimée lorsque le point de pixel n'est pas sur un instrument.

B.3 Conclusion

Nous concluons en présentant le résumé des travaux réalisés dans cette thèse, nos contributions uniques, leurs applications cliniques et leurs perspectives.

B.3.1 Résumé et contribution

Étant motivés par la nécessité d'optimiser et de prendre en charge le flux de travail chirurgical au bloc opératoire à l'aide de solutions d'IA, nous développons des méthodes d'apprentissage en profondeur pour la détection et la reconnaissance d'outils chirurgicaux et d'activités à granularité fine dans les vidéos laparoscopiques. Notre première contribution est l'innovation d'une méthode d'apprentissage en profondeur qui peut apprendre la position spatiale et le mouvement des outils via une supervision faible en utilisant des annotations de présence binaires. Nous formulons également des activités chirurgicales sous forme de triplets ⟨*instrument, verbe, cible*⟩ et, comme contribution supplémentaire, proposons la première méthode d'apprentissage en profondeur pour reconnaître ces triplets directement à partir de vidéos.

Nous innovons une méthode s'appuyant sur l'activation instrumentale (CAG) et l'attention formée par celle-ci (CAGAM) pour mieux détecter les verbes et les cibles dans les triplets. Nous concevons également une méthode (3Dis et MHMA) pour résoudre l'association des composants car il peut y avoir plusieurs triplets par trame. Nous avons présenté des résultats de pointe sur des ensembles de données à grande échelle (localisation spatiale Cholec80, CholecT40 et CholecT50) que nous avons générés en tant que contribution supplémentaire à nos travaux dans cette thèse.

B.3.2 Applications cliniques et perspectives

Les modèles résultant de nos expériences possèdent le potentiel d'être déployés dans les systèmes CAI au bloc opératoire pour soutenir positivement la chirurgie, ainsi que d'être utilisés dans le développement d'applications médicales. Le modèle de suivi d'instrument peut prendre en charge des signaux d'interaction outil-tissu en temps réel, en particulier ceux nécessaires pour mettre en garde contre l'utilisation d'outils dans les régions dangereuses du corps. Les modèles de reconnaissance de triplet peuvent aider à la surveillance de la sécurité et au retour d'informations par la nature détaillée de sa reconnaissance d'activité. Les informations de triplet peuvent servir de déclencheurs utiles pour observer les points de contrôle de sécurité. Il peut également être utilisé dans l'anticipation des actions, l'estimation des risques, l'interface utilisateur adaptative et la génération de rapports. Avec des applications potentielles très médiatisées telles que la surveillance de la sûreté clinique, l'évaluation des compétences et les rapports objectifs, notre méthode proposée, ainsi que la publication de notre ensemble de données, apportent une valeur considérable au domaine de la compréhension de l'activité chirurgicale.

Les recherches futures amélioreraient probablement les performances de la tâche de reconnaissance et, de la même manière, identifieraient les emplacements où se situent chaque tâche d'activité. Une approche plus large serait de modéliser la scène complète de la chirurgie, dans le temps et dans l'espace, en utilisant des graphes et en capturant toutes les formes de relation, et même en s'étendant au-delà du site opératoire pour inclure les patients, les appareils et les cliniciens dans la salle d'opération. Cela aurait un impact énorme sur la chirurgie et le monde en général et faciliterait l'automatisation des parties de l'intervention chirurgicale.

References

[Ahmadi 2006] S.-A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner and N. Navab. *Recovery of surgical workflow without explicit models.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 420–428. Springer, 2006. (Cited on pages 11, 12, 47, 51, and 183)

[Aksamentov 2017] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux and N. Padoy. *Deep neural networks predict remaining surgery duration from cholecystectomy videos.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 586–593. Springer, 2017. (Cited on pages 11 and 180)

[Al Hajj 2017] H. Al Hajj, M. Lamard, K. Charrière, B. Cochener and G. Quellec. *Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network.* In 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pages 2002–2005. IEEE, 2017. (Cited on page 34)

[Al Hajj 2018] H. Al Hajj, M. Lamard, P.-H. Conze, B. Cochener and G. Quellec. *Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks.* Medical image analysis, vol. 47, pages 203–218, 2018. (Cited on pages 32, 33, 47, 174, and 182)

[Al Hajj 2019] H. Al Hajj, M. Lamard, P.-H. Conze, S. Roychowdhury, X. Hu, G. Maršalkaitė, O. Zisimopoulos, M. A. Dedmari, F. Zhao, J. Prellberg*et al. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery.* Medical image analysis, vol. 52, pages 24–41, 2019. (Cited on page 171)

[Allan 2012] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly and D. Stoyanov. *Toward detection and localization of instruments in minimally invasive surgery.* IEEE Transactions on Biomedical Engineering, vol. 60, no. 4, pages 1050–1058, 2012. (Cited on page 34)

[Allan 2019] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt*et al. 2017 robotic instrument segmentation challenge.* arXiv preprint arXiv:1902.06426, 2019. (Cited on pages xvi and 38)

[Allan 2020] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen*et al. 2018 robotic scene segmentation challenge.* arXiv preprint arXiv:2001.11190, 2020. (Cited on page 38)

[Almushyti 2019] M. Almushyti and F. W. Li. *Recognising human-object interactions using attention-based LSTMs.* In CGVC, pages 135–139, 2019. (Cited on page 45)

[Alshirbaji 2018] T. A. Alshirbaji, N. A. Jalal and K. Möller. *Surgical tool classification in laparoscopic videos using convolutional neural network.* Current Directions in Biomedical Engineering, vol. 4, no. 1, pages 407–410, 2018. (Cited on page 32)

[Bae 2014] S.-H. Bae and K.-J. Yoon. *Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1218–1225, 2014. (Cited on pages 40 and 51)

[Bahdanau 2014] D. Bahdanau, K. Cho and Y. Bengio. *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473, 2014. (Cited on pages 46 and 106)

[Baldas 2010] V. Baldas, L. Tang, P. Bountris, G. Saleh and D. Koutsouris. *A real-time automatic instrument tracking system on cataract surgery videos for dexterity assessment.* In Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, pages 1–4. IEEE, 2010. (Cited on page 43)

[Ballantyne 2002] G. H. Ballantyne. *The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery.* Surgical Laparoscopy Endoscopy & Percutaneous Techniques, vol. 12, no. 1, pages 1–5, 2002. (Cited on pages 5, 8, 9, 179, and 180)

[Banerjee 2019] N. Banerjee, R. Sathish and D. Sheet. *Deep neural architecture for localization and tracking of surgical tools in cataract surgery.* In Computer Aided Intervention and Diagnostics in Clinical and Medical Images, pages 31–38. Springer, 2019. (Cited on pages 43 and 183)

[Baumhauer 2008] M. Baumhauer, M. Feuerstein, H.-P. Meinzer and J. Rassweiler. *Navigation in endoscopic soft tissue surgery: perspectives and limitations.* Journal of endourology, vol. 22, no. 4, pages 751–766, 2008. (Cited on page 17)

[Bawa 2021] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo*et al. The SARAS Endoscopic Surgeon Action Detection (ESAD) dataset: Challenges and methods.* arXiv preprint arXiv:2104.03178, 2021. (Cited on pages 20, 49, and 52)

[Bendersky 2014] M. Bendersky, L. Garcia-Pueyo, J. Harmsen, V. Josifovski and D. Lepikhin. *Up next: retrieval methods for large scale related video suggestion.* In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1769–1778, 2014. (Cited on page 44)

[Bergmann 2019] P. Bergmann, T. Meinhardt and L. Leal-Taixe. *Tracking without bells and whistles.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 941–951, 2019. (Cited on pages 40 and 51)

[Bernardin 2008] K. Bernardin and R. Stiefelhagen. *Evaluating multiple object tracking performance: the clear mot metrics.* EURASIP Journal on Image and Video Processing, vol. 2008, pages 1–10, 2008. (Cited on pages 40, 51, and 72)

[Bertinetto 2016] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. Torr. *Fully-convolutional siamese networks for object tracking.* In European conference on computer vision, pages 850–865. Springer, 2016. (Cited on page 40)

[Bewley 2016] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft. *Simple online and realtime tracking.* In 2016 IEEE international conference on image processing (ICIP), pages 3464–3468. IEEE, 2016. (Cited on page 39)

[Birkmeyer 2013] J. D. Birkmeyer, J. F. Finks, A. O'Reilly, M. Oerline, A. M. Carlin, A. R. Nunn, J. Dimick, M. Banerjee and N. J. Birkmeyer. *Surgical skill and complication rates after bariatric surgery.* New England Journal of Medicine, vol. 369, no. 15, pages 1434–1442, 2013. (Cited on page 5)

[Blank 2005] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri. *Actions as space-time shapes.* In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, volume 2, pages 1395–1402. IEEE, 2005. (Cited on pages 44 and 49)

[Blum 2008] T. Blum, N. Padoy, H. Feußner and N. Navab. *Modeling and online recognition of surgical phases using hidden markov models.* In International conference on medical image computing and computer-assisted intervention, pages 627–635. Springer, 2008. (Cited on page 47)

[Blum 2010] T. Blum, H. Feußner and N. Navab. *Modeling and segmentation of surgical workflow from laparoscopic video.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 400–407, 2010. (Cited on pages 26, 47, and 183)

[Bodenstedt 2017] S. Bodenstedt, M. Wagner, D. Katić, P. Mietkowski, B. Mayer, H. Kenngott, B. Müller-Stich, R. Dillmann and S. Speidel. *Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis.* arXiv preprint arXiv:1702.03684, 2017. (Cited on page 22)

[Bodenstedt 2018a] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov*et al. Comparative* *evaluation of instrument segmentation and tracking methods in minimally invasive surgery.* arXiv preprint arXiv:1805.02475, 2018. (Cited on pages 42 and 43)

[Bodenstedt 2018b] S. Bodenstedt, A. Ohnemus, D. Katic, A.-L. Wekerle, M. Wagner, H. Kenngott, B. Müller-Stich, R. Dillmann and S. Speidel. *Real-time image-based instrument classification for laparoscopic surgery.* arXiv preprint arXiv:1808.00178, 2018. (Cited on page 34)

[Bouget 2015] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele and P. Jannin. *Detecting surgical tools by modelling local appearance and global shape.* IEEE transactions on medical imaging, vol. 34, no. 12, pages 2603–2617, 2015. (Cited on page 34)

[Bouget 2017] D. Bouget, M. Allan, D. Stoyanov and P. Jannin. *Vision-based and marker-less surgical tool detection and tracking: a review of the literature.* Medical image analysis, vol. 35, pages 633–654, 2017. (Cited on pages 11, 34, 42, 43, and 180)

[Brasó 2020] G. Brasó and L. Leal-Taixé. *Learning a neural solver for multiple object tracking.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6247–6257, 2020. (Cited on page 39)

[Bucholz 1995] R. D. Bucholz. *Introduction to journal of image guided surgery.* Journal of image guided surgery, vol. 1, no. 1, pages 1–3, 1995. (Cited on page 7)

[Cadene 2016] R. Cadene, T. Robert, N. Thome and M. Cord. *M2cai workflow challenge: convolutional neural networks with time smoothing and hidden Markov model for video frames classification.* arXiv preprint arXiv:1610.05541, 2016. (Cited on page 47)

[Chakraborty 2013] I. Chakraborty, A. Elgammal and R. S. Burd. *Video based activity recognition in trauma resuscitation.* In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–8, 2013. (Cited on page 47)

[Chao 2015] Y.-W. Chao, Z. Wang, Y. He, J. Wang and J. Deng. *Hico: A benchmark for recognizing human-object interactions in images.* In Proceedings of the IEEE International Conference on Computer Vision, pages 1017–1025, 2015. (Cited on pages xvii, 44, and 81)

[Chao 2018] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng and J. Deng. *Learning to detect human-object interactions.* In 2018 ieee winter conference on applications of computer vision (wacv), pages 381–389. IEEE, 2018. (Cited on pages xvii, 44, 49, 81, and 98)

[Charriere 2014] K. Charriere, G. Quellec, M. Lamard, G. Coatrieux, B. Cochener and G. Cazuguel. *Automated surgical step recognition in normalized cataract surgery videos.* In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 4647–4650. IEEE, 2014. (Cited on pages 47 and 183)

[Charriere 2016] K. Charriere, G. Quelled, M. Lamard, D. Martiano, G. Cazuguel, G. Coatrieux and B. Cochener. *Real-time multilevel sequencing of cataract surgery videos.* In 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pages 1–6. IEEE, 2016. (Cited on page 48)

[Chen 2017] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille. *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.* IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pages 834–848, 2017. (Cited on page 172)

[Chen 2018] W. Chen, J. Feng, J. Lu and J. Zhou. *Endo3d: online workflow analysis for endoscopic surgeries based on 3d cnn and lstm.* In OR 2.0 Context-Aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis, pages 97–107. Springer, 2018. (Cited on pages 35, 36, and 51)

[Chen 2021] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille and Y. Zhou. *Transunet: Transformers make strong encoders for medical image segmentation.* arXiv preprint arXiv:2102.04306, 2021. (Cited on page 123)

[Cho 2014] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio. *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* arXiv preprint arXiv:1406.1078, 2014. (Cited on page 33)

[Choi 2017] B. Choi, K. Jo, S. Choi and J. Choi. *Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery.* In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1756–1759. Ieee, 2017. (Cited on pages 34, 35, and 182)

[Chollet 2017] F. Chollet. *Xception: Deep learning with depthwise separable convolutions.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017. (Cited on page 172)

[Colleoni 2019] E. Colleoni, S. Moccia, X. Du, E. De Momi and D. Stoyanov. *Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers.* IEEE Robotics and Automation Letters, vol. 4, no. 3, pages 2714–2721, 2019. (Cited on pages 42, 50, and 183)

[Czempiel 2020] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim and N. Navab. *TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 343–352. Springer, 2020. (Cited on pages xvi, 47, 48, and 68)

[Czempiel 2021] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam and N. Navab. *OperA: Attention-Regularized Transformers for Surgical Phase Recognition.* arXiv preprint arXiv:2103.03873, 2021. (Cited on page 47) [da Costa Rocha 2019] C. da Costa Rocha, N. Padoy and B. Rosa. *Self-supervised surgical tool segmentation using kinematic information.* In 2019 International Conference on Robotics and Automation (ICRA), pages 8720–8726. IEEE, 2019. (Cited on pages 22, 35, and 38)

[Dawoud 2020] Y. Dawoud, J. Hornauer, G. Carneiro and V. Belagiannis. *Few-Shot Microscopy Image Cell Segmentation.* arXiv preprint arXiv:2007.01671, 2020. (Cited on page 22)

[Delaitre 2011] V. Delaitre, J. Sivic and I. Laptev. *Learning person-object interactions for action recognition in still images.* In NIPS 2011: Twenty-Fifth Annual Conference on Neural Information Processing Systems, 2011. (Cited on page 44)

[Dendorfer 2020] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler and L. Leal-Taixé. *Mot20: A benchmark for multi object tracking in crowded scenes.* arXiv preprint arXiv:2003.09003, 2020. (Cited on pages 40 and 51)

[Deng 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. *Imagenet: A large-scale hierarchical image database.* In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. (Cited on pages 32, 71, and 94)

[Dergachyova 2016] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi and P. Jannin. *Automatic data-driven real-time segmentation and recognition of surgical workflow.* International journal of computer assisted radiology and surgery, vol. 11, no. 6, pages 1081–1089, 2016. (Cited on pages 26, 47, and 183)

[Dergachyova 2018] O. Dergachyova, X. Morandi and P. Jannin. *Knowledge transfer for surgical activity prediction.* International journal of computer assisted radiology and surgery, vol. 13, no. 9, pages 1409–1417, 2018. (Cited on page 22)

[Ding 2020] N. Ding, N. A. Jalal, T. Alshirbaji and K. Möller. *The evaluation of synthetic datasets on training AlexNet for surgical tool detection.* Current Directions in Biomedical Engineering, vol. 6, no. 3, pages 319–321, 2020. (Cited on pages 21 and 22)

[DiPietro 2016] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee and G. D. Hager. *Recognizing surgical activities with recurrent neural networks.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 551–558, 2016. (Cited on pages 26, 47, 51, 180, and 182)

[DiPietro 2019] R. DiPietro, N. Ahmidi, A. Malpani, M. Waldram, G. I. Lee, M. R. Lee, S. S. Vedula and G. D. Hager. *Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks.* International journal of computer assisted radiology and surgery, vol. 14, no. 11, pages 2005–2020, 2019. (Cited on pages 47 and 183)

[Do 2017] T. Do and J. Pustejovsky. *Fine-grained event learning of human-object interaction with lstm-crf.* arXiv preprint arXiv:1710.00262, 2017. (Cited on pages 45 and 53)

[Doignon 2005] C. Doignon, P. Graebling and M. De Mathelin. *Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature.* Real-Time Imaging, vol. 11, no. 5-6, pages 429–442, 2005. (Cited on page 34)

[Dollár 2009] P. Dollár, C. Wojek, B. Schiele and P. Perona. *Pedestrian detection: A benchmark.* In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 304–311. IEEE, 2009. (Cited on pages 40 and 51)

[Dong 2016] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu and H. Huang. *Occlusion-aware real-time object tracking.* IEEE Transactions on Multimedia, vol. 19, no. 4, pages 763–771, 2016. (Cited on page 40)

[Dong 2019] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling and F. Porikli. *Dynamical hyperparameter optimization via deep reinforcement learning in tracking.* IEEE transactions on pattern analysis and machine intelligence, 2019. (Cited on page 40)

[Dosovitskiy 2020] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly*et al. An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020. (Cited on pages xxii, 123, 126, and 130)

[Du 2016] X. Du, M. Allan, A. Dore, S. Ourselin, D. Hawkes, J. D. Kelly and D. Stoyanov. *Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery.* International journal of computer assisted radiology and surgery, vol. 11, no. 6, pages 1109–1119, 2016. (Cited on pages 42, 50, and 183)

[Du 2018] X. Du, T. Kurmann, P.-L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly and D. Stoyanov. *Articulated multi-instrument 2-D pose estimation using fully convolutional networks.* IEEE transactions on medical imaging, vol. 37, no. 5, pages 1276–1287, 2018. (Cited on pages 42, 43, 50, and 183)

[Du 2019] X. Du, M. Allan, S. Bodenstedt, L. Maier-Hein, S. Speidel, A. Dore and D. Stoyanov. *Patch-based adaptive weighting with segmentation and scale (PAWSS) for visual tracking in surgical video.* Medical image analysis, vol. 57, pages 120–135, 2019. (Cited on page 43)

[Durand 2017] T. Durand, T. Mordan, N. Thome and M. Cord. *Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation.* In CVPR, volume 2, 2017. (Cited on pages xvi, 64, 68, and 185)

[Ellis 2010] A. Ellis and J. Ferryman. *Pets2010: Dataset and challenge.* AVSS, 00 (undefined), pages 143–150, 2010. (Cited on pages 40 and 51)

[Everingham 2010] M. Everingham, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman. *The pascal visual object classes (voc) challenge.* International journal of computer vision, vol. 88, no. 2, pages 303–338, 2010. (Cited on page 44)

[Fang 2015] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt*et al.* *From captions to visual concepts and back.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015. (Cited on pages 44 and 49)

[Fang 2018] K. Fang, Y. Xiang, X. Li and S. Savarese. *Recurrent autoregressive networks for online multi-object tracking.* In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 466–475. IEEE, 2018. (Cited on page 41)

[Felli 2019] E. Felli, P. Mascagni, T. Wakabayashi, D. Mutter, J. Marescaux and P. Pessaux. *Feasibility and value of the critical view of safety in difficult cholecystectomies.* Annals of surgery, vol. 269, no. 4, page e41, 2019. (Cited on page 5)

[Fitzek 2021] F. H. Fitzek, S.-C. Li, S. Speidel, T. Strufe, M. Simsek and M. Reisslein. Tactile internet: With human-in-the-loop. Academic Press, 2021. (Cited on pages 10 and 180)

[Fotouhi 2011] M. Fotouhi, A. Gholami and S. Kasaei. *Particle filter-based object tracking using adaptive histogram.* In 2011 7th Iranian Conference on Machine Vision and Image Processing, pages 1–5. IEEE, 2011. (Cited on page 39)

[Fried 1997] M. P. Fried, J. Kleefield, H. Gopal, E. Reardon, B. T. Ho and F. A. Kuhn. *Image-guided endoscopic surgery: results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system.* The Laryngoscope, vol. 107, no. 5, pages 594–601, 1997. (Cited on page 30)

[Fu 2019] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu. *Dual attention network for scene segmentation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3146–3154, 2019. (Cited on pages 46, 106, and 126)

[Fuchs 2002] K. Fuchs. *Minimally invasive surgery.* Endoscopy, vol. 34, no. 02, pages 154–159, 2002. (Cited on page 8)

[Fuentes-Hurtado 2019] F. Fuentes-Hurtado, A. Kadkhodamohammadi, E. Flouty, S. Barbarisi, I. Luengo and D. Stoyanov. *EasyLabels: weak labels for scene segmentation in laparoscopic videos.* International journal of computer assisted radiology and surgery, vol. 14, no. 7, pages 1247–1257, 2019. (Cited on pages 37 and 183)

[Funke 2018] I. Funke, A. Jenke, S. T. Mees, J. Weitz, S. Speidel and S. Bodenstedt. *Temporal coherence-based self-supervised learning for laparoscopic workflow analysis.* In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pages 85–93. 2018. (Cited on pages 22, 47, and 183)

[Gaab 2013] M. R. Gaab. *Instrumentation: endoscopes and equipment.* World neurosurgery, vol. 79, no. 2, pages S14–e11, 2013. (Cited on page 8)

[Galloway 2015] R. L. Galloway. *Chapter 1 - Introduction and Historical Perspectives on Image-Guided Surgery.* In A. J. Golby, editor, Image-Guided Neurosurgery, pages 1–22. Academic Press, Boston, 2015. (Cited on page 7)

[Gao 2018] C. Gao, Y. Zou and J.-B. Huang. *ican: Instance-centric attention network for human-object interaction detection.* arXiv preprint arXiv:1808.10437, 2018. (Cited on page 46)

[Gao 2021] X. Gao, Y. Jin, Y. Long, Q. Dou and P.-A. Heng. *Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer.* arXiv preprint arXiv:2103.09712, 2021. (Cited on page 47)

[García-Peraza-Herrera 2016] L. C. García-Peraza-Herrera, W. Li, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren and S. Ourselin. *Real-time segmentation of non-rigid surgical tools based on deep learning and tracking.* In International Workshop on Computer-Assisted and Robotic Endoscopy, pages 84–95. Springer, 2016. (Cited on page 42)

[Garrow 2021] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel*et al. Machine learning for surgical phase recognition: a systematic review.* Annals of Surgery, vol. 273, no. 4, pages 684–693, 2021. (Cited on pages 6, 11, and 180)

[Geiger 2012] A. Geiger, P. Lenz and R. Urtasun. *Are we ready for autonomous driving? the kitti vision benchmark suite.* In 2012 IEEE conference on computer vision and pattern recognition, pages 3354–3361. IEEE, 2012. (Cited on pages 40 and 51)

[Gibson 2018] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu*et al. NiftyNet: a deep-learning platform for medical imaging.* Computer methods and programs in biomedicine, vol. 158, pages 113–122, 2018. (Cited on pages 5 and 179)

[Gkioxari 2014] G. Gkioxari, B. Hariharan, R. Girshick and J. Malik. *R-cnns for pose estimation and action detection.* arXiv preprint arXiv:1406.5212, 2014. (Cited on page 44)

[Gkioxari 2015] G. Gkioxari, R. Girshick and J. Malik. *Contextual action recognition with r* cnn.* In Proceedings of the IEEE international conference on computer vision, pages 1080–1088, 2015. (Cited on page 44)

[Gkioxari 2018] G. Gkioxari, R. Girshick, P. Dollár and K. He. *Detecting and recognizing human-object interactions.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8359–8367, 2018. (Cited on pages xvii, 44, 52, 90, 91, and 98)

[Grammatikopoulou 2019] M. Grammatikopoulou, E. Flouty, A. Kadkhodamohammadi, G. Quellec, A. Chow, J. Nehme, I. Luengo and D. Stoyanov. *CaDIS: Cataract dataset for image segmentation.* arXiv preprint arXiv:1906.11586, 2019. (Cited on pages xix and 171)

[Grimson 1999] W. Grimson, R. Kikinis, F. Jolesz and P. Black. *Image-guided surgery.* Sci Am, 1999. (Cited on page 7)

[Gupta 2015] S. Gupta and J. Malik. *Visual semantic role labeling.* arXiv preprint arXiv:1505.04474, 2015. (Cited on pages xvii, 45, and 81)

[Haase 2013] S. Haase, J. Wasza, T. Kilgus and J. Hornegger. *Laparoscopic instrument localization using a 3-D Time-of-Flight/RGB endoscope.* In 2013 IEEE Workshop on Applications of Computer Vision (WACV), pages 449–454. IEEE, 2013. (Cited on pages 18 and 34)

[Hager 1995] G. D. Hager, W.-C. Chang and A. S. Morse. *Robot hand-eye coordination based on stereo vision.* IEEE Control Systems Magazine, vol. 15, no. 1, pages 30–39, 1995. (Cited on pages 10 and 180)

[Hager 1996] G. Hager. *Special section on vision-based control of robot manipulators.* IEEE Transactions on Robotics and Automation, vol. 12, no. 5, pages 649–650, 1996. (Cited on page 10)

[Hariyono 2014] J. Hariyono, V.-D. Hoang and K.-H. Jo. *Moving object localization using optical flow for pedestrian detection from a moving vehicle.* The Scientific World Journal, vol. 2014, 2014. (Cited on page 39)

[Haro 2012] B. B. Haro, L. Zappella and R. Vidal. *Surgical gesture classification from video data.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 34–41. Springer, 2012. (Cited on page 48)

[He 2016] K. He, X. Zhang, S. Ren and J. Sun. *Deep residual learning for image recognition.* In CVPR, pages 770–778, 2016. (Cited on pages 32, 62, 63, and 175)

[Henriques 2014] J. F. Henriques, R. Caseiro, P. Martins and J. Batista. *High-speed tracking with kernelized correlation filters.* IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 3, pages 583–596, 2014. (Cited on page 40)

[Hochreiter 1997] S. Hochreiter and J. Schmidhuber. *Long short-term memory.* Neural computation, vol. 9, no. 8, pages 1735–1780, 1997. (Cited on page 33)

[Horenko 2020] I. Horenko. *On a scalable entropic breaching of the overfitting barrier for small data problems in machine learning.* Neural Computation, vol. 32, no. 8, pages 1563–1579, 2020. (Cited on page 21)

[Hu 2009] Y. Hu, H. U. Ahmed, C. Allen, D. Pendsé, M. Sahu, M. Emberton, D. Hawkes and D. Barratt. *MR to ultrasound image registration for guiding prostate biopsy and interventions.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 787–794. Springer, 2009. (Cited on page 30)

[Hu 2013] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong and T. Xiang. *Recognising human-object interaction via exemplar based modelling.* In Proceedings of the IEEE international conference on computer vision, pages 3144–3151, 2013. (Cited on page 44)

[Hu 2017] X. Hu, L. Yu, H. Chen, J. Qin and P.-A. Heng. *AGNet: Attention-guided network for surgical tool presence detection.* In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 186–194. Springer, 2017. (Cited on page 33)

[Huang 2019] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu. *Ccnet: Criss-cross attention for semantic segmentation.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 603–612, 2019. (Cited on page 126)

[Huang 2020] Y. Huang, F. Zhu, Z. Zeng, X. Qiu, Y. Shen and J. Wu. *Sqe: a self quality evaluation metric for parameters optimization in multi-object tracking.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8306–8314, 2020. (Cited on pages 40 and 51)

[Huber 1992] P. J. Huber. *Robust estimation of a location parameter.* In Breakthroughs in statistics, pages 492–518. Springer, 1992. (Cited on page 33)

[Hwang 2016] S. Hwang and H.-E. Kim. *Self-transfer learning for weakly supervised lesion localization.* In International conference on medical image computing and computer-assisted intervention, pages 239–246. Springer, 2016. (Cited on pages 37 and 61)

[Iandola 2014] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell and K. Keutzer. *Densenet: Implementing efficient convnet descriptor pyramids.* arXiv preprint arXiv:1404.1869, 2014. (Cited on page 32)

[Idrees 2017] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar and M. Shah. *The THUMOS challenge on action recognition for videos "in the wild".* Computer Vision and Image Understanding, vol. 155, pages 1–23, 2017. (Cited on page 44)

[Jannin 2001] P. Jannin, M. Raimbault, X. Morandi, E. Seigneuret and B. Gibaud. *Design of a neurosurgical procedure model for multimodal image-guided surgery.* In International Congress Series, volume 1230, pages 102–106. Elsevier, 2001. (Cited on page 11)

[Ji 2019] Z. Ji, H. Wang, J. Han and Y. Pang. *Saliency-guided attention network for image-sentence matching.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5754–5763, 2019. (Cited on page 106)

[Jia 2017] Z. Jia, X. Huang, I. Eric, C. Chang and Y. Xu. *Constrained deep weak supervision for histopathology image segmentation.* IEEE transactions on medical imaging, vol. 36, no. 11, pages 2376–2388, 2017. (Cited on pages 19, 25, 37, and 61)

[Jiang 2003] B.-a. Jiang and H.-z. LU. *Particle Filter for target tracking [J].* Radar Science and Technology, vol. 3, 2003. (Cited on page 39)

[Jiang 2007] H. Jiang, S. Fels and J. J. Little. *A linear programming approach for multiple object tracking.* In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007. (Cited on page 40)

[Jiang 2017] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen and Y. Wang. *Artificial intelligence in healthcare: past, present and future.* Stroke and vascular neurology, vol. 2, no. 4, 2017. (Cited on pages 10 and 12)

[Jiang 2021] H. Jiang, Z. Yu, W. Nie, Y. Zhu and A. Anandkumar. *Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions.* 2021. (Cited on pages 45 and 49)

[Jin 2018] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein and L. Fei-Fei. *Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks.* In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 691–699. IEEE, 2018. (Cited on pages xv, 10, 15, 34, 35, 36, 42, 50, 180, and 182)

[Jin 2020] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu and P.-A. Heng. *Multi-task recurrent convolutional network with correlation loss for surgical video analysis.* Medical image analysis, vol. 59, page 101572, 2020. (Cited on pages 32, 33, 88, and 182)

[Kadkhodamohammadi 2014] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin and N. Padoy. *Temporally consistent 3D pose estimation in the interventional room using discrete MRF optimization over RGBD sequences.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 168–177. Springer, 2014. (Cited on pages 11 and 180)

[Kannan 2019] S. Kannan, G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *Future-state predicting LSTM for early surgery type recognition.* IEEE transactions on medical imaging, vol. 39, no. 3, pages 556–566, 2019. (Cited on pages 11, 46, 180, and 183)

[Karpathy 2015] A. Karpathy and L. Fei-Fei. *Deep visual-semantic alignments for generating image descriptions.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015. (Cited on pages 44 and 49)

[Katić 2014] D. Katić, A.-L. Wekerle, F. Gärtner, H. Kenngott, B. P. Müller-Stich, R. Dillmann and S. Speidel. *Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 158–167. Springer, 2014. (Cited on pages 6, 14, 16, 26, 48, 49, 81, 138, 182, and 184)

[Katić 2015] D. Katić, C. Julliard, A.-L. Wekerle, H. Kenngott, B. P. Müller-Stich, R. Dillmann, S. Speidel, P. Jannin and B. Gibaud. *LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase recognition.* International journal of computer assisted radiology and surgery, vol. 10, no. 9, pages 1427–1434, 2015. (Cited on pages 6, 26, 48, 49, 81, and 184)

[Kay 2017] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev*et al. The kinetics human action video dataset.* arXiv preprint arXiv:1705.06950, 2017. (Cited on pages 44 and 46)

[Kendall 2018] A. Kendall, Y. Gal and R. Cipolla. *Multi-task learning using uncertainty to weigh losses for scene geometry and semantics.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7482–7491, 2018. (Cited on page 111)

[Keuper 2016] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox and B. Schiele. *A multi-cut formulation for joint segmentation and tracking of multiple objects.* arXiv preprint arXiv:1607.06317, 2016. (Cited on pages 40 and 51)

[Khatibi 2020] T. Khatibi and P. Dezyani. *Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos.* Multimedia Tools and Applications, vol. 79, no. 41, pages 30111–30133, 2020. (Cited on pages 13, 20, 47, 52, 182, and 184)

[Kim 2016] D.-S. Kim and J. Kwon. *Moving object detection on a vehicle mounted back-up camera.* Sensors, vol. 16, no. 1, page 23, 2016. (Cited on page 39)

[Kim 2021] B. Kim, J. Lee, J. Kang, E.-S. Kim and H. J. Kim. *HOTR: End-to-End Human-Object Interaction Detection with Transformers.* arXiv preprint arXiv:2104.13682, 2021. (Cited on page 46)

[Kitaguchi 2019] D. Kitaguchi, N. Takeshita, H. Matsuzaki, H. Takano, Y. Owada, T. Enomoto, T. Oda, H. Miura, T. Yamanashi, M. Watanabe*et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach.* Surgical Endoscopy, pages 1–8, 2019. (Cited on pages 47 and 183)

[Kletz 2017] S. Kletz, K. Schoeffmann, B. Münzer, M. J. Primus and H. Husslein. *Surgical action retrieval for assisting video review of laparoscopic skills.* In Proceedings of the 2017 ACM Workshop on Multimedia-based Educational and Knowledge Technologies for Personalized and Social Online Training, pages 11–19, 2017. (Cited on page 20)

[Kokkinos 2017] I. Kokkinos. *Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6129–6138, 2017. (Cited on page 41)

[Kolesnikov 2019] A. Kolesnikov, A. Kuznetsova, C. Lampert and V. Ferrari. *Detecting visual relationships using box attention.* In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019. (Cited on page 46)

[Kondo 2020] S. Kondo. *LapFormer: surgical tool detection in laparoscopic surgical video using transformer architecture.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pages 1–6, 2020. (Cited on pages 33 and 36)

[Kranzfelder 2013a] M. Kranzfelder, A. Schneider, A. Fiolka, E. Schwan, S. Gillen, D. Wilhelm, R. Schirren, S. Reiser, B. Jensen and H. Feussner. *Real-time instrument detection in* *minimally invasive surgery using radiofrequency identification technology.* journal of surgical research, vol. 185, no. 2, pages 704–710, 2013. (Cited on pages 34 and 42)

[Kranzfelder 2013b] M. Kranzfelder, C. Staub, A. Fiolka, A. Schneider, S. Gillen, D. Wilhelm, H. Friess, A. Knoll and H. Feussner. *Toward increased autonomy in the surgical OR: needs, requests, and expectations.* Surgical endoscopy, vol. 27, no. 5, pages 1681–1688, 2013. (Cited on pages 10 and 12)

[Kranzfelder 2014] M. Kranzfelder, A. Schneider, A. Fiolka, S. Koller, S. Reiser, T. Vogel, D. Wilhelm and H. Feussner. *Reliability of sensor-based real-time workflow recognition in laparoscopic cholecystectomy.* International journal of computer assisted radiology and surgery, vol. 9, no. 6, pages 941–948, 2014. (Cited on page 11)

[Kristan 2015] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay and R. Pflugfelder. *The visual object tracking vot2015 challenge results.* In Proceedings of the IEEE international conference on computer vision workshops, pages 1–23, 2015. (Cited on pages 40 and 51)

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever and G. E. Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, vol. 25, pages 1097–1105, 2012. (Cited on pages 31 and 32)

[Kuhn 1955] H. W. Kuhn. *The Hungarian method for the assignment problem.* Naval research logistics quarterly, vol. 2, no. 1-2, pages 83–97, 1955. (Cited on pages 39 and 70)

[Kumar 2013] S. Kumar, M. S. Narayanan, P. Singhal, J. J. Corso and V. Krovi. *Product of tracking experts for visual tracking of surgical tools.* In 2013 IEEE International Conference on Automation Science and Engineering (CASE), pages 480–485. IEEE, 2013. (Cited on page 42)

[Kushwaha 2012] A. K. S. Kushwaha, O. Prakash, A. Khare and M. H. Kolekar. *Rule based human activity recognition for surveillance system.* In 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pages 1–6. IEEE, 2012. (Cited on page 44)

[Lahanas 2016] V. Lahanas, C. Loukas and E. Georgiou. *A simple sensor calibration technique for estimating the 3D pose of endoscopic instruments.* Surgical endoscopy, vol. 30, no. 3, pages 1198–1204, 2016. (Cited on page 30)

[Lalys 2014] F. Lalys and P. Jannin. *Surgical process modelling: a review.* International journal of computer assisted radiology and surgery, vol. 9, no. 3, pages 495–511, 2014. (Cited on pages xv, 11, 17, 24, 30, 34, 46, and 180)

[Laptev 2008] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld. *Learning realistic human actions from movies.* In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. (Cited on pages 44 and 49)

[Lea 2016a] C. Lea, J. H. Choi, A. Reiter and G. Hager. *Surgical phase recognition: from instrumented ORs to hospitals around the world.* In Medical image computing and computer-assisted intervention M2CAI—MICCAI workshop, pages 45–54, 2016. (Cited on page 47)

[Lea 2016b] C. Lea, R. Vidal, A. Reiter and G. D. Hager. *Temporal convolutional networks: A unified approach to action segmentation.* In European Conference on Computer Vision, pages 47–54. Springer, 2016. (Cited on page 113)

[Leal-Taixé 2014] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn and S. Savarese. *Learning an image-based motion context for multiple people tracking.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3542–3549, 2014. (Cited on page 40)

[Leal-Taixé 2015] L. Leal-Taixé, A. Milan, I. Reid, S. Roth and K. Schindler. *Motchallenge 2015: Towards a benchmark for multi-target tracking.* arXiv preprint arXiv:1504.01942, 2015. (Cited on pages 40 and 51)

[Lecuyer 2020] G. Lecuyer, M. Ragot, N. Martin, L. Launay and P. Jannin. *Assisted phase and step annotation for surgical videos.* International journal of computer assisted radiology and surgery, pages 1–8, 2020. (Cited on pages 47 and 183)

[Lee 2016] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung and P. K. Rhee. *Multi-class multi-object tracking using changing point detection.* In European Conference on Computer Vision, pages 68–83. Springer, 2016. (Cited on pages 40 and 51)

[Lee 2019] E.-J. Lee, W. Plishker, X. Liu, S. S. Bhattacharyya and R. Shekhar. *Weakly supervised segmentation for real-time surgical tool tracking.* Healthcare technology letters, vol. 6, no. 6, pages 231–236, 2019. (Cited on page 38)

[Lemke 2005] H. U. Lemke, O. M. Ratib and S. C. Horii. *Workflow in the operating room: A summary review of the Arrowhead 2004 Seminar on Imaging and Informatics.* In International Congress Series, volume 1281, pages 862–867. Elsevier, 2005. (Cited on pages 5, 10, 12, and 180)

[Li 2004] X. Li and N. Zheng. *Adaptive target color model updating for visual tracking using particle filter.* In 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), volume 4, pages 3105–3109. IEEE, 2004. (Cited on page 39)

[Li 2014] Y. Li, C. Chen, X. Huang and J. Huang. *Instrument tracking via online learning in retinal microsurgery.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 464–471. Springer, 2014. (Cited on page 43)

[Li 2019] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, M. Chen, Z. Ma, S. Wang, H.-S. Fang and C. Lu. *Hake: Human activity knowledge engine.* arXiv preprint arXiv:1904.06539, 2019. (Cited on page 44)

[Li 2020a] J. Li, X. Gao and T. Jiang. *Graph networks for multiple object tracking.* In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 719–728, 2020. (Cited on page 39)

[Li 2020b] Y.-L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li and C. Lu. *Detailed 2d-3d joint representation for human-object interaction.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10166–10175, 2020. (Cited on pages 45 and 49)

[Li 2020c] Z. Li, S. Cai, X. Wang, Z. Liu and N. Xue. *GAKP: GRU Association and Kalman Prediction for Multiple Object Tracking.* arXiv preprint arXiv:2012.14314, 2020. (Cited on page 40)

[Liew 2018] C. Liew. *The future of radiology augmented with artificial intelligence: a strategy for success.* European journal of radiology, vol. 102, pages 152–156, 2018. (Cited on pages 10 and 12)

[Lin 2008] W. Lin, M.-T. Sun, R. Poovandran and Z. Zhang. *Human activity recognition for video surveillance.* In 2008 IEEE international symposium on circuits and systems, pages 2737–2740. IEEE, 2008. (Cited on page 44)

[Lin 2014] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick. *Microsoft coco: Common objects in context.* In European conference on computer vision, pages 740–755. Springer, 2014. (Cited on pages 45 and 49)

[LIN 2019] X.-g. LIN, Y.-w. CHEN, B.-l. QI, W. Peng and K.-h. ZHONG. *Presence Detection of Surgical Tool Via Densely Connected Convolutional Networks.* DEStech Transactions on Computer Science and Engineering, no. icaic, 2019. (Cited on page 32)

[Litynski 1999] G. S. Litynski. *Endoscopic surgery: the history, the pioneers.* World journal of surgery, vol. 23, no. 8, pages 745–753, 1999. (Cited on page 8)

[Liu 2002] B. Liu, D. Maier, M. A. Schill and R. Männer. *Robust Real-time Tracking of Surgical Instruments in the Eye Surgery Simulator (EyeSi).* 2002. (Cited on page 43)

[Liu 2016a] S. Liu, T. Zhang, X. Cao and C. Xu. *Structural correlation filter for robust visual tracking.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4312–4320, 2016. (Cited on page 40)

[Liu 2016b] X. Liu, S. Kang, W. Plishker, G. Zaki, T. D. Kane and R. Shekhar. *Laparoscopic stereoscopic augmented reality: toward a clinically viable electromagnetic tracking solution.* Journal of Medical Imaging, vol. 3, no. 4, page 045001, 2016. (Cited on page 42)

[Liu 2018] M. Liu and M. Zhu. *Mobile video object detection with temporally-aware feature maps.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5686–5695, 2018. (Cited on pages 40 and 51)

[Liu 2019] M. Liu, M. Zhu, M. White, Y. Li and D. Kalenichenko. *Looking fast and slow: Memory-guided mobile video object detection.* arXiv preprint arXiv:1903.10172, 2019. (Cited on pages 40 and 51)

[Liu 2020] Y. Liu, R. Li, R. T. Tan, Y. Cheng and X. Sui. *Object Tracking Using Spatio-Temporal Future Prediction.* arXiv preprint arXiv:2010.07605, 2020. (Cited on pages 40 and 51)

[Lo 2003] B. P. Lo, A. Darzi and G.-Z. Yang. *Episode classification for the analysis of tissue/instrument interaction with multiple visual cues.* In Int. conference on medical image computing and computer-assisted intervention, pages 230–237, 2003. (Cited on pages 12, 20, 47, 48, and 183)

[Loukas 2015] C. Loukas and E. Georgiou. *Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events.* The International Journal of Medical Robotics and Computer Assisted Surgery, vol. 11, no. 1, pages 80–94, 2015. (Cited on pages 26, 47, 51, 182, and 183)

[Luo 2016] H. Luo, Q. Hu and F. Jia. *Surgical tool detection via multiple convolutional neural networks*, 2016. (Cited on page 32)

[Ma 2016] X. Ma and E. Hovy. *End-to-end sequence labeling via bi-directional lstm-cnns-crf.* arXiv preprint arXiv:1603.01354, 2016. (Cited on page 22)

[Ma 2018] C. Ma, C. Yang, F. Yang, Y. Zhuang, Z. Zhang, H. Jia and X. Xie. *Trajectory factory: Tracklet cleaving and re-connection by deep siamese bi-gru for multiple object tracking.* In 2018 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2018. (Cited on page 40)

[MacKenzie 2001] L. MacKenzie, J. Ibbotson, C. Cao and A. Lomax. *Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment.* Minimally Invasive Therapy & Allied Technologies, vol. 10, no. 3, pages 121–127, 2001. (Cited on page 11)

[Madani 2021] A. Madani, B. Namazi, M. S. Altieri, D. A. Hashimoto, A. M. Rivera, P. H. Pucher, A. Navarrete-Welton, G. Sankaranarayanan, L. M. Brunt, A. Okrainec*et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy.* Annals of Surgery, 2021. (Cited on page 149)

[Mahmoudi 2019] N. Mahmoudi, S. M. Ahadi and M. Rahmati. *Multi-target tracking using CNN-based features: CNNMTT.* Multimedia Tools and Applications, vol. 78, no. 6, pages 7077–7096, 2019. (Cited on page 41)

[Maier-Hein 2017] L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou*et al. Surgical data science: enabling next-generation surgery.* arXiv preprint arXiv:1701.06482, 2017. (Cited on pages xv, 5, 9, 11, 19, 29, 47, and 180)

[Maier-Hein 2020] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni*et al.* *Surgical Data Science–from Concepts to Clinical Translation.* arXiv preprint arXiv:2011.02284, 2020. (Cited on pages 11 and 21)

[Mallya 2016] A. Mallya and S. Lazebnik. *Learning models for actions and person-object interactions with transfer to question answering.* In European Conference on Computer Vision, pages 414–428. Springer, 2016. (Cited on page 44)

[Malpani 2016] A. Malpani, C. Lea, C. C. G. Chen and G. D. Hager. *System events: readily accessible features for surgical phase detection.* International journal of computer assisted radiology and surgery, vol. 11, no. 6, pages 1201–1209, 2016. (Cited on pages 26, 47, and 183)

[Mascagni 2018] P. Mascagni, F. Longo, M. Barberio, B. Seeliger, V. Agnus, P. Saccomandi, A. Hostettler, J. Marescaux and M. Diana. *New intraoperative imaging technologies: Innovating the surgeon's eye toward surgical precision.* Journal of surgical oncology, vol. 118, no. 2, pages 265–282, 2018. (Cited on page 7)

[Mascagni 2020] P. Mascagni, C. Fiorillo, T. Urade, T. Emre, T. Yu, T. Wakabayashi, E. Felli, S. Perretta, L. Swanstrom, D. Mutter*et al.* *Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety.* Surgical endoscopy, vol. 34, no. 6, pages 2709–2714, 2020. (Cited on pages 9 and 16)

[Mascagni 2021a] P. Mascagni, D. Alapatt, T. Urade, A. Vardazaryan, D. Mutter, J. Marescaux, G. Costamagna, B. Dallemagne and N. Padoy. *A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy.* Annals of Surgery, vol. 274, no. 1, pages e93–e95, 2021. (Cited on page 13)

[Mascagni 2021b] P. Mascagni, A. Vardazaryan, D. Alapatt, T. Urade, T. Emre, C. Fiorillo, P. Pessaux, D. Mutter, J. Marescaux, G. Costamagna*et al.* *Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning.* Annals of Surgery, 2021. (Cited on pages 9, 12, 16, and 180)

[Massarweh 2007] N. N. Massarweh and D. R. Flum. *Role of intraoperative cholangiography in avoiding bile duct injury.* Journal of the American College of Surgeons, vol. 204, no. 4, pages 656–664, 2007. (Cited on pages xv, 9, and 179)

[Meireles 2021] O. R. Meireles, G. Rosman, M. S. Altieri, L. Carin, G. Hager, A. Madani, N. Padoy, C. M. Pugh, P. Sylla, T. M. Ward*et al.* *SAGES consensus recommendations on an annotation framework for surgical video.* Surgical endoscopy, pages 1–12, 2021. (Cited on page 20)

[Milan 2016] A. Milan, L. Leal-Taixé, I. Reid, S. Roth and K. Schindler. *MOT16: A benchmark for multi-object tracking.* arXiv preprint arXiv:1603.00831, 2016. (Cited on pages 40 and 51)

[Milan 2017] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid and K. Schindler. *Online multi-target tracking using recurrent neural networks.* In Thirty-First AAAI Conference on Artificial Intelligence, 2017. (Cited on pages 40 and 51)

[Mills-Tettey 2007] G. A. Mills-Tettey, A. Stentz and M. B. Dias. *The dynamic hungarian algorithm for the assignment problem with changing costs.* Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27, 2007. (Cited on page 39)

[Mirota 2011] D. J. Mirota, M. Ishii and G. D. Hager. *Vision-based navigation in image-guided interventions.* Annual review of biomedical engineering, vol. 13, 2011. (Cited on page 10)

[Mishra 2017] K. Mishra, R. Sathish and D. Sheet. *Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 58–65, 2017. (Cited on pages 33 and 182)

[Misra 2017] S. K. Misra. *Message of ASI President.* Indian J Surg, vol. 79, pages 1–3, 2017. (Cited on page 10)

[Mohla 2020] S. Mohla, S. Pande, B. Banerjee and S. Chaudhuri. *Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 92–93, 2020. (Cited on pages 46, 106, and 127)

[Mondal 2019] S. S. Mondal, R. Sathish and D. Sheet. *Multitask learning of temporal connectionism in convolutional networks using a joint distribution loss function to simultaneously identify tools and phase in surgical videos.* arXiv preprint arXiv:1905.08315, 2019. (Cited on pages 32, 33, 88, and 182)

[Morita 2020] S. Morita, H. Tabuchi, H. Masumoto, H. Tanabe and N. Kamiura. *Real-Time Surgical Problem Detection and Instrument Tracking in Cataract Surgery.* Journal of Clinical Medicine, vol. 9, no. 12, page 3896, 2020. (Cited on page 43)

[Munkres 1957] J. Munkres. *Algorithms for the assignment and transportation problems.* Journal of the society for industrial and applied mathematics, vol. 5, no. 1, pages 32–38, 1957. (Cited on page 39)

[Münzer 2013] B. Münzer, K. Schoeffmann and L. Böszörmenyi. *Relevance segmentation of laparoscopic videos.* In 2013 IEEE international symposium on multimedia, pages 84–91. IEEE, 2013. (Cited on page 183)

[Najafzadeh 2015] N. Najafzadeh, M. Fotouhi and S. Kasaei. *Object tracking using Kalman filter with adaptive sampled histogram.* In 2015 23rd Iranian Conference on Electrical Engineering, pages 781–786. IEEE, 2015. (Cited on page 39)

[Namazi 2019] B. Namazi, G. Sankaranarayanan and V. Devarajan. *LapTool-Net: a contextual detector of surgical tools in laparoscopic videos based on recurrent convolutional neural networks.* arXiv preprint arXiv:1905.08983, 2019. (Cited on pages 33 and 182)

[Nathan 2012] M. Nathan, J. M. Karamichalis, H. Liu, S. Emani, C. Baird, F. Pigula, S. Colan, R. R. Thiagarajan, E. A. Bacha and P. Del Nido. *Surgical technical performance scores are predictors of late mortality and unplanned reinterventions in infants after cardiac surgery.* The Journal of thoracic and cardiovascular surgery, vol. 144, no. 5, pages 1095–1101, 2012. (Cited on page 5)

[Navab 1999] N. Navab. *Method and apparatus for calibrating an intra-operative X-ray system,* Juillet 13 1999. US Patent 5,923,727. (Cited on pages 10 and 180)

[Navab 2002] N. Navab and Y. Genc. *Method and system for computer assisted localization and navigation in industrial environments,* Janvier 24 2002. US Patent App. 09/741,581. (Cited on page 10)

[Navab 2007] N. Navab, X. Zhang, Y. Genc and V. Kumar. *Augmented reality system,* Septembre 25 2007. US Patent 7,274,380. (Cited on pages 10 and 180)

[Navab 2012] N. Navab, T. Blum, L. Wang, A. Okur and T. Wendler. *First deployments of augmented reality in operating rooms.* Computer, vol. 45, no. 7, pages 48–55, 2012. (Cited on pages 10 and 180)

[Neimark 2021] D. Neimark, O. Bar, M. Zohar, G. D. Hager and D. Asselmann. *" Train one, Classify one, Teach one"–Cross-surgery transfer learning for surgical step recognition.* arXiv preprint arXiv:2102.12308, 2021. (Cited on page 22)

[Neumuth 2006] T. Neumuth, G. Strauß, J. Meixensberger, H. U. Lemke and O. Burgert. *Acquisition of process descriptions from surgical interventions.* In International Conference on Database and Expert Systems Applications, pages 602–611, 2006. (Cited on pages 6, 48, 81, 182, and 184)

[Neumuth 2009] T. Neumuth, P. Jannin, G. Strauss, J. Meixensberger and O. Burgert. *Validation of knowledge acquisition for surgical process models.* Journal of the American Medical Informatics Association, vol. 16, no. 1, pages 72–80, 2009. (Cited on page 11)

[Neumuth 2010] T. Neumuth, B. Kaschek, D. Neumuth, M. Ceschia, J. Meixensberger, G. Strauss and O. Burgert. *An observation support system with an adaptive ontology-driven user interface for the modeling of complex behaviors during surgical interventions.* Behavior research methods, vol. 42, no. 4, pages 1049–1058, 2010. (Cited on pages 81 and 184)

[Nordsjo 2004] A. E. Nordsjo. *A constrained extended Kalman filter for target tracking.* In Proceedings of the 2004 IEEE Radar Conference (IEEE Cat. No. 04CH37509), pages 123–127. IEEE, 2004. (Cited on page 39)

[Nwoye 2019] C. I. Nwoye, D. Mutter, J. Marescaux and N. Padoy. *Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos.* International journal of computer assisted radiology and surgery, vol. 14, no. 6, pages 1059–1067, 2019. (Cited on pages 15, 19, 25, 27, 58, 92, 94, 174, 175, 181, and 183)

[Nwoye 2020] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux and N. Padoy. *Recognition of instrument-tissue interactions in endoscopic videos via action triplets.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 364–374. Springer, 2020. (Cited on pages xvii, 15, 26, 27, 80, 81, 83, 102, 103, 106, 107, 110, 112, 113, 132, 168, 181, 184, and 188)

[Nwoye 2021] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux and N. Padoy. *Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos.* arXiv preprint arXiv:2109.03223, 2021. (Cited on pages 20, 26, 27, 103, 122, 125, 128, 132, 170, and 181)

[Oktay 2018] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz*et al. Attention u-net: Learning where to look for the pancreas.* arXiv preprint arXiv:1804.03999, 2018. (Cited on page 106)

[Olsen 1991] D. O. Olsen. *Laparoscopic cholecystectomy.* The American journal of surgery, vol. 161, no. 3, pages 339–344, 1991. (Cited on pages 8 and 179)

[Otsu 1979] N. Otsu. *A threshold selection method from gray-level histograms.* IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pages 62–66, 1979. (Cited on pages 64 and 70)

[Padoy 2007] N. Padoy, T. Blum, I. Essa, H. Feussner, M.-O. Berger and N. Navab. *A boosted segmentation method for surgical workflow analysis.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 102–109. Springer, 2007. (Cited on page 47)

[Padoy 2008] N. Padoy, T. Blum, H. Feussner, M.-O. Berger and N. Navab. *On-line Recognition of Surgical Activity for Monitoring in the Operating Room.* In AAAI, pages 1718–1724, 2008. (Cited on pages 10 and 17)

[Padoy 2012] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger and N. Navab. *Statistical modeling and recognition of surgical workflow.* Medical image analysis, vol. 16, no. 3, pages 632–641, 2012. (Cited on pages 8 and 15)

[Padoy 2019] N. Padoy. *Machine and deep learning for workflow recognition during surgery.* Minimally Invasive Therapy & Allied Technologies, vol. 28, no. 2, pages 82–90, 2019. (Cited on pages xvi and 47)

[Pakhomov 2019] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian and N. Navab. *Deep residual learning for instrument segmentation in robotic surgery.* In International Workshop on Machine Learning in Medical Imaging, pages 566–573. Springer, 2019. (Cited on page 38)

[Park 2021] J. Park and C. H. Park. *Recognition and Prediction of Surgical Actions Based on Online Robotic Tool Detection.* IEEE Robotics and Automation Letters, vol. 6, no. 2, pages 2365–2372, 2021. (Cited on pages 47, 52, and 183)

[Pérez 2002] P. Pérez, C. Hue, J. Vermaak and M. Gangnet. *Color-based probabilistic tracking.* In European Conference on Computer Vision, pages 661–675. Springer, 2002. (Cited on page 39)

[Petscharnig 2018a] S. Petscharnig and K. Schöffmann. *Learning laparoscopic video shot classification for gynecological surgery.* Multimedia Tools and Applications, vol. 77, no. 7, pages 8061–8079, 2018. (Cited on page 183)

[Petscharnig 2018b] S. Petscharnig, K. Schöffmann, J. Benois-Pineau, S. Chaabouni and J. Keckstein. *Early and late fusion of temporal information for classification of surgical actions in laparoscopic gynecology.* In 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pages 369–374. IEEE, 2018. (Cited on page 20)

[Pfeiffer 2019a] M. Pfeiffer, I. Funke, M. R. Robu, S. Bodenstedt, L. Strenger, S. Engelhardt, T. Roß, M. J. Clarkson, K. Gurusamy, B. R. Davidson*et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 119–127. Springer, 2019. (Cited on page 22)

[Pfeiffer 2019b] M. Pfeiffer, C. Riediger, J. Weitz and S. Speidel. *Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks.* International journal of computer assisted radiology and surgery, vol. 14, no. 7, pages 1147–1155, 2019. (Cited on pages 10 and 180)

[Poormehdi Ghaemmaghami 2017] M. Poormehdi Ghaemmaghami. *Tracking of Humans in Video Stream Using LSTM Recurrent Neural Network*, 2017. (Cited on page 40)

[Pucher 2018] P. H. Pucher, L. M. Brunt, N. Davies, A. Linsk, A. Munshi, H. A. Rodriguez, A. Fingerhut, R. D. Fanelli, H. Asbun and R. Aggarwal. *Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis.* Surgical endoscopy, vol. 32, no. 5, pages 2175–2183, 2018. (Cited on pages 8 and 179)

[Qi 2018] S. Qi, W. Wang, B. Jia, J. Shen and S.-C. Zhu. *Learning human-object interactions by graph parsing neural networks.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 401–417, 2018. (Cited on pages 45, 52, 53, 90, 91, and 98)

[Quellec 2014a] G. Quellec, K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener and G. Cazuguel. *Real-time recognition of surgical tasks in eye surgery videos.* Medical image analysis, vol. 18, no. 3, pages 579–590, 2014. (Cited on page 42)

[Quellec 2014b] G. Quellec, M. Lamard, B. Cochener and G. Cazuguel. *Real-time segmentation and recognition of surgical tasks in cataract surgery videos.* IEEE transactions on medical imaging, vol. 33, no. 12, pages 2352–2360, 2014. (Cited on page 42)

[Raju 2016] A. Raju, S. Wang and J. Huang. *M2CAI surgical tool detection challenge report.* In Workshop and Challenges on Modeling and Monitoring of Computer Assisted Intervention (M2CAI), Athens, Greece, Technical report, 2016. (Cited on pages 32 and 182)

[Ramesh 2021] S. Ramesh, D. Dall'Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorini and N. Padoy. *Multi-Task Temporal Convolutional Networks for Joint Recognition of Surgical Phases and Steps in Gastric Bypass Procedures.* arXiv preprint arXiv:2102.12218, 2021. (Cited on pages 13, 47, 51, 180, 182, and 183)

[Rautaray 2010] S. S. Rautaray and A. Agrawal. *A novel human computer interface based on hand gesture recognition using computer vision techniques.* In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, pages 292–296, 2010. (Cited on page 44)

[Reiley 2011] C. E. Reiley, H. C. Lin, D. D. Yuh and G. D. Hager. *Review of methods for objective surgical skill evaluation.* Surgical endoscopy, vol. 25, no. 2, pages 356–366, 2011. (Cited on pages 11 and 180)

[Reiter 2010] A. Reiter and P. K. Allen. *An online learning approach to in-vivo tracking using synergistic features.* In 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3441–3446. IEEE, 2010. (Cited on pages 18, 42, and 183)

[Reiter 2012a] A. Reiter, P. K. Allen and T. Zhao. *Articulated surgical tool detection using virtually-rendered templates.* In Computer Assisted Radiology and Surgery (CARS), pages 1–8, 2012. (Cited on pages 30, 34, and 42)

[Reiter 2012b] A. Reiter, P. K. Allen and T. Zhao. *Feature classification for tracking articulated surgical tools.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 592–600. Springer, 2012. (Cited on pages 18, 42, and 183)

[Richa 2011] R. Richa, M. Balicki, E. Meisner, R. Sznitman, R. Taylor and G. Hager. *Visual tracking of surgical tools for proximity detection in retinal surgery.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 55–66. Springer, 2011. (Cited on page 43)

[Rieke 2015] N. Rieke, D. J. Tan, M. Alsheakhali, F. Tombari, C. A. di San Filippo, V. Belagiannis, A. Eslami and N. Navab. *Surgical tool tracking and pose estimation in retinal micro-*

surgery. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 266–273. Springer, 2015. (Cited on page 42)

[Rieke 2016] N. Rieke, D. J. Tan, C. A. di San Filippo, F. Tombari, M. Alsheakhali, V. Belagiannis, A. Eslami and N. Navab. *Real-time localization of articulated surgical instruments in retinal microsurgery.* Medical Image Analysis, vol. 34, pages 82–100, 2016. (Cited on page 25)

[Robu 2020] M. Robu, A. Kadkhodamohammadi, I. Luengo and D. Stoyanov. *Towards real-time multiple surgical tool tracking.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pages 1–7, 2020. (Cited on pages 42, 43, 51, and 183)

[Rodas 2015] N. L. Rodas and N. Padoy. *Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, Monte Carlo simulations and wireless dosimeters.* International journal of computer assisted radiology and surgery, vol. 10, no. 8, pages 1181–1191, 2015. (Cited on pages 10 and 180)

[Ross 2020] T. Ross, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. M. Filimon, P. Scholz, T. N. Tran*et al. Robust medical instrument segmentation challenge 2019.* arXiv preprint arXiv:2003.10299, 2020. (Cited on page 38)

[Roychowdhury 2017] S. Roychowdhury, Z. Bian, A. Vahdat and W. G. Macready. *Identification of Surgical Tools using Deep Neural Networks.* In Technical Report. D-Wave Systems Inc, 2017. (Cited on page 32)

[Rupprecht 2016] C. Rupprecht, C. Lea, F. Tombari, N. Navab and G. D. Hager. *Sensor substitution for video-based action recognition.* In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5230–5237. IEEE, 2016. (Cited on pages 13, 47, and 184)

[Ryu 2013] J. Ryu, J. Choi and H. C. Kim. *Endoscopic vision-based tracking of multiple surgical instruments during robot-assisted surgery.* Artificial organs, vol. 37, no. 1, pages 107–112, 2013. (Cited on pages 34, 42, and 50)

[Sahu 2016] M. Sahu, A. Mukhopadhyay, A. Szengel and S. Zachow. *Tool and phase recognition using contextual CNN features.* arXiv preprint arXiv:1610.08854, 2016. (Cited on pages 32 and 182)

[Sahu 2017] M. Sahu, A. Mukhopadhyay, A. Szengel and S. Zachow. *Addressing multi-label imbalance problem of surgical tool detection using CNN.* International journal of computer assisted radiology and surgery, vol. 12, no. 6, pages 1013–1020, 2017. (Cited on page 32)

[Sahu 2020] M. Sahu, A. Szengel, A. Mukhopadhyay and S. Zachow. *Surgical phase recognition by learning phase transitions.* Current Directions in Biomedical Engineering, vol. 6, no. 1, 2020. (Cited on page 47)

[Sankaran 2016] B. Sankaran, H. Mi, Y. Al-Onaizan and A. Ittycheriah. *Temporal attention model for neural machine translation.* arXiv preprint arXiv:1608.02927, 2016. (Cited on pages 46 and 106)

[Sarikaya 2017] D. Sarikaya, J. J. Corso and K. A. Guru. *Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection.* IEEE transactions on medical imaging, vol. 36, no. 7, pages 1542–1549, 2017. (Cited on pages 34 and 35)

[Sarikaya 2020] D. Sarikaya and P. Jannin. *Towards generalizable surgical activity recognition using spatial temporal graph convolutional networks.* arXiv preprint arXiv:2001.03728, 2020. (Cited on pages 47 and 183)

[Sastry 2017] A. V. Sastry, J. H. Swet, K. J. Murphy, E. H. Baker, D. Vrochides, J. B. Martinie, I. H. McKillop and D. A. Iannitti. *A novel 3-dimensional electromagnetic guidance system increases intraoperative microwave antenna placement accuracy.* HPB, vol. 19, no. 12, pages 1066–1073, 2017. (Cited on page 42)

[Schirmer 1991] B. D. Schirmer, S. B. Edge, J. Dix, M. J. Hyser, J. B. Hanks and R. S. Jones. *Laparoscopic cholecystectomy. Treatment of choice for symptomatic cholelithiasis.* Annals of surgery, vol. 213, no. 6, page 665, 1991. (Cited on page 8)

[Sestini 2021] L. Sestini, B. Rosa, E. De Momi, G. Ferrigno and N. Padoy. *A Kinematic Bottleneck Approach For Pose Regression of Flexible Surgical Instruments directly from Images.* IEEE Robotics and Automation Letters, vol. 6, no. 2, pages 2938–2945, 2021. (Cited on pages 35 and 38)

[Shaffer 2006] E. A. Shaffer. *Epidemiology of gallbladder stone disease.* Best practice & research Clinical gastroenterology, vol. 20, no. 6, pages 981–996, 2006. (Cited on page 8)

[Shen 2018] L. Shen, S. Yeung, J. Hoffman, G. Mori and L. Fei-Fei. *Scaling human-object interaction recognition through zero-shot learning.* In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1568–1576. IEEE, 2018. (Cited on pages xvi, xvii, 45, 52, 90, 91, 92, 93, and 187)

[Shi 2015] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-c. Woo. *Convolutional LSTM network: A machine learning approach for precipitation nowcasting.* arXiv preprint arXiv:1506.04214, 2015. (Cited on page 65)

[Shi 2021] X. Shi, Y. Jin, Q. Dou and P.-A. Heng. *Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition.* Medical Image Analysis, vol. 73, page 102158, 2021. (Cited on page 22)

[Shvets 2018] A. A. Shvets, A. Rakhlin, A. A. Kalinin and V. I. Iglovikov. *Automatic instrument segmentation in robot-assisted surgery using deep learning.* In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 624–628. IEEE, 2018. (Cited on pages xvi and 38)

[Simonyan 2014] K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014. (Cited on page 32)

[Simpson 2019] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze*et al.* *A large annotated medical image dataset for the development and evaluation of segmentation algorithms.* arXiv preprint arXiv:1902.09063, 2019. (Cited on page 5)

[Singh 2017] K. K. Singh and Y. J. Lee. *Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization.* In ICCV, 2017. (Cited on pages 64, 94, and 185)

[Singh 2018] K. K. Singh, H. Yu, A. Sarmasi, G. Pradeep and Y. J. Lee. *Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond.* arXiv preprint arXiv:1811.02545, 2018. (Cited on pages xvi and 64)

[Soleimanitaleb 2019] Z. Soleimanitaleb, M. A. Keyvanrad and A. Jafari. *Object Tracking Methods: A Review.* In 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), pages 282–288. IEEE, 2019. (Cited on page 39)

[Soomro 2012] K. Soomro, A. R. Zamir and M. Shah. *UCF101: A dataset of 101 human actions classes from videos in the wild.* arXiv preprint arXiv:1212.0402, 2012. (Cited on pages 44 and 46)

[Speidel 2006] S. Speidel, M. Delles, C. Gutt and R. Dillmann. *Tracking of instruments in minimally invasive surgery for surgical skill analysis.* In International Workshop on Medical Imaging and Virtual Reality, pages 148–155. Springer, 2006. (Cited on pages 10, 15, and 42)

[Speidel 2008] S. Speidel, G. Sudra, J. Senemaud, M. Drentschew, B. P. Müller-Stich, C. Gutt and R. Dillmann. *Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling.* In Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling, volume 6918, page 69180X. International Society for Optics and Photonics, 2008. (Cited on pages 11, 12, 34, 42, 50, and 183)

[Speidel 2009] S. Speidel, J. Benzko, S. Krappe, G. Sudra, P. Azad, B. P. Müller-Stich, C. Gutt and R. Dillmann. *Automatic classification of minimally invasive instruments based on endoscopic image sequences.* In Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling, volume 7261, page 72610A. International Society for Optics and Photonics, 2009. (Cited on pages 10, 38, 48, 81, 180, and 184)

[Speidel 2014] S. Speidel, E. Kuhn, S. Bodenstedt, S. Röhl, H. Kenngott, B. Müller-Stich and R. Dillmann. *Visual tracking of da vinci instruments for laparoscopic surgery.* In Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling, volume 9036, page 903608. International Society for Optics and Photonics, 2014. (Cited on pages 10, 18, 42, and 180)

[Stauder 2014] R. Stauder, A. Okur, L. Peter, A. Schneider, M. Kranzfelder, H. Feussner and N. Navab. *Random forests for phase detection in surgical workflow analysis.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 148–157. Springer, 2014. (Cited on page 15)

[Stefan 2020] P. Stefan, J. Traub, C. Hennersperger, M. Esposito and N. Navab. *Challenges in Computer Assisted Interventions: Challenges in design, development, evaluation, and clinical deployment of Computer Assisted Intervention solutions.* In Handbook of Medical Image Computing and Computer Assisted Intervention, pages 979–1012. Elsevier, 2020. (Cited on page 10)

[Stoyanov 2012] D. Stoyanov. *Surgical vision.* Annals of biomedical engineering, vol. 40, no. 2, pages 332–345, 2012. (Cited on pages 10 and 180)

[Strasberg 1995] S. M. Strasberg, M. Hertl and N. J. Soper. *An analysis of the problem of biliary injury during laparoscopic cholecystectomy.* Journal of the American College of Surgeons, vol. 180, no. 1, pages 101–125, 1995. (Cited on page 16)

[Sun 2017] C. Sun, A. Shrivastava, S. Singh and A. Gupta. *Revisiting unreasonable effectiveness of data in deep learning era.* In Proceedings of the IEEE international conference on computer vision, pages 843–852, 2017. (Cited on page 21)

[Sun 2019] S. Sun, N. Akhtar, H. Song, A. Mian and M. Shah. *Deep affinity network for multiple object tracking.* IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 1, pages 104–119, 2019. (Cited on page 41)

[Szegedy 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. *Going deeper with convolutions.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015. (Cited on page 32)

[Szegedy 2016] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. *Rethinking the inception architecture for computer vision.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016. (Cited on page 32)

[Sznitman 2011] R. Sznitman, A. Basu, R. Richa, J. Handa, P. Gehlbach, R. H. Taylor, B. Jedynak and G. D. Hager. *Unified detection and tracking in retinal microsurgery.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 1–8. Springer, 2011. (Cited on pages 10, 43, and 180)

[Sznitman 2012a] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager and P. Fua. *Data-driven visual tracking in retinal microsurgery.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 568–575. Springer, 2012. (Cited on pages 18, 25, 42, 43, and 50)

[Sznitman 2012b] R. Sznitman, R. Richa, R. H. Taylor, B. Jedynak and G. D. Hager. *Unified detection and tracking of instruments during retinal microsurgery.* IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 5, pages 1263–1273, 2012. (Cited on pages 18, 42, and 183)

[Sznitman 2014] R. Sznitman, C. Becker and P. Fua. *Fast part-based classification for instrument detection in minimally invasive surgery.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 692–699. Springer, 2014. (Cited on pages 25 and 34)

[Tang 2016] S. Tang, B. Andres, M. Andriluka and B. Schiele. *Multi-person tracking by multicut and deep matching.* In European Conference on Computer Vision, pages 100–111. Springer, 2016. (Cited on pages 40 and 51)

[Tang 2017] S. Tang, M. Andriluka, B. Andres and B. Schiele. *Multiple people tracking by lifted multicut and person re-identification.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3539–3548, 2017. (Cited on page 40)

[Tao 2016] R. Tao, E. Gavves and A. W. Smeulders. *Siamese instance search for tracking.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1420–1429, 2016. (Cited on page 41)

[Twinanda 2014] A. P. Twinanda, J. Marescaux, M. De Mathelin and N. Padoy. *Towards better laparoscopic video database organization by automatic surgery classification.* In International Conference on Information Processing in Computer-Assisted Interventions, pages 186–195. Springer, 2014. (Cited on pages 46 and 183)

[Twinanda 2015] A. P. Twinanda, E. O. Alkan, A. Gangi, M. de Mathelin and N. Padoy. *Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms.* Int. journal of computer assisted radiology and surgery, vol. 10, no. 6, pages 737–747, 2015. (Cited on pages 26 and 47)

[Twinanda 2016a] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin and N. Padoy. *Single-and multi-task architectures for tool presence detection challenge at M2CAI 2016.* arXiv preprint arXiv:1610.08851, 2016. (Cited on pages 19, 32, 88, and 182)

[Twinanda 2016b] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin and N. Padoy. *Endonet: a deep architecture for recognition tasks on laparoscopic videos.* IEEE transactions on medical imaging, vol. 36, no. 1, pages 86–97, 2016. (Cited on pages xv, 12, 19, 22, 26, 31, 32, 47, 59, 71, 83, 103, 167, 174, 182, and 183)

[Twinanda 2017] A. P. Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos.* PhD thesis, Strasbourg, 2017. (Cited on pages 10, 20, and 49)

[Twinanda 2018] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations.* IEEE transactions on medical imaging, vol. 38, no. 4, pages 1069–1078, 2018. (Cited on page 22)

[Ulutan 2020] O. Ulutan, A. Iftekhar and B. S. Manjunath. *Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13617–13626, 2020. (Cited on pages 46, 53, and 106)

[van Amsterdam 2019] B. van Amsterdam, H. Nakawala, E. De Momi and D. Stoyanov. *Weakly supervised recognition of surgical gestures.* In 2019 International Conference on Robotics and Automation (ICRA), pages 9565–9571. IEEE, 2019. (Cited on pages 37, 47, and 61)

[Vander Poorten 2020] E. Vander Poorten, C. N. Riviere, J. J. Abbott, C. Bergeles, M. A. Nasseri, J. U. Kang, R. Sznitman, K. Faridpooya and I. Iordachita. *Robotic retinal surgery.* In Handbook of Robotic and Image-Guided Surgery, pages 627–672. Elsevier, 2020. (Cited on pages 10 and 180)

[Vardazaryan 2018] A. Vardazaryan, D. Mutter, J. Marescaux and N. Padoy. *Weakly-supervised learning for tool localization in laparoscopic videos.* In Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pages 169–179. Springer, 2018. (Cited on pages xvi, 19, 25, 33, 34, 36, 37, 50, 61, 64, 69, 70, 77, 107, and 183)

[Vaswani 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin. *Attention is all you need.* arXiv preprint arXiv:1706.03762, 2017. (Cited on pages 46, 106, 107, 123, 126, and 127)

[Velanovich 2000] V. Velanovich. *Laparoscopic vs open surgery.* Surgical endoscopy, vol. 14, no. 1, pages 16–21, 2000. (Cited on pages 5, 8, and 179)

[Vercauteren 2019] T. Vercauteren, M. Unberath, N. Padoy and N. Navab. *Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions.* Proceedings of the IEEE, vol. 108, no. 1, pages 198–214, 2019. (Cited on pages 5, 17, 18, 38, and 48)

[Voigtlaender 2019] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger and B. Leibe. *Mots: Multi-object tracking and segmentation.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7942–7951, 2019. (Cited on page 41)

[Wagenaar 2010] D. A. Wagenaar and W. B. Kristan. *Automated video analysis of animal movements using Gabor orientation filters.* Neuroinformatics, vol. 8, no. 1, pages 33–42, 2010. (Cited on page 39)

[Wagner 2021a] M. Wagner, A. Bihlmaier, H. G. Kenngott, P. Mietkowski, P. M. Scheikl, S. Bodenstedt, A. Schiepe-Tiska, J. Vetter, F. Nickel, S. Speidel*et al.* *A learning robot for cognitive camera control in minimally invasive surgery.* Surgical Endoscopy, pages 1–10, 2021. (Cited on page 10)

[Wagner 2021b] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek*et al.* *Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark.* arXiv preprint arXiv:2109.14956, 2021. (Cited on pages 20, 47, and 184)

[Wang 2017] S. Wang, A. Raju and J. Huang. *Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos.* In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pages 620–623. IEEE, 2017. (Cited on page 32)

[Wang 2018] X. Wang, R. Girshick, A. Gupta and K. He. *Non-local neural networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7794–7803, 2018. (Cited on page 126)

[Wang 2019a] S. Wang, Z. Xu, C. Yan and J. Huang. *Graph convolutional nets for tool presence detection in surgical videos.* In International Conference on Information Processing in Medical Imaging, pages 467–478. Springer, 2019. (Cited on pages 32, 36, and 51)

[Wang 2019b] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao and J. Laaksonen. *Deep contextual attention for human-object interaction detection.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5694–5702, 2019. (Cited on page 46)

[Wang 2020] Z. Wang, L. Zheng, Y. Liu, Y. Li and S. Wang. *Towards real-time multi-object tracking.* In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pages 107–122. Springer, 2020. (Cited on page 41)

[Weese 1997] J. Weese, G. P. Penney, P. Desmedt, T. M. Buzug, D. L. Hill and D. J. Hawkes. *Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery.* IEEE transactions on information technology in biomedicine, vol. 1, no. 4, pages 284–293, 1997. (Cited on page 30)

[Weng 2020] X. Weng, Y. Wang, Y. Man and K. M. Kitani. *Gnn3dmot: Graph neural network for 3d multi-object tracking with 2d-3d multi-feature learning.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6499–6508, 2020. (Cited on page 39)

[Westebring-van der Putten 2008] E. P. Westebring-van der Putten, R. H. Goossens, J. J. Jakimowicz and J. Dankelman. *Haptics in minimally invasive surgery–a review.* Minimally Invasive Therapy & Allied Technologies, vol. 17, no. 1, pages 3–16, 2008. (Cited on page 8)

[Wojke 2017] N. Wojke, A. Bewley and D. Paulus. *Simple online and realtime tracking with a deep association metric.* In 2017 IEEE international conference on image processing (ICIP), pages 3645–3649. IEEE, 2017. (Cited on pages 39, 41, and 59)

[Wolf 2011] R. Wolf, J. Duchateau, P. Cinquin and S. Voros. *3D tracking of laparoscopic instruments using statistical and geometric modeling.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 203–210. Springer, 2011. (Cited on page 42)

[Xingjian 2015] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-c. Woo. *Convolutional LSTM network: A machine learning approach for precipitation nowcasting.* In Advances in neural information processing systems, pages 802–810, 2015. (Cited on page 66)

[Xu 2019] B. Xu, Y. Wong, J. Li, Q. Zhao and M. S. Kankanhalli. *Learning to detect human-object interactions with knowledge.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. (Cited on pages 52, 90, 91, and 98)

[Xu 2021] M. Xu, M. Islam, C. M. Lim and H. Ren. *Learning Domain Adaptation with Model Calibration for Surgical Report Generation in Robotic Surgery.* arXiv preprint arXiv:2103.17120, 2021. (Cited on pages 20, 49, and 52)

[Yao 2020] Q. Yao and X. Gong. *Saliency guided self-attention network for weakly and semi-supervised semantic segmentation.* IEEE Access, vol. 8, pages 14413–14423, 2020. (Cited on page 106)

[Ye 2016] M. Ye, L. Zhang, S. Giannarou and G.-Z. Yang. *Real-time 3d tracking of articulated tools for robotic surgery.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 386–394. Springer, 2016. (Cited on pages 22, 42, 50, and 183)

[Yengera 2018] G. Yengera, D. Mutter, J. Marescaux and N. Padoy. *Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks.* arXiv preprint arXiv:1805.08569, 2018. (Cited on pages 22 and 68)

[Yoon 2020] J. Yoon, J. Lee, S. Park, W. J. Hyung and M.-K. Choi. *Semi-supervised Learning for Instrument Detection with a Class Imbalanced Dataset.* In Interpretable and Annotation-Efficient Learning for Medical Image Computing, pages 266–276. Springer, 2020. (Cited on pages 37 and 61)

[Yu 2016] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi and J. Yan. *Poi: Multiple object tracking with high performance detection and appearance feature.* In European Conference on Computer Vision, pages 36–42. Springer, 2016. (Cited on page 41)

[Yu 2018] T. Yu, D. Mutter, J. Marescaux and N. Padoy. *Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition.* arXiv preprint arXiv:1812.00033, 2018. (Cited on pages 22, 47, 68, and 183)

[Zamir 2012] A. R. Zamir, A. Dehghan and M. Shah. *Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs.* In European conference on computer vision, pages 343–356. Springer, 2012. (Cited on page 40)

[Zhang 2008] L. Zhang, Y. Li and R. Nevatia. *Global data association for multi-object tracking using network flows.* In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. (Cited on page 40)

[Zhang 2020a] B. Zhang, S. Wang, L. Dong and P. Chen. *Surgical tools detection based on modulated anchoring network in laparoscopic videos.* IEEE Access, vol. 8, pages 23748–23758, 2020. (Cited on pages 34, 36, 50, and 182)

[Zhang 2020b] Y. Zhang, C. Wang, X. Wang, W. Zeng and W. Liu. *Fairmot: On the fairness of detection and re-identification in multiple object tracking.* arXiv preprint arXiv:2004.01888, 2020. (Cited on pages xvi and 41)

[Zhao 2009] Z. Zhao, S. Yu, X. Wu, C. Wang and Y. Xu. *A multi-target tracking algorithm using texture for real-time surveillance.* In 2008 IEEE International Conference on Robotics and Biomimetics, pages 2150–2155. IEEE, 2009. (Cited on page 39)

[Zhao 2017] Z. Zhao, S. Voros, Y. Weng, F. Chang and R. Li. *Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method.* Computer Assisted Surgery, vol. 22, no. sup1, pages 26–35, 2017. (Cited on pages 43, 51, and 183)

[Zhou 2014] J. Zhou and S. Payandeh. *Visual tracking of laparoscopic instruments.* Journal of Automation and Control Engineering Vol, vol. 2, no. 3, pages 234–241, 2014. (Cited on page 42)

[Zhou 2019] P. Zhou and M. Chi. *Relation parsing neural network for human-object interaction detection.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 843–851, 2019. (Cited on page 46)

[Zhu 2018] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang and M.-H. Yang. *Online multi-object tracking with dual matching attention networks.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 366–382, 2018. (Cited on page 41)

[Zhuang 2018] B. Zhuang, Q. Wu, C. Shen, I. Reid and A. van den Hengel. *Hcvrd: a benchmark for large-scale human-centered visual relationship detection.* In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018. (Cited on pages 45 and 49)

[Zia 2016] A. Zia, D. Castro and I. Essa. *Fine-tuning deep architectures for surgical tool detection.* In Workshop and Challenges on Modeling and Monitoring of Computer Assisted Intervention (M2CAI), Athens, Greece, Technical report, 2016. (Cited on pages 32 and 182)

[Zia 2018] A. Zia, A. Hung, I. Essa and A. Jarc. *Surgical activity recognition in robot-assisted radical prostatectomy using deep learning.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 273–280, 2018. (Cited on page 47)

[Zisimopoulos 2017] O. Zisimopoulos, E. Flouty, M. Stacey, S. Muscroft, P. Giataganas, J. Nehme, A. Chow and D. Stoyanov. *Can surgical simulation be used to train detection and classification of neural networks?* Healthcare technology letters, vol. 4, no. 5, pages 216–222, 2017. (Cited on page 38)

[Zisimopoulos 2018] O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow and D. Stoyanov. *Deepphase: surgical phase recognition in cataracts videos.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 265–272, 2018. (Cited on pages 47 and 183)

[Zou 2021] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei*et al.* *End-to-end human object interaction detection with hoi transformer.* arXiv preprint arXiv:2103.04503, 2021. (Cited on pages 46 and 53)

Université de Strasbourg

Chinedu Innocent Nwoye

msii ÉCOLE DOCTORALE

Deep Learning Methods for the Detection and Recognition of Surgical Tools and Activities in Laparoscopic Videos

Summary

In this thesis, we address the two problem of tool detection and fine-grained activity recognition in the operating room (OR), which are key ingredients in the development of surgical assistance applications. Leveraging weak supervision for temporal modeling and spatial localization, we propose a joint detection and tracking model for surgical instruments, circumventing the lack of spatially annotated dataset on this task. For a more helpful AI assistance in the OR, we formalize surgical activities as triplets of ⟨*instrument, verb, target*⟩, and propose several deep learning methods, that leverages instrument's activation, spatial attention, and semantic attention mechanisms, to recognize these triplets directly from surgical videos. Evaluation is performed on large scale datasets, which we introduce in this thesis, obtaining state-of-the-art results for these tasks.

Keywords: Deep learning, tool detection, tool tracking, tool-tissue interaction, action triplet recognition, CholecT50, weak supervision, attention mechanism, transformer.

Résumé

Dans cette thèse, nous abordons les deux problèmes de détection d'outils et de reconnaissance d'activité à grain fin en salle d'opération, qui sont des ingrédients clés dans le développement d'applications d'assistance chirurgicale. En tirant parti d'une supervision faible pour la modélisation temporelle et la localisation spatiale, nous proposons un modèle de détection et de suivi conjoint pour les instruments chirurgicaux, contournant le manque de jeu de données annotées spatialement sur cette tâche. Pour une assistance plus utile de l'IA dans la salle d'opération, nous formalisons les activités chirurgicales sous forme de triplets de ⟨*instrument, verbe, cible*⟩, et proposons plusieurs méthodes d'apprentissage en profondeur, qui tirent parti des mécanismes d'activation, d'attention spatiale et d'attention sémantique de l'instrument, pour reconnaître les triplés directement à partir de vidéos chirurgicales. L'évaluation est effectuée sur des ensembles de données à grande, que nous introduisons dans cette thèse, et donne des résultats de pointe pour ces tâches.

Mots-clés: Apprentissage profond, détection d'outils, suivi d'outils, interaction outil-tissu, reconnaissance de triplet d'action, CholectT50, supervision faible, mécanisme d'attention, transformateur.

Special thanks to our sponsors and clinical collaborators.