Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269 Université de Strasbourg

ÉCOLE DOCTORALE MATHEMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGENIEUR – ED269

ICube Laboratory (UMR 7357)

THÈSE

présentée par:

Sanat RAMESH

soutenue le: 23 May 2023

pour obtenir le grade de: **Docteur de l'université de Strasbourg**

Discipline/ Spécialité: Image and Vision

Multi-level Surgical Activity Recognition

THÈSE dirigée par:

Prof. Paolo FIORINI	Professor, Università di Verona
Prof. Nicolas PADOY	Professor, Université de Strasbourg

RAPPORTEURS:

Dr. Sandrine VOROS	Director of Research, TIMC-IMAG
Dr. Stamatia GIANNAROU	Senior Lecturer, Imperial College London

AUTRES MEMBRES DU JURY:

Prof. Marco CRISTANI	Professor, Università di Verona
Dr. Thomas LAMPERT	Assistant Professor, Université de Strasbourg

UNIVERSITY OF VERONA

DEPARTMENT OF

Computer Science

DOCTORAL SCHOOL

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Computer Science

Cycle XXXV, 2019

Multi-level Surgical Activity Recognition

IN CO-TUTELLE DE THÉSE WITH THE UNIVERSITY OF STRASBOURG

Sanat Ramesh

Supervisor:
Prof. Paolo Fiorini
Prof. Nicolas Padoy (Co-Supervisor, UNISTRA)

PhD program chair:
Prof. Ferdinando Cicalese

Multi-level Surgical Activity Recognition

Sanat Ramesh

Ph.D. Thesis

Verona, Italy, May 23, 2023

Thesis presented by

Sanat Ramesh

Defended publicly on **23rd May 2023**

For obtaining the degree of **Doctor of Philosophy**
From the **University of Verona** and the **University of Strasbourg**

Multi-level Surgical Activity Recognition

Thesis Directors

Prof. Paolo Fiorini
Professor of Computer Science,
University of Verona, Italy

Prof. Nicolas Padoy
Professor of Computer Science,
University of Strasbourg,
IHU Strasbourg, France

Thesis Reviewers

Dr. Sandrine Voros
Director of Research,
TIMC-IMAG, France

Dr. Stamatia Giannarou
Senior Lecturer,
Imperial College London, UK

Thesis Examiners

Prof. Marco Cristani
Professor of Computer Science,
University of Verona, Italy

Dr. Thomas Lampert
Assistant Professor,
University of Strasbourg, France

ತಾಳಿದವನು ಬಾಳಿಯಾನು

(Pronunciation: Taalidavanu baaliyaanu)
Patience is a virtue for triumphing in life

(Kannada Proverb)

Dedicated to my parents

Mr. Thunganalli Satyanarayana Rao **Ramesh**
&
Mrs. Sampige Anantharamaiah **Pushpalatha**

Abstract

The demand for therapeutic care based on Minimally Invasive Surgery (MIS) is accelerating due to technological innovations that have improved patient outcomes. Although these technological advances have enabled information systems to provide rich intra-operative data in the Operating Room (OR), they have increased the complexity of the surgical workflow along with surgeons' cognitive workload. Optimizing workflow besides reducing surgeons' workload via intelligent systems that could provide clinical decision support or context-aware assistance is a growing necessity. A main component of workflow optimization is the ability to automatically recognize the current state of the surgery. This is effectively accomplished by modeling workflows as a set of activities that could be defined at different levels of detail: procedure, phase, step, action triplet, gesture, etc. Despite the vast literature on activity recognition in the surgical community, a majority of research efforts have been on coarse-grained phase recognition. Developing more detailed activity recognition methods is essential to better model surgical workflows and advance the capabilities of Context-Aware Systems (CAS) in modern ORs.

This thesis aims to develop multi-level (i.e. phase and step) activity recognition methods from endoscopic videos based on deep learning. We focus on the analysis of a high-volume surgical procedure to treat obesity that exhibits a complex workflow, called Laparoscopic Roux-en-Y Gastric Bypass (LRYGB). We introduce a large video dataset of LRYGB procedures fully annotated with phase and step labels. We then target joint phase and step recognition and develop a multi-task model based on temporal convolutional networks. Next, to alleviate the difficulty of manually annotating large datasets with fine-grained step labels, we propose a novel weakly-supervised learning method using easier-to-annotate phase labels as weak signals for step recognition. Subsequently, we investigate data augmentation for spatio-temporal activity recognition models as it is an essential component for optimal training of deep learning models. We propose a simplified augmentation method designed to incorporate the temporal dimension present in the task and videos. Finally, we study the generalization of the proposed activity recognition models on a large dataset of LRYGB procedures constructed from surgeries performed at two medical centers.

Sommario

La richiesta di cure terapeutiche basate sul Minimally Invasive Surgery (MIS) sta accelerando grazie alle innovazioni tecnologiche che hanno migliorato i risultati dei pazienti. Sebbene questi progressi tecnologici abbiano permesso ai sistemi informatici di fornire ricchi dati intraoperatori nel Operating Room (OR), hanno aumentato la complessità del flusso di lavoro chirurgico e il carico di lavoro cognitivo dei chirurghi. L'ottimizzazione del flusso di lavoro, oltre alla riduzione del carico di lavoro dei chirurghi, attraverso sistemi intelligenti in grado di fornire supporto decisionale clinico o assistenza context-aware, è una necessità crescente. Una componente principale dell'ottimizzazione del flusso di lavoro è la capacità di riconoscere automaticamente lo stato attuale dell'intervento. Ciò si ottiene efficacemente modellando i flussi di lavoro come un insieme di attività che possono essere definite a diversi livelli di dettaglio: procedura, fase, passo, terna di azioni, gesto, ecc. Nonostante la vasta letteratura sul riconoscimento delle attività nella comunità chirurgica, la maggior parte degli sforzi di ricerca si è concentrata sul riconoscimento delle fasi a grana grossa. Lo sviluppo di metodi di riconoscimento dell'attività più dettagliati è essenziale per modellare meglio i flussi di lavoro chirurgici e far progredire le capacità dei Context-Aware Systems (CAS) nei moderni OR.

Questa tesi mira a sviluppare metodi di riconoscimento dell'attività a più livelli (cioè fase e passo) da video endoscopici basati sul deep learning. Ci concentriamo sull'analisi di una procedura chirurgica ad alto volume per il trattamento dell'obesità che presenta un flusso di lavoro complesso, chiamato Laparoscopic Roux-en-Y Gastric Bypass (LRYGB). Introduciamo un ampio set di video di procedure LRYGB completamente annotate con etichette di fase e passo. In seguito, ci occupiamo del riconoscimento congiunto di fase e passo e sviluppiamo un modello multi-task basato su reti convoluzionali temporali. Per ovviare alla difficoltà di annotare manualmente grandi insiemi di dati con etichette di fase a grana fine, proponiamo un nuovo metodo di apprendimento debolmente supervisionato che utilizza le etichette di fase più facili da annotare come segnali deboli per il riconoscimento dei passi. Successivamente, analizziamo l'aumento dei dati per i modelli di riconoscimento dell'attività spazio-temporale, in quanto è una componente essenziale per l'addestramento ottimale dei modelli di apprendimento profondo. Proponiamo un metodo di incremento semplificato, progettato per incorporare la dimensione temporale presente nell'attività e

nei video. Infine, studiamo la generalizzazione dei modelli di riconoscimento dell'attività proposti su un ampio set di dati di procedure LRYGB costruite a partire da interventi chirurgici eseguiti in due centri medici.

Acknowledgments

This thesis is the outcome of my research across three years at the *Altair Robotics Laboratory* at the University of Verona, Italy and *CAMMA* group of ICube Laboratory at the University of Strasbourg, France. I would like to acknowledge all the support and teachings I have received from numerous people who contributed to the successful completion of my thesis.

First and foremost, I would like to thank **Dr. Sandrine Voros** and **Dr. Stamatia Giannarou** for their time dedicated to reviewing my dissertation and for being part of my defense committee. I would like to thank **Prof. Marco Cristani**, **Dr. Thomas Lampert**, and **Prof. Elena De Momi** for being part of my defense committee. It was a great honor to have my work evaluated by such respected members of the research community and to have received valuable feedback.

I would like to express my gratitude to both my thesis directors **Prof. Paolo Fiorini** and **Prof. Nicolas Padoy**. Thank you **Paolo** for giving me the freedom to explore the field of surgical robotics and choose the research direction through my thesis. Your enormous experience in the field and passion have very much motivated me over the years. Thank you **Nicolas** for the time, effort, and resources that you dedicated to guiding me to build painstaking and rigorous research throughout my thesis. I appreciate his brilliance and strong emphasis on quality over quantity which motivated me to always strive for excellence. In particular, I highly value the direct, and sometimes intimidating, feedback he has provided in the past 4 years. My profound appreciation to **Dr. Diego Dall'Alba** for his cheerful attitude and continuous support which has been fundamental for every achievement over the past 4 years.

I appreciate the immense time and effort dedicated by Dr. **Cristians Gonzalez** and Dr. **Joël L. Lavanchy** to devise ontology and annotate large video datasets needed for research. This clinical collaboration with them has been beneficial in my understanding of the surgical domain allowing me to effectively communicate my results in a clinically sensible way. I thank them for answering all my questions on the subject matter.

I would like to thank all the friends I have gained over the years. First, I would like to thank all my friends in the **ALTAIR** lab. I am particularly grateful to **Giacomo De Rossi**, **Giovanni Menegozzo**, **Eleonora Tagliabue**, and **Andrea Roberti** for

welcoming me and making Verona a home away from home. I will cherish the time spent together traveling and partying with **Ameya Pore**, **Eleonora Tagliabue**, and **Giovanni Menegozzo**. I would like to thank **Giacomo De Rossi**, **Nicola Piccinelli**, **Daniele Meli**, **Luca Pasetto**, **Damiano Rigo**, and **Federico Vesentini** for all the aperitivos at Alpini/Clipper and discussions on various topics: robotics, sports, TV series, psychology, philosophy, and most importantly food. I have unconsciously imbibed the "Italianness" that will stay with me for the rest of my life.

I would to thank all the people in **Strasbourg**. My secondment period wouldn't have been memorable without all the coffee, lunch, and after-work discussions with the team members, in particular, **Deepak Alapatt**, **Vinkle Srivastav**, **Luca Sestini**, **Pietro Mascagni**, **Chinedu Nwoye**, and **Tong Yu**. The knowledge, tips, and motivation from all of you have been very helpful in the success of my doctoral thesis. I would like to thank **Suzy** for all her support and valuable life lessons.

I would like to thank all my colleagues from the **ATLAS** project. Working together in a multi-disciplinary team has enabled me to learn different facets of autonomous surgical robotics. I would also like to thank all my friends from **home** who have helped me grow at different stages of my life and for being there throughout my journey. In particular, I would like to thank **Shashank Gupta** and **Ananth Murthy** for instilling in me this hunger for knowledge, imprinting open nature to new experiences, and bringing out the best in me.

Finally, I would like to thank my family. I am particularly indebted to my parents for imparting their moral values and ambitions, unwavering support, and numerous sacrifices. I cannot imagine any success without their love and support.

Carrying out my research wouldn't have been possible without the financial support from the ATLAS project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813782. Additionally, my work was supported by French state funds managed within the Investissements d'Avenir program by BPI France (project CONDOR) and by the ANR (ANR-16-CE33-0009, ANR-10-IAHU-02). I also would like to acknowledge the privilege of access to the HPC resources from GENCI-IDRIS (Grant 2021-AD011012832, 2022-AD011012832R1) and Unistra Mesocentre.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACC Accuracy. xvii–xix, 50–52, 62–64, 66, 89, 90, 118, 120, 121, 124

BernBypass70 BernBypass70. xv, xviii, 15, 85–90, 92, 93, 107, 124

biLSTM Bidirectional Long Short-Term Memory. 26

BY40 Bypass40. xv, 14, 15, 30, 36, 39, 44, 49, 56, 60, 68, 73, 80, 86, 89, 106, 107, 114, 117, 122, 123, 125

CA50 CATARACTS. 56, 73

CAI Computer-Assisted Intervention. 4, 5, 7, 11, 47

CAS Context-Aware Systems. iii, v, 6, 7, 10, 11, 13, 80, 99, 100, 102

CNN Convolutional Neural Network. 19–23, 26, 27, 48, 49, 87

CRF Conditional Random Field. 26

DTW Dynamic Time Warping. 18

F1 F1-score. xvii–xix, 50, 51, 53, 62–66, 88–90, 120, 121, 124

fps frames-per-second. 39, 49, 60, 73, 86

HMM Hidden Markov Model. 18, 19, 23

HOG histogram of oriented gradients. 22

LC Laparoscopic Cholecystectomy. xiii, 8–10, 19, 22, 27, 30, 102

LRYGB Laparoscopic Roux-en-Y Gastric Bypass. iii, v, vi, xiv, xviii, 12, 14, 15, 30, 36, 37, 39, 40, 44, 48, 60, 68, 73, 80–82, 84–87, 91, 100, 106, 107

LSTM Long Short-Term Memory. 20–22, 27, 50, 52

MIS Minimally Invasive Surgery. iii, v, 4, 5, 8, 11, 102

ML Machine Learning. 11, 23, 24

MS-TCN Multi-Stage Temporal Convolutional Networks. 30, 48, 58, 87

MTMS-TCN Multi-Task Multi-Stage Temporal Convolutional Networks. xvii–xix, 14, 30, 48–54, 57, 73, 76, 86–89, 114, 118, 123, 124

MultiBypass140 MultiBypass140. 80, 84, 87, 89, 124

OR Operating Room. iii, v, 4, 6–9, 11, 13, 14, 17, 80, 89, 99, 105, 107

PR Precision. xvii–xix, 50, 51, 53, 62–64, 66, 89, 90, 120, 121, 124

RAS Robot-Assisted Surgery. 4, 5, 7

RE Recall. xvii–xix, 50, 51, 53, 62–64, 66, 89, 90, 120, 121, 124

RFID radio-frequency identification. 8

RGB Red-Green-Blue. 48, 58, 72

RNN Recurrent Neural Network. 19, 23, 72

SDS Surgical Data Science. 5, 6, 10, 13, 17–19, 22, 30, 44, 107, 108

SIFT scale-invariant feature transform. 22

SOTA state-of-the-art. 11, 12

SS-TCN Single-Stage Temporal Convolutional Networks. xiv, 57, 58, 61, 68, 72, 73

StrasBypass70 StrasBypass70. xv, 15, 85–90, 92, 93, 107, 124

SVM Support Vector Machine. 18, 22

TCN Temporal Convolutional Network. xiii, 21, 23, 30, 48–50, 52, 58, 72, 73, 87

TeCNO TeCNO. xvii, 48, 50–54, 118

TRandAugment Temporal Random Augmentations. 15, 70, 74, 77, 107, 115, 122

Part I
Introduction and related work

1 Introduction

ಆರೊಗ್ಯವೇ ಭಾಗ್ಯ
(Pronunciation: Aarogyave bhaagya)
Health is wealth

(Kannada Proverb)

Chapter Summary

Surgery is a specialty field of medicine that directs attention to treating pathological conditions such as disease or injury by use of manual and instrumental operative techniques on a person. Surgery can have many benefits for patients, like improved bodily

functions, enriching physical appearance, or repairing ruptured areas. Thus, as old as humanity, surgery has been in practice and reached a heightened level of advancement in different ancient civilizations in China, Egypt, India, and Greece. For instance, in the Indian subcontinent, one of the oldest known surgical texts usually placed around 1200–600 BC called Sushruta Samhita describes in detail the diagnosis and treatment for various forms of cosmetic surgery, plastic surgery, and rhinoplasty. Similarly, ancient Greeks also performed some surgical procedures including setting broken bones, bloodletting, draining the lungs of patients with pneumonia, and amputations. However, surgery was not taught in many universities until the United Company of Barber Surgeons of London was formed in 1540[1]. This paved way for the establishment of control over the qualifications of those who performed operations.

The term *modern surgery* was introduced in the 18th century to highlight the marked progress made by surgery thanks to the introduction of experimental and empirical scientific approaches. In the following centuries, the constant inventions and innovations in various fields of science have transformed modern surgery to the present day. This transformation in the surgery can be observed through the advancement of the modern Operating Room (OR). A glimpse of the transition of the OR can be seen in Figure 1.1. Some of the advances in surgery are asepsis, different anesthesia techniques, antibiotics, hemostat for hemostasis, suturing, blood transfusions, grafts, organ transplants, etc. Concurrently with surgical advances, the OR transformation includes many technological advances: projectional radiographs, computed tomography, fluoroscopy & C-arm, autoclaves, blood pressure & pulse rate monitoring systems, electrocardiograms for observing heart contractions, electroencephalograms to watch brain activity, heart-lung machines, and others. All of these advances, coupled with the availability of better and more specialized surgical tools, have allowed for the introduction of less invasive and more effective surgical techniques. All of this encapsulates modern surgery.

In the last few decades, developments have been focused on the shift from traditional open surgery to Minimally Invasive Surgery (MIS). The distinct motivation for this change is the benefits of less pain, shorter recovery time, and fewer complications which MIS provides to the patients. Laparoscopy - surgery performed in the abdomen or pelvis using small incisions and the aid of a camera - is one of the first types of MIS to be accomplished. The first laparoscopic appendectomy was performed in 1981 [Meljnikov 2009] which encouraged the first laparoscopic cholecystectomy in 1985 [Reynolds 2001]. The success of these surgeries led to rapid acceptance and spread of MIS to other complex surgical procedures such as adrenalectomy (adrenal gland removal), brain surgery, colectomy (colon), gastrointestinal surgery (esophagus, stomach, small intestine, large intestine, rectum), heart surgery, hiatal hernia (stomach), kidney transplant, nephrectomy (kidney removal), splenectomy (spleen removal).

The popularity of MIS has led to the innovation of Computer-Assisted Intervention (CAI) and Robot-Assisted Surgery (RAS). CAI is the field that deals with developing computer systems and technology to continuously support physicians in making the

[1] https://www.britannica.com/science/surgery-medicine#ref253436

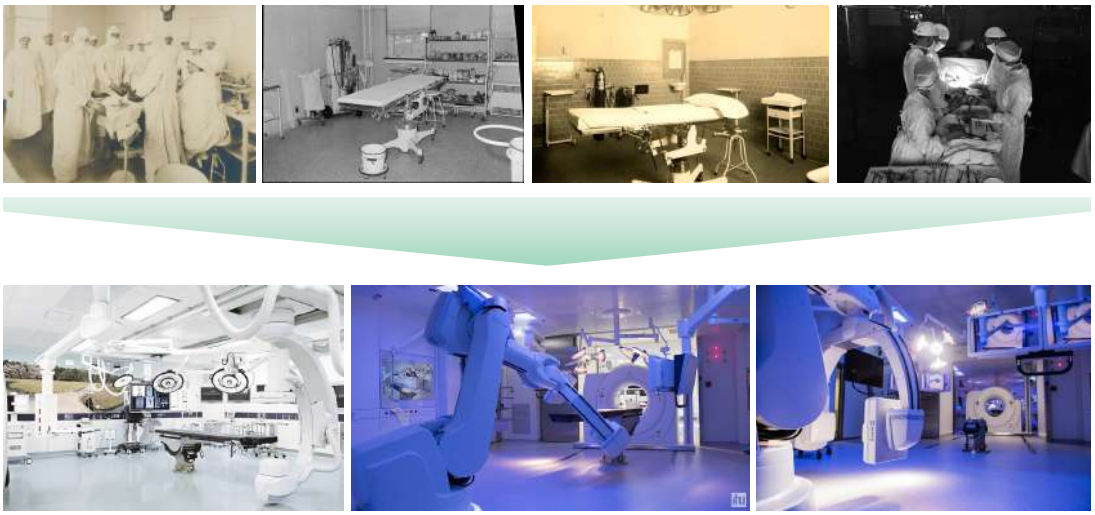


Figure 1.1: The transformation of the operating room in the last few centuries.

right decisions and executing the right actions. It includes advanced assistance via imaging, robotics, machine learning, and augmented reality during surgery as well as for the training of physicians and the surgical team. On the other hand, RAS employs robots to support clinicians with precise control over the instrumentation. Although the advances in MIS provide many benefits [Darzi 2004, Mohiuddin 2013], they introduce new challenges such as steep learning curves for new perioperative staff members & surgeons, restricted view of the anatomy, longer duration of some procedures compared to open surgery, limited range of motion of the instruments, limited sensory inputs of depth & touch, etc. Addressing these challenges is the primary goal of an emerging scientific discipline, Surgical Data Science (SDS).

The following sections present the aim of this thesis. First, the context of this study is introduced in Section 1.1 & 1.2. These sections also discuss some of the popular directions of research in the community which serves as the motivation for the research of this thesis. Next, Section 1.3 presents different challenges associated with surgical workflow analysis. Lastly, the contribution and outline of this thesis are summarized in Section 1.4 and 1.5, respectively.

1.1 The rise of Surgical Data Science (SDS)

The technological advances in CAI have enabled a vast array of data sources that can be recorded effortlessly. Surgical Data Science (SDS) has emerged as a scientific field that aims to improve the quality of capturing, organizing, analyzing, and modeling of pre-, intra-, and post-operative data [Maier-Hein 2017]. This pertains to a broad spectrum of data collected concerning patients, caretakers, and technology utilized in clinical care. The data ranging from patients' initial presentation to long-term outcomes, information from clinical guidelines, experiences, practices, patient preferences, and medical devices or sensors are analyzed and contextualized as generic domain-specific knowledge.

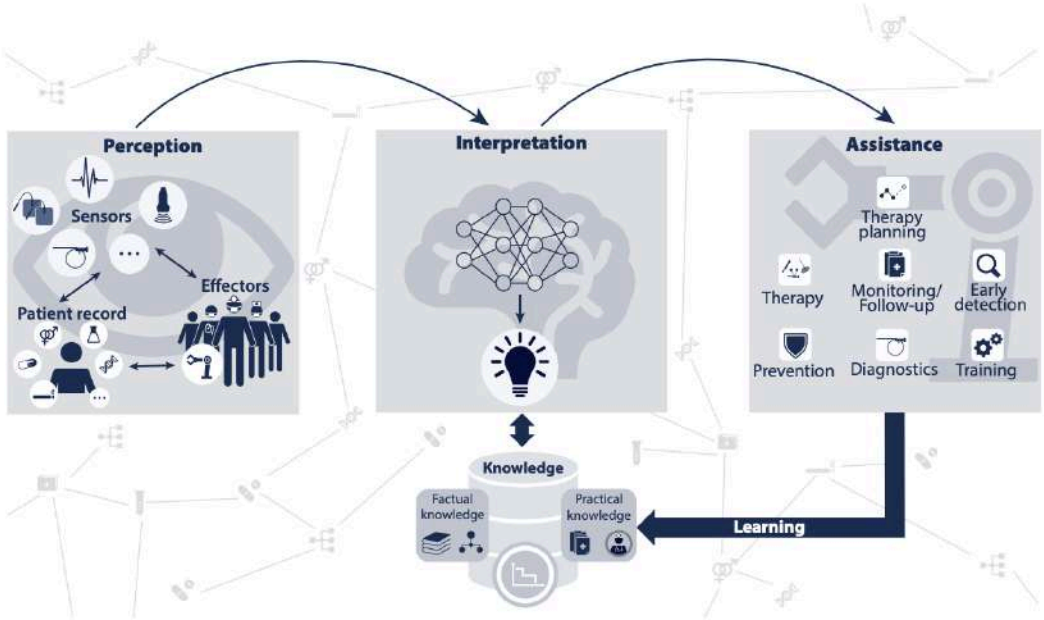


Figure 1.2: Overview of surgical data science (SDS) system. Image credits: [Maier-Hein 2017]

Through this knowledge, SDS targets assisting physicians in decision-making and planning execution, predicting events and clinical outcomes, specialized training of clinicians, advancing preventive methodologies, and/or assessing the quality of care provided to the patients. Figure 1.2 provides an overview of the SDS system.

A key clinical application of SDS is the development of Context-Aware Systems (CAS) which aims at providing contextual support to clinicians by exploiting the various sensory information available in the OR [Lemke 2005, Bricon-Souf 2007, Kranzfelder 2012, Maier-Hein 2017, Vercauteren 2020]. The contextual support could range from a simple display of relevant information effectively to aiding clinicians with suggestions on the course of action while performing difficult surgical tasks. **To design an effective CAS system, this thesis spotlights on one of its key components, which is the automatic analysis of a surgical workflow.**

1.2 Surgical Activity Recognition

Research in advancing the modern OR has proposed to develop Context-Aware Systems (CAS) [Lemke 2005]. CAS are advanced support systems that have the ability to draw context from the available data encompassing patients' health record, surgery type & its historical record, clinicians' experience, hospital facilities, intra-operative signals in the OR, postoperative complications, etc. The context could then enable these systems to adapt to the changing circumstances, both in and out of an OR, and act accordingly by presenting relevant information & services to a user, executing a service, and storing the context for effective retrieval [Bricon-Souf 2007]. Note, many surgeons and even engineers are skeptical about realizing CAS with such a high level of situa-

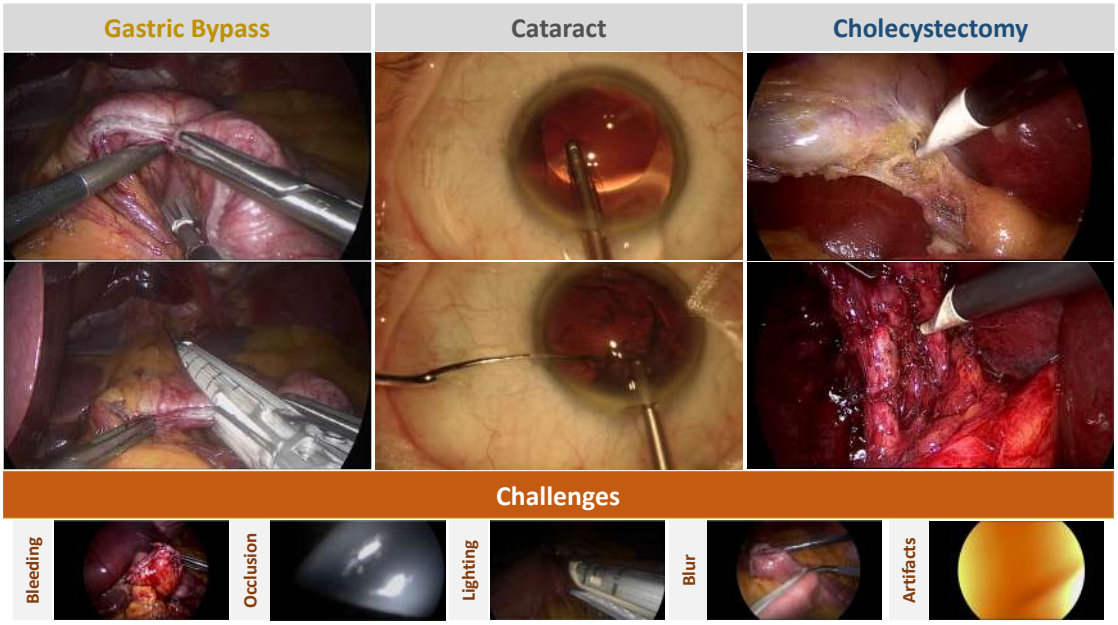


Figure 1.3: Sample images from endoscopic cameras of three different procedures.

tional awareness [Cobianchi 2023]. However, the rise of Computer-Assisted Intervention (CAI) and Robot-Assisted Surgery (RAS) has also increased the complexity of executing surgical procedures, consequently affirming the need for CAS due to their potential benefits. The contextual support of a CAS could contribute to simplifying surgical workflows, improving human-machine communications, and faster execution of surgical maneuvers; resulting in decreased surgical workload and strain, thus reducing surgical errors, increasing patient safety, and improving overall safety, quality, and efficiency of care [Maier-Hein 2017, Vercauteren 2020].

One of the primary functions of a CAS is the ability to automatically analyze the surgical workflows, by means of reliable recognition of the surgical activities [Kranzfelder 2012]. By examining the comprehensive online data from the OR, if systems could recognize the current state of the procedure, then they could also be capable of predicting the progression of the procedure. This capability could provide active support to surgeons helping in their clinical decision-making which successively could induce autonomy in RAS. Additionally, understanding workflows would allow these systems to automatically generate surgical reports and annotate data appropriately for effortless retrospective studies. This semantic information is at the core of the cognitive understanding of the surgery.

1.2.1 Digital signals in the modern OR

Recognition of surgical activities requires collection and analysis of the information available in the OR. Especially in the modern OR, the use of digital monitoring systems provides many useful signals which can be easily recorded and analyzed. Initial works

proposed the use of tool usage information collected manually [Ahmadi 2006, Padoy 2007, Blum 2008b, Padoy 2012], using radio-frequency identification (RFID) based tool tracking, or additional camera-equipped trocar [Kranzfelder 2009, Toti 2014, Joerger 2017] to recognize surgical activities. With the rise of robotic systems such as the da Vinci surgical robot, additional system events [Malpani 2016] and dexterous human motion signals [van Amsterdam 2021], through kinematics of tools, could also be easily collected for supporting activity recognition. However, these signals by themselves are complex to interpret and pose acquisition challenges as new devices or modifications to the existing systems are required to extract data during the procedure.

The **endoscopic camera** is the primary source of information used by clinicians during MIS procedures. Since it **captures the detailed interaction of the surgical instruments with the underlying anatomical structures, it is a powerful (and readily available) source of information in the OR.** A couple of images from endoscopic cameras of different surgeries is illustrated in Figure 1.3. Furthermore, recent breakthroughs in computer vision driven by deep learning methods have provided strong incentives to use visual signals. Hence, recording from camera devices has been utilized extensively in research studies tackling the problem of surgical activity recognition [Garrow 2020, Demir 2022]. Note, other signals acquired in the OR could also be integrated with endoscopic images favoring adequate interpretation of the information processed from them.

1.2.2 Types of activities

Surgical activities that describe a surgical workflow can be defined at different levels of granularity: procedure, phase, stage, step, action, and other low-level information (Figure 1.4) [Katić 2015, Meireles 2021]. This hierarchical subdivision of surgical activities has been proposed to develop a common ontology for surgical workflows which could improve the translation of results and facilitate multi-institutional research efforts [Gibaud 2018]. Additionally, it enables workflow modeling with a high degree of detail, favoring a standardization of their execution and the definition of accurate and easily applicable guidelines in clinical practice. The following sections detail popular types of activities in the literature.

1.2.2.1 Phase

At a coarser level, a surgical workflow can be described by phases. Phases are a set of fundamental surgical aims of a procedure that needs to be performed to complete it successfully. For example, Laparoscopic Cholecystectomy (LC) has been captured using different sets of phase or surgical aims [Garrow 2020], the most popular being 7 phases - Preparation, Calot triangle dissection, Clipping and cutting, Gallbladder dissection, Gallbladder packaging, Cleaning and coagulation, Gallbladder retraction - presented by [Twinanda 2017a]. The inception of phase recognition could be attributed to [Jannin 2001] who proposed to model surgical procedures as a sequence of tasks. The

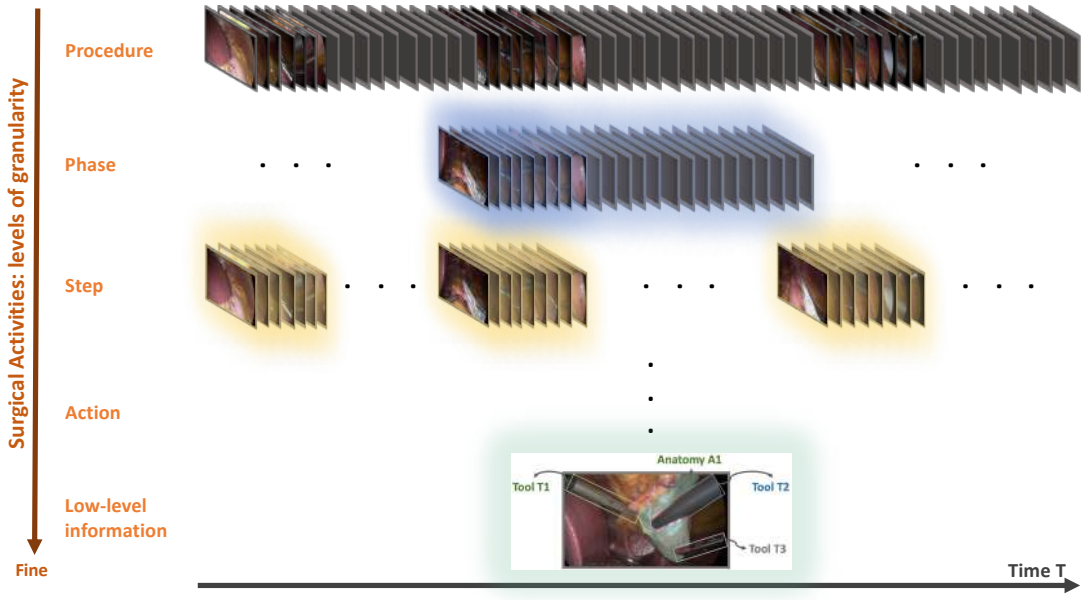


Figure 1.4: Types of surgical activities based on level of granularity.

formalization allowed to effectively encapsulates the information in the OR based on prior knowledge of the surgical workflow. Each phase of a surgery encapsulates information on the current state of the procedure, the tools involved in the task, human and technological resources required, the next task to perform, and suitable visualization with interactive features [Jannin 2001]. The 'Calot triangle dissection' phase of LC, for instance, implies that two tools (grasper and hook) are used to perform tissue dissection on the gallbladder. Additionally, this phase could also indicate that deploying a visualization of safe ("Go") and dangerous ("No-Go") zones during dissection [Laplante 2022] and warning when an instrument is approaching a risk structure [Speidel 2008] is needed to potentially reduce injuries to anatomy (bile duct). This initiated studies on phase recognition which in the last two decades has flourished immensely and is at the forefront of research on workflow analysis [Garrow 2020, Demir 2022].

1.2.2.2 Step

At a finer level than phases, steps represent activities that have to be carried out to complete each phase of a surgical workflow. The earliest work on step recognition was in 2014 for workflow analysis of cataract procedures [Quellec 2014]. A few later works continued this direction of research and even attempted to recognize both phases and steps of cataract surgery [Charrière 2014, Charrière 2017]. The growing interest in steps could be due to the natural hierarchical definition of different types of activities introduced in the literature from both informatics [Katić 2015, Gibaud 2018] and medical community [Meireles 2021]. This help clinicians break down a surgical workflow and formulate ontology with varying level of detail that facilitates the standardized execution of procedures. Since steps represent a workflow in more detail than phases, they

could further enhance capabilities of CAS that could monitor the successful completion of each phase or surgical aim of a procedure. The potential benefits for the surgeons are two-fold: 1) reducing their cognitive workload by presenting the information pertinent to the current state of the surgery, and 2) assisting in training novice surgeons on both simulators [Agha 2015, Meling 2020] and live surgeries. Thus, this thesis focuses strongly on step recognition alongside recognizing phases.

1.2.2.3 Other activities

Procedure. At the coarse level, above phase, procedure recognition aims at classifying the type of surgery from the available signals. Information at this level enables CAS to analyze and offer a user interface specific to the surgery. While this information can be obtained manually, it could cause disruption in the surgical workflow. On the other hand, the information recorded using proprietary systems is not easily accessible. Hence, automatic recognition of procedure could be helpful in overcoming the above two restrictions. Additionally, procedure recognition is valuable for efficient database indexing and fast data retrieval from a large database of surgical recordings [Kannan 2020].

Surgical action triplets. Formulated as a triple of the instrument, action, and target anatomy, surgical action triplets comprehensively capture the activities involved in a workflow [Katić 2014]. The action triplets provide fine-grained information on the instrument-tissue interactions: what instrument is being used, the action performed by the instrument, and the anatomy acted upon. In a LC for example, an automated surgical safety system would benefit from automatically detecting individual actions, such as clipping, when performed on critical anatomies, such as a cystic artery or other blood vessels [Mascagni 2020]. Recently, automatic recognition of action triplets has witnessed a surge of research [Nwoye 2020, Nwoye 2022b]. To foster research in this direction, two editions of the challenge on surgical action triplet recognition [Nwoye 2022a, Nwoye 2023] have been organized as part of the EndoVis grand challenge [2]. In spite of its potential, enumerating all the possible action triplets and annotating large datasets requires immense effort. LC alone can have 100+ triplets that are composed from 6 instruments, 8+ actions, and 15+ targets [Nwoye 2020, Nwoye 2022b].

Gestures. Gestures are another popular type of activity that has received significant interest from the SDS community [van Amsterdam 2021]. At the finest level, surgical gestures capture the dexterous motion made with a specific purpose such as "G1: Reaching for needle with right hand", "G3: Pushing needle through tissue", or "G13: Making C loop around right hand" [Gao 2014, Ahmidi 2017]. Gesture recognition is beneficial for devising objective criteria for training surgeons and qualitative evaluation of surgical skills [Rosen 2001, Reiley 2009, Vaughan 2016]. Furthermore, by

[2] https://endovis.grand-challenge.org/

linking the gesture information back to the robot control loop, an advanced robotic system could achieve surgical automation.

In conclusion, surgical workflow analysis, one of the most essential components for achieving context-awareness in the Operating Room (OR), can be performed at different levels of detail. Although tremendous research efforts can be found across different types of activities, they are primarily carried out independently. Subsequent studies should aim at multi-level activity recognition and capitalize on the inherent hierarchical relationship between them.

1.3 Challenges

Despite the community's great interest in surgical activity recognition, the objective of most of these studies has been phase recognition. Step, and multi-level activity, recognition has been limited due to challenges in the availability of data and methodology.

1.3.1 Multi-level activity recognition dataset

Recognition of phase, step, action triplet, and gesture have all been studied to some extent. However, they have all been researched independently with very few works attempting to recognize activities at multiple levels. Furthermore, these data-driven research activities utilize large datasets with annotations of a specific activity. In the surgical vision community, popular datasets include Cholec80 [Twinanda 2017a] & M2CAI [Twinanda 2017a, Stauder 2016] with phase and tool presence annotation, CholecT40 [Nwoye 2020] & CholecT50 [Nwoye 2022b] with surgical action triplets labels, CATARACTS [Hajj 2019] & cataract-101 [Schoeffmann 2018] which contains step and tool presence labels for cataract surgical procedure, and JIGSAWS [Gao 2014, Ahmidi 2017] that contains gesture annotations for bench-top training exercises in robotic surgery. This lack of datasets for multi-level activity recognition could be a consequence of a primary bottleneck from the medical society: the lack of standard and reusable representation of surgical knowledge, particularly of the surgical workflow. Recent efforts from both the informatics and the medical community have presented hierarchical characterization of a workflow [Katić 2015, Meireles 2021]. Yet formulation of an ontology is surgery and activity specific which demands expert knowledge of the respective domain. Despite these challenges, research attempt to generate datasets of different surgical workflows with multi-level annotations is paramount for the technological advancement of CAS.

1.3.2 Labeled data scarcity

Owing to the technological transformation in the OR and the rise of MIS and CAI, accumulating the data from the OR and constructing large video databases can be achieved effortlessly. Unfortunately, these databases need to be extended with annotations since the state-of-the-art (SOTA) methods proposed in the literature follow the

Machine Learning (ML) paradigm called supervised learning that requires large datasets with supervisory signals for learning and also evaluation. Thus, large-scale datasets are of utmost importance for the design and validation of the SOTA activity recognition models. Alongside formulating an ontology that effectively defines the surgical workflow of interest, annotating a dataset with activity labels is a very challenging task [Ward 2021]. First and foremost the task demands annotators with experience mainly with surgery and also video annotation. Selecting experienced surgeons as annotators is costly from a financial standpoint and additionally from an opportunity perspective due to time spent away from treating patients. Next, ensuring consistency of annotations across the dataset is necessary. Although consistency can be achieved by utilizing a single expert annotator, incorporating multiple clinical expert annotators helps in reducing the burden of annotation on an individual. However, this raises the question of inter-annotator reliability which could be impacted considerably depending on the surgical activity being annotated. In the case of Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) procedure, the inter-annotator reliability could drop from $\sim$ 96% (phase) to $\sim$ 81% (step) when annotating a finer activity (step) between two expert clinicians on just 10 videos [Lavanchy 2022].

To reduce the dependency on labeled datasets, recent studies in both general computer vision and surgical vision communities have proposed different methodological approaches: weakly-, semi-, and self-supervised learning. Weakly-supervised learning methods aim to reduce the annotation cost by utilizing other easy-to-annotate "weak" labels such as global statistics, incomplete labels which provide access to partial knowledge on each class, or noisy labels from non-expert annotators from a crowd-sourcing platform[3,4]. On the other hand, semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data to achieve performance comparable to supervised learning methods. Self-supervised learning takes it a step further by utilizing auto-generated labels and eliminating the need for expert annotators. For instance, using tool presence labels as "weak" signals for the task of tool segmentation or localization & tracking [Nwoye 2019]; labeling less than 25% of the dataset with semi-supervised learning [Yu 2019]; and auto-generating supervisory signals using inherent information of remaining surgical duration for phase recognition [Yengera 2018]. In this regard, a part of this thesis focuses on weakly-supervised learning for surgical activity recognition (Chapter 5).

1.3.3 Optimal model training

Recent state-of-the-art methods addressing the problem of surgical activity recognition from endoscopic videos are based on deep learning. These deep learning models, consisting of millions of parameters, approximate the underlying function that maps input images/videos to corresponding activity labels utilizing large datasets. Effectively training the model parameters (i.e., finding their optimal values) requires careful tuning of

[3]https://www.mturk.com/
[4]https://scalehub.com/

various hyperparameters, such as optimizer, learning rate, data augmentation, batch size, number of epochs, momentum, model weight initialization, and others. Popular techniques to find optimal values of the hyperparameters are manual search, grid search, random search, and Bayesian optimization [Bergstra 2011, Bergstra 2013].

One of the most essential hyperparameters that demand closer examination is data augmentation [Shorten 2019, Mumuni 2022]. Data augmentation consists of an automatic process to extend the datasets by applying predefined transformations. A few popular image transformations used in the community are horizontal flip, rotation, random cropping, translation, scale, and color jitter. Models trained on extended datasets with variations introduced by data augmentation have been shown to impact model robustness [Lopes 2019] and performance on semi-supervised and self-supervised learning methods [Qian 2021, Pan 2021, Shi 2021]. To design optimal augmentation policies, it is important to incorporate prior knowledge of each domain. This calls for expertise in both the domain and data augmentation and involves strenuous manual work. Thus, data augmentation methods are difficult to extend to other domains and applications. Few papers in the general computer vision community have proposed simplified methods [Cubuk 2020] and learning optimal augmentation policies on a subset of the data on a proxy task [Cubuk 2019, Lim 2019]. Similar research attempts are required but missing in the SDS community for the task of activity recognition.

1.3.4 Robustness and generalizability - multi-center validation

To realize the application of activity recognition in the OR via CAS, the recognition module must possess characteristics of reliability, portability, and integrity. The three characteristics together loosely state that the recognition module should be safe to use and perform consistently under different working conditions for a specific period of time. Given that the latest methods for recognition are based on deep learning [Garrow 2020], these characteristics are studied in terms of robustness and generalizability where robustness ensures the integrity of the module while generalizability ensures reliability and portability. One of the key challenges to developing robust and generalizable deep learning methods is their susceptibility to overfitting and memorization because of the complexity of the number of parameters involved [Geirhos 2018, Feng 2019]. Popular approaches to prevent overfitting are data augmentation, feature selection, weight regularization, early stopping, adding more training data, etc. While data augmentation enables the generation of additional training data by perturbing the input data, weight regularization attempts to simplify the model complexity by adding a penalty term, based on the number of parameters, to the cost function.

Adding more training data is an expensive approach to tackle overfitting. Nevertheless, it is a crucial way to add natural variations of a domain to a dataset. In the medical and surgical domains, patients' age, height, weight, gender, race, ethnicity, and many other factors contribute to the variability in the data. Additionally, variances present in the surgical domain are owed to the changes in the surgical workflow across surgeons, medical centers, communities, nations, etc. Hence, an

ideal surgical activity recognition module is required to be robust in its capability across all these variances in both anatomy and workflow. However, most of the research in the community has utilized datasets from a single center for experimentation [Twinanda 2017a, Hajj 2019, Hong 2020, Nwoye 2022b, Schoeffmann 2018]. And it remains speculative whether these methods would generalize to other centers. Hence multi-center validation of surgical activity recognition methods is quintessential for its rapid adaptation in the OR.

1.4 Contribution of the thesis

The fundamental aim of this thesis is to address the problem of surgical activity recognition, at different levels of granularity, from endoscopic videos by developing recognition models based on deep learning. Specifically, the contributions of the thesis revolve around online recognition of phases and steps.

The first contribution presents a method consisting of spatial and temporal models for joint phase and step recognition, differently from works in the literature that have strongly focused on developing methods to recognize one specific level of granularity from video data: phases [Garrow 2020, Demir 2022], steps [Quellec 2014, Charrière 2014, Charrière 2017], action triplets [Nwoye 2020, Nwoye 2022b], and robotic gestures [van Amsterdam 2021]. To achieve this, we first introduce a new large-scale dataset called Bypass40 (BY40) consisting of 40 videos of complex Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) procedures fully annotated with both phase and step labels. Subsequently, we introduce MTMS-TCN, Multi-Task Multi-Stage Temporal Convolutional Networks, for joint recognition of phases and steps extending MS-TCNs that were proposed for action segmentation. The motivation for this method stems from the fact that both activities are hierarchically related and recognizing them jointly could allow the model to implicitly learn the hierarchical relationship and benefit from it.

For the second contribution, we focus our attention on weakly supervised learning for fine-grained activity, i.e., step recognition. Although collecting large datasets of endoscopic videos is automated, annotating them is a manual process that is difficult and time-consuming as these tasks require domain-specific medical knowledge. Furthermore, as steps define a surgical workflow at a more fine-grained level than phases, the time required to annotate a dataset with steps is significantly higher than with phase annotations. For example, in LRYGB procedures, the workflow consists of 44 steps and 11 phases (Figure 3.3) and precisely defining and annotating all the steps requires a considerably longer time due to the number of steps and more importantly lower inter-class variances between steps. To reduce this reliance on fully annotated datasets, especially for fine-grained activities like steps, we present a weakly supervised learning method that uses phase labels as weak signals to assist in step recognition in settings of label scarcity.

As a third contribution, we inspect one of the most essential component of the training pipeline of the deep learning methods: Data Augmentation. Data augmentation has

shown the potential to improve the generalization of deep learning models which has spurred research on automated and simplified augmentation strategies for image classification and object detection on datasets of still images. Extending such augmentation methods to videos is not straightforward, as the temporal dimension needs to be considered. Furthermore, surgical videos pose additional challenges as they are composed of multiple, interconnected, and long-duration activities. To address this need, we introduce a new simplified augmentation method, called Temporal Random Augmentations (TRandAugment), specifically designed for training Spatio-temporal models on long surgical videos. TRandAugment treats each video as an assemble of temporal segments and applies consistent but random transformations to each segment. The validation of TRandAugment on different tasks and datasets opens new avenues for research on the impact of temporal data augmentation methods on model robustness [Lopes 2019] or weakly-/semi-/self-supervised learning [Qian 2021, Pan 2021, Shi 2021, Yu 2019, Ramesh 2022, Ramesh 2023b].

Our fourth and final contribution, building on previous contributions, presents a study on the generalization of activity recognition methods on data from different medical centers. As part of this study, we introduce two new datasets, namely StrasBypass70 and BernBypass70, consisting of 70 videos of LRYGB procedures fully annotated with phase and step labels. Subsequently, we study the performance of both fully and weakly supervised learning methods on these datasets demonstrating to the community the challenges and shortcomings while transitioning from research to clinical generalization.

1.5 Thesis outline

The thesis is organized into three parts:

- The first part consists of two chapters that present the clinical context and motivation in chapter 1 followed by a review of the related works existing in the literature in chapter 2.

- The second part presents the contribution of this thesis and spans chapters 3-7. Chapter 3 introduces the new Bypass40 (BY40) dataset of complex Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) procedures constructed with phase and step annotations. Chapter 4 presents a fully supervised method consisting of spatial and temporal models for joint phase and step recognition. The Bypass40 dataset along with the method presented in chapter 4 has been published in [Ramesh 2021]. Chapter 5 presents a weakly supervised learning method for recognition of fine-grained step recognition utilizing coarser phase labels as weak signals. Some of the results presented in this chapter have been published in [Ramesh 2023b]. Chapter 6 presents a simple and automated data augmentation method called TRandAugment for training Spatio-temporal activity recognition models, specifically improving the performance on both the phase and step recognition tasks. The method presented in this chapter has been published in [Ramesh 2023a]. Lastly, chapter 7

demonstrates the generalizability of the above methods by means of a cross-center study on new datasets of LRYGB procedures from Strasbourg and Bern medical centers. Some of the results presented in this chapter have been planned for submission to a medical journal.

- The third, and final, part of this thesis discusses the potential applications of the proposed methods in chapter 8 and a summary of the thesis in chapter 9. chapter 9 also provides a discussion on future perspectives for advancing research in surgical activity recognition.

2 Related Works

Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning.

- Albert Einstein

Chapter Summary

Surgical activity recognition is a burgeoning research direction of SDS with earlier works emerging from the year 2000. Initial work on this topic was established by Jannin et al. [Jannin 2001] with the conceptualization of surgical activities. In the following two decades, a large body of research works emerged tackling the problem of surgical activity recognition. These works have utilized various digital signals available in the OR: tool usage [Ahmadi 2006], system events [Malpani 2016], tool kinematics through dexterous human signals [Lin 2006], and endoscopic videos [Klank 2008]. Out of these signals, endoscopic videos have become the popular choice as they capture important information about the tool-tissue interaction and are readily available in the OR without requiring any additional modifications.

Hence, in this chapter, we review related works on surgical activity recognition from endoscopic videos. The review also centers heavily on works based on deep learning covering fully-, semi-, and weakly-supervised learning paradigms. When necessary, we also include works from the general computer vision community to present a suitable context.

2.1 Supervised Learning

Supervised learning is a predominant paradigm of both general computer vision and SDS communities that learn from large volumes of data with additional labels as supervisory signals. This section presents research on surgical activity recognition utilizing large labeled datasets at both coarse and fine levels of granularity.

2.1.1 Phase & step recognition

The prevalent objective studied in the literature is the automatic recognition of phases in endoscopic videos [Ahmadi 2006, Blum 2010, Dergachyova 2016, Twinanda 2016, Funke 2018, Zisimopoulos 2018]. Initial works used endoscopic videos to manually collect tool usage information as input for phase recognition [Ahmadi 2006, Padoy 2008, Blum 2008a, Blum 2010, Padoy 2012, Dergachyova 2016]. For instance, [Ahmadi 2006] constructed a series of multi-dimensional state vectors over time of 17 different laparoscopic instruments used in cholecystectomy. The intuition behind this was that in minimally-invasive surgeries the laparoscopic instruments used by the surgeon strongly correlate with the underlying workflow. Many works using tool usage signals have proposed using Dynamic Time Warping (DTW) [Ahmadi 2006, Blum 2010, Padoy 2012] and Hidden Markov Model (HMM) [Padoy 2008, Padoy 2012, Dergachyova 2016]. DTW is an algorithm proposed by [Sakoe 1978] that calculates an optimal match between two given temporal sequences which may vary in speed and consequently measures similarity between them. On the other hand, a HMM [Rabiner 1989] is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobservable ("hidden") states.

Unlike tool signals that capture the usage of instruments, images/videos from endoscopic cameras are high-dimensional as an image can be treated as a collection of signals of $h \times w$ pixels. Here, h and w stand for the height and width representing an image's resolution. For instance, an image from an endoscopic camera of resolution 720×480 can be treated as an input signal of 345,600 dimensions. Learning and organizing data in such high-dimensional spaces is extremely challenging due to the phenomenon of the curse of dimensionality[1]. Reducing the high dimensionality of images via feature extraction is a primary focus of the computer vision community.

One of the initial works using camera features proposed a genetic algorithm that learns the optimal feature extraction method that helps a Support Vector Machine (SVM) to predict the surgical phases [Klank 2008]. [Padoy 2008] derived two signals based on camera images: if the camera is inside the body and a metallic clip is visible

[1] https://en.wikipedia.org/wiki/Curse_of_dimensionality

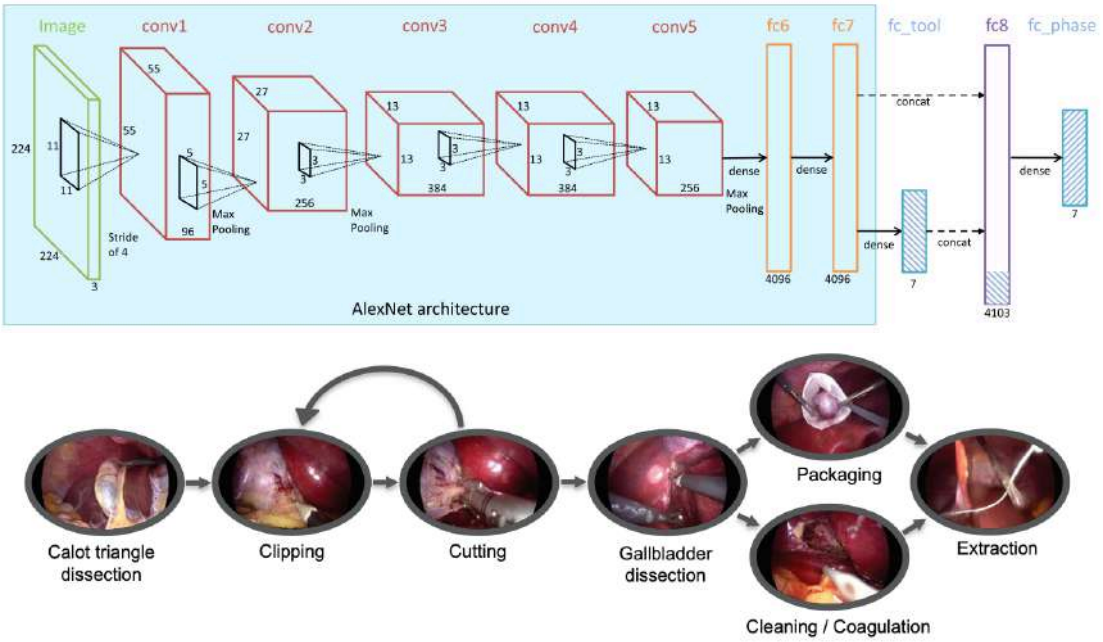


Figure 2.1: EndoNet architecture for recognizing the 7 phases of LC procedure. Image credits: [Twinanda 2017a, Padoy 2019]

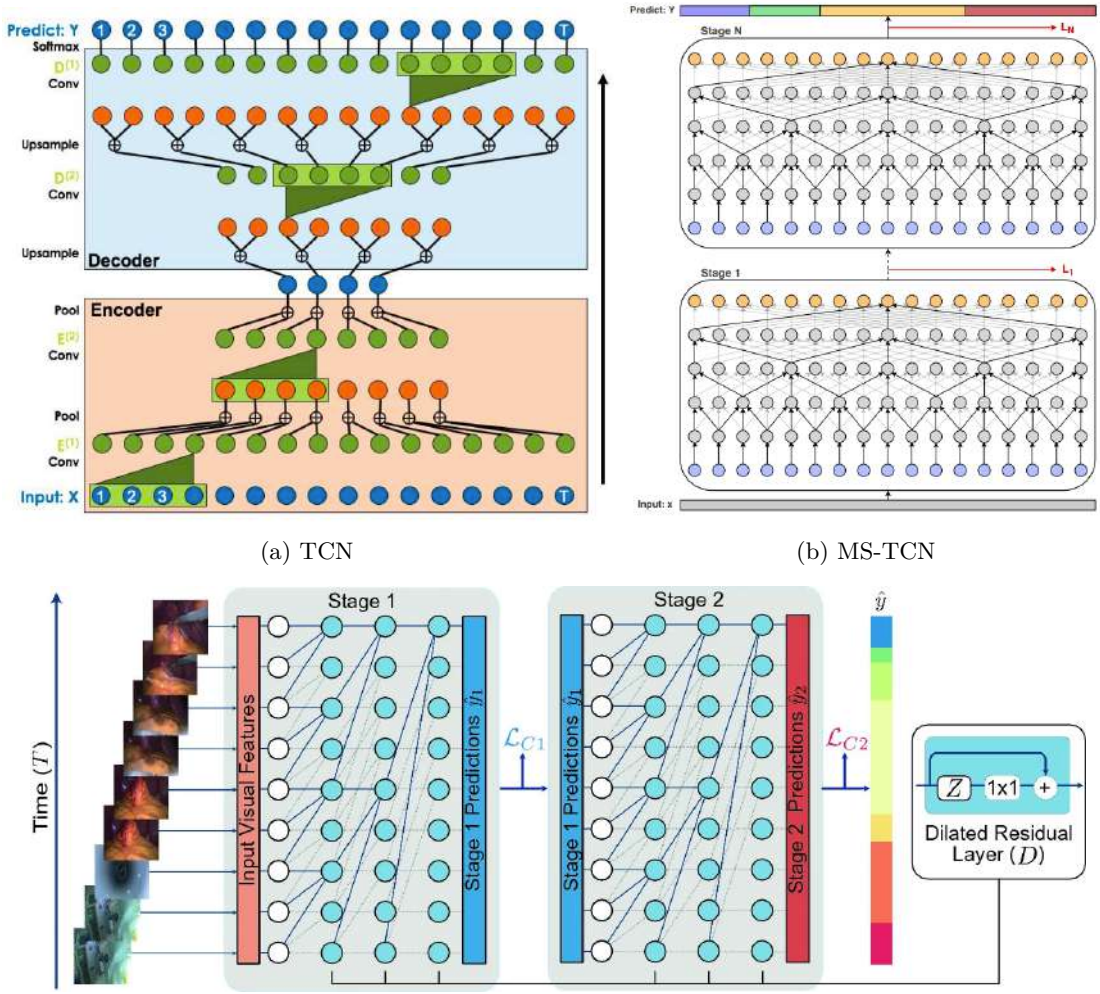
in the image. They employed image processing techniques such as color histograms, and color & shape classification. These two signals were used along with instrument usage signals and a HMM was trained for online recognition of surgical phases in Laparoscopic Cholecystectomy (LC). [Blum 2010] extracted 1932 features consisting of horizontal & vertical gradients, histograms, and pixel values of a 16x16 version of the image. These features were further reduced to a lower dimension using canonical correlation analysis to obtain features correlated with semantic meaningful signals. Finally, a HMM was used for phase segmentation. A drawback of these initial methods was that they required image features to be extracted using handcrafted methods which was a painstaking process.

The rise of deep learning for computer vision enabled the automatic extraction of visual features directly from images without any human intervention [Krizhevsky 2017]. In SDS, EndoNet [Twinanda 2017a] and DeepPhase [Zisimopoulos 2018] are early works that employed deep learning for surgical workflow analysis, particularly for phase recognition. EndoNet [Twinanda 2017a], presented in Figure 2.1, was proposed to recognize the 7 phases of LC along with instrument detection from endoscopic videos. The model consisted of a Convolutional Neural Network (CNN) for learning visual features followed by a hierarchical HMM for modeling temporal information. Similarly, DeepPhase [Zisimopoulos 2018] proposed a CNN followed by a Recurrent Neural Network (RNN) as a temporal model for recognizing 14 phases of cataract surgeries. These works laid the foundation for using CNNs for automatic visual feature learning.

The following works focused on introducing different temporal models to extract useful information from the temporal dimension available in endoscopic videos. One

of the popular temporal model used in the literature is Long Short-Term Memory (LSTM). EndoLSTM [Twinanda 2017b] was one the first works that evolved from EndoNet, combining a CNN for feature extraction and an LSTM for temporal refinement. SV-RCNet [Jin 2018] trained an end-to-end CNN and LSTM model utilizing ResNet [He 2016a] as the CNN architecture. Additionally, a prior knowledge inference scheme was proposed to further improve the consistency of the model's predictions. MTRCNet-CL [Jin 2020] proposed a multi-task model to detect tool presence and phase recognition. The features from the CNN were used to detect tool presence and also served as input to an LSTM model for phase prediction. Additionally, a correlation loss was introduced to enhance the synergy between the two tasks. In [Shi 2020], phase recognition was approached in an active learning framework where only video clips that contain richer information were subsampled for annotation. They proposed an NL-RCNet model consisting of an end-to-end CNN + LSTM model with an additional non-local block. The non-local block was used for long-range temporal dependency which provided criteria to subsample video clips for annotation. [Jin 2021] also trained an end-to-end CNN + LSTM model and proposed an additional memory bank for relating long-range and multi-scale temporal patterns to augment the present features. The long-range memory bank served as a memory cell that stored the rich supportive information and the temporal variation layer further enhanced this information using multi-scale temporal convolutions. To effectively incorporate the supportive cues a non-local bank operator was introduced to attentively relate the past to the present. While various methods use LSTMs, these models retain memory for a limited sequence. Since the average duration of a surgery can range from tens of minutes to a couple of hours, it makes it challenging for LSTM-based models to leverage temporal information for surgical phase recognition.

Temporal Convolutional Networks (TCNs) [Lea 2016b] were introduced to hierarchically process videos for action segmentation. An encoder-decoder architecture could encode both high- and low-level features in contrast to RNNs. Furthermore, dilated convolutions [van den Oord 2016] were utilized in TCNs for action segmentation that showed performance improvements due to a large receptive field for higher temporal resolution. Besides dilated convolutions that enable large receptive fields, MS-TCN [Farha 2019] consisted of a multi-stage predictor architecture with each stage consisting of multi-layer TCN that incrementally refined the previous stage's prediction. Recently, TeCNO [Czempiel 2020] adapted the MS-TCN architecture for online surgical phase prediction by implementing causal convolutions [van den Oord 2016]. These different TCN variants are visualized in Figure 2.2. More recent works have proposed transformer-based models to improve the performance of phase prediction models [Czempiel 2021,Gao 2021]. [Czempiel 2021] presented a transformer-based model with a novel attention regularization loss that encourages the model to focus on high-quality frames during training. The high-quality frames for each surgical phase are identified for summarizing a surgery using the attention weights. Parallelly, an aggregation Transformer that fuses spatial and temporal embeddings was proposed in [Gao 2021]. The spatial embeddings from a ResNet backbone and the temporal embeddings from a TeCNO model were aggregated by a



(a) TCN (b) MS-TCN

(c) TeCNO

Figure 2.2: Different temporal convolution networks proposed in the literature. (a) The architecture of an encoder-decoder Temporal Convolutional Network (TCN) for action segmentation proposed in [Lea 2016b]; (b) MS-TCN: a multi-stage TCN model for action segmentation that recursively refines predictions from the previous stage [Farha 2019]. Additionally, each stage consists of dilated residual layers to increase the receptive field; (c) TeCNO: TCN-based model architecture for online surgical phase recognition proposed in [Czempiel 2020]. Along with dilated residual layers, causal convolutions were proposed to achieve online recognition.

two-layer transformer model.

Alongside this large body of work on phase recognition, few works have attempted to recognize steps. [Charrière 2014] was the first work that aimed at real-time step recognition from cataract surgical videos. The proposed method was a Content-Based Video Retrieval (CBVR) system utilizing a novel pupil center and scale tracking method as pre-processing of motion features. In [Charrière 2017], the CBVR system along with surgical tool presence information was used as input to statistical models consisting of Bayesian Network and HMMs for multi-level online recognition of steps and phases. Recently, [Xia 2021] trained an end-to-end CNN + LSTM for the task of step recognition.

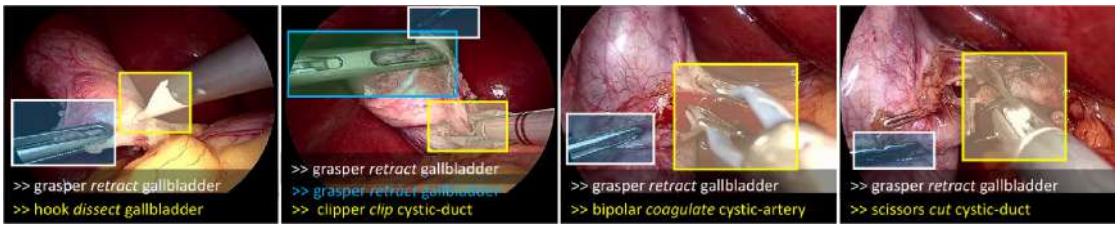


Figure 2.3: Few examples of action triplets in LC surgical videos. Modified image: [Nwoye 2022b]

A step-phase branch was designed to capture both global coarse-grained features and local fine-grained features. Furthermore, a contrastive branch was proposed to enlarge the distances of the features from different surgical phases and steps to handle the spatial-temporal discrepancy problem.

2.1.2 Recognition of other activities

Procedure. At the highest granularity, recognizing the type of surgical procedure has also received attention from the community. [Twinanda 2014] proposed to classify the type of laparoscopic surgery based on videos which could be useful for organizing large video databases automatically. [Twinanda 2014] evaluated SVM on different visual features - color, scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG) - and a combination of different visual features with multiple kernel learning. Recently, [Kannan 2020] approached the problem by proposing a CNN for learning visual features followed by an LSTM that captures temporal information. Furthermore, the LSTM was trained in a multi-task manner to predict future visual representations along with the surgery type to aid the model in early recognition of the type of surgery.

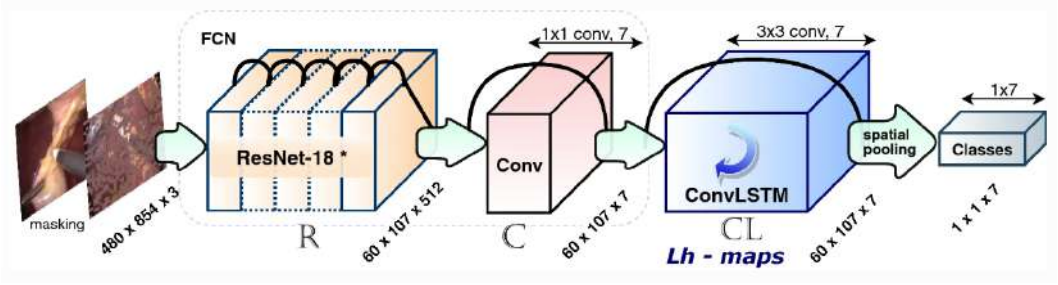
Surgical action triplets. Surgical action triplets capture surgical activities as a triplet consisting of the used instrument, the performed action, and the organ acted upon. The surgical action triplet was formulated by [Katić 2014, Katić 2015] as it provides a deeper understanding of the image contents in videos. Recognizing activities in this detail could be crucial in automating safety warnings in CAI [Vercauteren 2020]. For this reason, [Katić 2014, Katić 2015] leverage the triplet formulation manually provided as input signals for recognizing surgical phases. However, since then, the research in modeling surgical activities as triplets or using triplets for surgical workflow analysis has been hindered due to the difficulty in generating a large annotated dataset [Twinanda 2017a]. The success of phase recognition (Section 2.1.1) renewed interest in surgical action triplets in the SDS community [Nwoye 2020,Nwoye 2022b,Nwoye 2022a,Sharma 2022,Nwoye 2023]. [Nwoye 2020] aimed to recognize fine-grained activities as action triplets *⟨instrument, verb, target⟩* and introduced a new laparoscopic dataset, CholecT40, consisting of 40 videos from the public dataset Cholec80. All the frames in the dataset have been annotated using 128 triplet classes. Examples of the triplet annotations can be seen in Figure

2.3. This was the first work that proposed to recognize triplets directly from the videos. The recognition model presented in [Nwoye 2020] consisted of a multitask learning (MTL) network with three branches for the instrument, verb, and target recognition. Additionally, a class activation guide (CAG) module that uses the weak localization information from the instrument activation maps to guide the recognition of the verbs and targets was introduced. In the following work, [Nwoye 2022b] presented a new model called Rendezvous (RDV) that leverages transformer attention mechanism at two different levels for triplet recognition from surgical videos. A spatial attention module called Class Activation Guided Attention Mechanism (CAGAM) captured individual action triplet components in a scene. While a semantic attention module called Multi-Head of Mixed Attention (MHMA) solved the association problem between instruments, verbs, and targets. Recently, Rendezvous in Time (RiT) was introduced, which modified RDV to incorporate temporal cues in the surgical videos [Sharma 2022]. In particular, RiT focused more on the verbs to learn temporal attention-based features for enhanced triplet recognition. Research on surgical action triplets is gaining traction thanks to the works of [Nwoye 2020, Nwoye 2022b]. Recently, two endoscopic vision challenges have been organized at MICCAI 2021 [Nwoye 2022a] and MICCAI 2022 [Nwoye 2023] for the recognition of surgical action triplets in laparoscopic videos.

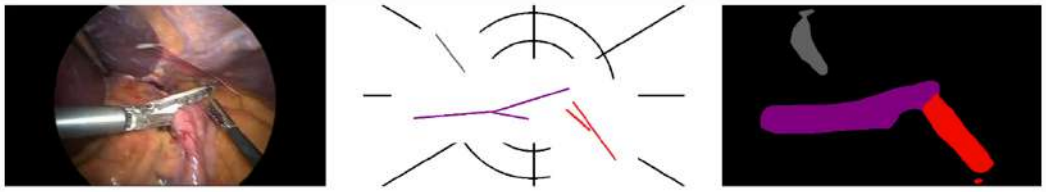
Gestures. With the rise of robotic systems such as da Vinci surgical robot, gesture recognition has seen a large body of research in the last decade [van Amsterdam 2021]. This onset can be attributed to JIGSAWS [Gao 2014], the first open-source dataset for surgical gesture recognition. Similar to phase recognition (Section 2.1.1), initial works proposed graphical models for recognizing gestures from kinematic data. These methods proposed the use of linear discriminant analysis [Lin 2006, Reiley 2008, Varadarajan 2009a], HMMs [Reiley 2008, Varadarajan 2009a, Tao 2012, Sefati 2015], Naive Bayes [Lin 2006], conditional random fields [Tao 2013, Lea 2015, Lea 2016c, Lea 2016a, Rupprecht 2016, Mavroudi 2018], and linear dynamical systems [Varadarajan 2011a, Varadarajan 2011b]. With the rise of deep learning, many works have recognized gestures from raw video data captured through endoscopic cameras. 3D CNNs [Funke 2019], RNNs [DiPietro 2016, DiPietro 2019, Gurcan 2019, van Amsterdam 2020], TCNs [Menegozzo 2019, Wang 2020, Zhang 2020, van Amsterdam 2022], and attention mechanism [van Amsterdam 2022] have all been explored in the last couple of years. Gestures recognition has also been modeled as a sequential decision-making process that can be learned with Reinforcement Learning [Liu 2018b, Gao 2020].

2.2 Semi- and Weakly-supervised Learning

Semi- and weakly-supervised learning are other ML paradigms that aim at learning from limited labeled data to alleviate the issue of vast amounts of labeled data needed for training. Semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data during training. While weakly-supervised learning utilizes



(a) Instrument segmentation and tracking.



(b) Surgical scene segmentation.

(c) Image classification at a finer level.

Figure 2.4: A few weakly-supervised learning methods in the literature. (a) The architecture of instrument segmentation and tracking method trained using only weak instrument presence labels [Nwoye 2019]; (b) EasyLabels: a segmentation method that uses weak stripe annotations to perform full surgical scene segmentation [Fuentes-Hurtado 2019]; (c) Coarse-to-fine few-shot learning problem tackled by [Bukchin 2020] where the training classes (e.g. animals) are of much coarser granularity than the target classes (e.g. breeds).

noisy, limited, or imprecise labels that are inexpensive to annotate large amounts of training data for a supervised learning setting. This section reviews the literature on these two paradigms of ML.

2.2.1 Learning in other domains

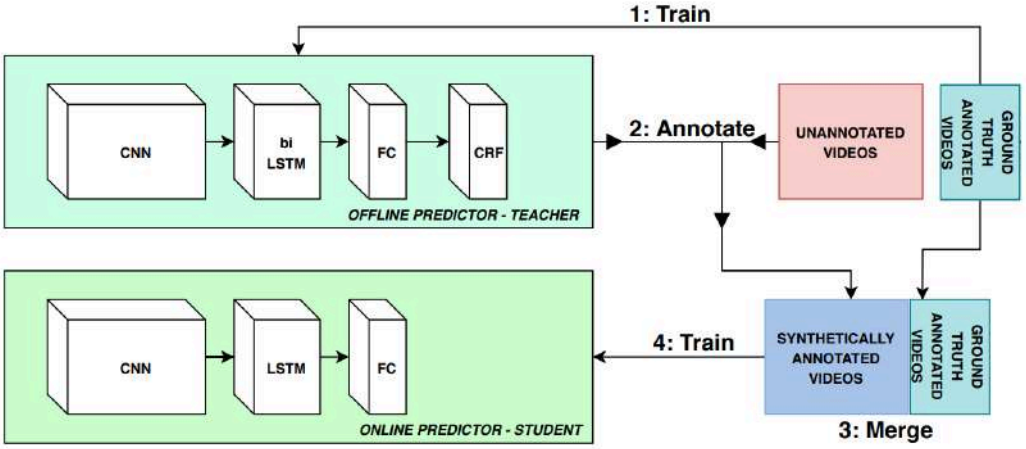
Weak supervision has seen a great interest in the medical computer vision community to tackle the need for high-volume annotated datasets that are difficult to generate.

Some of the interesting applications of weak supervision are seen in the detection of the region of interest in chest X-rays and mammograms [Hwang 2016], cancerous tissue segmentation [Jia 2017], gesture recognition [van Amsterdam 2019], surgical tool localization [Vardazaryan 2018, Nwoye 2019], and surgical scene segmentation [Fuentes-Hurtado 2019]. [Vardazaryan 2018, Nwoye 2019, Fuentes-Hurtado 2019] are particularly interesting as these approaches are developed on surgical datasets, especially laparoscopic videos. Two examples of weakly supervised learning methods are illustrated in Figure 2.4a & 2.4b. With the aim of localizing surgical instruments in laparoscopic images, [Vardazaryan 2018] proposed utilizing instrument presence labels as weak supervision. The output of a fully convolutional network was passed through a global pooling operation that constrains the network activations to focus on the most salient features needed to localize surgical instruments. Similarly, [Nwoye 2019] proposed using weak instrument presence labels for spatial localization and tracking of surgical instruments. In their work, a convolutional LSTM (ConvLSTM) was employed to learn the spatio-temporal cues across the surgical video frames. The trained ConvLSTM was successful in spatially localizing the instruments and tracking them over time. However, these methods focused specifically on instruments. Surgical scene segmentation of laparoscopic images was tackled in [Fuentes-Hurtado 2019]. Easy labels, annotated as stripes over different objects in the images as shown in Figure 2.4b, combined with partial cross-entropy loss were utilized to obtain dense pixel-level segmentation.

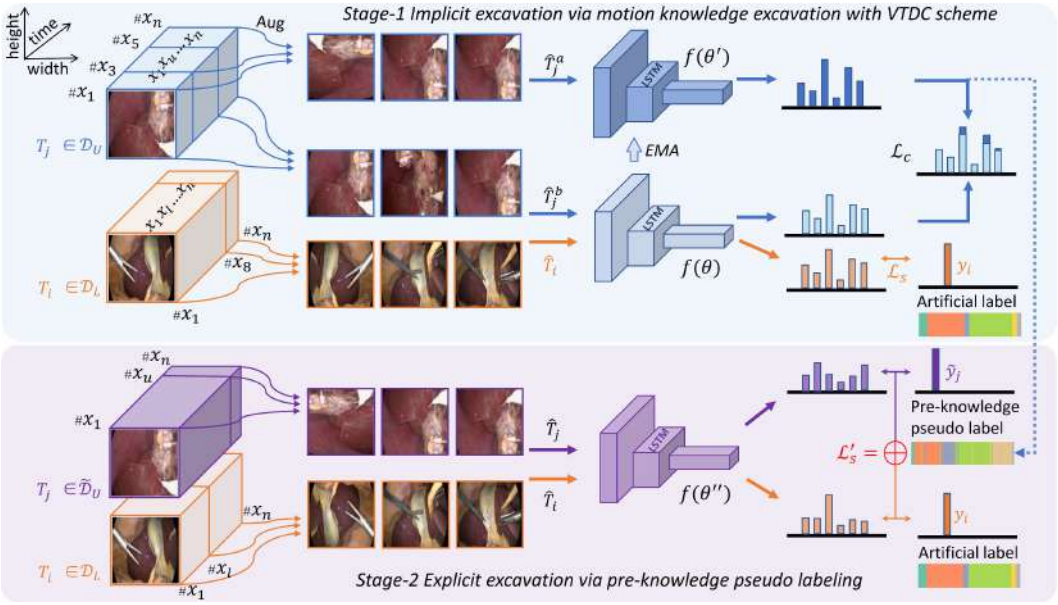
In the computer vision community, weakly supervised coarse-to-fine methods have received considerable interest for image classification [Taherkhani 2019, Bukchin 2020, Touvron 2021, Su 2021]. For example, coarse labels such as 'dog' or 'cat' is used to learn finer classification such as 'beagle', 'pug', 'ragdoll cat', 'persian cat', etc (Figure 2.4c). [Taherkhani 2019] proposed an image-based weakly supervised end-to-end model for object classification consisting of a CNN followed by two self-expressive layers. One self-expressive layer captures the global structures through coarse labels and the other captures the local structures for fine-grained classification. [Bukchin 2020] tackled the problem of Coarse-to-Fine Few-Shot (C2FS) learning and proposed a novel 'angular normalization' module that effectively combines supervised and self-supervised contrastive pre-training for C2FS. [Touvron 2021] tackled the problem of learning finer representations from coarser labels without any fine-grained labels. Their proposed method consists of CNN-based trunk and target networks that learn coarse representations from labels and finer representations with a self-supervised nearest-neighbor classifier. During training, the trunk gradients were used to update the target network weights as a moving average. Recently, [Su 2021] combined semi-supervised learning incorporating hierarchical coarse labels as weak signals to improve fine-grained image classification.

2.2.2 Surgical Workflow Analysis

To reduce the number of labeled videos, most of the recent research works in phase recognition have proposed approaches based on semi-supervised learning over weakly-supervised learning [Bodenstedt 2017, Funke 2018, Yengera 2018, Yu 2019, Shi 2021].



(a) Teacher-student approach for surgical phase recognition.



(b) SurgSSL for phase recognition from laparoscopic videos.

Figure 2.5: Semi-supervised learning for surgical workflow analysis. (a) The architecture of the semi-supervised learning method proposed in [Yu 2019] based on the teacher-student approach; (b) The architecture of the SurgSSL method proposed in [Shi 2021]. The method consists of two stages where stage I extracts knowledge from motion in the unlabeled data and generates pseudo labels for them while stage II retrains the model with pseudo labels previously generated and a small set of labeled data.

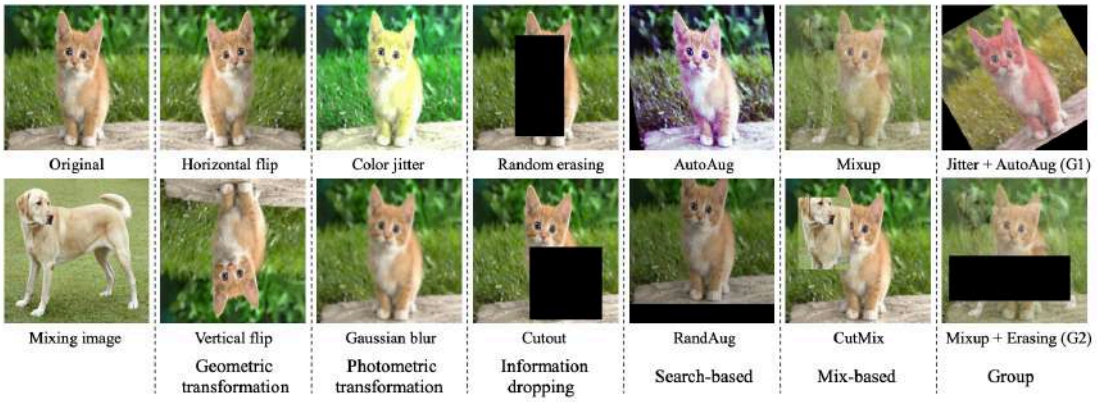
These approaches follow a similar strategy of pre-training the models on different proxy tasks of frame-sorting [Bodenstedt 2017], predicting the temporal distance between multiple frames [Funke 2018], and predicting the remaining surgery duration [Yengera 2018]. One of the works of particular interest is [Yu 2019], which proposed a teacher/student approach for phase recognition in scenarios of extreme manual annotation scarcity ($\leq 25\%$ of the training set). As can be seen in Figure 2.5a, the teacher model (CNN+biLSTM+CRF

trained on a small set) generated synthetic phase annotations for a large number of videos on which the student model (CNN+LSTM) was then trained. Another semi-supervised approach called SurgSSL [Shi 2021] presented a two-stage semi-supervised learning method for phase recognition that leveraged unlabeled data via motion knowledge excavation and pre-knowledge pseudo labeling (Figure 2.5b). In the first stage, a novel intra-sequence Visual and Temporal Dynamic Consistency (VTDC) scheme is proposed for mining motion knowledge. In the second stage, the pre-knowledge learned in the first stage is used to generate pseudo labels for unlabeled data and the model is re-trained on these pseudo labels along with the small labeled data. Recently semi-supervised learning across multiple centers in a federated learning setting was explored to recognize phases of LC [Kassem 2022]. Alongside learning task-specific knowledge from the labeled data, a contrastive loss was introduced for supervised learning with a temporal cycle consistency loss on the unlabeled data for learning temporal patterns found in the videos. The learning pipeline was followed at each center independently and the information was aggregated through a federated learning setup to preserve the privacy of data from each center.

2.3 Data augmentation

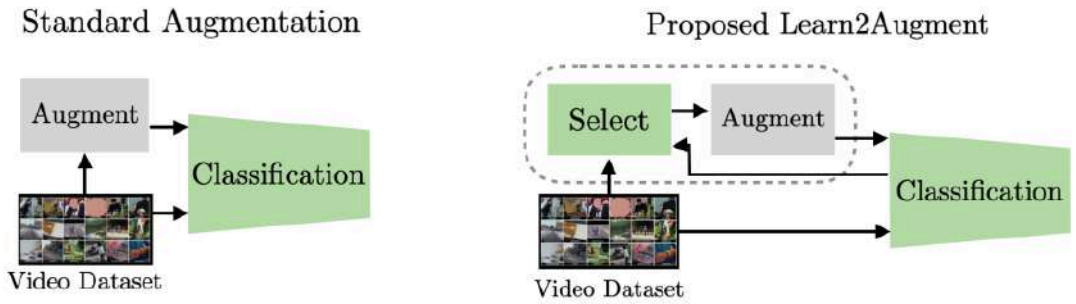
Selecting adequate hyperparameters is crucial to effectively train deep learning models. The different hyperparameters include optimizer, learning rate, number of epochs, batch size, data augmentation, and weight decay among others. Data augmentation is one hyperparameter that plays a critical role in model training by enabling ways to extend datasets with variations without requiring additional labeling processes. They have been extensively studied for improving the training of deep learning models for image classification [Ho 2019, Cubuk 2019, Lim 2019], object detection [He 2017, Kimata 2022], instance segmentation [He 2017, Fang 2019], etc. Data augmentations can be divided into five categories: geometric transformation, photometric transformation, information dropping, mix-based, and search-based [Han 2022]. The two most representative image augmentations for each category are illustrated in Figure 2.6a.

Geometric transformation introduces variants of an image attempting to mimic the effect of viewing a scene from different viewpoints. Popular geometric transformations include horizontal flip, vertical flip, horizontal translation, vertical translation, rotation, crop (zoom in), sheer (horizontal or vertical), and their combination. On the other hand, photometric transformation adds variations in the appearance of an image. Color jitter (achieved by changes in the brightness, contrast, saturation, and hue of an image), distortion, grayscale, gaussian blur, image invert, posterize, solarize, and equalize together constitute a common set of photometric transformations. Information dropping, through obscuring a portion of an image, as data augmentation has received considerable attention in the last couple of years [DeVries 2017b, Zhong 2020]. Cutout [DeVries 2017b] proposed to simply apply a square fixed-size zero-mask to a random location of each input image during each epoch of training. This simple augmentation was observed to



(a) Illustrations of image augmentations used for image classification. Image credits: [Han 2022]

(b) Inducing temporal variations via augmentations for video recognition. Image credits: [Kim 2020]



(c) Learn2augment: Learning-based augmentation method. Image credits: [Gowda 2022]

Figure 2.6: A few examples of data augmentations methods proposed in the literature.

be complementary to existing forms of data augmentation and regularization achieving state-of-the-art performance on the CIFAR10, CIFAR-100, and SVHN vision benchmarks. Random Erasing [Zhong 2020] extended Cutout to select a rectangle mask of random size filled with random values and applied it to a random location for each input image in an epoch. Additionally, object-aware random erasing that selected erasing

regions in the bounding box of each object was proposed to improve object detection networks. Mix-based augmentation strategy generally corrupts an image by blending it with another image [Zhang 2017, Yun 2019, Hendrycks 2020]. Mixup [Zhang 2017] was the first work that proposed to construct additional data generated by convex combinations of pairs of examples. CutMix [Yun 2019] cut a random rectangle from one image (inspired by Cutout [DeVries 2017b]) and pasted it on another image. Unlike previous mix-based methods, AugMix [Hendrycks 2020] generates different augmentations of the same image and blends them to produce a high diversity of augmented images. The different augmentations applied on the same image to produce multiple variants for blending are stochastically sampled. These various methods, however, have been manually designed with domain expertise.

To tackle the challenge of manually designing augmentation policies, the latest research has focused on automated search-based methods [Cubuk 2019,Lim 2019,Cubuk 2020]. AutoAugment [Cubuk 2019] used reinforcement learning on proxy tasks to select an optimal sequence of augmentations along with the magnitude and probability of applications. FastAutoAugment [Lim 2019] improved upon AutoAugment by optimizing the search strategy based on density matching. FastAutoAugment gaining speed up in search time by orders of magnitude while achieving comparable performances on image recognition tasks. But, these automated data augmentation methods introduce new difficulties, e.g., defining a proxy task and training on it or searching over 30 parameters. To address these difficulties, a simplified and more practical method, called RandAugment [Cubuk 2020], was recently proposed. RandAugment considerably simplified the search process to a grid search over two interpretable hyperparameters (M, N). Although the advances in automated data augmentation methods have been significant, these methods have been specifically developed for still images.

Recently, a few augmentation methods specifically designed for video have been proposed in the literature [Kim 2020, Gowda 2022, Kim 2022, Kimata 2022]. These methods have proposed inserting temporal perturbations successionally to the video frames [Kim 2020] or objects (obtained through instance segmentation) from one video onto another [Kimata 2022]. A learning-based method has been proposed in [Gowda 2022] that finds a pair of similar videos and then places objects from one video onto another video's background. In [Kim 2022], augmentation is applied to video frames ensuring smooth changes in its magnitude based on Fourier sampling. Examples of augmentations for videos are presented in Figure 2.6b & 2.6c. In the surgical vision community, the training pipeline of the video-based surgical activity recognition methods has used manually selected augmentations. Horizontal flip [Jin 2020,Ramesh 2021,Ramesh 2023b], rotations [Czempiel 2020,Ramesh 2021,Ramesh 2023b], random cropping [Jin 2020], translation [Czempiel 2020], scale [Czempiel 2020], and color jitter [Gao 2021, Ramesh 2021, Ramesh 2023b] are common augmentations used in the community.

2.4 Thesis setting

The primary objective of this thesis is to recognize surgical activities at multiple levels of granularity. We begin the thesis with the recognition of activities at two different granularity: phase and step. Despite large efforts in surgical activity recognition [Garrow 2020, Demir 2022, Charrière 2014, van Amsterdam 2021, Nwoye 2022b], three limitations that we can observe are: (1) methods recognize only one type of activity: phase, step, gesture, or action triplets, etc, (2) large labeled video datasets that represent other complex procedures like LRYGB is unexplored with most works analyzing LC or cataract surgery, (3) effective methods to exploit the temporal information present in the videos and tasks. Some effort has been made by [Charrière 2017, Czempiel 2020] to address some of the limitations. Recognition of multiple activities, phases and steps, of cataracts surgical procedure was undertaken by [Charrière 2017] while [Czempiel 2020] proposed TCN to exploit temporal information for the task of recognizing phases of LC workflow. To address all three limitations we propose to jointly recognize phases and steps of the complex workflow of LRYGB procedure. First, we introduce a novel dataset with multi-level surgical activity annotations, called Bypass40. Next, we introduce Multi-Task Multi-Stage Temporal Convolutional Networks (MTMS-TCN), extending MS-TCNs [Farha 2019], to jointly learn the tasks of phase and step recognition. We benchmark MTMS-TCN with other state-of-the-art deep learning models examined in Section 2.1.1 on the new Bypass40 dataset for surgical activity recognition, demonstrating the effectiveness of the joint modeling of phases and steps.

Besides multi-level activity recognition, we tackle fine-grained activity recognition under limited and imprecise labels. Most of the existing works in the SDS community have proposed semi-supervised learning to tackle limited labels (Section 2.2.2). Weakly-supervised learning is another direction of research that tackles the problem of limited and imprecise labels (Section 2.2.1). The previous works [Taherkhani 2019, Touvron 2021, Bukchin 2020] in the computer vision community propose weakly supervised learning methods exploiting hierarchical structures. However, the focus solely lies on object recognition in natural images containing a single object in each image. Additionally, similar approaches for surgical activity recognition are missing in the literature. Drawing inspirations from [Nwoye 2019, Fuentes-Hurtado 2019], in this thesis, we target weakly-supervised learning from videos instead of images. We aim to recognize the fine-grained activity, as opposed to an object, exploiting the temporal information available in videos. In particular, we target fine-grained surgical activity recognition on videos from endoscopic procedures.

The review in Section 2.3 shows that data augmentation is critical to improving the robustness and generalizability of deep learning models. However, adequate augmentation strategies are manually designed which requires domain expertise. Few approaches have attempted to automatically search for the optimal augmentation policy [Cubuk 2019, Lim 2019, Cubuk 2020, Gowda 2022]. For surgical activity recognition, existing methods completely rely on manually designed policies. Furthermore, these

specific augmentation policies have been applied at the image level to train backbone CNNs. On the other hand, no effort has been made to propose augmentation approaches for surgical videos. The temporal dimension in videos assumes particular importance in activity recognition and needs to be considered, and exploited, while designing augmentation policies for training spatio-temporal models. To this end, this thesis introduces a new simplified and automated data augmentation method, called *TRandAugment*, that aims to incorporate the essential temporal dimension. Inspired by work [Cubuk 2020], the *TRandAugment* method proposes a compact and simple parameterization consisting of only 3 parameters, where one parameter is dedicated to the temporal dimension. *TRandAugment* is extensively evaluated on the task of surgical activity recognition at two levels of granularity, i.e., phase and step.

Thesis Contribution Part II

3 Bypass40 Dataset

ಎರಡೂ ಕೈ ತಟ್ಟಿದರೆ ಸದ್ದು
(Pronunciation: Eradu kai tattidare saddu)
To clap you need both hands

(Kannada Proverb)

Chapter Summary

Deep learning models are function approximators that operate on historical or available observations. The number of observations required to estimate the underlying function hinges on the complexity of the task at hand which, subsequently, influences the depth of the deep learning models. For instance, state-of-the-art vision-based models consist of millions of parameters to estimate a complex function for the underlying challenging domain of computer vision. Consequently, these models require large labeled datasets to facilitate the model parameters to effectively capture valuable representations of the domain and achieve the best performance on the desired task. This effect can be witnessed in a sizable collection of literature frequently introducing new datasets. In the computer vision community, various large datasets have been introduced: ImageNet [Deng 2009], COCO [Lin 2014], Kinetics [Kay 2017, Li 2020, Carreira 2019, Smaira 2020], CityScapes [Cordts 2015], Berkley DeepDrive [Yu 2020], LSUN [Yu 2015],

and others. These various datasets have been curated to tackle different tasks such as image classification, face recognition, human pose estimation, semantic segmentation, instance segmentation, scene understanding, object detection and tracking, etc. Similar research directions have been pursued in the medical vision community with datasets for Radiology, Histopathology, Microscopy, Dermatology, etc [Li 2021].

Efforts to curate large datasets have been made in the surgical vision community. Various datasets have been released covering different tasks: Cholec80 [Twinanda 2017a], CATARACTS [Hajj 2019], & HeiSurF [Wagner 2021] for phase recognition; Cholec80, CATARACTS, HeiSurF, & EndoVis [Allan 2019, Allan 2020] for instrument presence detection; RMIT [Sznitman 2012], EndoVis, ROBUST-MIS2019 [Roß 2021], & CholecSeg8k [Hong 2020] for instrument segmentation; CholecT50 [Nwoye 2022b] for action triplet recognition; SARAS-MESAD [Bawa 2021] for action detection, etc. However, datasets to study the recognition of more than one type of surgical activity are missing in the surgical vision community. Furthermore, the public datasets cover procedures such as Laparoscopic Cholecystectomy, Cataract, Radical Prostatectomy, Rectal Resection, Laparoscopic Hysterectomy, or Proctocolectomy which represent a small set of high-volume surgeries. To this end, in this chapter, we present a new Bypass40 (BY40) dataset of high-volume Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) surgical procedures annotated with two types of surgical activities, phase and step.

3.1 Medical Lexicon

All the glossaries relevant to the LRYGB procedure are presented below (Figure 3.1):

Anastomosis. Anastomosis is the connection between two anatomical structures, usually between tubular structures such as blood vessels or loops of the intestine, that is surgically created.

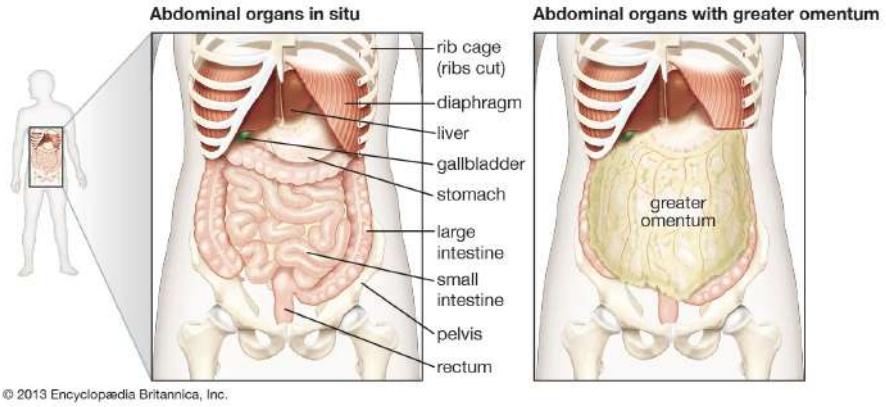
Alimentary limb. The alimentary limb (also jejunum) is the distal part of the small intestine which is connected to the gastric pouch during LRYGB procedure to create a path for transferring the incoming food to the colon (large intestine).

Biliary limb. The biliary limb (also biliopancreatic limb, duodenum) is the proximal part of the small intestine that carries digestive juices from the "remnant" stomach, bile, and pancreas to the alimentary limb after the LRYGB procedure.

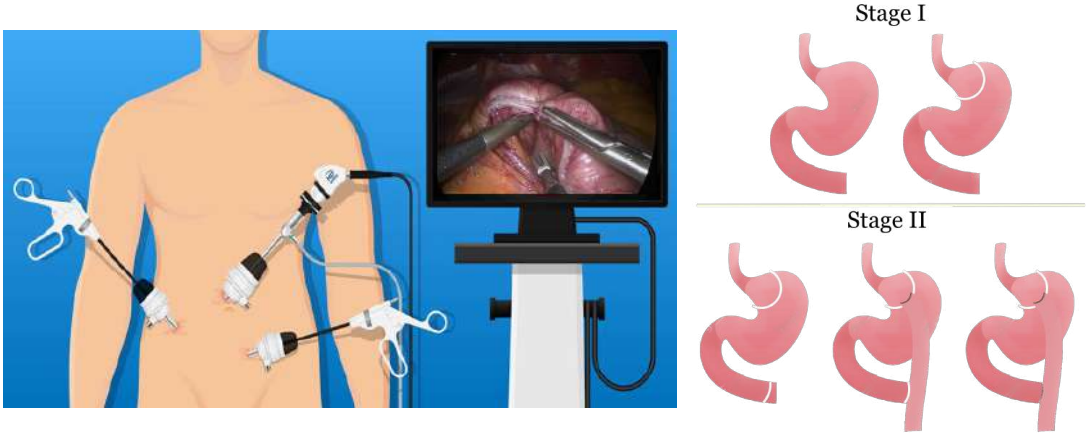
Gastrojejunal. The gastric pouch (gastro) and the alimentary limb (jejunum) collectively are referred to as the gastrojejunal.

Jejunojejunal. The distal part of the biliary limb (jejunum) and the alimentary limb (jejunum) collectively are referred to as jejunojejunal.

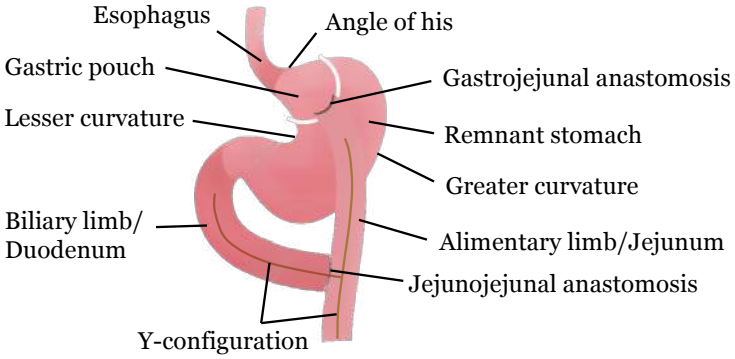
Lesser and greater curvature. The inner and outer borders of the stomach are called lesser and greater curvature, respectively (Figure 3.1c).



(a) Abdominal organs. Image credits: Encyclopaedia Britannica[1]



(b) Different steps of a LRYGB procedure.



(c) Anatomical structures of interest during a LRYGB procedure.

Figure 3.1: Illustrations of anatomy and surgical technique of Laparoscopic Roux-en-Y Gastric Bypass (LRYGB).

Mesentery. Mesentery, similar to the omentum, is a tissue layer of visceral fat that connects the intestine to the abdominal wall.

[1] https://www.britannica.com/science/abdominal-cavity#/media/1/852/68663

Mesenteric defect. Mesenteric defect refers to the space created in the mesentery due to the change in the anatomical structures during the surgery where the small intestine (alimentary limb) is raised to connect it to the gastric pouch.

Omentum. Omentum (also greater omentum) is a large smooth tissue of visceral fat that surrounds the abdominal organs. It pads and insulates the organs, helps to hold them in place, and reduces friction between the organs by secreting a lubricating fluid. As depicted in Figure 3.1a, the omentum has to be dissected to gain access to the small intestine.

Petersen space. Petersen space, similar to mesenteric defect, is the defect space created in the mesentery between the alimentary limb and the lower part of the transverse colon after the limb is connected to the gastric pouch.

Transverse mesocolon. Transverse mesocolon refers to a fold of fat tissue that surrounds the colon and connects it to the posterior abdominal wall.

Treitz angle. A sharp angle/bend in the small intestine between the duodenum and jejunum is called the Treitz angle.

Trocars. Trocars are medical devices used to make incisions into the abdominal cavity that are composed of an awl, a hollow tube, and a seal which essentially functions as a portal to introduce and manipulate surgical instruments inside the abdomen.

Y-configuration. When the biliary and alimentary limbs are connected during jejunojejunal anastomosis, together they form a Y shape which is referred to as the Y-configuration (Figure 3.1c).

3.2 Gastric Bypass for Obesity

Obesity is considered to be a global health epidemic by the World Health Organization [on Obesity 2000] due to its association with chronic diseases such as diabetes, cardiovascular diseases, and even some cancers in middle- and low-income countries. Around two billion people worldwide are facing obesity and these numbers are projected to continue to increase. By 2030, it is expected that a majority of the world's adult population will be either overweight or obese [Haththotuwa 2020]. According to [Ryan 2021], over three million people die each year due to obesity surpassing the number who die of being underweight. Furthermore, obese people are commonly more susceptible to infections and their complications; for example, a strong correlation is observed between obesity and increased risk of COVID-19 [Hepatology 2021]. This is a major world health concern that requires the rapid development of advanced approaches to efficiently manage obesity in the coming decades.

Bariatric surgery is a type of surgery that involves altering an individual's digestive system to treat obesity. The surgery is performed when diet and exercise haven't aided in weight loss or when a person faces serious health problems due to their weight. Bariatric

Figure 3.2: Sample images from Bypass40 dataset. Each column presents similar steps.

surgery is another high-volume procedure, like cholecystectomy, with approximately 500,000 surgeries performed laparoscopically every year worldwide [Angrisani 2015]. One of the most performed types of bariatric surgery, and also considered as the gold standard, is the Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) procedure. The LRYGB surgical technique requires $3-5$ skin incisions of $5-15$ mm to be made using trocars and involves two crucial stages. The first stage involves dividing the stomach into a small gastric pouch along with a much larger lower remnant stomach. In the second stage, the small intestine is divided into biliary and alimentary limbs that are then rearranged to form a Y-configuration (Figure 3.1c). The alimentary limb facilitates the passage of food from the gastric pouch while the biliary limb transports the stomach acids and enzymes from the remnant stomach to the colon. A graphical representation of the surgery can be seen in Figure 3.1b.

3.3 Dataset

With the aim to advance research on automatic workflow recognition by analyzing the complex LRYGB surgical procedures, we introduce a new large-scale dataset, called Bypass40 (BY40), in collaboration with Cristians Gonzalez, MD. The dataset, initially created as part of the CONDOR project, consists of 40 endoscopic videos of LRYGB surgeries performed by 7 expert surgeons at IHU Strasbourg, France. The recordings have been captured with patients' consent at 25 frames-per-second (fps) with a resolution of 854×480 or 1920×1080 and anonymized for privacy. All 40 videos are temporally annotated with two types of activities: phase and step. This section presents the definition of the activities of interest and statistics on the BY40 dataset.

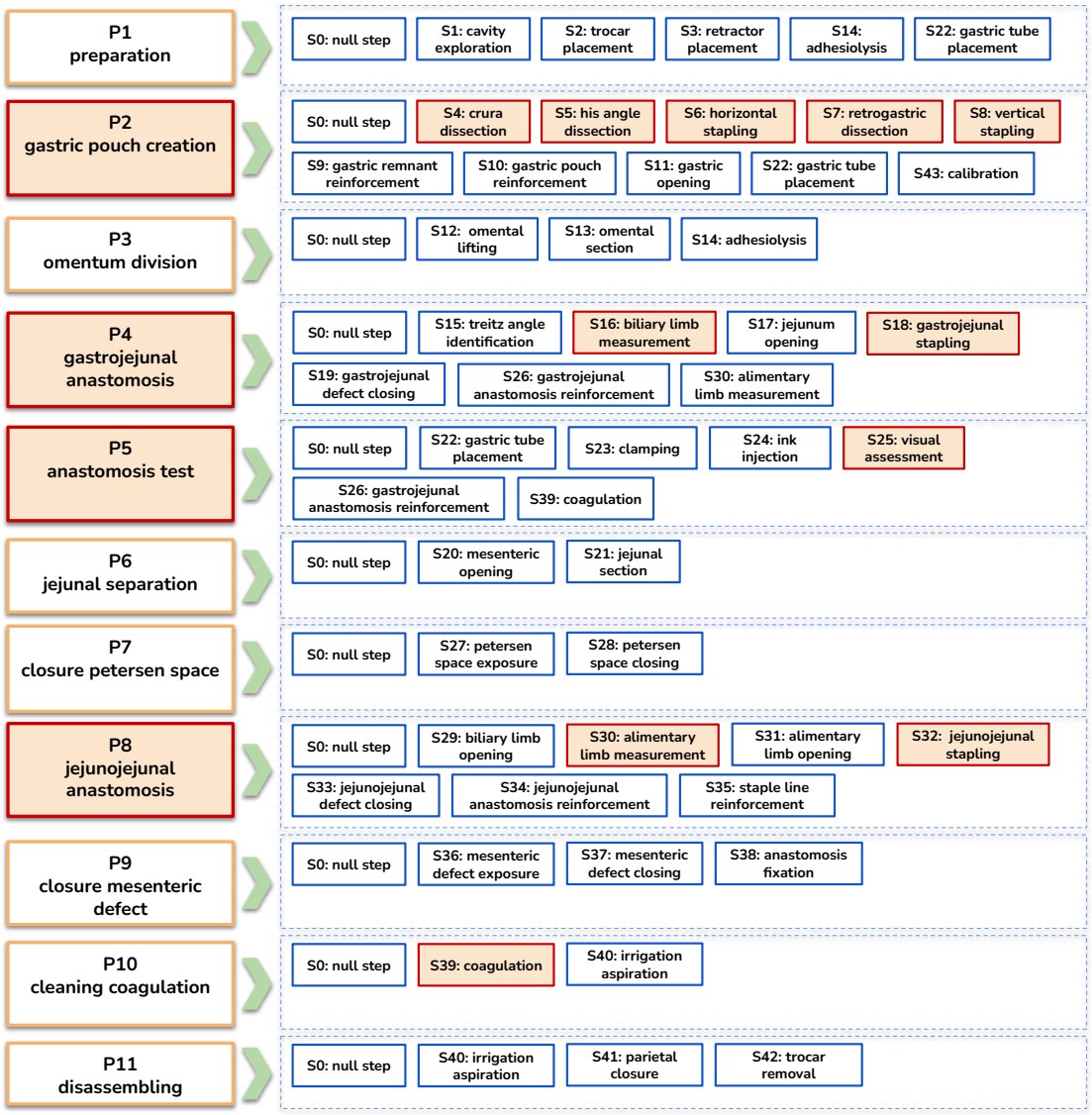


Figure 3.3: List of all the phases and steps defined in the dataset with their hierarchical relationship. The surgically critical activities are highlighted with a red box.

3.3.1 Phase and Step Definitions

As presented in Section 1.2.2, phase and step are two types of activities that represent a surgical workflow at two different levels of granularity with steps defined at a higher granularity than phases. The complex workflow of the LRYGB procedure followed at our partnering hospital is described using 11 phases and 44 steps. The first phase, plainly named 'preparation', captures the beginning of the procedure after the camera is inserted into the first trocar following the placements of other trocars. Similarly, the last phase, called 'disassembling', involves the removal of all the trocars and the closure of the ports. The second phase, called 'gastric pouch creation', is one of the most important parts of the procedure that accomplishes stage I showed in Figure 3.1b. The phases

Table 3.1: Definitions of all the proposed 11 phases for the gastric bypass procedure.

Phase ID	Phase Name	Description
P1	preparation	Phase of access to the abdominal cavity, installation of the ports for the camera and surgical instruments, and exposure of the operating field
P2	**gastric pouch creation**	Phase in which the small part of the stomach that is connected with the esophagus is separated from the rest to make a smaller gastric pouch
P3	omentum division	Vertical section of the omentum to facilitate the ascent of the small intestine to the gastric pouch
P4	**gastrojejunal anastomosis**	Connection of the distal small intestine with the gastric pouch
P5	**anastomosis test**	Verification that the gastrojejunostomy does not leak
P6	jejunal separation	Separation between the biliary and the alimentary limb
P7	closure petersen space	Closure of the space created between the mesentery and the mesocolon as the small intestine rises to make the bypass
P8	**jejunojejunal anastomosis**	Connection of the biliary limb with the alimentary limb
P9	closure mesenteric defect	Closure of the space created in the mesentery as the small intestine rises to make the bypass
P10	cleaning coagulation	Verification of the absence of bleeding, hemostasis, and aspiration of the remaining liquid in the abdominal cavity
P11	disassembling	Removal of surgical instruments and camera

from three to nine describe stage II of the procedure, i.e., division of the small intestine and rearranging anatomy to form a Y-configuration by connecting part of the small intestine to the newly created small stomach pouch. Regarding steps, the first step of the procedure, and also of the first phase, begins with the exploration of the abdominal cavity for evaluating the feasibility of the operation plan. Inherently, both these types of activity are hierarchically related with the possibility of several steps occurring in a given phase. For example, steps four to eleven describe all the tasks to be performed for successfully completing the 'gastric pouch creation' phase. A detailed list of all the phases is presented in Table 3.1 while steps are in Table 3.2. Additionally, a subset of 4 phases and 11 steps that are critical for a successful surgery is highlighted in the two tables. Sample images with respective phase and step labels are shown in Figure 3.2 while the relationship between them can be seen in Figure 3.3.

Table 3.2: Definitions of all the proposed 44 steps for the gastric bypass procedure.

Step ID	Step Name	Description
S0	null step	The camera is static and no actions are performed by the surgeon
S1	cavity exploration	The entire abdominal cavity is evaluated to verify the absence of alterations that could prevent or modify the planned surgery and to determine the technical feasibility of performing it
S2	trocar placement	The accessory work ports (usually four) are introduced into the abdominal cavity
S3	retractor placement	Introduction of the instrument to retract the liver and expose the esophagogastric junction
S4	**crura dissection**	The fatty tissue surrounding the esophagogastric junction is dissected to clearly expose the angle of his and separate the adhesions with the spleen
S5	**his angle dissection**	Opening of a retrogastric window at the level of the lesser curvature of the stomach to facilitate the passage of the stapling machine
S6	**horizontal stapling**	Horizontal section of the stomach at the level of the lesser curvature with the stapling machine
S7	**retrogastric dissection**	Dissection of the fatty and vascular tissue in the posterior part of the stomach
S8	**vertical stapling**	Vertical section of the stomach with the stapling machine
S9	gastric remnant reinforcement	Verification and reinforcement of the gastric remnant stapling with suture thread
S10	gastric pouch reinforcement	Verification and reinforcement of the gastric pouch stapling with suture thread
S11	gastric opening	Opening of the hole in the gastric pouch where the connection to the small intestine will be made
S12	omental lifting	Clamping and lifting of the omentum
S13	omental section	Full section of omentum to divide it into two parts
S14	adhesiolysis	Section of the connective tissue fibers between the structures
S15	treitz angle identification	Exposure of the transverse mesocolon to visualize the treitz angle
S16	**biliary limb measurement**	Measurement of the level at which the connection of the distal small intestine with the gastric reservoir will be made to perform the gastric bypass (around 70 cm)

Continued on next page

Table 3.2 – *Continued from previous page*

Step ID	Step Name	Description
S17	jejunum opening	Opening of the distal small intestine where it will be connected to the gastric reservoir to perform the gastric bypass
S18	**gastrojejunal stapling**	Connection of the gastric pouch to the distal small intestine (distal jejunum)
S19	gastrojejunal defect closing	Suture closure of the hole left by the stapling machine between the stomach and the jejunum
S20	mesenteric opening	Opening of the mesentery on the edge of the jejunum to facilitate the passage of the stapling machine
S21	jejunal section	Clamping and section of the jejunum proximal to the gastrojejunostomy
S22	gastric tube placement	Progression of the gastric tube from the stomach to the jejunum in order to calibrate the anastomosis and then verify that the connection does not leak
S23	clamping	Clamping of the jejunum distal to the gastrojejunostomy
S24	ink injection	Injection of the ink to detect any leakage
S25	**visual assessment**	Visual inspection of the gastrojejunostomy for any leakages
S26	gastrojejunal anastomosis reinforcement	Reinforcement and fixation of the connection between the stomach and the jejunum
S27	petersen space exposure	Traction of the mesocolon to expose the space created when the small intestine ascends towards the gastric pouch
S28	petersen space closing	Closing the petersen space with suture thread
S29	biliary limb opening	Opening of the hole in the proximal bowel where the connection between the biliary limb with the alimentary limb will be made
S30	**alimentary limb measurement**	Measurement of the level at which the connection of the biliary limb with the alimentary limb will be made (around 150 cm)
S31	alimentary limb opening	Opening of the hole in the distal bowel where the connection between the biliary limb with the alimentary limb will be made
S32	**jejunojejunal stapling**	Connection of the biliary limb to the alimentary limb

Continued on next page

Table 3.2 – *Continued from previous page*

Step ID	Step Name	Description
S33	jejunojejunal defect closing	Suture closure of the hole left by the stapling machine between the biliary and the alimentary limb
S34	jejunojejunal anastomosis reinforcement	Reinforcement and/or fixation of the jejunojejunal anastomosis with a suture thread
S35	staple line reinforcement	Reinforcement and/or fixation of the staple line with a suture thread
S36	mesenteric defect exposure	Traction of the anastomosis between the alimentary loop and the biliary loop and/or the mesentery to expose the space created when the small intestine ascends towards the gastric pouch
S37	mesenteric defect closing	Closing the space with suture thread
S38	anastomosis fixation	Reinforcement and/or fixation of the anastomosis with suture thread
S39	**coagulation**	Introduction of a cloth and/or hemostatic tool (bipolar grasper) and applying pressure to reduce bleeding
S40	irrigation aspiration	Suction of any remaining liquid in the abdominal cavity
S41	parietal closure	Closure of the abdominal port holes
S42	trocar removal	Removal of all the trocars (usually four) placed during the preparation phase under visual control
S43	calibration	Re-calibration and cleaning of the camera

3.3.2 Dataset statistics

LRYGB is a challenging surgical workflow due to its complexity which is depicted using 11 phases and 44 steps. The BY40 dataset poses further challenges due to the imbalance in the class distribution of both phases and steps, as can be seen in Figure 3.4. On average, phases can take 3 to 26 minutes with only 3 out of 11 phases of a duration of more than 10 minutes. A similar but stark trend is present in steps due to their higher granularity than phases. A step can last on average between 1 to 16 minutes with only 8 steps lasting more than 5 minutes.

Additionally, a class imbalance can be witnessed in the occurrences of the steps across the 40 surgeries. The class occurrences of phases and steps are presented in Figure 3.5. Out of the 44 steps, 6 steps occur in less than 10 surgeries, 8 steps below 20 surgeries, and 14 steps below 30 surgeries. These challenges need to be tackled by future works in the SDS community.

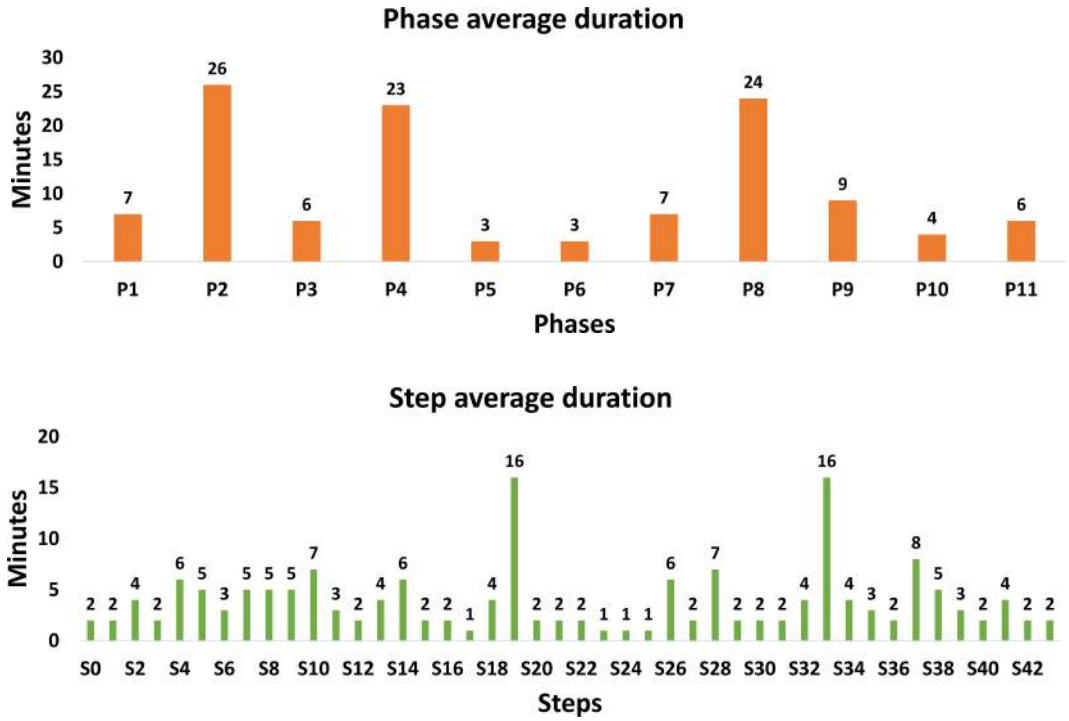


Figure 3.4: Average duration of phases and steps across videos in the dataset.

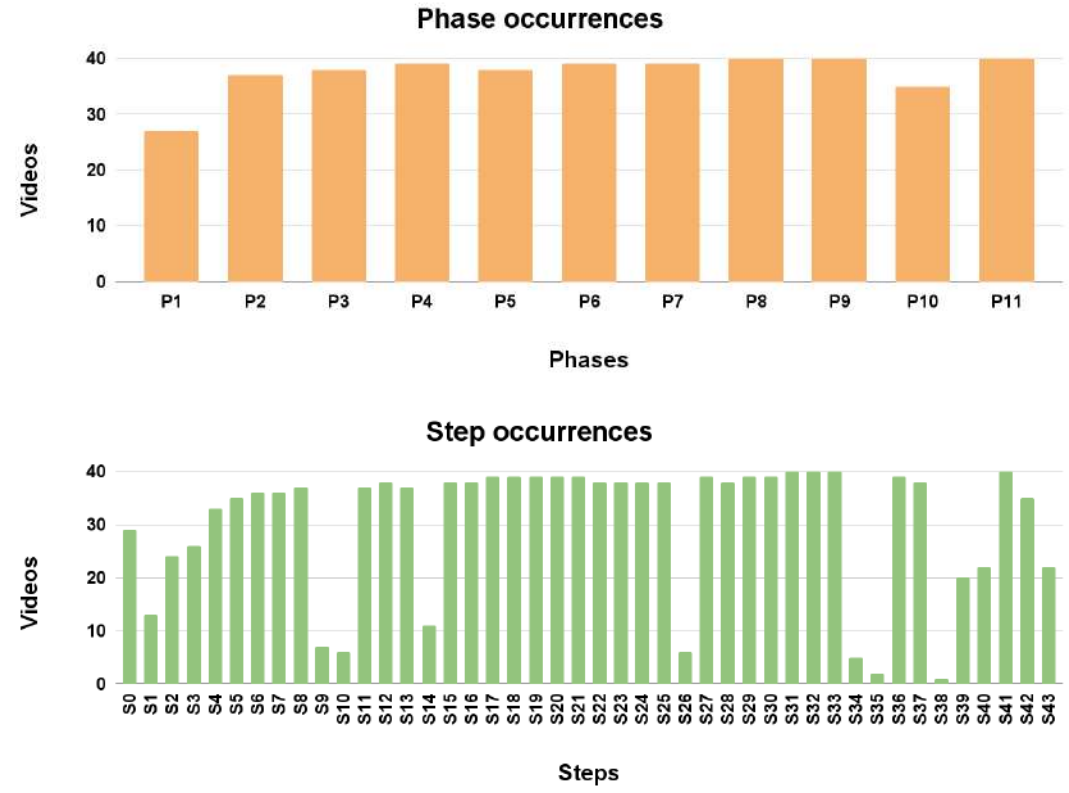


Figure 3.5: Total occurrences of phases and steps across videos in the dataset.

4 Multi-Task Multi-Stage Temporal Convolution Networks

Two heads are better than one, not because either is infallible, but because they are unlikely to go wrong in the same direction.

- C.S. Lewis

Chapter Summary

4.1 Research Objective

Surgical workflow can be described with activities defined at different levels of granularity, such as procedure, phase, step, action, gestures, etc. Recognizing these activities is an important research direction of the surgical vision community due to its potential in developing advanced support technologies in CAI. Yet, only a few works have attempted to recognize activities at multiple levels with most of the works focusing on only one level.

This study aims to design deep learning based recognition models that jointly recognize activities at different levels of granularity. To this end, in this chapter, we introduce MTMS-TCN, Multi-Task Multi-Stage Temporal Convolutional Networks to jointly recognize the phase and step of another complex LRYGB procedures [Ramesh 2021].

4.2 Methodology

With the aim of joint online recognition of phases and steps, we propose an online surgical activity recognition pipeline consisting of the following steps: 1) A multi-task ResNet-50 is employed as a visual feature extractor. 2) A multi-task multi-stage causal Temporal Convolutional Network (TCN) model refines the extracted feature of the current frame by encoding temporal information deduced by analyzing the history. We propose this two-step approach so that the temporal model training is independent of the backbone Convolutional Neural Network (CNN) feature extraction models. The overview of the model setup is depicted in Figure 4.1.

4.2.1 Feature Extraction Architecture

ResNet-50 [He 2016b] has been successfully employed in many works for phase segmentation [Yu 2019, Czempiel 2020, Jin 2018, Jin 2020]. In this work, we utilize the same architecture as our backbone visual feature extraction model. The model maps $224 \times 224 \times 3$ RGB images to a feature space of size $N_f = 2048$. The model is trained on frames extracted from the videos, without any temporal context, in a multi-task setup to predict both phases and steps as shown in Figure 1 (a). Since both activities are multi-class classification problems that exhibit an imbalance in the class distribution, softmax activations and class-weighted cross-entropy loss are utilized. The class weights for both activities are calculated using the median frequency balancing [Eigen 2015]. The total loss, $\mathcal{L}_{total} = \mathcal{L}_{phase} + \mathcal{L}_{step}$, is obtained by combining equally weighted contributions of class-weighted cross-entropy loss for phases ($\mathcal{L}_{phase}$) and steps ($\mathcal{L}_{step}$).

4.2.2 Temporal Modeling

For the joint temporal surgical activity recognition task, we propose MTMS-TCN, a multi-task extension of a Multi-Stage Temporal Convolutional Networks (MS-TCN). The model takes an input video consisting of $x_{1:t}$, $t \in [1, T]$ frames, where T is the total number of frames, and predicts $y_{1:t}$ where y_t is the class label for the current timestamp t. Following the design of MS-TCN, MTMS-TCN contains neither pooling layers nor fully connected layers and it is only constructed with temporal convolutional layers. Our temporal model consists of only temporal convolutional layers, in particular, they are dilated residual layers performing dilated convolutions. Since our aim is to segment surgical activities online, similar to TeCNO [Czempiel 2020], we perform causal convolutions [van den Oord 2016] at each layer which depends only on n past frames and does not rely on any future frames. The dilation factor is increased by a factor of 2 for each consecutive layer which increases exponentially the temporal receptive field of

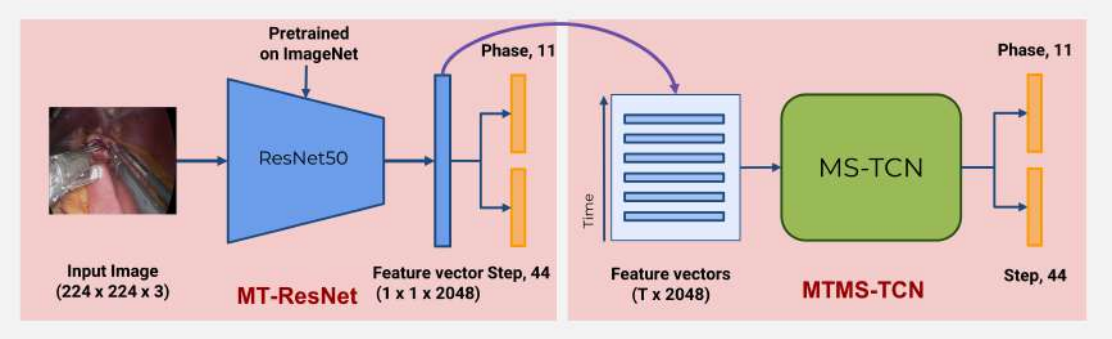


Figure 4.1: Overview of our model setup. The multi-task architecture of the ResNet-50 feature extractor backbone is on the left and the multi-task setup of the TCN temporal model is on the right.

the network without introducing any pooling layer. Additionally, the multi-stage model recursively refines the output of the previous stage.

Similar to our setup for CNN, we train our MTMS-TCN in a multi-task fashion to jointly predict the two activities by attaching two heads at the end of a stage. Softmax activations with cross-entropy loss for phase and step are applied and the total loss is similar to the loss utilized for training the backbone CNN (Eq. 4.2.1). Please note that the cross-entropy loss is not class-weighted. This is done to allow the temporal model to learn implicitly the duration and occurrence of each class in both phases and steps.

4.3 Experimental Setup

4.3.1 Dataset

We evaluate our method on the BY40 dataset described in Section 3. We split the 40 videos in the dataset into 4 subsets of 10 videos each to perform 4-fold cross-validation. Each subset was used as a test set, while the other subsets were combined together and divided into training and validation tests consisting of 24 and 6 videos respectively. The dataset was subsampled at 1 fps, amounting to approximately 149,000 frames for training, 41,000 for validation, and 66,000 for testing in each fold. The frames are resized to ResNet-50's input dimension of $224 \times 224 \times 3$ and the training dataset is augmented by applying horizontal flip, saturation, and rotation.

4.3.2 Model Training

The ResNet-50 model is initialized with weights pre-trained on ImageNet. The model is then trained for phase and step recognition in a single-task setup called ResNet and jointly in a multi-task setup called MT-ResNet, described in Section 4.2.1. In all the experiments, the model is trained for 30 epochs with a learning rate of 1e-5, weight regularization of 5e-5, and a batch size of 32. The test results presented are from the best performing model on the validation set. The baseline TCN model is trained in a single-task setup utilizing the features extracted from backbone ResNet (Figure 4.2).

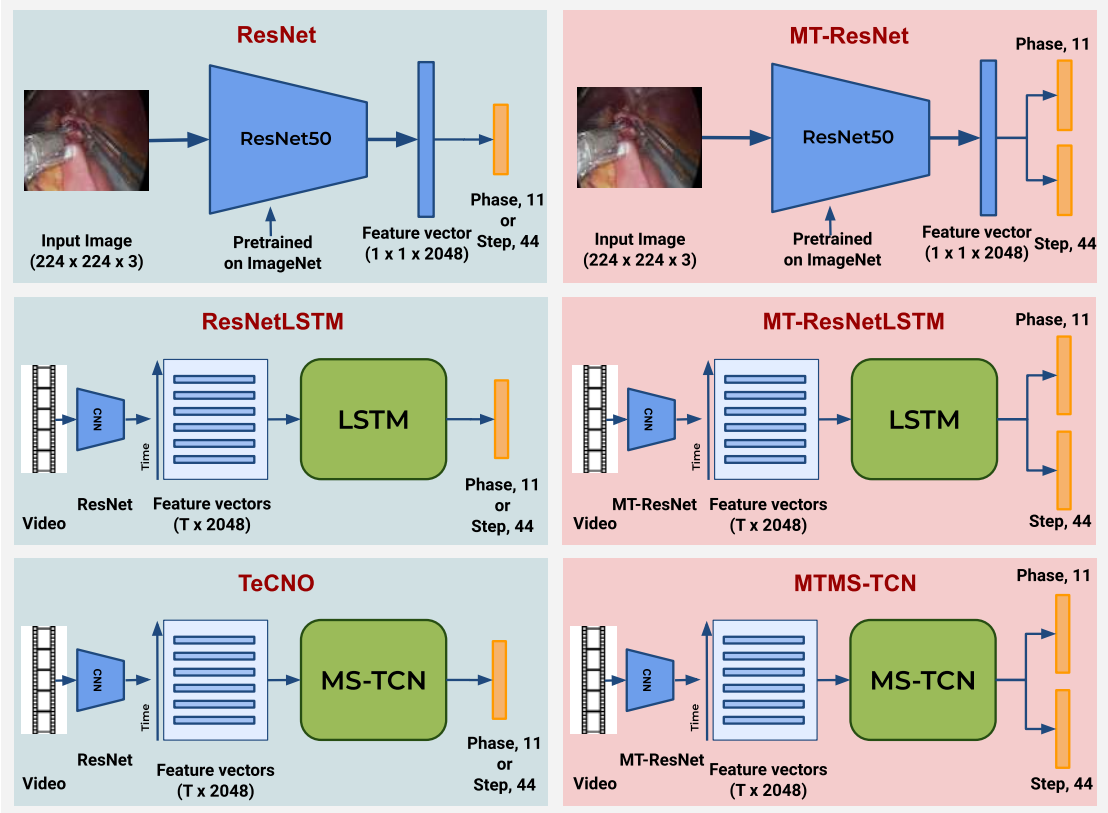


Figure 4.2: Overview of all the models used for evaluation. All the models trained in a single-task setup are shown on the left, while all the models trained in a multi-task setup are shown on the right.

This is effectively achieved by training TeCNO separately for the two activity recognition tasks. The MTMS-TCN model is trained in a multi-task setup utilizing the backbone MT-ResNet trained in a similar fashion. All models are trained with different TCN stages to identify the effect of the number of stages on long temporal associations. In all the experiments, the model is trained for 200 epochs with a learning rate of 3e-4. The feature representations of augmented data for CNN are also utilized for training the TCN model (Figure 4.2). Our CNN backbone was implemented in Tensorflow while the temporal models (TCN and LSTM) were implemented in PyTorch. Our models were trained on NVIDIA GeForce RTX 2080 Ti GPUs.

4.3.3 Evaluation Metrics

We follow the same evaluation metrics used in other related publications [Czempiel 2020, Jin 2018, Jin 2020], where Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) are used to effectively compare the results. Accuracy quantifies the total correct classification of activity in the whole video. PR, RE, and F1 are computed class-wise,

Table 4.1: Baseline comparison on the dataset for phase recognition. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported across all the 4-fold cross-validation.

	Models	Phase			
		ACC	PR	RE	F1
No TCN	ResNet	82.1 ± 3.3	73.9 ± 3.3	72.2 ± 3.4	72.5 ± 3.6
	MT-ResNet	81.7 ± 2.7	73.1 ± 2.8	72.1 ± 2.3	72.1 ± 2.6
	ResNetLSTM	**89.1 ± 2.8**	**82.1 ± 3.6**	**82.3 ± 3.5**	**81.7 ± 3.5**
	MT-ResNetLSTM	88.6 ± 2.7	81.4 ± 3.9	81.1 ± 3.5	80.7 ± 3.8
Stage I	TeCNO	89.8 ±3.5	85.4 ± 4.0	82.3 ± 4.5	83.0 ± 4.1
	MTMS-TCN	**91.2 ±2.9**	**86.1 ± 3.7**	**83.8 ± 4.0**	**84.4 ± 3.5**
Stage II	TeCNO	89.9 ± 3.3	84.4 ± 4.3	83.3 ± 3.9	83.5 ± 4.0
	MTMS-TCN	**90.9 ± 3.2**	**85.6 ± 4.5**	**84.0 ± 4.2**	**84.2 ± 4.2**

Table 4.2: Baseline comparison on the dataset for step recognition. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported across all the 4-fold cross-validation.

	Models	Step			
		ACC	PR	RE	F1
No TCN	ResNet	65.5 ± 2.0	45.3 ± 3.0	43.2 ± 2.7	42.6 ± 2.3
	MT-ResNet	66.6 ± 2.4	46.0 ± 3.1	44.7 ± 3.1	43.8 ± 2.9
	ResNetLSTM	71.3 ± 2.3	47.8 ± 4.1	47.7 ± 2.8	45.8 ± 2.7
	MT-ResNetLSTM	**72.2 ± 2.0**	**51.0 ± 3.3**	**49.3 ± 1.8**	**47.9 ± 2.1**
Stage I	TeCNO	75.1 ± 2.4	54.7 ± 2.6	50.9 ± 2.4	49.9 ± 1.8
	MTMS-TCN	**76.1 ±2.7**	**56.4 ± 3.6**	**52.5 ± 3.3**	**51.9 ± 2.9**
Stage II	TeCNO	74.8 ± 2.5	53.2 ± 2.5	50.8 ± 3.3	49.9 ± 3.7
	MTMS-TCN	**75.5 ± 3.1**	**54.9 ± 4.4**	**52.6 ± 4.2**	**51.8 ± 4.1**

defined as:

$$PR = \frac{|GT \cap P|}{|P|},\ RE = \frac{|GT \cap P|}{|GT|},\ F1 = \frac{2}{\frac{1}{PR} + \frac{1}{RE}},$$

where GT and P represent the ground truth and prediction for one class, respectively. These values are averaged across all the classes to obtain PR, RE, and F1 for the entire test set. We perform 4-fold cross-validation and report the results as mean and standard deviation across all the folds.

4.3.4 Baseline Comparison

The overview of all evaluated models is depicted in Figure 4.2. MTMS-TCN is evaluated against popular surgical phase recognition networks, ResNetLSTM [Jin 2018], and TeCNO [Czempiel 2020]. Both these networks are trained in a two-step process for the

Table 4.3: Baseline comparison on the dataset for joint phase and step recognition. Accuracy (ACC) is reported after 4-fold cross-validation

	Models	Phase ACC	Step ACC	Phase-Step ACC
No TCN	ResNet	82.1 ± 3.3	65.5 ± 2.0	54.9 ± 2.6
	MT-ResNet	81.7 ± 2.7	66.6 ± 2.4	64.8 ± 2.0
	ResNetLSTM	**89.1 ± 2.8**	71.3 ± 2.3	68.5 ± 2.3
	MT-ResNetLSTM	88.6 ± 2.7	**72.2 ± 2.0**	**70.7 ± 1.9**
Stage I	TeCNO	89.8 ± 3.5	75.1 ± 2.4	72.3 ± 3.0
	MTMS-TCN	**91.2 ± 2.9**	**76.1 ± 2.7**	**75.1 ± 2.8**
Stage II	TeCNO	89.9 ± 3.3	74.8 ± 2.5	71.9 ± 2.7
	MTMS-TCN	**90.9 ± 3.2**	**75.5 ± 3.1**	**75.1 ± 2.8**

single task of phase and step separately. Furthermore, ResNetLSTM is extended to get MT-ResNetLSTM where the ResNetLSTM model is trained in a multi-task setup. Since causal convolutions are used in the model for online recognition of activities, for fair comparison unidirectional LSTM is utilized. The LSTM, with 64 hidden units, is trained using the video features extracted from the CNN backbone with a sequence length equal to the length of the videos for 200 epochs with a learning rate of 3e-4.

4.4 Results and Discussions

Comparison of MTMS-TCN (Stage I) with other state-of-the-art methods, utilizing both LSTMs and TCNs, is presented in Table 4.1 and Table 4.2 on both phase and step recognition tasks. TeCNO which utilizes TCNs outperforms both ResNetLSTM and MT-ResNetLSTM models by 1% and 3% in terms of accuracy. MTMS-TCN outperforms TeCNO, ResNetLSTM and MT-ResNetLSTM models for by 2% the phase recognition.

Similarly, for step recognition, TeCNO outperforms both LSTM-based models by 3-4% with respect to Accuracy, and 3-6% in terms of precision. MTMS-TCN improves over TeCNO by 1% in accuracy and outperforms it by 2% and 1.5% in terms of precision and recall, respectively. In turn, MTMS-TCN outperforms LSTM-based models by 4-5% in terms of accuracy and 3-8% in terms of precision and recall.

Table 4.3 presents the performance of all the models on joint recognition of phase and step. We present joint phase-step prediction accuracy which is computed as the average number of instances where both the phase and step are correctly recognized by the model. All the multi-task models outperform their single-task counterparts. In particular, MTMS-TCN outperforms TeCNO by 3%. Moreover, the joint-recognition accuracy of MTMS-TCN is very close to its step recognition accuracy which indicated that the model has implicitly learned the hierarchical relationship and benefited from it.

The improvement achieved by both MTMS-TCN and TeCNO in both recognition tasks over LSTM-based models is attributed to the higher temporal resolution and large receptive field of the underlying TCN module. On the other hand, the improvement

Table 4.4: TeCNO vs MTMS-TCN: 4-fold cross-validation average precision, recall, and F1-score (%) reported for the critical steps.

Step	TeCNO			MTMS-TCN		
ID	PR	RE	F1	PR	RE	F1
S4	84.2±5.7	**90.0±3.8**	85.6±4.1	**86.4±10.8**	88.3±3.9	**86.1±6.6**
S5	**87.7±1.7**	**80.4±9.4**	**80.8±7.6**	**87.5±4.3**	77.4±6.7	79.2±6.8
S6	**77.4±7.8**	64.7±22.3	**63.0±16.3**	76.4±15.8	**66.9±22.5**	62.5±13.6
S7	**77.2±10.1**	**64.7±11.8**	**67.8±9.3**	72.1±8.0	64.0±10.7	66.4±9.8
S8	**78.0±8.3**	**77.1±10.5**	**72.8±4.0**	75.6±7.0	**77.1±9.8**	**72.7±3.4**
S16	76.4±7.1	**69.1±6.5**	**68.7±4.2**	**79.1±3.2**	67.7±4.0	**68.6±4.4**
S18	**92.4±2.3**	**83.1±5.3**	**86.6±2.3**	89.8±4.9	80.5±3.1	83.4±3.6
S25	**55.1±12.4**	39.4±18.6	40.6±16.1	47.6±6.6	**49.5±18.3**	**45.2±10.7**
S30	62.3±4.8	62.0±13.5	57.5±10.3	**65.3±6.7**	**71.2±5.2**	**64.8±5.6**
S32	**87.9±3.8**	85.4±4.4	**84.0±6.6**	85.1±5.4	**86.3±3.3**	**83.7±2.9**
S39	46.2±27.1	**47.8±25.4**	39.0±22.2	**49.6±33.9**	42.9±27.2	**40.6±25.5**

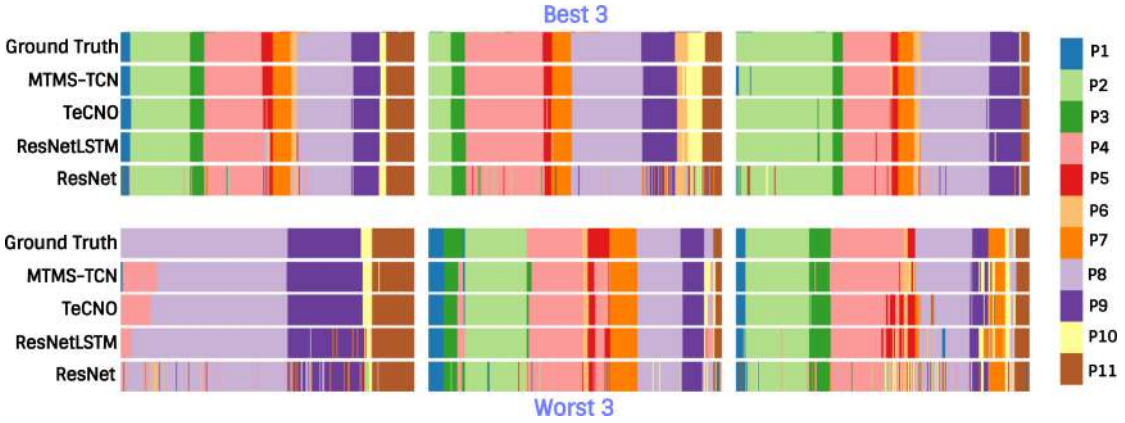


Figure 4.3: Phase recognition on complete videos in Bypass40 for quality assessment. The top row shows 3 videos in which our model performs best and the bottom row shows 3 videos with the worst performance.

of MTMS-TCN over TeCNO is attributed to the multi-task setup. Additionally, MT-ResNet, the backbone of our MTMS-TCN, achieves improved performance in steps with a small decrease in performance for phase recognition compared to ResNet, the backbone of TeCNO.

A set of surgically critical steps along with their average precision, recall, and F1-score are presented in Table 4.4. MTMS-TCN performs better than TeCNO in recognizing many of the steps. Moreover, even short-duration steps such as S25, S30, and S39 that are harder to recognize, are significantly better recognized by our MTMS-TCN over TeCNO. All these results validate our model trained in a multi-task setup for joint recognition of phases and steps.

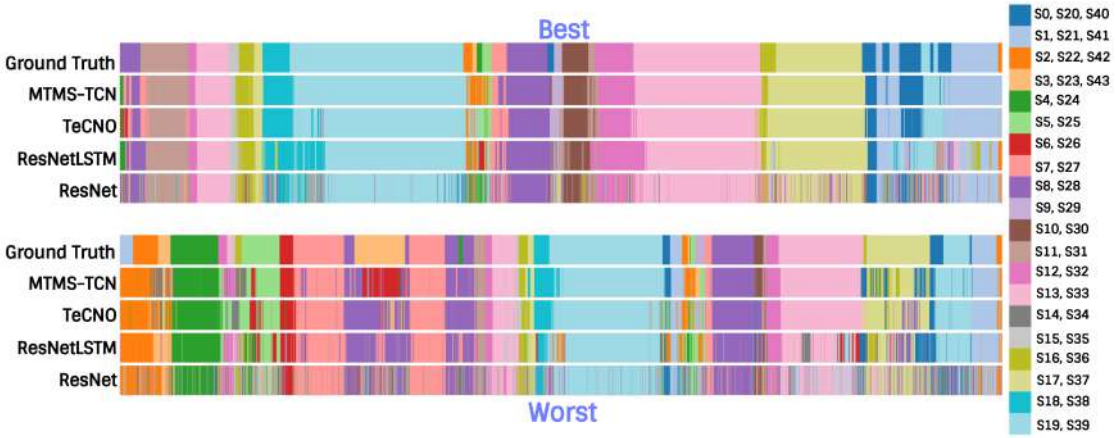


Figure 4.4: Step recognition on complete videos in Bypass40 for quality assessment. The figure shows the best (top) and worst (bottom) performance of our model. The 44 distinct steps are mapped to the same 20 categorical colormap.

Phase and Step Recognition Consistency Figure 4.3 visualizes a video set of 3 best and 3 worst performances of MTMS-TCN for phase recognition. The predictions of MTMS-TCN in some cases perform better than TeCNO in recognizing smaller phases, such as P5, P7, P9, and P10. MTMS-TCN is also able to recognize phase transitions better than TeCNO in some instances (e.g. P3, P4, P9). Additionally, both the methods outperform ResNet and ResNetLSTM models.

Figure 4.4 visualizes the complete video set of one best and one worst performance of MTMS-TCN for step recognition. Since there are 44 steps, visualizing all of them is quite challenging and clutters the plot. To effectively show the results, we look at one video instead of 3 in each best and worst category. Furthermore, for better visualization, we use a 20-categorical colormap and all 44 steps are mapped onto this colormap. The results clearly show that MTMS-TCN is able to better capture smaller steps and step transitions in comparison to TeCNO and ResNetLSTM.

4.5 Conclusion

In this study, we introduce new multi-level surgical activity annotations for the LRYGB procedures, namely phases and steps. We proposed MTMS-TCN, a multi-task multi-stage temporal convolutional network that was successfully deployed for joint online phase and step recognition. The model is evaluated on a new dataset and compared to state-of-the-art methods in both single-task and multi-task setups and demonstrates the benefits of modeling jointly the phases and steps for surgical workflow recognition.

5 Weakly Supervised Fine-grained Surgical Activity Recognition

ಹನಿ ಹನಿ ಕೂಡಿದರೆ ಹಳ್ಳ ತೆನೆ ತೆನೆ ಕೂಡಿದರೆ ಬಳ್ಳ
(Pronunciation: Hani hani koodidare halla, thene thene koodidare balla)
Every drop of water contributes to the formation of a pond/lake, similarly, every small work contributes to success

(Kannada Proverb)

Chapter Summary

Figure 5.1: Sample images from Bypass40 and CATARACTS datasets. Each column of Bypass40 images presents similar steps.

In this chapter, we draw attention to the importance of weak supervision of fine-grained activity, i.e., step recognition. We present, in Section 5.2, the weakly supervised learning methodology along with the end-to-end spatio-temporal model utilized [Ramesh 2023b]. In Section 5.3, we present the different experiments carried out on two datasets: Bypass40 (BY40) and CATARACTS (CA50). Finally, we discuss the significance of the experimental results in Section 5.4, highlighting the need to reduce the reliance on manual annotations.

5.1 Aim of the Study

Deep learning models require large labeled datasets to achieve top-class performance on relevant tasks. Annotating large datasets is difficult, time-consuming, and requires domain-specific medical knowledge. Moreover, the effort required to define and anno-

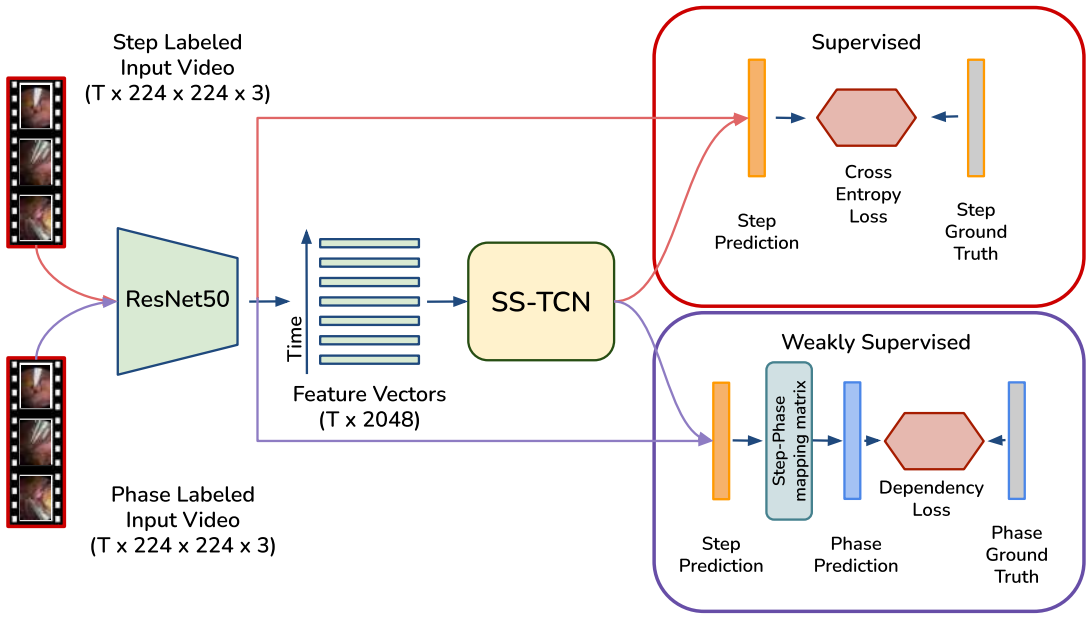


Figure 5.2: Overview of our end-to-end spatio-temporal model setup: ResNet50 + SS-TCN (Single-Stage Temporal Convolutional Networks). When step labels are available, the model is trained through the supervised pathway (red) and weakly supervised pathway (purple) utilizing phase labels. The model is trained end-to-end in a single learning stage.

tate a dataset with fine-grained activity such as steps is significantly higher than with phases. The challenges can be seen in the sample images presented in Figure 5.1. For instance, in the Bypass40 dataset, similar actions are performed across different steps belonging to different phases. Dissection is performed in at least 7 steps spread across 3 different phases. Similarly, Stapling is performed in 5 steps across 4 different phases. Designing and training a deep learning model to distinguish between these similar steps poses a great challenge. Even the state-of-the-art method, MTMS-TCN [Ramesh 2021], trained on a fully annotated dataset achieves an accuracy of ∼76% with a precision of ∼56%, accentuating the difficulty of the problem. The class imbalance further creates a challenge for training deep learning models that require large datasets with plenty of samples for each class. Hence, it is crucial to reduce this dependency of deep learning models on large labeled datasets. In this chapter, we aim to address this bottleneck for the task of step recognition by employing phase as weak supervision.

5.2 Methodology

The overview of our proposed method is presented in Figure 5.2. In this section, we first present our end-to-end spatio-temporal (ResNet-50 + SS-TCN) model for the task of fine-grained activity, i.e, step, recognition. Then we introduce the phase-step dependency loss for weak supervision of step recognition using phase annotation.

5.2.1 Spatio-temporal Model

Our weakly supervised step recognition network consists of a ResNet-50 model for visual feature extraction followed by an SS-TCN for modeling the recognition problem temporally. The complete model is trained in an end-to-end fashion. The overview of the model setup is depicted in Figure 5.2.

For phase segmentation, ResNet-50 [He 2016b] has been successfully employed as the backbone in many previous works [Yu 2019, Czempiel 2020, Jin 2018, Jin 2020]. In this work, we utilize the same architecture for visual feature extraction. We use a Single-Stage Temporal Convolutional Networks (SS-TCN), a single-stage variant of MS-TCN, to learn the spatial coherence across video frames. The choice of SS-TCN was motivated by the work of [Ramesh 2021] where MS-TCN did not provide a significant improvement over SS-TCN for both the step and phase recognition. Following the design of MS-TCN, the SS-TCN contains neither pooling layers nor fully connected layers and is constructed with only temporal convolutional layers, specifically dilated residual layers performing dilated convolutions. With the aim of online activity segmentation, we perform at each layer causal convolutions [van den Oord 2016, Czempiel 2020, Ramesh 2021] that depend only on the current frame and n previous frames.

The complete model takes an input video consisting of T frames $x_{1:T}$. The ResNet-50 maps $224 \times 224 \times 3$ RGB images to a feature space of size $N_f = 2048$. These frame-wise features are collected over time and are inputs to the TCN model that predicts $\hat{y}^s_{1:T}$ where $\hat{y}^s_t$ is the class label for the current timestamp t, $t \in [1, T]$. Since step recognition is a multi-class classification problem that exhibits an imbalance in the class distribution, softmax activation and class-weighted cross-entropy loss are utilized. Additionally, the dependency loss used when step labels are not available also relies on softmax activation and weighted cross-entropy loss, utilizing phase labels instead. The class weights for both steps and phases are calculated using the median frequency balancing [Eigen 2015] on the training set. The total loss is given by:

$$\mathcal{L}_{total} = \delta_{step} \cdot \mathcal{L}_{step} + (1 - \delta_{step}) \cdot \mathcal{L}_{dep},$$

where $\mathcal{L}_{step}$ represents weighted cross-entropy loss for steps, $\mathcal{L}_{dep}$ is the step-phase dependency loss (subsection 5.2.2), and δ_{step} is a binary variable that indicates if the video contains step labels.

5.2.2 Weak Supervision: Step-Phase dependency loss

Steps and phases are two types of activities describing the surgical workflow that are defined at different levels of granularity and possess an inherent hierarchical relationship [Katić 2015, Ramesh 2021]. Steps are defined at a higher level of detail compared to phases. This brings about lower inter-class variances between steps, compared to phases, making it a more complex task to clearly define and distinguish between them.

In the scenario presented in this study where the number of annotations is scarce, the recognition difficulties increase drastically. To overcome some of the challenges,

Table 5.1: Phases and steps for the cataract procedure.

Phases	Idle	Opening	Phacoemulsification	Implantation	Closure
Steps	Idle	Idle Toric Marking Implant Ejection Incision Viscodilatation Capsulorhexis Hydrodissection	Idle Nucleus Breaking Phacoemulsification Vitrectomy Irrigation/Aspiration	Idle Incision Viscodilatation Preparing Implant Manual Aspiration Implantation Positioning OVD Aspiration	Idle Suturing Sealing Control Wound Hydration

this chapter proposes a weakly supervised approach that utilizes labels of less granular activities, i.e., phases. Phase information alone could help the model in two ways. Firstly, phase information could help the model reduce errors related to recognizing similar-looking steps, e.g., 'S6: horizontal stapling' and 'S18: gastrojejunal stapling', belonging to two different phases. Secondly, we can gather a smaller subset of probable steps that could occur in a given phase eliminating the rest. For example, given the phase to be 'Phacoemulsification' of cataract surgery, only 5 out of 19 steps are likely to occur (Table 5.1). Similarly, a phase such as 'P5: anastomosis test' in the Bypass40 dataset, reduces the possible steps to 7 out of 44 (Figure 3.3). Here, the phase information provides cues to the model to learn to distinguish between steps belonging to the subset rather than the whole set. Thus we hypothesize that the additional available weak phase information could be very beneficial for step recognition in the low data regime.

We propose to represent the relationship as a step-phase mapping matrix $M_{s\to p}$, where the elements m_{ij} of the matrix are binary indicator variables which are 1 if step s_i occurs in phase p_j. The matrix encodes the weak information about which steps can occur in a particular phase and does not provide details of their occurrence, duration, and/or order. To enforce this weak link between steps and phases, the step predictions $\hat{y}_t^s$ of our spatio-temporal model (as described earlier) are linearly transformed by $M_{s\to p}$ into the phase space. Then a weighted cross-entropy loss ($\mathcal{L}_{CE}$) captures the similarity between the phase labels (y_t^p) and the transformed predictions ($M_{s\to p} \times \hat{y}_t^s$) of the model. The dependency loss ($\mathcal{L}_{dep}$) is given by:

$$\mathcal{L}_{dep} = \mathcal{L}_{CE}(y_t^p, M_{s\to p} \times \hat{y}_t^s).$$

5.3 Experimental Setup

In this section, we discuss the experimental setup of our method. We first present the datasets used for evaluation. Next, we discuss the experimental study followed by the training setup and evaluation metrics.

5.3.1 Datasets

For completeness, we briefly present the two datasets used in this study.

5.3.1.1 Bypass40

The BY40 dataset [Ramesh 2021] consists of 40 videos of LRYGB procedures with resolution 854×480 or 1920×1080 pixels recorded at 25 fps. Each frame is manually assigned to one of the 11 phases and one of the 44 steps. For example, steps such as *gastric opening, gastric tube placement, horizontal stapling*, and *vertical stapling* occur in *gastric pouch creation* phase. A detailed list of phases and steps along with their hierarchical relationship is presented in Chapter 3. We split the 40 videos into 24, 6, and 10 videos for training, validation, and test sets, respectively, and sub-sampled them at 1 fps. This amounts to 150k, 40k, and 65k images in each set. The images are resized to ResNet-50's input dimension of 224×224, and the training dataset is augmented by applying horizontal flip, saturation, and rotation. A few statistics of the two datasets are presented in Table 5.2.

5.3.1.2 CATARACTS

The CATARACTS dataset, proposed in [Hajj 2019], contains 50 videos of cataract surgery. With the recent CATARACTS2020 challenge, the dataset has been released with step annotations. Similar to [Charrière 2017], we define a phase ontology for available step labels. Cataract surgery consists of 5 phases and 19 steps that are summarized in Table 5.1. The dataset is extended with phase labels that are automatically generated using the available step annotations and the ontology presented in Table 5.1. For each frame in a video, the phase label is obtained by a simple lookup of the step label in Table 5.1. The only constraint while generating phase labels is when there are steps that can occur in several phases. In this case, the phase of the immediately preceding frame is assigned to the current frame. Since the only steps that occur in more than one phase are Idle, Incision, and Viscodilatation, and they do not occur at the beginning or at the end of a phase, it is therefore always possible to identify the correct phase by checking the phase of the previous step. Since very few steps occur in multiple phases, the automatically generated phase labels by table lookup are accurate and do not require expert knowledge or verification from a clinical expert.

We split the 50 videos (following the challenge[1]) into 25, 5, and 20 videos for training, validation, and test sets, respectively. Each set consists of 66k, 3.5k, and 11.8k frames extracted at 1 fps from the videos. The frames are resized from 1920×1080 to 224×224, and the training set is augmented with horizontal flip, saturation, and rotation.

[1] https://www.synapse.org/#!Synapse:syn21680292/wiki/601563

Table 5.2: Statistics of the two datasets considered in this chapter.

Dataset	Bypass40	CATARACTS
# Phases	11	5
# Steps	44	19
# Train videos	24	25
# Test videos	10	20
# Validation videos	6	5
# Train images	150K	66K
# Test images	65K	11.8K
# Validation images	40K	3.5K

5.3.2 Study

To demonstrate the effectiveness of our approach, we train and evaluate different configurations of the model. Given n videos, of which k are annotated with steps and the rest $(n-k)$ are weakly annotated with phases, the spatio-temporal model is trained in the proposed weakly supervised setting utilizing the dependency loss, presented as 'DEP'. To analyze the efficacy of 'DEP', we compare it against the spatio-temporal model trained only on k videos in a fully-supervised approach for the task of step recognition, which we refer to as 'FSA'. Additionally, we add a state-of-the-art semi-supervised learning method proposed by [Yu 2019] to our results. [Yu 2019], proposed a teacher/student semi-supervised learning method where both the teacher and student models consisted of spatial and temporal components, CNN-biLSTM-CRF and CNN-LSTM respectively. As noted in Section 2, [Yu 2019] is a closely related work in the literature to the work presented in this paper. Hence, we have implemented and adapted the method of [Yu 2019] for the task of step recognition. We repeat all the experiments for different values of $k \in \{3, 6, 12, 18\}$.

Furthermore, to analyze the influence of the number of additional videos with phase labels on the model performance, we conduct experiments where we fix k videos with step annotations and vary the number of videos with phase annotations from 0 to $n-k$ (i.e., 3, 6, 12, etc.).

5.3.3 Training

The ResNet-50 model is initialized with weights pre-trained on ImageNet. The complete ResNet-50 + SS-TCN model is then trained end-to-end for the task of step recognition. Since SS-TCN models the temporal information in an online setup, features from all the past frames in the video must be cached. To achieve this, a feature buffer is maintained to store features from the spatial model of the past frames. The feature buffer is reset at the end of the video. In all the experiments, the model is trained for 50 epochs with a learning rate of 1e-5, weight regularization of 5e-4, and a batch size of 64. The test results presented are from the best performing model on the validation set. The models were implemented in PyTorch and trained on NVIDIA RTX 2080 Ti.

Table 5.3: Bypass40: Effect of weak supervision on varying amount of step labeled videos. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported. 'FSA' denotes the model trained for step recognition without any phase annotations. 'DEP' denotes the dependency loss added for weak supervision using phase labels on the remaining videos.

Model	# Videos Step	Phase	ACC	PR	RE	F1
FSA	3 (12%)	-	45.02 ± 9.96	26.62 ± 5.32	21.87 ± 4.70	19.44 ± 5.31
[Yu 2019]	3 (12%)	-	43.27 ± 11.8	23.63 ± 4.41	23.91 ± 5.71	19.77 ± 4.89
DEP	3 (12%)	21	$\mathbf{57.20 \pm 8.31}$	$\mathbf{33.44 \pm 6.04}$	$\mathbf{33.16 \pm 6.37}$	$\mathbf{29.38 \pm 6.11}$
FSA	6 (25%)	-	59.80 ± 10.17	37.19 ± 8.52	35.93 ± 7.31	32.15 ± 8.03
[Yu 2019]	6 (25%)	-	62.55 ± 10.09	40.63 ± 7.85	43.71 ± 8.35	37.68 ± 8.54
DEP	6 (25%)	18	$\mathbf{68.03 \pm 9.04}$	$\mathbf{50.05 \pm 6.82}$	$\mathbf{45.86 \pm 6.46}$	$\mathbf{42.05 \pm 7.44}$
FSA	12 (50%)	-	68.26 ± 8.31	47.57 ± 7.84	44.74 ± 7.59	41.30 ± 8.44
[Yu 2019]	12 (50%)	-	67.89 ± 11.04	46.26 ± 9.97	50.11 ± 8.20	43.41 ± 10.33
DEP	12 (50%)	12	$\mathbf{73.43 \pm 8.43}$	$\mathbf{53.40 \pm 7.43}$	$\mathbf{51.19 \pm 8.20}$	$\mathbf{48.34 \pm 8.85}$
FSA	18 (75%)	-	72.82 ± 6.76	50.60 ± 7.90	48.98 ± 8.33	46.08 ± 8.61
[Yu 2019]	18 (75%)	-	73.33 ± 10.15	$\mathbf{54.78 \pm 11.05}$	$\mathbf{57.21 \pm 8.51}$	$\mathbf{51.72 \pm 10.59}$
DEP	18 (75%)	6	$\mathbf{73.88 \pm 8.11}$	54.33 ± 6.38	51.79 ± 7.10	48.62 ± 7.49
FSA	24 (100%)	-	76.12 ± 7.39	54.23 ± 8.24	50.94 ± 7.53	48.17 ± 8.02

5.3.4 Evaluation Metrics

To effectively analyze our models, we observe the Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) metrics used in related publications [Czempiel 2020, Jin 2018, Jin 2020]. Accuracy quantifies the total correct classification of activity in the whole video. PR, RE, and F1 are computed class-wise, defined as:

$$PR = \frac{|GT \cap P|}{|P|}, \; RE = \frac{|GT \cap P|}{|GT|}, \; F1 = \frac{2}{\frac{1}{PR} + \frac{1}{RE}},$$

where GT and P represent the ground truth and prediction for one class, respectively. These values are averaged across all the classes to obtain PR, RE, and F1 for each video in the test set. All four metrics, computed per video, are averaged across all the videos in the test set. Furthermore, where applicable, standard deviations are also computed across all the videos in the test set.

5.4 Results

5.4.1 Bypass40

5.4.1.1 Effect of weak supervision

To quantitatively evaluate our method, the results of step recognition on the test set are presented in Table 5.3. The table contains the results of our model with a varying number of videos in the training set labeled with steps (3, 6, 12, and 18) along with the rest of the training set containing phase annotations. The introduction of dependency

Table 5.4: Bypass40: Effect of the number of phase annotated videos for step recognition using 'DEP' loss for weak supervision. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported for setups with 6, 12, and 24 videos fully annotated with steps.

	# Videos					
Model	Step	Phase	ACC	PR	RE	F1
FSA	6	-	59.80	37.19	35.93	32.15
DEP	6	3	62.15	40.48	37.15	33.48
DEP	6	6	67.94	46.17	42.61	39.67
DEP	6	12	68.07	47.18	43.18	40.42
DEP	6	18	68.03	50.05	45.86	42.05
FSA	12	-	68.26	47.57	44.74	41.30
DEP	12	3	72.79	50.10	48.39	45.06
DEP	12	6	72.43	53.02	51.20	47.26
DEP	12	12	73.43	53.40	51.19	48.34
FSA	24	-	76.12	54.23	50.94	48.17

loss 'DEP' for weak supervision significantly improves the performance over the model (FSA) trained only on the step labeled subset of the dataset. We notice a 10-13% improvement of the model trained with 'DEP' loss containing only 3 videos annotated with steps. Similarly, we see a 10-13% and 5-7% increase in performance in all the metrics of the 'DEP' model in experiments corresponding to 6 and 12 step annotated videos, respectively. Interestingly, our 'DEP' model, trained on a dataset with 50% of step and 50% of phase annotated videos, achieves performance close to the upper baseline 'FSA' model trained on the whole fully labeled dataset.

Moreover, the results of [Yu 2019] semi-supervised method are also presented in Table 5.3 for different step annotated videos (3, 6, 12, and 18) used to train both teacher and student model. The student model's performance increases by 3-8% over 'FSA' in all the metrics for 6 videos with step annotations. Furthermore, an increase of 6% and 2% is noticed in recall and F1-score above 'FSA' with 12 step annotated videos. However, the method falls short of our proposed 'DEP' method. We notice a 10-15%, 2-6%, and 1-6% increase in performance in all the metrics of the 'DEP' model over [Yu 2019] with 3, 6, and 12 step annotated videos, respectively. Although both methods use 100% of the training videos for the task of step recognition, [Yu 2019] aim at exploiting the knowledge learned by an offline teacher model to generate pseudo labels for additional videos without step annotations while 'DEP' aims to use weak supervision through phase annotations. Hence, the method of [Yu 2019] is limited by the knowledge learned by the teacher model which uses only k step annotated videos although it learns from both current and future frames. On the other hand, the superior performance of the 'DEP' model indicates the additional cues present in phase annotated videos, although weak, are advantageous and that the proposed method effectively utilizes this information in the lower data settings.

Table 5.5: CATARACTS: Effect of weak supervision on varying amount of step labeled videos. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported. 'FSA' denotes the model trained for step recognition without any phase annotations. 'DEP' denotes the dependency loss added for weak supervision using phase labels on the remaining videos.

Model	# Videos Step	# Videos Phase	ACC	PR	RE	F1
FSA	3 (12%)	-	48.47 ± 10.62	51.32 ± 11.91	37.44 ± 9.85	37.12 ± 10.15
[Yu 2019]	3 (12%)	-	59.61 ± 10.67	56.02 ± 14.31	**61.82 ± 14.45**	53.26 ± 13.61
DEP	3 (12%)	22	**66.78 ± 12.21**	**64.29 ± 12.50**	59.73 ± 11.93	**58.31 ± 12.73**
FSA	6 (25%)	-	69.51 ± 11.16	71.05 ± 14.13	56.70 ± 12.67	59.28 ± 13.50
[Yu 2019]	6 (25%)	-	74.62 ± 8.22	67.71 ± 11.48	**75.93 ± 12.48**	67.67 ± 12.46
DEP	6 (25%)	19	**75.28 ± 11.50**	**71.84 ± 14.30**	69.19 ± 12.72	**68.09 ± 13.97**
FSA	12 (50%)	-	78.02 ± 9.05	79.02 ± 13.20	69.55 ± 12.04	71.18 ± 13.04
[Yu 2019]	12 (50%)	-	77.84 ± 12.55	71.48 ± 13.41	**79.92 ± 15.28**	72.96 ± 14.46
DEP	12 (50%)	13	**79.94 ± 9.17**	**80.52 ± 12.93**	72.62 ± 11.91	**73.52 ± 13.29**
FSA	18 (75%)	-	82.5 ± 8.07	**82.58 ± 11.91**	76.05 ± 11.62	77.39 ± 12.12
[Yu 2019]	18 (75%)	-	78.59 ± 10.71	74.55 ± 14.17	**78.16 ± 12.64**	73.55 ± 13.67
DEP	18 (75%)	7	**82.64 ± 9.72**	82.20 ± 13.70	77.32 ± 12.70	**77.67 ± 13.56**
FSA	25 (100%)	-	83.37 ± 9.50	85.29 ± 12.05	78.96 ± 11.93	80.09 ± 13.34

5.4.1.2 Effect of the amount of phase annotated videos

In Table 5.4, we present the results of our model with a varying number of phase annotated videos. Utilizing 6 videos containing step annotations, the addition of phase labeled videos as weak supervision improves all metrics: accuracy, F1-score, precision, and recall. With 6 videos annotated with phases, the model performance increases by 7-8% in all metrics over the baseline 'FSA' model. The addition of more videos does not affect the accuracy but further improves both precision and recall by 4%. This is due to our weakly-supervised method, which only provides supervision information if a step can occur in the given phase. This information helps to distinguish steps belonging to different phases, as opposed to steps belonging to the same phase. Therefore, the precision and recall of the model improve with more phase annotated videos, and no significant improvement in accuracy is seen. We see a similar trend when using 12 videos annotated with steps and increasing the number of videos annotated with phase labels. Thus, ultimately it is beneficial to train our method utilizing all additional videos in the dataset with phase annotations for weak supervision.

5.4.2 CATARACTS

5.4.2.1 Effect of weak supervision

We quantitatively evaluate our method and present the results of step recognition in Table 5.5. The table contains the results of our model, on a similar set of experiments as with *Bypass40*, by varying the number of videos in the training set labeled with steps (3, 6, 12, and 18) along with the rest of the training set containing phase annotations. We see

a similar trend as with bypass where the 'DEP' model outperforms 'FSA'. We notice a 13-22% improvement 'DEP' model considering only 3 step annotated videos. Furthermore, we see a 6-13% and 1-3% increase in performance in all the metrics of the 'DEP' model in experiments corresponding to 6 and 12 step annotated videos, respectively. We see that our method achieves a similar performance improvement on a relatively easier surgical workflow, such as cataracts, consistently surpassing the FSA in all labeled ratios. The semi-supervised method of [Yu 2019] achieves performance improvement of 16%, 8%, and 1.5% over 'FSA' in F1-score for experiments corresponding to 3, 6, and 12 videos, respectively. However, as seen earlier, it falls short of 'DEP' by 5%, 0.5%, and 0.5% in the F1-score for experiments corresponding to 3, 6, and 12 videos. Interestingly, [Yu 2019] achieve high recall on both datasets (Table 5.3 & 5.5). On CATARACTS, it even outperforms the 'DEP' model in recall in all the experiments but falls short significantly in precision. This could be credited to the student model which learns from imperfect pseudo labels generated by the teacher model. Since our proposed 'DEP' model learns from true phase labels on additional videos its performance increases in both precision and recall. This validates the applicability of our approach to different surgical workflows.

5.4.2.2 Effect of the amount of phase annotated videos

We present the results of our experiments, with a varying number of phase annotated videos, on CATARACTS in Table 5.6. We notice that utilizing 6 step annotated videos with additional phase labeled videos improves all the metrics by 6-13%. In particular, with 6 videos annotated with phases, we see a performance increase of 5% in accuracy and F1-score and 8% in recall of the 'DEP' model over the baseline 'FSA'. The addition of more videos provides a fractional improvement in accuracy but further improves both recall and F1-score by 1-4%. We see a similar trend when using 12 videos with step annotations reaffirming our hypothesis that it is beneficial to train our method utilizing all additional videos in the dataset with phase annotations for weak supervision.

5.4.3 Weak supervision on step predictions

To visualize the effectiveness of our method, we visualize the step predictions of our method on the CATARACTS dataset which contains fewer phases and steps thereby enabling us to render a simple and clearer graphical diagram. We compare the step predictions of our 'DEP' model against 'FSA' for 2 best and 2 worst videos in CATARACTS in Figure 5.3 for different labeled ratios (3, 6, and 12 videos with step annotations). Along with the step predictions we present the errors in the phase predictions for both models. The phase prediction error plot is computed as the errors in phase predictions derived from step predictions, using the step-phase mapping matrix, against ground truth phase predictions. Figure 5.3 clearly depicts the effectiveness of our method for different labeled ratios. By correcting for the phase labels through dependency loss, our 'DEP' model is able to correct for corresponding step labels without explicit supervision

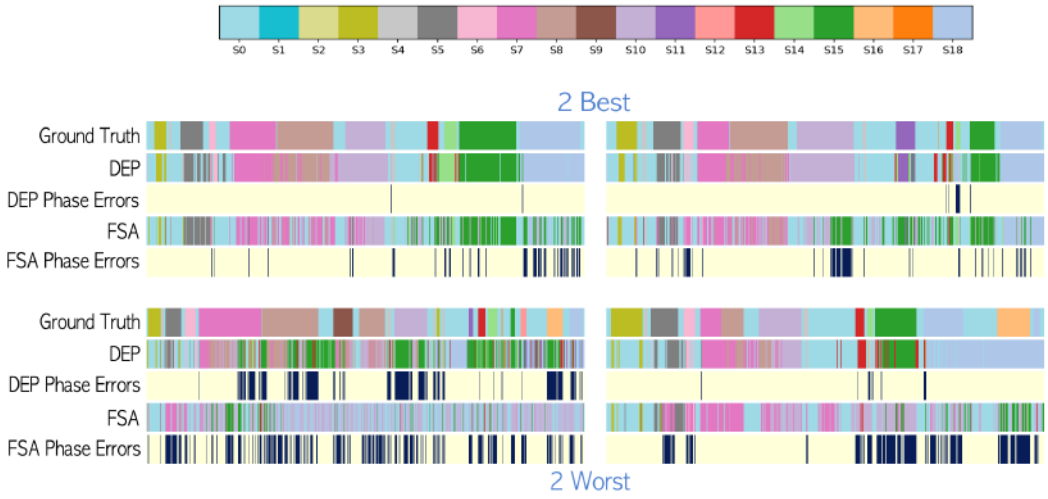
Table 5.6: CATARACTS: Effect of the number of phase annotated videos for step recognition using 'DEP' loss for weak supervision. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported for setups with 6, 12, and 25 videos fully annotated with steps.

	# Videos					
Model	Step	Phase	ACC	PR	RE	F1
FSA	6	-	69.51	71.05	56.70	59.28
DEP	6	3	71.34	67.84	62.27	62.01
DEP	6	6	74.30	71.70	64.18	64.96
DEP	6	12	73.57	70.88	65.68	66.03
DEP	6	19	75.28	71.84	69.19	68.09
FSA	12	-	78.02	79.02	69.55	71.18
DEP	12	3	77.60	78.26	68.60	69.87
DEP	12	6	80.11	81.60	72.46	73.98
DEP	12	13	79.94	80.52	72.62	73.52
FSA	25	-	83.37	85.29	78.96	80.09

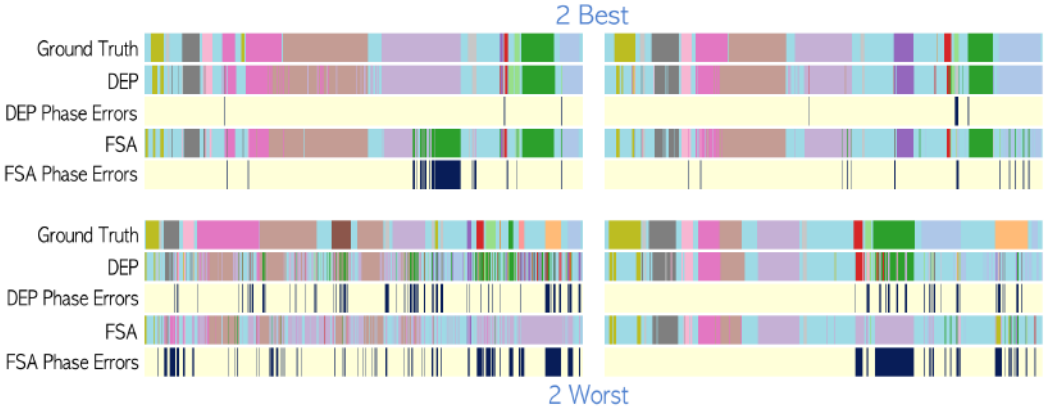
for step recognition (e.g. S10, S15, S18). The top row of Figure 5.3a shows this effect where we see a marked improvement in recognition of steps S18 (first video) and S10 (second video) by correcting for phase errors.

5.4.4 Limitations

In some cases, for example, S16 (Figure 5.3a, 5.3b, 5.3c), correcting for phase errors does not improve step recognition. The step is misrecognized with another step that occurs in the same phase. This is an expected outcome due to the intrinsic limitations of our weakly supervised method using coarser phase labels. Given the phase to be 'P2: gastric pouch creation' (Figure 3.1b), it is impossible for a model to differentiate between 'crura dissection' and 'his angle dissection' or between 'horizontal stapling' and 'vertical stapling'. As can be seen in Figure 3.2, the steps are quite similar in appearance and perform similar actions on the same anatomy (i.e., stomach or small intestine). This makes it challenging for a model to learn even when all the annotations are available. Furthermore, the phase information is too weak and does not provide any cues to better distinguish between the steps because both are valid steps in the current phase. Another limitation of our method is that adding more videos with phase annotations is not always beneficial. This limitation also stems from weak phase signals. If the fully supervised 'FSA' model learns to separate steps belonging to different phases, i.e., it has no or few phase-step correspondence errors, then additional videos with phase labels add no significant value as the model, during training, makes no/few errors in phase-step correspondence that helps improve feature learning. The significant errors by the model would be the inter-class separation of steps belonging to the same phase. Learning good representations to reduce these errors without supervision is a challenging task that needs to be tackled in future works.



(a) FSA vs DEP: 3 videos with step annotations.



(b) FSA vs DEP: 6 videos with step annotations.

(c) FSA vs DEP: 12 videos with step annotations.

Figure 5.3: Step predictions on two best and two worst videos on the CATARACTS dataset for different labeled ratios. For each video, we visualize the step prediction of ground truth, DEP model predictions, DEP model phase prediction errors, FSA model predictions, and phase prediction errors of the FSA model.

Meanwhile, the effect of utilizing more phase annotated videos as weak supervision for improving the model performance on step recognition is presented in Tables 5.6 & 5.6. As observed in Sections 5.4.1.2 & 5.4.2.2, it is beneficial to train the 'DEP' model utilizing all the additional phase annotated videos in the dataset for weak supervision. We also observe that in the lower data setting (6 videos with step annotations) model performance improves even when the phase annotated videos are increased from 12 to 18 (19 for cataracts). However, our study doesn't provide insights as to how many phase annotated videos are truly required to achieve the best performance by our proposed 'DEP' model. This is another limitation of our study, irrespective of the complexity of the procedure, that is hindered by the size of the available labeled datasets (24 in Bypass40 & 25 in CATARACTS). Understanding the extent of the 'DEP' model would require extending these datasets which is an important direction that needs to be pursued in future studies.

5.5 Conclusion

In this chapter, we introduce a weakly-supervised learning method for surgical step recognition utilizing less demanding phase annotations. To model the weak supervision between steps and phases, we introduce a step-phase dependency loss and train a ResNet-50 + SS-TCN model end-to-end. The proposed method is extensively evaluated on a BY40 dataset consisting of 40 LRYGB procedures and on the CATARACTS dataset containing 50 cataract surgeries. The proposed 'DEP' model significantly improves the step recognition metrics over the baseline 'FSA' model for all the amounts of step annotations available. We hope that this work will inspire and foster future research in weak supervision for surgical workflow analysis utilizing multi-level descriptions of the workflow.

6 TRandAugment: Temporal Random Augmentation Strategy for Surgical Activity Recognition

Mais pareille aux kaléidoscopes qui tournent de temps en temps, la société place successivement de façon différente des éléments qu'on avait crus immuables et compose une autre figure.

Like the kaleidoscopes that turn from time to time, society successively places in a different way elements that we had thought immutable and composes a new pattern.

- Marcel Proust

Chapter Summary

In the previous chapter, we presented a method for step recognition in a weakly supervised learning paradigm. An important component of the method was the spatio-temporal model. In this chapter, we investigate a key building block of similar spatio-temporal models, i.e, Data augmentation [Ramesh 2023a].

6.1 Objective of Research

While a large body of work has proposed various deep learning model architectures, a significant amount of research has examined the different components of these models. One such component is Data augmentation. As discussed in Section 1.3.3, data augmentation plays an important role in the optimal training of deep learning models and improves their robustness. Designing effective augmentation policies requires expertise. Besides, methods in the surgical vision community have utilized augmentation policies that are designed manually. Various different augmentation methods have been explored in the computer vision community with recent works proposing to learn augmentation policies on a proxy task. Yet, no effort has been made to examine new data augmentation methods for surgical videos. The temporal dimension in videos assumes particular importance in activity recognition as intraoperative surgical videos are of longer duration compared to videos examined in the computer vision community and they capture the complete surgical procedure composed of multiple complex activities. This temporality present in both surgical videos and activities needs to be considered and exploited while designing augmentation policies for training spatio-temporal models. To this end, this chapter introduces a new simplified and automated data augmentation method, called TRandAugment, that aims to incorporate the essential temporal dimension.

6.2 Methodology

Automated activity recognition methods aim to segment endoscopic videos into surgical activities, i.e., phase or step. To improve the generalizability of activity recognition methods based on deep learning, this section introduces the proposed augmentation method, called TRandAugment, and the spatio-temporal model used to evaluate the method.

6.2.1 TRandAugment

The goal of TRandAugment is to incorporate the temporal dimension present in surgical videos into the data augmentation methods for improving the generalization of activity recognition models. In pursuing this goal, we also want to propose a simplified and auto-

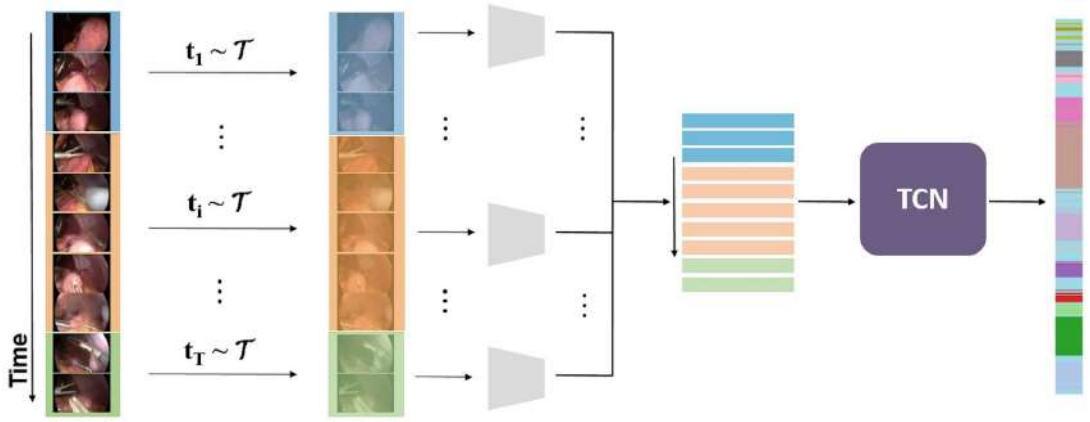


Figure 6.1: Pictographical representation of TRandAugment. A video is segmented into T clips and a random augmentation t_i, sampled from a list of transforms τ, is applied to clip i. The augmented clips are merged back to form a new video which is passed as input while training an end-to-end CNN+TCN network that predicts phases or steps.

mated data augmentation method. Given that a recent method [Cubuk 2020] operates only on a two-parameter space (M, N) compared to learned augmentation methods with over 30 parameters [Cubuk 2019, Lim 2019], TRandAugment is designed to require only 3 parameters, where the first two adopt the same parameterization used in [Cubuk 2020], while the third additional parameter T is used to characterize the temporal dimension. Similar to previous works [Cubuk 2019, Cubuk 2020], a set τ of 10 transformations is utilized and applied with uniform probability $\frac{1}{|\tau|}$:

- identity
- color
- brightness
- sharpness
- autoContrast
- rotate
- shear-x
- shear-y
- translate-x
- translate-y

The choice of $|\tau|= 10$ transformations is selected based on the domain knowledge of possible transformations that occur in endoscopic videos. Thus, we have excluded all the augmentations that, when applied, result in drastically different looking images that are highly unlikely to arise in surgical videos, such as posterize, solarize and equalize used in [Cubuk 2020] and other novel augmentations proposed in the literature: YOCO [Han 2022], CutOut [DeVries 2017b], MixUp [Zhong 2020], CutMix [Yun 2019] or AugMix [Hendrycks 2020].

As schematically represented in Figure 6.1, the idea of TRandAugment is to apply different transformations to different temporal video segments. Thus, parameter T is introduced to control the number of temporal segments. Each video is split into a

random $T' \in [1, T]$ segments and for each segment, i ($i \in [1, T']$), a random set of N transformation $\{t_{i,1}, ..., t_{i,N} |\ t_{i,j} \sim \tau\}$ is applied uniformly on all the frames of that segment. The strength of each transformation is represented by magnitude M and linearly scaled between its minimum and maximum values mapped to an arbitrarily chosen integer scale from 0 to 30.

To maintain a notation consistent with previous methods, in particular [Cubuk 2020], the proposed method is parameterized as (M, N, T), where M and N are defined as the magnitude and number of transformations to apply per segment, and T is the maximum number of temporal segments.

6.2.2 Spatio-temporal Model

The spatio-temporal model is comprised of a ResNet-50 backbone, for visual feature learning, followed by a Temporal Convolutional Network (TCN), for temporal modeling. The presented model is a powerful architecture comparable to other recent state-of-the-art methods [Czempiel 2020, Czempiel 2021, Ramesh 2021, Gao 2021]. Furthermore, it is modular and can easily accommodate new spatial and temporal models that could be proposed for activity recognition. This model is used in all the experiments and is trained end-to-end for the task of surgical activity recognition considering both phases and steps.

ResNet-50 [He 2016b] has been a popular model of choice in many recent works on phase/step recognition [Yu 2019, Jin 2020, Czempiel 2020, Ramesh 2021]. The model is also employed in this work for visual feature learning. For long temporal modeling, TCNs have been shown to outperform RNNs [Czempiel 2020, Ramesh 2021]. A single-stage model is employed over a multi-stage. This is motivated by the work of [Ramesh 2021] where the multi-stage did not show improvements over the single-stage for both phase and step recognition. SS-TCN consists of only temporal convolutional layers that perform causal convolutions which depend only on the current and n previous frames designed for online recognition.

The spatio-temporal model takes as input a video containing Υ frames $x_{1:\Upsilon}$. ResNet-50 extracts visual features of size $f = 2048$ from $224 \times 224 \times 3$ RGB images. The frame-wise features are stacked over time for the TCN model, which outputs predictions $\hat{y}_{1:\Upsilon}$, where $\hat{y}_i$ is the class label for the current timestamp i, $i \in [1, \Upsilon]$. Since both the tasks at hand (phase and step) are multi-class classification problems with imbalance in class distribution, following [Czempiel 2020, Ramesh 2021], class-weighted cross-entropy loss is used.

6.3 Experimental Setup

6.3.1 Datasets

For simplicity we briefly reintroduce the two datasets used in the this study.

6.3.1.1 Bypass40

The Bypass40 (BY40) dataset [Ramesh 2021], presented in Chapter 3, comprises 40 LRYGB procedures with an average video duration of 1 hour and 45 minutes. The complex workflow of LRYGB surgeries is represented with 11 phases and 44 steps and the dataset is fully annotated with both these types of activities defined at different levels of granularity. All the videos have a resolution of 854×480 or 1920×1080 pixels and are recorded at 25 fps. Following the same data split as Chapter 4, the dataset has been segregated into 24, 6, and 10 videos for training, validation, and test sets, respectively. The frames have been extracted at 1 fps and resized to ResNet-50's input size of 224×224.

6.3.1.2 CATARACTS

The CATARACTS (CA50) dataset[1] [Charrière 2017, Hajj 2019] consists of 50 videos of cataract surgical procedures. The dataset is annotated per frame with only steps as part of the CATARACTS2020 challenge. A complete list of all 19 steps is tabulated on the challenge website[2]. The 50 videos are split into 25, 5, and 20 subsets for training, validation, and test sets, respectively. Frames are extracted at 1 fps and resized from 1920×1080 to 224×224.

6.3.2 Training and Evaluation

6.3.2.1 Baselines

TRandAugment, or TRA, is compared against different augmentation methods as baselines. RandAugment [Cubuk 2020], referred to as RA, is the first comparison where the augmentations are applied independently for each image in a video. Next, RandAugment is extended to UniformRandAugment, called URA, where augmentation is applied uniformly on all the frames in a video. TRA is a more generalized method encapsulating both RA and URA, where setting $T = 1$ reduces TRA to URA while $T = \Upsilon$ (Υ: number of frames in a video) transforms TRA to RA. Finally, all the methods are compared against the state-of-the-art MTMS-TCN [Ramesh 2021] that used a manually designed 'Custom' set of augmentations (flip, saturation, rotation) for surgical activity recognition.

6.3.2.2 Training

In all the experiments, the ResNet-50 backbone model is initialized with ImageNet pretrained weights. Then the complete ResNet-50 + SS-TCN model is trained in an end-to-end fashion for the task of phase/step recognition. To train the TCN, which requires temporal information, features from all the past frames in the video are cached by utilizing a feature buffer. This feature buffer is reset at the end of the video. The

[1] https://ieee-dataport.org/open-access/cataracts
[2] https://www.synapse.org/#!Synapse:syn21680292/wiki/601563

spatio-temporal model is trained for 50 epochs with a learning rate of 1e-5 and a batch size of 64. The proposed method and model have been implemented in PyTorch and the experiments ($\sim$ 3500 GPU hours) were trained on NVIDIA RTX 6000 and V100 GPUs.

6.3.2.3 Evaluation

The effectiveness of the method is measured using accuracy (ACC), precision (PR), recall (RE), and F1-score (F1) metrics. The metrics are computed per video (averaged across classes) and are averaged across all the videos in the given set, following the same evaluation protocol as [Czempiel 2020, Czempiel 2021, Ramesh 2021, Shi 2021].

6.4 Results

In this section, we analyze the different components that influence the design of *TRandAugment*. Initially, we study the importance of temporally consistent augmentations in Section 6.4.1, then we analyze the impact of parameter M in Section 6.4.2, the number of transformations in Section 6.4.3 and impact of the parameter T in Section 6.4.4. Finally, we present the performance of the proposed method considering the optimal parameters on both datasets (Section 6.4.5) and discuss the limitations of TRandAugment in Section 6.4.6.

6.4.1 Do temporally consistent augmentations matter?

One of the key differences between videos and images is the additional temporal dimension. An obvious question is to study the importance of temporally consistent augmentations when training models on videos. To study the effect of temporal consistency, Table 6.1 compares the image-based augmentation method, RA, against the temporally consistent URA method on the CATARACTS dataset. The comparison is carried out at different settings ($M = \{15, 30\}$, $N = 1$, $\tau' \subset \tau : |\tau'| = \{3, 5, 9\}$). URA consistently performs better than RA in all the settings. Furthermore, the mean of RA, when averaged across $|\tau'|$ at both settings of $M = \{15, 30\}$, is $\sim$3-7% below the best-performing model compared to URA ($\sim$1%). This indicates the instability of RA due to its policy of independent frame-wise augmentation which breaks temporal visual consistency. Interestingly, the best RA model is obtained by utilizing a smaller set of augmentations $|\tau| = 3$, which indicates that the model can learn significantly better when there is less variance in image appearance temporally. All the observations confirm that temporally consistent augmentations are important when training spatio-temporal models.

6.4.2 Effect of magnitude (M)

To study the effect of augmentation magnitude, Table 6.2 compares model performance over various settings of $M = \{5, 10, 15, 20, 30\}$ for URA and TRA while keeping all other parameters fixed ($|\tau'| = 5$, $N = 1$, $T = 5$). Both URA and TRA show higher performance at higher magnitudes with the best results obtained at $M = 30$ on both

Table 6.1: The use of temporally consistent augmentations does matter: RA vs URA. All results are reported on the validation set on the CA50 dataset for step recognition.

M	$\|\tau'\|$	RA		URA	
		ACC	F1	ACC	F1
15	3	74.63	58.75	76.81	63.73
15	5	70.10	54.35	75.75	64.43
15	9	73.31	61.21	76.20	62.80
15	avg	72.68	58.10	76.25	63.65
30	3	**77.31**	**64.62**	78.05	66.88
30	5	69.66	54.48	78.45	66.99
30	9	70.70	53.87	**79.74**	**68.07**
30	avg	72.55	57.66	78.75	67.31

Table 6.2: Effect of magnitude M. All results are reported on the F1-score metric.

M	CA50 - step		BY40 - Phase		BY40 - Step	
	URA	TRA	URA	TRA	URA	TRA
5	64.23	60.59	85.06	85.02	54.55	53.78
10	63.75	63.40	82.72	84.59	54.39	54.62
15	64.43	63.67	84.83	85.64	56.64	56.38
20	61.61	62.22	84.54	82.70	57.39	56.06
30	**66.99**	**64.56**	**87.71**	**86.18**	**58.70**	**59.34**

tasks and datasets. Irrespective of the augmentation method used, higher magnitudes seem to have a direct effect on the performance of the model for different tasks and datasets. However, we notice that TRA performance is below URA at $M = 30$. This is not a valid comparison as the other parameters $|\tau'|$, N, and T are fixed and sub-optimal. Hence we perform these experiments to solely study the effect of magnitude on URA and TRA independently. The full comparison of TRA against other methods is discussed in Section 6.4.5.

6.4.3 Do all augmentations help?

To study the importance of using all the augmentations, Table 6.3 lists different experiments in terms of F1-score on the validation set, with $N = 1$ and $T = 5$, where subsets of transforms ($\tau' \subset \tau : |\tau'| = \{3, 5, 9\}$) are randomly sampled from τ. For the task of step recognition on both datasets, the best model performances are obtained when all transforms are utilized. On the other hand, the model performs best at an intermediate $|\tau'| = 5$ for recognizing phases for both settings of $M = \{15, 30\}$. However, at a higher magnitude ($M = 30$), the model performs equally well at $|\tau'| = 10$ compared to $|\tau'| = 5$ for phase recognition. In short, TRA benefits by utilizing all the transforms τ.

Table 6.3: Influence of the set of augmentations. All results report the F1-score metric.

$\|\tau'\|$	M	TRA		
		CA50 - Step	BY40 - Phase	BY40 - Step
3	15	65.92	83.21	56.36
5	15	63.67	**85.64**	56.38
9	15	**66.81**	82.99	**57.65**
3	30	62.93	83.27	59.85
5	30	64.56	**86.18**	59.34
9	30	**68.66**	**86.10**	**60.92**

Table 6.4: Impact of the number of temporal segments T with different augmentations on TRA. All results are reported on the F1-score metric on the validation set.

T	M	F1		
		CA50 - Step	BY40 - Phase	BY40 - Step
3	15	66.11	85.53	56.94
5	15	66.81	84.98	55.69
8	15	67.10	85.49	55.66
3	30	65.21	**86.16**	59.05
5	30	**68.66**	**86.22**	**60.47**
8	30	66.74	85.92	59.13

6.4.4 Impact of parameter T on TRA

The key component of the proposed TRA method is the parameter T that captures the variance in the appearance of the frames across a video. TRA is inspected with different settings of parameter $T = \{1, 3, 5, 8\}$ at two different magnitudes $M = \{15, 30\}$ while fixing $N = 1$ and $|\tau'|= 10$. The results in Table 6.4 show that at $T = 5, M = 30$ the model achieves the best performance on all the different tasks and across the two datasets. This indicates that augmenting at the clip level benefits the training of activity recognition models and the proposed TRA parameterization (M, N, T) allows us to easily find optimal parameters.

6.4.5 TRandAugment

Table 6.5 compares different augmentations methods on the test set with optimal parameters. As noticed earlier, temporally consistent augmentations are beneficial and hence both URA and TRA, which enforce this consistency, outperform image-level augmentation method RA by 1-2% in F1 and ∼3% in accuracy for the task of step recognition on CATARACTS. Additionally, URA and TRA both show improvement over the state-of-the-art MTMS-TCN model which utilized a 'Custom' set of augmentations by 1-5% across all the metrics for phase recognition on Bypass40. We can further notice a significant improvement of 5-11% across all the metrics for recognizing steps on Bypass40. This improvement could be attributed to the larger set of transforms $|\tau|= 10$.

Table 6.5: Comparison of different methods on BY40 and CA50 test sets. * denotes models trained in a multi-task setup requiring additional phase/step labels.

Dataset Task	Method	$\|\tau'\|$	M, N, T	ACC	PR	RE	F1
CA50 Step	Custom	-	-, -, -	81.79±12.30	77.82±13.61	82.25±14.69	78.21±14.90
	RA	3	30, 1, -	80.45±10.33	76.48±13.00	81.34±13.56	76.87±14.01
	URA (ours)	10	30, 1, -	83.24±10.64	77.04±14.20	82.33±14.68	78.02±14.98
	TRA (ours)	10	30, 1, 5	**83.64±10.67**	**78.38±14.11**	**84.06±14.18**	**79.43±15.09**
BY40 Phase	Custom*	-	-, -, -	90.26 ± 6.44	84.74 ± 7.71	81.75 ± 9.12	81.31 ± 9.07
	URA (ours)	10	30, 3, -	**93.55 ± 3.24**	83.25 ± 7.80	86.07 ± 7.61	83.51 ± 7.93
	TRA (ours)	10	30, 2, 5	93.17 ± 4.27	**86.42 ± 8.50**	**86.70 ± 6.72**	**85.20 ± 8.40**
BY40 Step	Custom*	-	-, -, -	75.46 ± 9.34	55.58 ± 9.88	52.78 ± 9.22	50.35 ± 9.75
	URA (ours)	10	30, 2, -	80.55 ± 6.61	61.32 ± 8.11	62.13 ± 7.74	58.52 ± 8.46
	TRA (ours)	10	30, 2, 5	**80.80 ± 7.90**	**63.66 ± 9.08**	**63.94 ± 8.31**	**60.06 ± 9.22**

TRA, on the other hand, outperforms URA on both the phase and step recognition tasks and both datasets. TRA achieves a 1-3% improvement in phase and step recognition on Bypass40 and CATARACTS, respectively. Moreover, for step recognition on Bypass40, TRA achieves a +2% and +1.5% improvement in precision and F1-score over URA. The performance improvement of the proposed TRA method over URA could be attributed to the temporally consistent augmentations applied at the clip level. TRA enables the extension of video datasets with videos composed of different segments augmented differently, which when used in training improves the generalization of deep learning models. Besides, the parameterization of TRA is independent of the underlying recognition task or dataset which enables the proposed method to be applicable to other surgical procedures and tasks.

6.4.6 Limitations

The (M, N, T) parameterization of TRandAugment simplifies the process of selecting a good augmentation policy, for training, that induces both spatial and temporal variations in the input videos. Yet, it does not completely eliminate the search for optimal parameters which adds computational expense. Further studies are required to better understand if or when datasets or tasks may require a separate search to achieve optimal performance. Another drawback of TRandAugment is that it works only in the input space. Few works in the literature have proposed adding variations in the model's feature space to improve generalizability [Liu 2018a, Chu 2020]. Unlike input space augmentations, designing feature space augmentations is extremely challenging because the domain or the noise characteristics of the feature space is not well-studied. Nevertheless, this could be an interesting extension to our proposed method, especially for training the temporal component of spatio-temporal models.

6.5 Conclusion

This chapter introduced a new augmentation method called *TRandAugment* that simplifies data augmentation pipelines. Given a video, creates pseudo videos with different clips augmented differently. The method is parameterized with magnitude (M), the number of augments (N), and the number of temporal augments (T). This parameterization provides a simple framework to search for optimal configuration and operates at a level with significantly reduced search space, in line with current research in data augmentation. The proposed method has been validated on two large surgical video datasets, considering both the phase and step recognition tasks, obtaining a boost in the performances thus demonstrating the impact of *TRandAugment*. New open questions arise on how this method may improve model robustness [Lopes 2019], federated learning [Kassem 2022], or weakly-/semi-/self-supervised learning [Ramesh 2023b, Pan 2021, Shi 2021, Yu 2019, Ramesh 2022]. Furthermore, the proposed method could be applicable to other tasks, such as tool localization and tracking [Nwoye 2019], action triplets [Nwoye 2020, Nwoye 2022b], and video semantic segmentation [Alapatt 2021]. Future work will study the value of TRandAugment in these different settings and tasks.

7 Cross-center Generalization Study

ಪ್ರತ್ಯಕ್ಷವಾಗಿ ಕಂಡರೂ ಪ್ರಮಾಣಿಸಿ ನೋಡು
(Pronunciation: pratyaksa kandaruu pramanisi nodu)
Even when something is evident, examine it

(Kannada Proverb)

Chapter Summary

Deep learning models have shown tremendous potential in recognizing surgical activities from endoscopic videos. Chapter 4, 5, & 6 studied their potential on phase and step

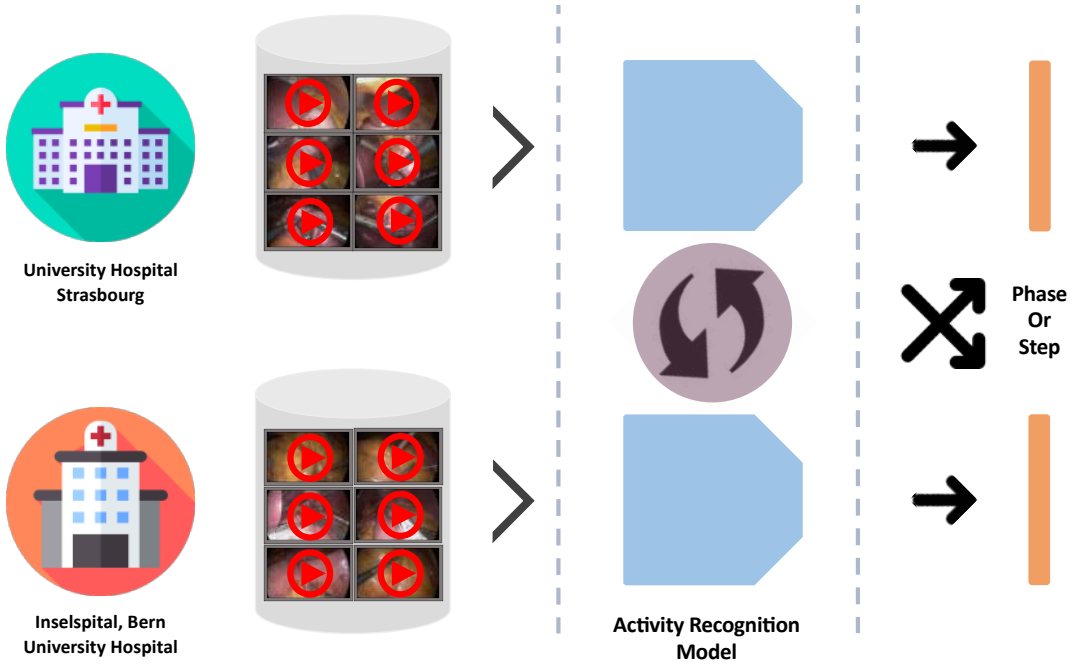


Figure 7.1: Setup of cross-center study of activity recognition models.

recognition using the BY40 dataset which consists of 40 videos of LRYGB procedures performed by expert surgeons at IHU Strasbourg, France. BY40 introduces diversity in surgical techniques through 7 expert surgeons that help activity recognition models to generalize. As discussed in Section 1.3.4, introducing diversity or variance in the training data is vital to address the overfitting and memorization of deep learning models.

Although useful, the diversity introduced in BY40 is only a small fraction of diversity existing in the domain owing to the changes in the surgical workflow across surgeons, medical centers, communities, nations, etc. Additional variance due to patients' age, height, weight, gender, race, and ethnicity also occurs in the medical domain. Characterizing the robustness and generalizability of deep learning models over these variations is a precursor to integrating them into a CAS in the OR. To this end, this chapter introduces two large video datasets of LRYGB procedures from two medical centers and analyzes the performance of activity recognition models across centers.

In this chapter, we first introduce in Section 7.1 a revised workflow of the LRYGB procedure integrating the different workflow followed at another medical center. We follow up with the description of the new multi-center dataset comprised of 70 videos from two different centers, called MultiBypass140, in Section 7.2. In Section 7.3, we present the different experimental studies and discuss their results in Section 7.4. Finally, we summarize the study of this chapter highlighting the important takeaways in Section 7.5.

Table 7.1: Definitions of all the proposed 12 phases for the gastric bypass procedure.

Phase ID	Phase Name	Description
P1	preparation	Phase of access to the abdominal cavity, installation of the ports for the camera and surgical instruments, and exposure of the operating field
P2	**gastric pouch creation**	Phase in which the small part of the stomach that is connected with the esophagus is separated from the rest to make a smaller gastric pouch
P3	omentum division	Vertical section of the omentum to facilitate the ascent of the small intestine to the gastric pouch
P4	**gastrojejunal anastomosis**	Connection of the distal small intestine with the gastric pouch
P5	**anastomosis test**	Verification that the gastrojejunostomy does not leak
P6	jejunal separation	Separation between the biliary and the alimentary limb
P7	closure petersen space	Closure of the space created between the mesentery and the mesocolon as the small intestine rises to make the bypass
P8	**jejunojejunal anastomosis**	Connection of the biliary limb with the alimentary limb
P9	closure mesenteric defect	Closure of the space created in the mesentery as the small intestine rises to make the bypass
P10	cleaning coagulation	Verification of the absence of bleeding, hemostasis, and aspiration of the remaining liquid in the abdominal cavity
P11	disassembling	Removal of surgical instruments and camera
P12	**Other interventions**	If additional intervention is performed (e.g. liver biopsy, cholecystectomy)

7.1 Phase and Step Definitions and Differences

The ontology consisting of phase and step presented in Section 3.3.1 captured the LRYGB workflow followed at IHU Strasbourg (France) and was defined by an expert surgeon. In order to annotate video datasets from other medical centers with surgical activities, the ontology may need to be revised to include variations in the workflow followed at other centers. In this work, we build another large-scale video dataset of LRYGB surgeries performed at Inselspital, Bern University Hospital, Switzerland, in collaboration with Dr. med. Joël Lavanchy. As a prerequisite, the previously defined ontology has been inspected by an expert surgeon from Inselspital and a few revisions have been introduced. Table 7.1 describes all the phases of LRYGB workflow followed at both Strasbourg and Bern centers while the steps are presented in Table 7.2. The new phases and steps added to the two tables are highlighted in magenta.

'P12: Other interventions' is a minor addition to the phases that capture other interventions carried out alongside gastric bypass. This phase has no associated steps as they belong to the ontology of the other interventions. In Table 3.2, two new steps - 'S44: drainage insertion' and 'S45: specimen retrieval' - are appended to the LRYGB workflow as these extra steps are performed as part of the workflow followed at Inselspital. With these modifications, the resulting ontology encapsulates the workflow from both medical centers.

Table 7.2: Definitions of all the proposed 46 steps for the gastric bypass procedure.

Step ID	Step Name	Description
S0	null step	The camera is static and no actions are performed by the surgeon
S1	cavity exploration	The entire abdominal cavity is evaluated to verify the absence of alterations that could prevent or modify the planned surgery and determine the technical feasibility of performing it
S2	trocar placement	The accessory work ports (usually four) are introduced into the abdominal cavity
S3	retractor placement	Introduction of the instrument to retract the liver and expose the esophagogastric junction
S4	**crura dissection**	The fatty tissue surrounding the esophagogastric junction is dissected to clearly expose the angle of his and separate the adhesions with the spleen
S5	**his angle dissection**	Opening of a retrogastric window at the level of the lesser curvature of the stomach to facilitate the passage of the stapling machine
S6	**horizontal stapling**	Horizontal section of the stomach at the level of the lesser curvature with the stapling machine
S7	**retrogastric dissection**	Dissection of the fatty and vascular tissue in the posterior part of the stomach
S8	**vertical stapling**	Vertical section of the stomach with the stapling machine
S9	gastric remnant reinforcement	Verification and reinforcement of the gastric remnant stapling with suture thread
S10	gastric pouch reinforcement	Verification and reinforcement of the gastric pouch stapling with suture thread
S11	gastric opening	Opening of the hole in the gastric pouch where the connection to the small intestine will be made
S12	omental lifting	Clamping and lifting of the omentum
S13	omental section	Full section of omentum to divide it into two parts

Continued on next page

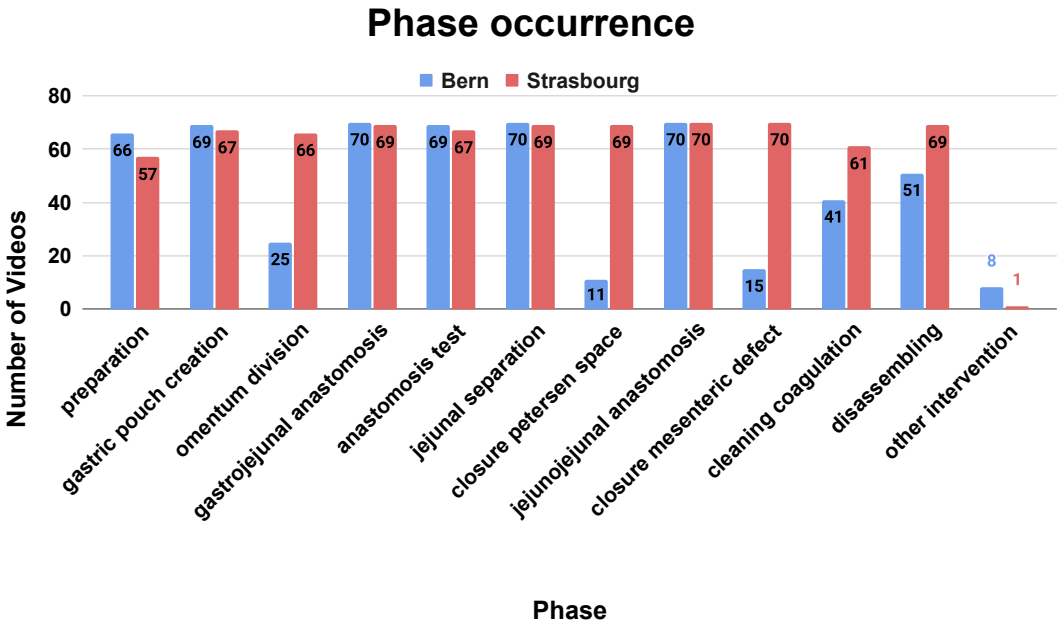
Table 7.2 – *Continued from previous page*

Step ID	Step Name	Description
S14	adhesiolysis	Section of the connective tissue fibers between the structures
S15	treitz angle identification	Exposure of the transverse mesocolon to visualize the treitz angle
S16	**biliary limb measurement**	Measurement of the level at which the connection of the distal small intestine with the gastric reservoir will be made to perform the gastric bypass (around 70 cm)
S17	jejunum opening	Opening of the distal small intestine where it will be connected to the gastric reservoir to perform the gastric bypass
S18	**gastrojejunal stapling**	Connection of the gastric pouch to the distal small intestine (distal jejunum)
S19	gastrojejunal defect closing	Suture closure of the hole left by the stapling machine between the stomach and the jejunum
S20	mesenteric opening	Opening of the mesentery on the edge of the jejunum to facilitate the passage of the stapling machine
S21	jejunal section	Clamping and section of the jejunum proximal to the gastrojejunostomy
S22	gastric tube placement	Progression of the gastric tube from the stomach to the jejunum in order to calibrate the anastomosis and then verify that the connection does not leak
S23	clamping	Clamping of the jejunum distal to the gastrojejunostomy
S24	ink injection	Injection of the ink to detect any leakage
S25	**visual assessment**	Visual inspection of the gastrojejunostomy for any leakages
S26	gastrojejunal anastomosis reinforcement	Reinforcement and fixation of the connection between the stomach and the jejunum
S27	petersen space exposure	Traction of the mesocolon to expose the space created when the small intestine ascends towards the gastric pouch
S28	petersen space closing	Closing the petersen space with suture thread
S29	biliary limb opening	Opening of the hole in the proximal bowel where the connection between the biliary limb with the alimentary limb will be made

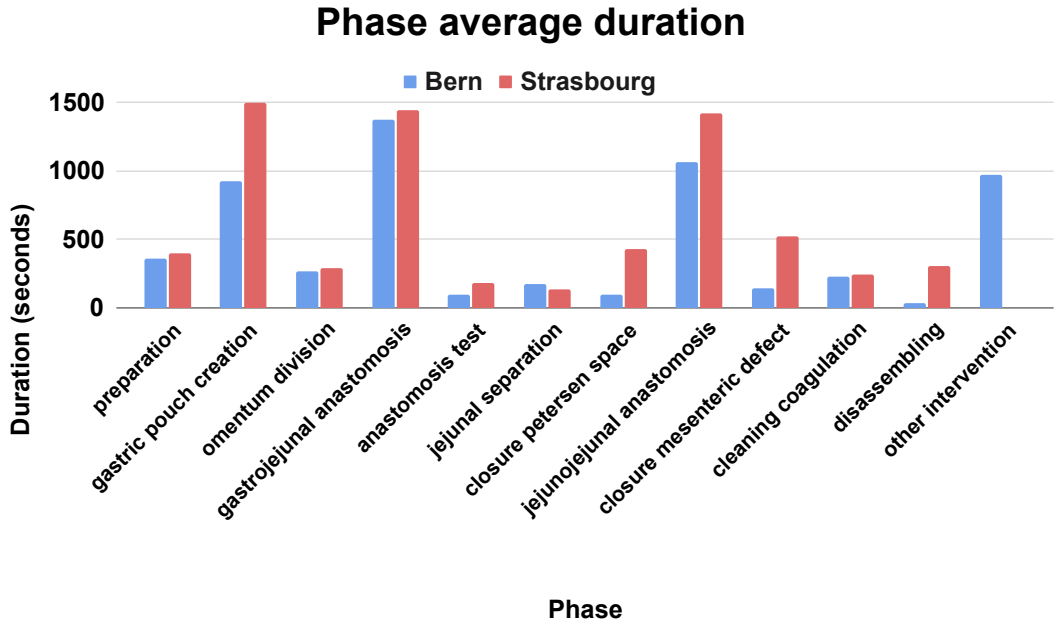
Continued on next page

Table 7.2 – *Continued from previous page*

Step ID	Step Name	Description
S30	**alimentary limb measurement**	Measurement of the level at which the connection of the biliary limb with the alimentary limb will be made (around 150 cm)
S31	alimentary limb opening	Opening of the hole in the distal bowel where the connection between the biliary limb with the alimentary limb will be made
S32	**jejunojejunal stapling**	Connection of the biliary limb to the alimentary limb
S33	jejunojejunal defect closing	Suture closure of the hole left by the stapling machine between the biliary and the alimentary limb
S34	jejunojejunal anastomosis reinforcement	Reinforcement and/or fixation of the jejunojejunal anastomosis with a suture thread
S35	staple line reinforcement	Reinforcement and/or fixation of the staple line with a suture thread
S36	mesenteric defect exposure	Traction of the anastomosis between the alimentary loop and the biliary loop and/or the mesentery to expose the space created when the small intestine ascends towards the gastric pouch
S37	mesenteric defect closing	Closing the space with suture thread
S38	anastomosis fixation	Reinforcement and/or fixation of the anastomosis with suture thread
S39	**coagulation**	Introduction of a cloth and/or hemostatic tool (bipolar grasper) and applying pressure to reduce bleeding
S40	irrigation aspiration	Suction of any remaining liquid in the abdominal cavity
S41	parietal closure	Closure of the abdominal port holes
S42	trocar removal	Removal of all the trocars (usually four) placed during the preparation phase under visual control
S43	calibration	Re-calibration and cleaning of the camera
S44	**drainage insertion**	Insertion of drainage into the abdominal cavity to drain fluids
S45	**specimen retrieval**	Removal of any spare tissue (e.g. omentum, small bowel, or stomach)



(a) Total occurrence of phases in the videos from the two medical centers.



(b) Average duration of phases across videos from the two medical centers.

Figure 7.2: BernBypass70 vs StrasBypass70: Total occurrence and average duration of phases across videos in the datasets.

7.2 Multi-center Dataset: MultiBypass140

MultiBypass140 is a multi-center dataset constructed from 140 videos of LRYGB surgeries operated in two medical centers: Bern and Strasbourg. The following sections

provide a detailed description of the dataset.

7.2.1 Bern Center

In collaboration with Dr. med. Joël Lavanchy, we present a new dataset, called BernBypass70, consisting of laparoscopic videos of LRYGB surgeries performed at Inselspital, Bern University Hospital, Switzerland. The recordings have been captured at 25 fps with a resolution of 720×576 or 1280×720 and anonymized for privacy. A few interesting statistics of the dataset are presented in Table 7.3. On overage, surgery at the Bern medical center lasts for 72 minutes. The dataset is fully annotated with both phases and steps by an expert surgeon using the ontology presented earlier. The distribution of the phases and steps across the dataset is graphed in Figure 7.2, 7.3, & 7.4. Not all phases and steps are carried out in each procedure. 3 out of 11 phases concerning the LRYGB procedure are performed in less than 25 surgeries. Similarly, 20 out of 46 steps are carried out in less than 25 surgeries.

7.2.2 Strasbourg Center

We extend the BY40 introduced in Chapter 3 with additional 30 videos and create a new large video dataset called StrasBypass70. Out of the 70 videos, 40 videos are annotated by a clinician at IHU Strasbourg as part of BY40 while the remaining 30 videos are annotated by a clinician from Bern University Hospital. The inter- and intra-rater reliability between the two clinicians on the multi-center dataset from Bern and Strasbourg has been studied in [Lavanchy 2022] showing excellent reliability in the phase and step annotations of the two datasets (BernBypass70 and StrasBypass70). The recordings have been captured with patients' consent at 25 fps with a resolution of 854×480 or 1920×1080 and anonymized for privacy. The average duration of a LRYGB procedure at IHU Strasbourg is 110 minutes. All 11 phases are routinely carried out with only 11 out of 46 steps performed in less than 25 surgeries (Figure 7.2 & 7.3).

7.2.3 Dataset setup

In the experimental study conducted in this chapter, we split the 70 videos in BernBypass70 and StrasBypass70 into 40, 10, and 20 videos for training, validation, and test sets, respectively. Images were extracted at 1 fps. This amounts to 166,431, 46,497, and 92,979 frames in the three sets of BernBypass70. Similarly, the total frames in each set of StrasBypass70 amount to 252,913, 72,555, and 138,326. The frames are resized to $224 \times 224 \times 3$ and the training dataset is augmented by applying horizontal flip, saturation, and rotation.

7.3 Study Design

In this chapter, we study the effectiveness of activity recognition methods on the two datasets presented above. The overview of the study design is presented in Figure 7.1.

Table 7.3: Statistics of the LRYGB datasets from two medical centers.

Dataset	videos	minimum duration (mins)	maximum duration (mins)	mean±std duration (mins)	frames @ 1 fps
StrasBypass70					
Train	40	41	171	106 ± 32	253,913
Validation	10	78	176	121 ± 33	72,555
Test	20	63	178	115 ± 32	138,326
BernBypass70					
Train	40	37	114	69± 19	166,431
Validation	10	54	116	77± 20	46,497
Test	20	52	145	77 ± 24	92,979
MultiBypass140					
Train	80	37	171	88 ± 33	420,344
Validation	20	54	176	99 ± 35	119,052
Test	40	52	178	96 ± 34	231,305

We analyze MTMS-TCN for joint phase and step recognition similar to the experimental study of Section 4. Additionally, we evaluate the weakly supervised learning method presented in Section 5. For completeness, we briefly describe the two methods below.

7.3.1 Multi-level activity recognition

MTMS-TCN architecture consists of two stages where first a multi-task CNN, i.e., ResNet-50, is employed for extracting visual features from images followed by a multi-task multi-stage causal TCN to refine the features and extracting temporal information for joint phase and step recognition. The ResNet-50 model is initialized with pre-trained weights on ImageNet and trained using adam optimizer for 30 epochs. While we use MTMS-TCN with a single TCN stage as the second stage does not improve the model performance. The temporal model was trained in a multi-task learning setup on video features extracted from ResNet-50 for 200 epochs.

We conduct a set of seven experimental setups to analyze the performance of MTMS-TCN on the multi-center dataset:

(a) Training and evaluation on StrasBypass70;

(b) Training and evaluation on BernBypass70;

(c) Training on StrasBypass70 and evaluation on BernBypass70;

(d) Training on BernBypass70 and evaluation on StrasBypass70;

(e) Training and evaluation on the joint MultiBypass140 dataset;

(f) Training on MultiBypass140 and evaluation on StrasBypass70;

(g) Training on MultiBypass140 and evaluated on BernBypass70;

7.3.2 Weakly-supervised learning

As presented in Chapter 5, an end-to-end spatio-temporal model was built utilizing ResNet-50 as the spatial model and a single-stage variant of MS-TCN as the temporal model. The model is designed for step recognition with limited step labels and a large amount of weaker phase labels. The spatio-temporal model is trained using the dependency loss (Section 5.2.2) for 30 epochs with a learning rate of 1e-5. We split the dataset into 40, 10, and 20 videos for training, validation, and test sets respectively. We study the 'FSA' vs 'DEP' performance for the different number of videos with step labels $k \in \{3, 6, 12, 24, 30\}$ and use all 40 videos of the training set with phase labels.

7.3.3 Metrics

The performance of the methods is evaluated using accuracy (ACC), precision (PR), recall (RE), and F1-score (F1) metrics. The metrics are computed per video (averaged across classes) and are averaged across all the videos in the given set, following the same protocol as previous chapters.

7.4 Results

7.4.1 Multi-level activity recognition

7.4.1.1 Quantitative analysis

The quantitative performance of MTMS-TCN on the seven different experimental studies for phase and step recognition are presented in Table 7.4 & 7.5. The model performs best in recognizing both types of activities when trained and evaluated on the same dataset ((a) StrasBypass70 or (b) BernBypass70). However, the performance of the model drops by significantly on BernBypass70 compared to StrasBypass70. Performance of MTMS-TCN drops by $\sim 5\%$ in accuracy and $\sim 21\%$ in F1-score on phase recognition while it drops by $\sim 11\%$ in accuracy and $\sim 10\%$ in F1-score on step recognition. Furthermore, a cross-center evaluation, captured by experimental study (c) & (d), shows a huge gap in the transferability of an activity recognition model. MTMS-TCN trained on StrasBypass70 achieves an accuracy $\sim 71\%$ and F1 $\sim 35\%$ when evaluated on BernBypass70 for phase recognition. While the model trained on BernBypass70 achieves $\sim 63\%$ accuracy and $\sim 33\%$ F1 when evaluated on StrasBypass70. Particularly, we observe a transferability gap of 18-22% in accuracy and 28-47% in F1 on phase recognition and 21-30% in accuracy and 26-37% in F1 on step recognition.

All the previously observed performance drops could be attributed to the various differences between the two datasets. As seen in Figure 7.2a & 7.3, both phases and steps occur evenly in all the 70 procedures of StrasBypass70. In contrast, only a subset of phases and steps occur in most of the procedures in BernBypass70 with very few procedures containing all the phases and steps. These differences in the distribution of phases and steps between StrasBypass70 and BernBypass70 exist due to the variations

Table 7.4: Performance of MTMS-TCN on different datasets on phase recognition.

Train	Test	ACC	PR	RE	F1
StrasBypass70	StrasBypass70	90.70 ± 6.92	82.32 ± 8.69	85.86 ± 7.70	82.31 ± 8.83
	BernBypass70	71.95 ± 13.98	38.38 ± 8.10	43.26 ± 10.44	35.69 ± 9.88
BernBypass70	StrasBypass70	63.63 ± 9.43	36.67 ± 5.00	38.44 ± 8.15	33.12 ± 5.52
	BernBypass70	85.01 ± 13.22	62.79 ± 10.99	66.41 ± 12.20	61.21 ± 11.73
MultiBypass140	StrasBypass70	90.14 ± 6.80	81.68 ± 7.93	83.79 ± 7.85	81.17 ± 8.09
	BernBypass70	85.97 ± 12.92	61.81 ± 10.92	67.04 ± 11.6	60.58 ± 11.32
	MultiBypass140	88.05 ± 10.53	71.75 ± 13.78	75.41 ± 12.97	70.88 ± 14.24

Table 7.5: Performance of MTMS-TCN on different datasets on step recognition.

Train	Test	ACC	PR	RE	F1
StrasBypass70	StrasBypass70	78.79 ± 10.28	62.12 ± 7.14	64.79 ± 8.68	60.53 ± 8.23
	BernBypass70	49.57 ± 14.39	24.50 ± 6.57	29.50 ± 6.95	23.00 ± 6.47
BernBypass70	StrasBypass70	46.04 ± 11.00	30.36 ± 4.82	29.67 ± 6.19	24.69 ± 4.91
	BernBypass70	67.61 ± 13.51	52.75 ± 9.50	55.81 ± 11.41	50.08 ± 10.67
MultiBypass140	StrasBypass70	78.16 ± 10.07	62.12 ± 6.79	63.54 ± 8.15	59.87 ± 7.71
	BernBypass70	68.60 ± 13.35	52.38 ± 8.17	55.01 ± 9.59	49.69 ± 9.43
	MultiBypass140	73.38 ± 12.75	57.25 ± 8.95	59.28 ± 9.87	54.78 ± 10.00

in the surgical technique and workflows followed in the two medical centers. For instance, phase 'P3: omentum division' or 'P7: closure petersen space' routinely carried out in the Strasbourg center is not regularly performed in the Bern center. Given the hierarchical structure of phases and steps, with every phase missing, corresponding steps are missing as well. On average, a procedure of BernBypass70 contains 2 phases and 6 steps less than the average StrasBypass70 procedure. This finding is also reflected by the average duration of a surgery which is 38 minutes shorter in BernBypass70 than in StrasBypass70. All these natural variations in the workflows across different centers must be introduced while training to improve the robustness and generalizability of activity recognition models. Moreover, multi-center validation of the methods is required before a large-scale deployment in the ORs.

Note, study (e) demonstrates that our model trained and evaluated on a MultiBypass140 (71% and 55% F1-score for phases and steps, respectively) has a performance in between the performance on the individual monocentric datasets. Interestingly, studies (f) and (g) reveal that when the model is evaluated separately on each center, its performance is close to monocentric training and evaluation. This illustrates that the model trained on MultiBypass140 is capable of learning the variations in the two datasets without compromising its efficiency on either dataset.

Table 7.6: BernBypass70: Effect of weak supervision on varying amount of step labeled videos. Accuracy (ACC), Precision (PR), Recall (RE), and F1-score (F1) (%) are reported. 'FSA' denotes the model trained for step recognition without any phase annotations. 'DEP' denotes the dependency loss added for weak supervision using phase labels on the remaining videos.

	# Videos					
Model	Step	Phase	ACC	PR	RE	F1
FSA	3 (8%)	-	42.38 ± 11.90	22.60 ± 5.33	24.96 ± 5.51	19.29 ± 5.33
DEP	3 (8%)	37	**49.10 ± 11.25**	**26.21 ± 6.52**	**25.26 ± 6.12**	**22.43 ± 6.42**
FSA	6 (15%)	-	47.94 ± 12.82	32.29 ± 7.95	35.01 ± 8.05	29.39 ± 7.79
DEP	6 (15%)	34	**52.50 ± 11.56**	**35.98 ± 6.96**	**36.18 ± 7.15**	**31.58 ± 7.14**
FSA	12 (30%)	-	59.79 ± 13.29	42.98 ± 8.48	41.84 ± 7.96	38.51 ± 8.29
DEP	12 (30%)	28	**61.27 ± 13.18**	**46.26 ± 7.81**	**45.39 ± 8.19**	**41.95 ± 8.35**
FSA	24 (60%)	-	66.99 ± 11.82	52.51 ± 7.73	51.33 ± 7.47	48.18 ± 8.12
DEP	24 (60%)	16	**67.92 ± 13.05**	**53.95 ± 7.47**	**53.98 ± 8.71**	**49.47 ± 8.94**
FSA	30 (75%)	-	**68.53 ± 10.93**	**53.73 ± 7.66**	**53.75 ± 9.11**	**49.91 ± 8.54**
DEP	30 (75%)	10	67.23 ± 10.87	49.61 ± 7.42	50.78 ± 7.64	46.47 ± 7.87
FSA	40 (100%)	-	68.60 ± 13.35	52.38 ± 8.17	55.01 ± 9.59	49.69 ± 9.43

7.4.1.2 Qualitative analysis

Figure 7.5 & 7.6 visualizes a video set of one best and one worst performance of MTMS-TCN for phase and step recognition on the three datasets: BernBypass70, StrasBypass70, and MultiBypass140. Similar to the results on BY40 in Section 4.4, MTMS-TCN model performs well in recognizing short duration phases and steps on all the three datasets. Interestingly, the activity recognition model fails when complications arise in the procedure resulting in deviations from the expected workflow. This can be noticed in the worst performing video on BernBypass70 in both Figure 7.5 & 7.6. An unexpectedly long time of the surgery is spent in 'P10: cleaning coagulation', and its corresponding step, followed by an unusual transition back to 'P8', 'P6', 'P5', and 'P4'.

7.4.2 Weakly-supervised learning

To study the transferability of the weakly-supervised learning method presented in Chapter 5, we present its results on BernBypass70 in Table 7.6. The table contains the results of our model with a varying number of videos in the training set labeled with steps (3, 6, 12, 24, and 30) along with the rest of the training set containing phase annotations. The 'DEP' model improves by 3-7% over 'FSA' when trained with only 3 videos annotated with steps. Similarly, we see a 2-5% and 2-4% increase in performance in all the metrics of the 'DEP' model in experiments corresponding to 6 and 12 step annotated videos, respectively. However, the performance gains of the 'DEP' model decrease with increasing step annotated videos. Despite the differences in the workflow between Bern and Strasbourg medical centers, exploiting the phase-step hierarchy in a weakly-supervised

learning setup does benefit step recognition.

7.5 Conclusion

In this chapter, we introduced multi-center video datasets consisting of 140 videos from two medical centers, called BernBypass70 and StrasBypass70. Additionally, we present a revised ontology of LRYGB procedure combining the workflows followed in the two centers. We study the transferability of surgical activity recognition methods on the two datasets. This study demonstrates the need to introduce variations in surgical techniques and workflow to deep learning models to avoid the generalization gap described in the literature [Bar 2020,Kitaguchi 2022]. With an extensive experimental study, the origin of the performance differences in our datasets has been investigated. It has been shown that dataset distribution and size due to different LRYGB techniques and workflows between centers have a major impact on model performance. This highlights the importance of multi-centric datasets for the training and evaluation of deep learning models in surgical video analysis.

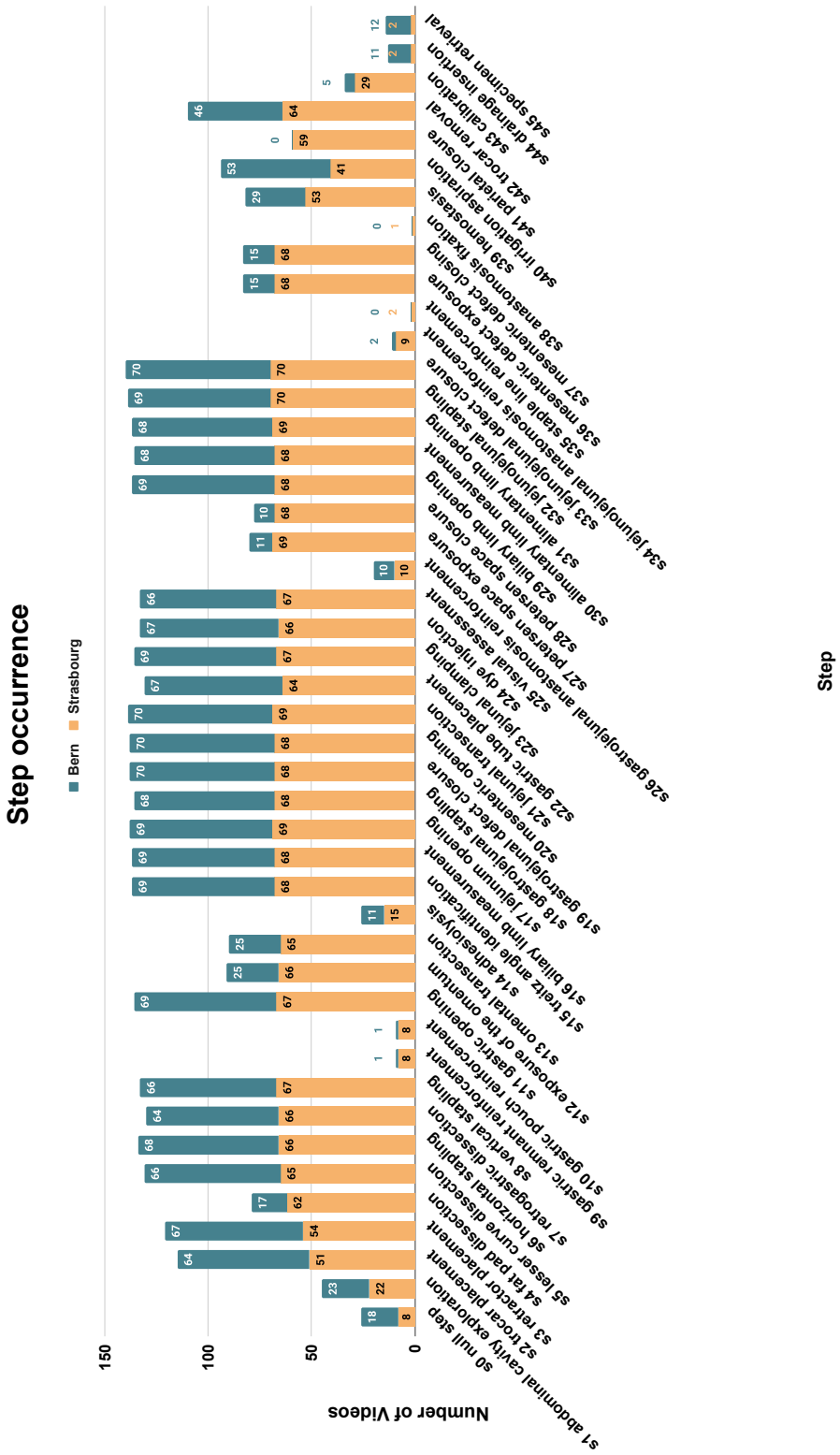


Figure 7.3: BernBypass70 vs StrasBypass70: Total occurrence of steps across videos.

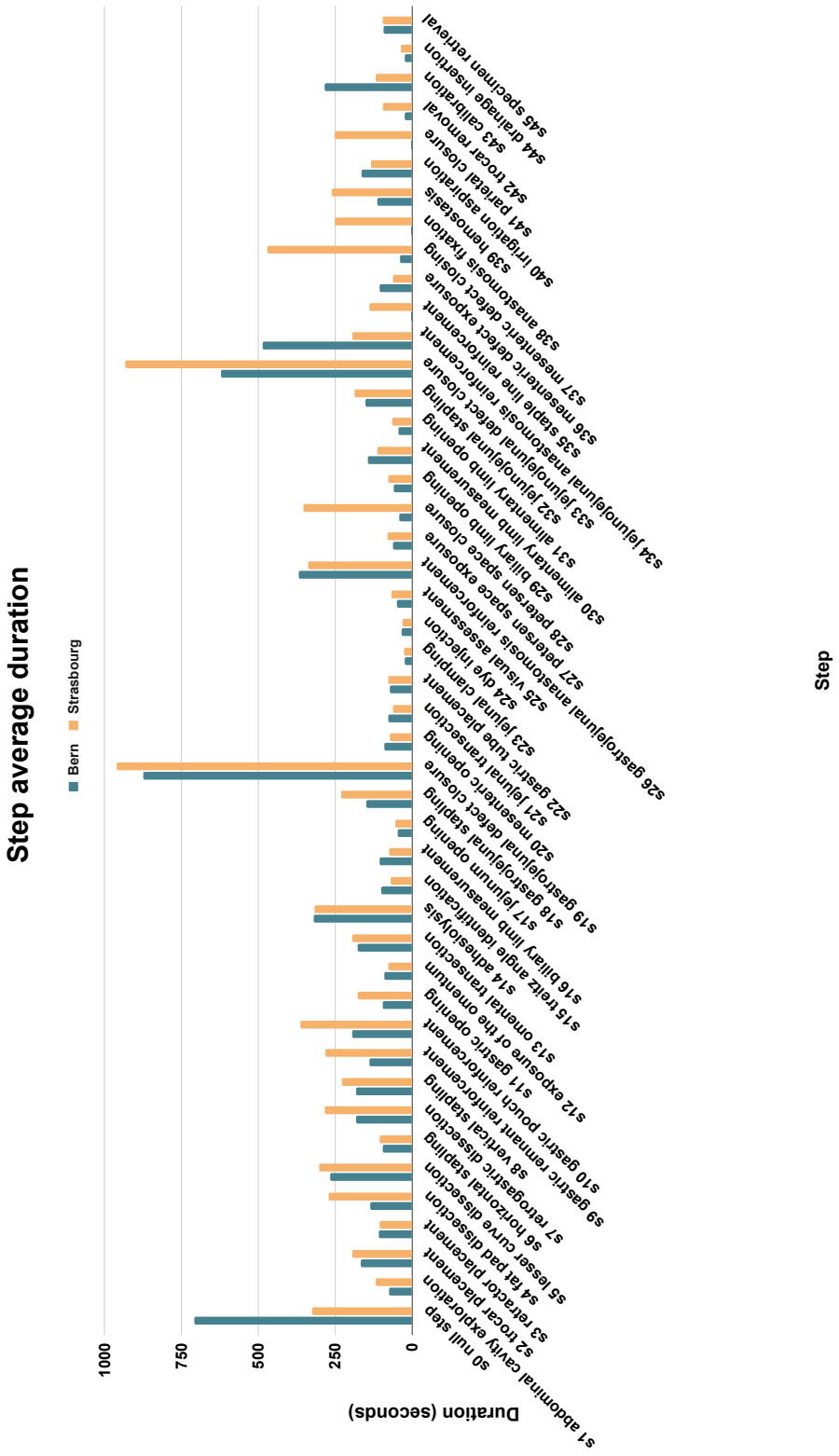


Figure 7.4: BernBypass70 vs StrasBypass70: Average duration of steps across videos.

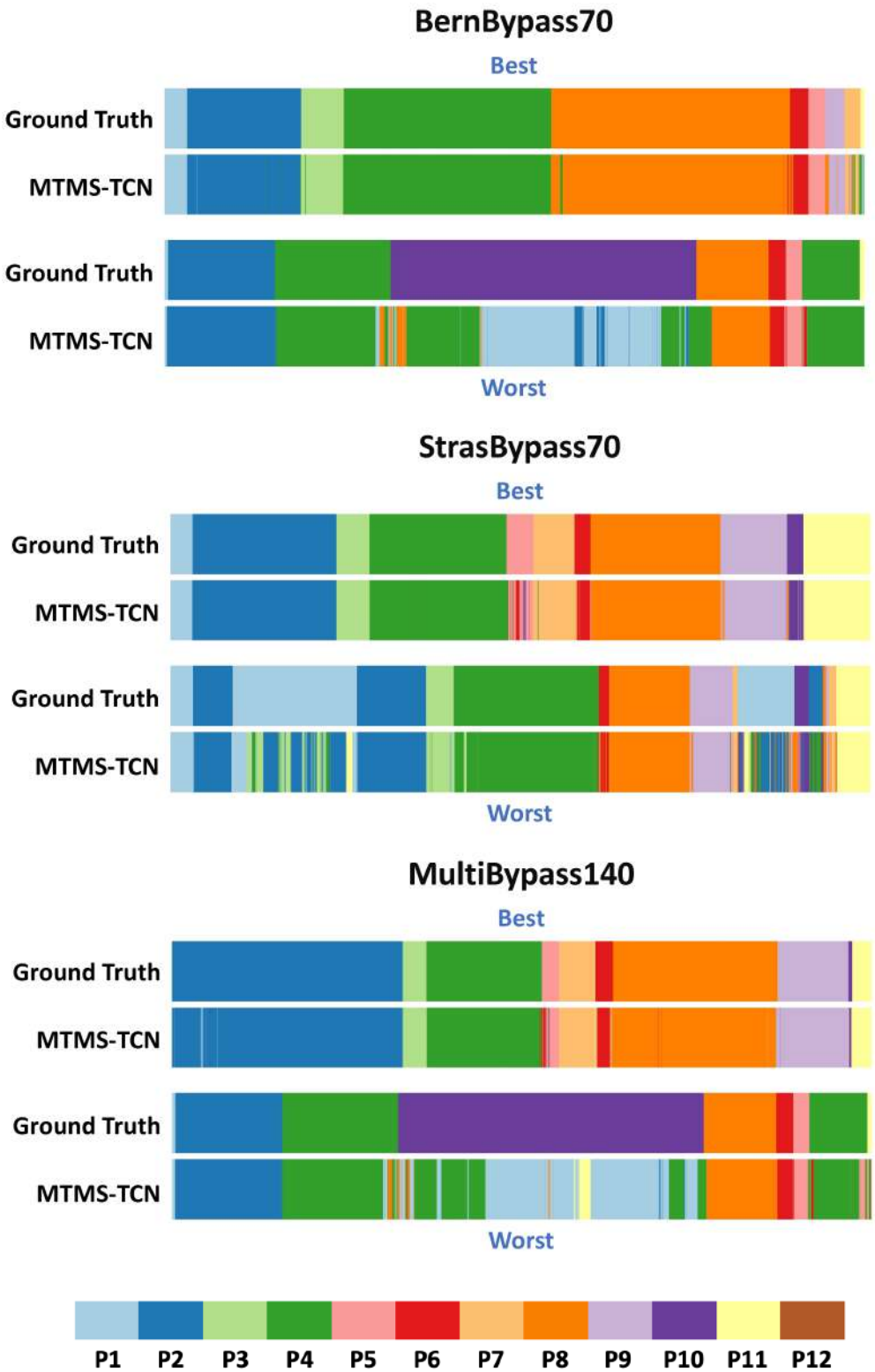


Figure 7.5: Phase predictions on one best and one worst video from the multi-center datasets.

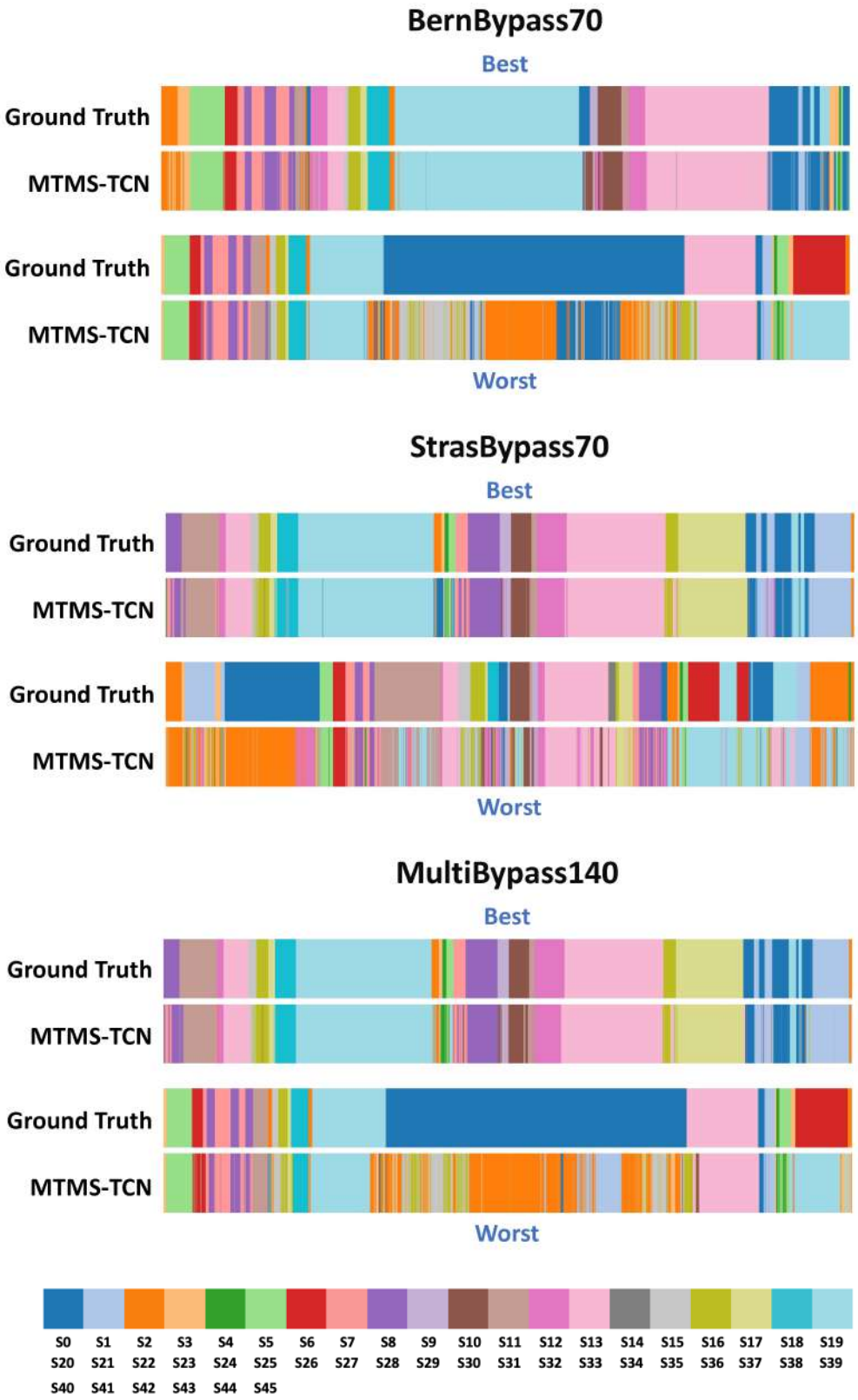


Figure 7.6: Step predictions on one best and one worst video from the multi-center datasets.

Part III

Applications, Conclusion, and Future Perspectives

8 Potential Applications

In the end we retain from our studies only that which we practically apply.

- Johann Wolfgang Von Goethe

Chapter Summary

The capability to automatically recognize surgical activities from endoscopic videos could empower CAS to be successfully deployed in the OR. Especially, these systems would be effective in many pre-, intra-, and post-operative applications targeting surgical training, safety monitoring, clinical decision support, data indexing and reporting, autonomous execution, and others. In this chapter, we present a few of the potential applications of activity recognition in the OR. An overview of applications of CAS is presented in Figure 8.1.

8.1 Automatic Report Generation

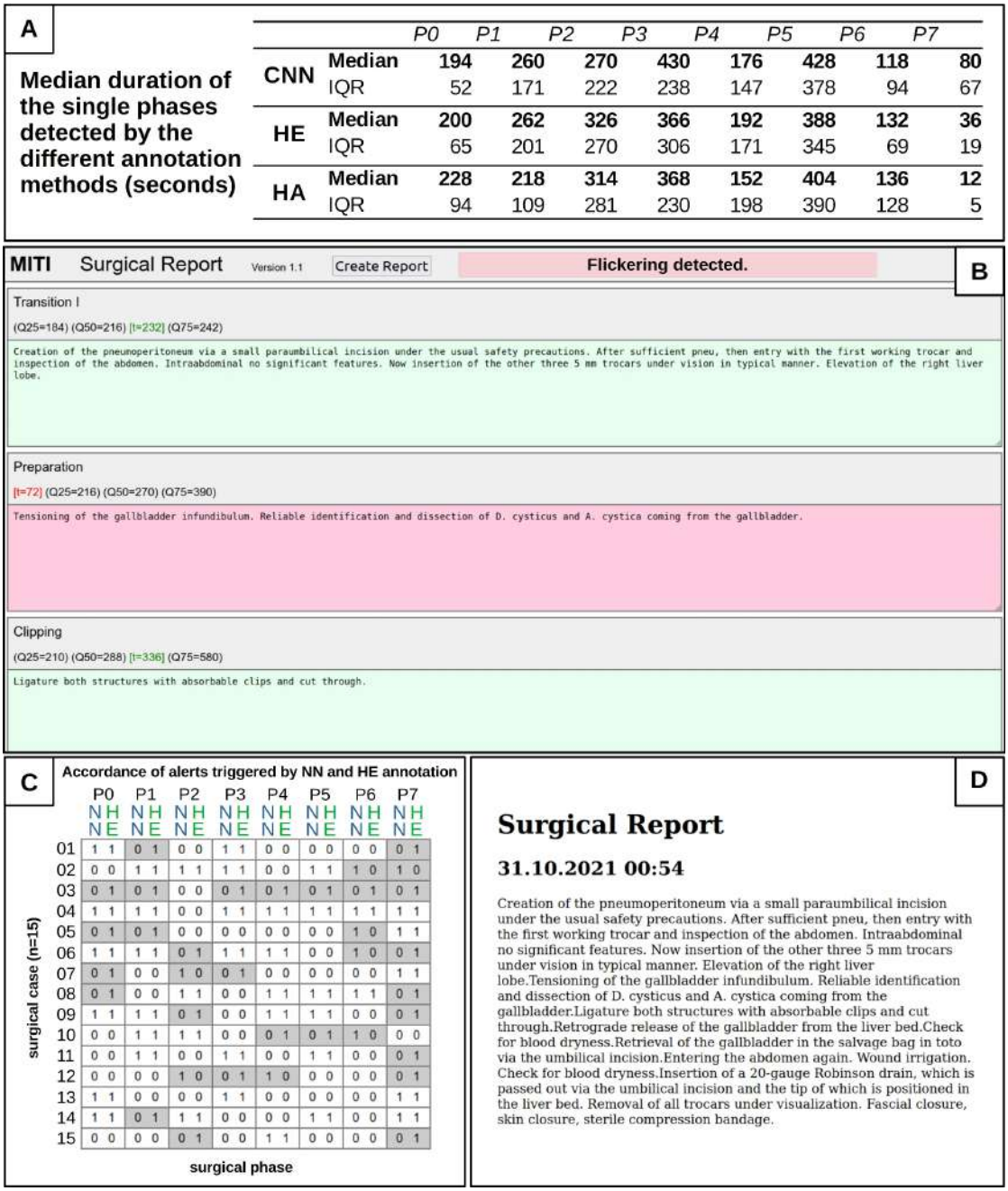
Documenting the complete process followed during each surgery is a fundamental part of the workflow that captures valuable information used in scientific, administrative, and judicial applications. The surgical reports are filed by surgeons by making use of structured templates with predefined wording to describe operative notes. Following this structured process for reporting empowers automated extraction of information for

Figure 8.1: An illustration of context-aware assistance that could be provided in and out of the operating room using some of the systems developed in this thesis.

efficient storage and quick retrieval. However, generating a surgical report is a tedious job carried out by surgeons worldwide. After completing a physically and cognitively intense surgery that could last from a few minutes to hours, writing a detailed report, even when structured, could feel like an arduous undertaking. Assisting in reporting or completely auto-generating them could be achieved via the activity recognition module of CAS. Reliably recognizing activities, both coarse and fine-grained, allows automatic recording of all the phases and steps achieved in the surgery along with the total time spent in each activity. Critical activities that must be carried out for ensuring safe surgery could also be tracked and reported automatically. For instance, key activities like 'P4: gastrojejunal anastomosis' or 'S6: horizontal stapling' of the Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) procedure are necessarily required to be completed to avoid unnecessary complications. Recording if these activities were achieved or not would be to understand and tackle post-operative complications. [Berlet 2022] is an exemplar of a research attempt to automatically generate surgical reports using a phase recognition model. A sample report generated by [Berlet 2022] is shown in Figure 8.2.

8.2 Surgical Skill Assessment and Training

Automatic surgical skill assessment has been studied in the community by recognizing gestures on bench-top surgical training tasks in videos and comparing between them [Varadarajan 2009b, Doughty 2018]. Similarly, other activities such as phases or steps can be recognized during training with advanced surgical simulation platforms



A

Median duration of the single phases detected by the different annotation methods (seconds)

		P0	P1	P2	P3	P4	P5	P6	P7
CNN	Median	194	260	270	430	176	428	118	80
	IQR	52	171	222	238	147	378	94	67
HE	Median	200	262	326	366	192	388	132	36
	IQR	65	201	270	306	171	345	69	19
HA	Median	228	218	314	368	152	404	136	12
	IQR	94	109	281	230	198	390	128	5

Figure 8.2: A sample report automatically generate for laparoscopic cholecystectomy based on phase recognition model. Image credit: [Berlet 2022]

and evaluated to assess the skill of a trainee. The list and order of activities carried out by a trainee alongside the time spent in completing each activity provide valuable insights into the surgeon's competency in performing them. By devising quantitative metrics to measure the skills, the system should be capable to rank surgeons on their competency and shortcomings. Besides, the training can be assisted by quickly searching activity relevant video clips from large video databases. Note that the large surgical

video databases could also be effectively indexed based on activity labels. Learners' engagement in the training process may be further increased by using gamification techniques to build training courses.

8.3 Decision Support and Monitoring Systems

Automatic recognition of phases or steps provides information on all the phases and steps carried out prior to the current phase/step. Combining domain knowledge with such information can be employed to guide the surgical decision-making process. Contextual knowledge extracted from such information can suggest probable next phase/step to undertake along with listing all the tools and technology required to successfully complete it. Furthermore, context-enabled situational awareness may be useful for real-time simulation of the complete surgery which could lead to early recognition of possible outcomes or complications and assist in early decision-making on the action plan.

Safety checklists can be automatically verified by a monitoring system utilizing the contextual information extracted through activity recognition. Critical phases, steps, or actions could all be automatically identified and warnings could be issued if any of the critical activities are not accomplished during the surgery. For example, in Laparoscopic Cholecystectomy the achievement of a critical view of safety could be monitored if the surgeon is executing the clipping and cutting phase [Mascagni 2020]. Furthermore, deviations from the expected surgical plan could be detected by analyzing phase or step transitions and appropriate alerts could be sent to both hospital administration and relevant clinicians to swiftly assess all the revisions in the surgical plan.

8.4 Autonomous Surgical Robots

The introduction and advances in MIS to minimize pain, improve recovery time, and reduce complications continuously increase the complexity of the workflow. This has brought about new challenges such as a steep learning curve, a restricted view of anatomy, and/or a longer duration of a procedure compared to open surgery. Longer duration of surgery implies longer exposure of the patient to anesthesia and longer application of CO_2 pressure in laparoscopic procedures. These have many side effects on the patients ranging from dizziness or headaches to gas embolism or cardiac arrest [Wu 2019, Dowdy 2021]. Reducing these side effects requires reducing the overall duration of the surgery. Although decision support systems could assist towards this aim, introducing partial or complete autonomy in performing surgeries may contribute significantly. A straightforward benefit of surgical autonomy is the reduction in the cognitive workload of the surgeons by offloading mundane or repetitive tasks. On a scale of 0 to 5, different level or degree of autonomy has been proposed for a robotic surgical system [Battaglia 2021]. This is illustrated in Figure 8.3. Here, 0 corresponds to no autonomy with the surgery managed fully by clinicians and 5 to a system fully capable of performing entire surgeries, with no human intervention. Across all these different levels of autonomy, a CAS plays a central role by presenting contextual knowledge use-

Figure 8.3: The six levels of autonomy in robotic surgery proposed in [Yang 2017].

ful for decision-making, intra-operative surgical planning, or monitoring the safety of autonomous executions. Particularly, modeling and recognizing phases and steps could enable or disable robotic platforms to execute actions autonomously. Moreover, continuous planning of the future set of surgical phases or steps to be performed can be achieved from the current state of the surgery. Finally, the autonomous execution of the surgery can be monitored by tracking the phases or steps and any divergence from the expected surgical plan could send alerts requesting immediate human interventions.

8.5 Conclusion

In this chapter, we outlined a few potential applications of the research addressed in this thesis, i.e., multi-level surgical activity recognition, covering pre-operative, intra-operative, and post-operative parts of a surgical workflow.

9 Conclusion and Future Perspectives

Se tu segui tua stella, non puoi fallire a glorioso porto
If you follow your own star, you cannot fail to reach a glorious harbor.

- Dante Alighieri

Chapter Summary

This chapter presents a summary of the works presented in this thesis. We tackle different objectives related to surgical activity recognition and highlight the key takeaways of these studies. We conclude this chapter by suggesting perspectives on future research.

9.1 Thesis Conclusion

Safe and efficient surgery that minimizes pain, allows fast recovery, and reduces complications are the primary goal of developing next-generation surgical interventions. Advances in the last few decades have introduced high-tech surgical systems for less invasive and more effective surgical techniques which have increased the complexity in the OR. This demands the development of advanced systems to support clinicians across the surgical workflow inspecting the vast array of data sources that can be recorded via the information systems. To enable context-aware assistance in the OR, this thesis addresses a primary research field, i.e., automatic analysis of surgical workflows by reliable recognition of the surgical activities from endoscopic videos. A large body of research in the community tackling activity recognition has strongly focused on developing methods to

recognize activities at one specific level of granularity from video data: phases, steps, action triplets, or robotic gestures. The principal theme of this thesis is multi-level recognition of surgical activities.

First, we present the Bypass40 dataset, a new large-scale dataset of 40 videos capturing the complex LRYGB procedures. The dataset is fully annotated with activities at two levels of granularity - phase and step - by an expert surgeon. Subsequently, we propose a multi-task temporal convolution network to jointly recognize phases and steps from videos. The results demonstrated the benefits of modeling jointly the phases and steps for surgical workflow recognition. Furthermore, designing temporal models that extract useful temporal information is key to improving activity recognition. A limitation of the proposed method is that it indirectly models the hierarchical relationship between the two activities through multi-task learning. More explicit modeling of the hierarchy may be required to fully exploit this property for improving the recognition of either task. Another minor limitation of the method is that it doesn't consider the class imbalance existing in phases and steps. For instance, 'S34: jejunojejunal anastomosis reinforcement' or 'S35: staple line reinforcement' occurs in less than 5 videos in the dataset. Although the class imbalance was incorporated using class-weighted cross-entropy loss, it would be beneficial to tackle this challenge in the data. Nevertheless, this study sets a foundation for research that advances deep learning models to recognize activities at multiple levels.

The following research tackled the problem of fine-grained activity recognition with fewer annotated videos. State-of-the-art activity recognition methods heavily rely on large-scale labeled datasets as they are based on deep learning. Generating large labeled datasets is a very challenging task as it is labor-intensive, time-consuming, and expensive. To reduce this dependency, we present a weakly-supervised learning method that recognizes steps utilizing phase labels as weak signals. We introduced a step-phase dependency loss that allows using phase labels as weak signals and extensively evaluated the method on two large datasets. We demonstrated that hierarchical knowledge present in the ontology of a surgical workflow is beneficial in fine-grained activity recognition with fewer labeled data. As discussed in chapter 5, a major limitation of the proposed method is that the coarser phase labels are insufficient to distinguish between steps belonging to the same phase. Especially, when two steps in a phase perform similar actions on the same anatomy like 'horizontal stapling' and 'vertical stapling'. Nonetheless, this study is one of the few works that tackled activity recognition through weakly-supervised learning. We hope that this study fosters future research in weak supervision for surgical workflow analysis utilizing multi-level descriptions of the workflow.

Automatic recognition of surgical activities although primarily revolves around developing new deep learning models, efficient training pipeline of these models is critical for their success. Data augmentation is one component of the training pipeline that needs a closer examination as it has been shown to impact model robustness and generalizability. However, devising effective augmentation policies is mostly carried out manually and could be domain- and/or task-specific. Additionally, most of the methods have proposed

augmentation strategies for still images and extending them to videos is not straightforward as the temporal dimension needs to be considered. We address this challenge by introducing TRandAugment, a new simplified augmentation method specially designed for training Spatio-temporal models on long surgical videos. TRandAugment simplifies the search space for optimal configuration as it is parameterized with magnitude (M), the number of augments (N), and the number of temporal augments (T). The performance boost seen in the results on two large video datasets, and on two different tasks, validate the impact of the proposed method. We speculate that TRandAugment could play an impactful role in federated learning or weakly-/semi-/self-supervised learning and on different tasks such as action triplets, video segmentation, gesture recognition, etc.

In this thesis, we take a step forward to understand the generalization of activity recognition methods on data from different medical centers. To this end, we build two large video datasets, namely StrasBypass70 and BernBypass70, of 70 videos of LRYGB procedures annotated with activities at two levels: phases and steps. Next, we study the performance of both fully and weakly supervised learning methods on these datasets. The results accentuate the need for a multi-centric study of deep learning models in surgical video analysis as variations in surgical techniques and data distribution between centers could majorly impact the model's performance. One limitation of this study is that datasets from only two centers were involved. Additional study centers would have increased the variability in surgical technique and dataset distributions further emphasizing the importance of multi-centric training and evaluation.

The problem addressed in this thesis has many potential applications both in and out of the OR, such as skill assessment, surgical report generation, safety monitoring, and eventually instigating autonomous surgical systems. We hope that the methods, along with datasets of Bypass40, BernBypass70, and StrasBypass70, will influence research in the field of SDS considerably, specifically, on activity recognition.

9.2 Perspectives on Future Research

Hierarchical Recognition of Activities. We aimed to jointly recognize activities at two levels of granularity: phase and step. While the results are encouraging, it could be interesting to develop specific model components that would explicitly capture the hierarchy between the two activities. Furthermore, the hierarchy could be further extended to include gestures, action triplets, instrument detection, etc.

Weakly-supervised Learning. We tackled the problem of fine-grained step recognition using phase labels as weak signals. Future studies could extend this by using other weak signals such as remaining surgical duration or using phase and step labels as weak supervision for gestures or action triplets. Furthermore, transcripts of activities or actions could also be an interesting source of weak supervision. Ultimately, attempts could be made to auto-generate transcripts from just the surgical ontology and its mean distribution.

Few-shot Learning. We studied the generalizability of deep learning models on two large video datasets from two medical centers. While the results show that the model is capable of learning variations in surgical workflow between the centers, it could be impractical to create large labeled datasets to introduce all the variations in the domain. This calls for interesting research using the few-shot learning paradigm.

Self-supervised Learning. In this thesis, we utilize some form of supervision while developing deep learning models. Generating large labeled datasets is an important bottleneck that needs to be addressed in the future. Self-supervised that aims to learn semantically meaningful representations from unlabeled data is a key enabler to tackle the bottleneck of label generation. The representations learned through self-supervision may be valuable in analyzing activities at different granularity. A crucial aspect that must be taken into consideration while studying self-supervised learning is the temporal dimension that exists both in activities and videos. New self-supervised approaches should target deriving both spatial and temporal representations from unlabeled data.

Multi-Centric Generalization Study. Our multi-center study delivers a strong motivation to continue efforts in this direction. However, our study is limited by the fact that it involved datasets from only two centers. Additional study centers would be required to capture the variability in surgical techniques and dataset distributions. Attempts at generating even a small labeled dataset per center are necessary as this could aid in defining a standardized ontology of a workflow to represent surgical knowledge. This then enables labeled datasets from multiple centers to be used in conjunction to develop robust methods for the automatic recognition of activities.

Multi-Surgery Activity Recognition Most of the research works in the SDS community, similar to this thesis, aim to recognize activities of a single type of surgery. Although surgeries treat pathological conditions concerning various anatomical structures, subsets of surgeries that operate in similar regions such as laparoscopic surgeries can be clubbed together. Interesting research works could undertake using these subsets of surgeries constructively to learn effective surgical knowledge.

List of Publications

International journals

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy, *Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures*, International Journal of Computer Assisted Radiology and Surgery (IJCARS), 16, 1111–1119, May 2021.

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy, *TRandAugment: temporal random augmentation strategy for surgical activity recognition from videos*, International Journal of Computer Assisted Radiology and Surgery, March 2023.

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy, *Weakly Supervised Temporal Convolutional Networks for Fine-grained Surgical Activity Recognition*, IEEE Transactions on Medical Imaging, pages 1–1, 2023.

International conferences with proceedings

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy, *Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures*, International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), 2021. **Long oral presentation**.

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy, *TRandAugment: temporal random augmentation strategy for surgical activity recognition from videos*, Accepted

at International Conference on Computer-Assisted Radiology and Surgery (CARS), 2023. **Long oral presentation**.

Others

Marco Bombieri, Diego Dall'Alba, Sanat Ramesh, Giovanni Menegozzo, Caitlin Schneider, and Paolo Fiorini, *Joints-space metrics for automatic robotic surgical gestures classification*, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3061-3066, 2020. **Long oral presentation**.

Martina Finocchiaro, Xuan Thao Ha, Jorge Lazo, Lai Chun-Feng, Sanat Ramesh, Albert Hernansanz, Gianni Borghesan, Diego Dall'Alba, Selene Tognarelli, Benoit Rosa, Alicia Casals, Nicolas Padoy, Paolo Fiorini, Dankelman Jenny, Emmanuel Vander Porten, Arianna Menciassi, and Elena De Momi, *Multi-level-assistance Robotic Platform for Navigation in the Urinary System: Design and Preliminary Tests*, Proceeding of the 11th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery, 90-91, 2022.

Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, Georgios Exarchakis, Alexandros Karargyris, and Nicolas Padoy, *Dissecting Self-Supervised Learning Methods for Surgical Computer Vision*, Submitted to Medical Image Analysis (MedIA), 2022. arXiv preprint arXiv:2207.00449.

A Résumé de thèse en français

Chapter Summary

A.1 Introduction

La chirurgie est un domaine spécialisé de la médecine qui s'attache à traiter des états pathologiques tels que des maladies ou des blessures en utilisant des techniques opératoires manuelles et instrumentales sur une personne. Au fil des siècles, les inventions et innovations constantes dans divers domaines scientifiques ont transformé la chirurgie jusqu'à aujourd'hui. Cette transformation de la chirurgie peut être observée à travers les progrès de la salle d'opération moderne (OR). La Figure A.1 donne un aperçu de la transition de la salle d'opération. Au cours des dernières décennies, les développements se sont concentrés sur le passage de la chirurgie ouverte traditionnelle à la chirurgie mini-

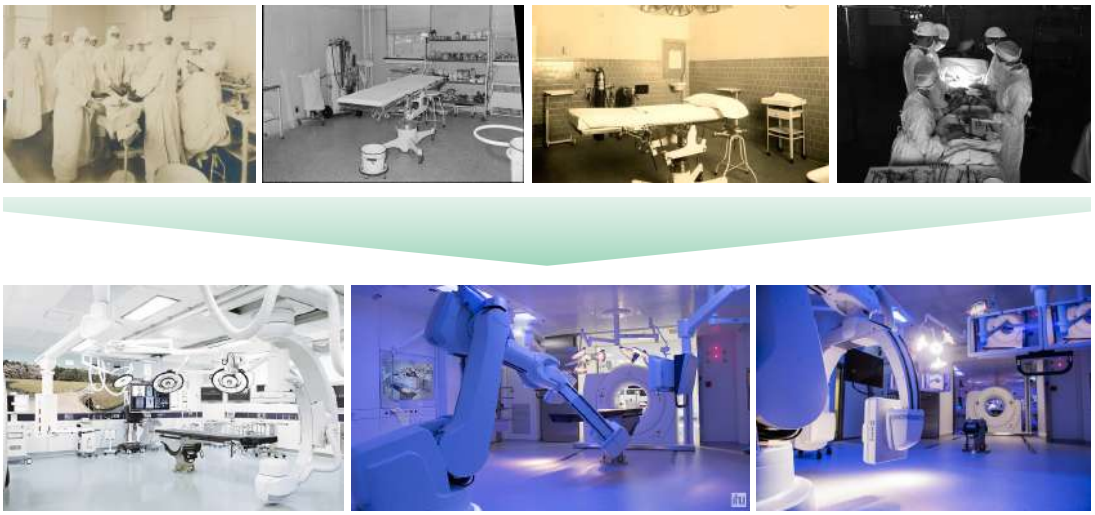


Figure A.1: La transformation du bloc opératoire au cours des derniers siècles.

invasive. Ce changement est motivé par les avantages que la chirurgie mini-invasive procure aux patients: moins de douleur, un temps de récupération plus court et moins de complications. Malgré ces avantages, les progrès de la chirurgie mini-invasive présentent de nouveaux défis, tels que des courbes d'apprentissage abruptes pour les nouveaux membres du personnel péri-opératoire et les chirurgiens, une vue limitée de l'anatomie, une durée plus longue de certaines procédures par rapport à la chirurgie ouverte, une amplitude de mouvement limitée des instruments, des entrées sensorielles limitées en termes de profondeur et de toucher, etc. Pour relever certains de ces défis, la communauté a proposé le développement de systèmes conscients du contexte (CAS) qui visent à fournir une aide contextuelle aux cliniciens en exploitant les diverses informations sensorielles disponibles dans la salle d'opération [Bricon-Souf 2007, Kranzfelder 2012, Maier-Hein 2017, Vercauteren 2020].

L'essor de l'intervention assistée par ordinateur et de la chirurgie assistée par robot (SAR) a également accru la complexité de l'exécution des procédures chirurgicales, ce qui confirme la nécessité des CAS en raison de leurs avantages potentiels. L'assistance contextuelle d'un CAS pourrait contribuer à simplifier les flux de travail chirurgicaux, à améliorer les communications homme-machine et à accélérer l'exécution des manœuvres chirurgicales, ce qui permettrait de réduire la charge de travail et la tension chirurgicales, et donc de réduire les erreurs chirurgicales, d'accroître la sécurité des patients et d'améliorer la sécurité, la qualité et l'efficacité globales des soins [Maier-Hein 2017, Vercauteren 2020]. **Pour concevoir un système CAS efficace, cette thèse met l'accent sur l'un de ses composants clés, à savoir l'analyse automatique d'un flux de travail chirurgical.** L'analyse automatique des flux de travail chirurgicaux est réalisée par une reconnaissance fiable des activités chirurgicales [Kranzfelder 2012]. En examinant les données en ligne complètes de la salle d'opération, si les systèmes pouvaient reconnaître l'état actuel de la procédure, ils pourraient également être capables

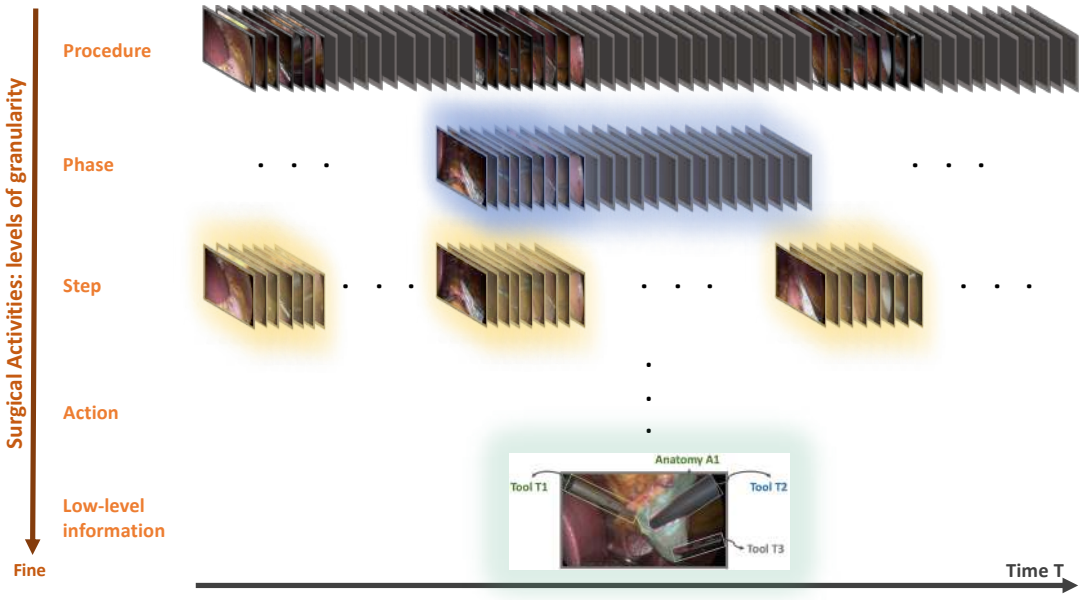


Figure A.2: Types d'activités chirurgicales en fonction du niveau de granularité.

de prédire la progression de la procédure. Cette capacité pourrait fournir un soutien actif aux chirurgiens en les aidant dans leur prise de décision clinique, ce qui pourrait successivement induire une autonomie dans les SAR. En outre, la compréhension des flux de travail permet à ces systèmes de générer automatiquement des rapports chirurgicaux et d'annoter les données de manière appropriée pour une rétrospection sans effort. Ces informations sémantiques sont au cœur de la compréhension cognitive de la chirurgie.

Selon le niveau de granularité, une procédure chirurgicale peut être décomposée en activités, telles que l'ensemble de la procédure, les phases, les étapes et les actions (Figure A.2) [Katić 2015, Meireles 2021]. La reconnaissance automatisée des phases a reçu beaucoup d'attention et constitue un domaine de recherche très actif dans la communauté de la vision chirurgicale [Garrow 2020, Demir 2022]. Parallèlement aux phases, d'importantes recherches ont été menées sur des activités à grain fin telles que les gestes robotiques [van Amsterdam 2021], les triplets d'action [Nwoye 2020, Nwoye 2022b, Sharma 2022], et la détection et le suivi d'instruments [Hajj 2018, Nwoye 2019, Jin 2020]. Récemment, un grand nombre de travaux de recherche se sont concentrés en particulier sur la reconnaissance des pas [Charrière 2014, Charrière 2017]. Cependant, ils ont tous fait l'objet de recherches indépendantes et très peu de travaux tentent de reconnaître des activités à plusieurs niveaux. Par conséquent, nous visons à reconnaître deux types d'activités à différents niveaux de granularité, c'est-à-dire les phases et les étapes. Plus précisément, nous nous concentrons sur la reconnaissance d'activités à plusieurs niveaux pour analyser une autre procédure à fort volume, à savoir le bypass gastrique, qui est assez intéressant en raison de la complexité de son flux de travail. Le bypass gastrique est une procédure visant à traiter l'obésité, considérée comme une épidémie mondiale par l'Organisation mondiale de la santé [on Obesity 2000], avec environ 500,000 procédures

bariatriques laparoscopiques réalisées chaque année dans le monde [Angrisani 2015]. La dérivation gastrique Roux-En-Y par laparoscopie (LRYGB), l'intervention de chirurgie bariatrique la plus pratiquée et la plus courante [Angrisani 2015], consiste à réduire l'estomac et à contourner une partie de l'intestin grêle.

Dans cette thèse, nous introduisons d'abord une représentation hiérarchique, similaire à [Kaijser 2018], pour la procédure LRYGB contenant des phases et des étapes représentant le workflow réalisé dans notre hôpital et visons à reconnaître ces deux types d'activités. À cet effet, nous construisons un nouveau jeu de données à grande échelle, appelé Bypass40, contenant 40 vidéos endoscopiques de procédures chirurgicales de bypass gastrique et annotons les vidéos avec la représentation hiérarchique introduite des phases et des étapes. Ensuite, nous introduisons Multi-Task Multi-Stage Temporal Convolutional Networks (MTMS-TCN) pour la reconnaissance conjointe des phases et des étapes, en étendant les MS-TCN qui ont été proposés pour la segmentation des actions. La motivation de cette méthode vient du fait que les activités et les vidéos contiennent une dimension temporelle inhérente, en plus du contenu spatial, qui nécessite des modèles temporels pour extraire des informations utiles.

Bien que le MTMS-TCN montre un grand potentiel dans la reconnaissance des deux types d'activités, le goulot d'étranglement du MTMS-TCN, et des méthodes similaires [Garrow 2020, Demir 2022], est le besoin de grands ensembles de données entièrement annotées pour l'entraînement des modèles d'apprentissage profond. La constitution de ces ensembles de données annotées est difficile et prend du temps, car ces tâches nécessitent des connaissances médicales spécifiques à un domaine. Pour résoudre ce problème, la deuxième contribution porte sur l'apprentissage faiblement supervisé pour les activités à grain fin, c'est-à-dire la reconnaissance des pas. Nous exploitons les relations hiérarchiques étape-phase et utilisons des annotations de phase faibles plus faciles à annoter sur des vidéos où les annotations d'étape sont manquantes. Nous introduisons une nouvelle perte de dépendance pour renforcer la supervision faible et encoder la relation hiérarchique étape-phase sous forme de matrice. En optimisant cette perte, nous encourageons le modèle à apprendre les séquences d'étapes et les transitions possibles à partir de vidéos contenant uniquement des annotations de phase.

Inspirés par le succès des deux méthodes précédentes, notre troisième contribution consiste à examiner l'un des composants les plus essentiels du pipeline de formation de ces méthodes d'apprentissage profond: L'augmentation des données. L'augmentation des données est une méthode couramment utilisée pour générer des données supplémentaires afin d'améliorer la formation des modèles d'apprentissage profond à forte intensité de données pour la classification d'images [DeVries 2017a, Cubuk 2019, Lim 2019], la détection d'objets [Girshick 2018], la segmentation d'instances [Fang 2019], etc. De plus, il a été démontré que l'augmentation a un impact sur la robustesse du modèle [Lopes 2019] et sur les performances des méthodes d'apprentissage semi-supervisées et auto-supervisées [Qian 2021, Pan 2021, Shi 2021]. Cependant, des politiques d'augmentation spécifiques doivent être conçues pour capturer les connaissances préalables pour chaque domaine, ce qui nécessite une expertise et un travail manuel, rendant les méthodes d'augmentation

des données difficiles à étendre à d'autres domaines et applications [Cubuk 2019,Lim 2019, Cubuk 2020]. De plus, puisque les méthodes de reconnaissance d'activité chirurgicale travaillent avec de longues vidéos chirurgicales, les vidéos ajoutent une dimension temporelle aux images qui doit être prise en compte lors de la conception des politiques d'augmentation. Pour répondre à ce besoin, nous introduisons une nouvelle méthode d'augmentation simplifiée, appelée Temporal Random Augmentations (TRandAugment), spécialement conçue pour l'entraînement de modèles spatio-temporels sur de longues vidéos chirurgicales. TRandAugment est largement évalué sur la tâche de reconnaissance de l'activité chirurgicale à deux niveaux de granularité, à savoir la phase et le pas.

Pour réaliser l'application de la reconnaissance d'activité dans le OR via le CAS, le module de reconnaissance doit posséder les caractéristiques de fiabilité, de portabilité et d'intégrité. Ces trois caractéristiques indiquent de manière générale que le module de reconnaissance doit pouvoir être utilisé en toute sécurité et fonctionner de manière constante dans différentes conditions de travail pendant une période donnée. Étant donné que les dernières méthodes de reconnaissance sont basées sur l'apprentissage profond, ces caractéristiques sont étudiées en termes de robustesse et de généralisation. L'un des principaux défis pour développer des méthodes d'apprentissage profond robustes et généralisables est leur susceptibilité à l'overfitting et à la mémorisation en raison de la complexité du nombre de paramètres impliqués [Geirhos 2018, Feng 2019]. L'ajout de données d'entraînement supplémentaires est une approche coûteuse pour lutter contre l'overfitting. Néanmoins, c'est un moyen crucial d'ajouter des variations naturelles d'un domaine à un ensemble de données. Les variations présentes dans le domaine chirurgical sont dues aux changements dans le flux de travail chirurgical entre les chirurgiens, les centres médicaux, les communautés, les nations, etc. Par conséquent, un module idéal de reconnaissance de l'activité chirurgicale doit être robuste et capable de faire face à toutes ces variations de l'anatomie et du flux de travail. Dans notre quatrième et dernière contribution, nous présentons une étude sur la généralisation des méthodes de reconnaissance d'activité sur des données provenant de différents centres médicaux. Dans le cadre de cette étude, nous introduisons deux nouveaux jeux de données, à savoir StraBypass70 et BernBypass70, chacun composé de 70 vidéos de procédures LRYGB entièrement annotées avec des étiquettes de phase et d'étape. Par la suite, nous étudions la performance des méthodes d'apprentissage entièrement et faiblement supervisées sur ces ensembles de données, démontrant à la communauté les défis et les lacunes lors de la transition de la recherche à la traduction clinique.

A.2 Contribution

A.2.1 Bypass40 Dataset

Nous introduisons deux activités chirurgicales hiérarchiquement définies appelées phases et étapes pour la procédure LRYGB. Ces deux éléments définissent le déroulement de la chirurgie à deux niveaux de granularité, les phases décrivant le déroulement de la

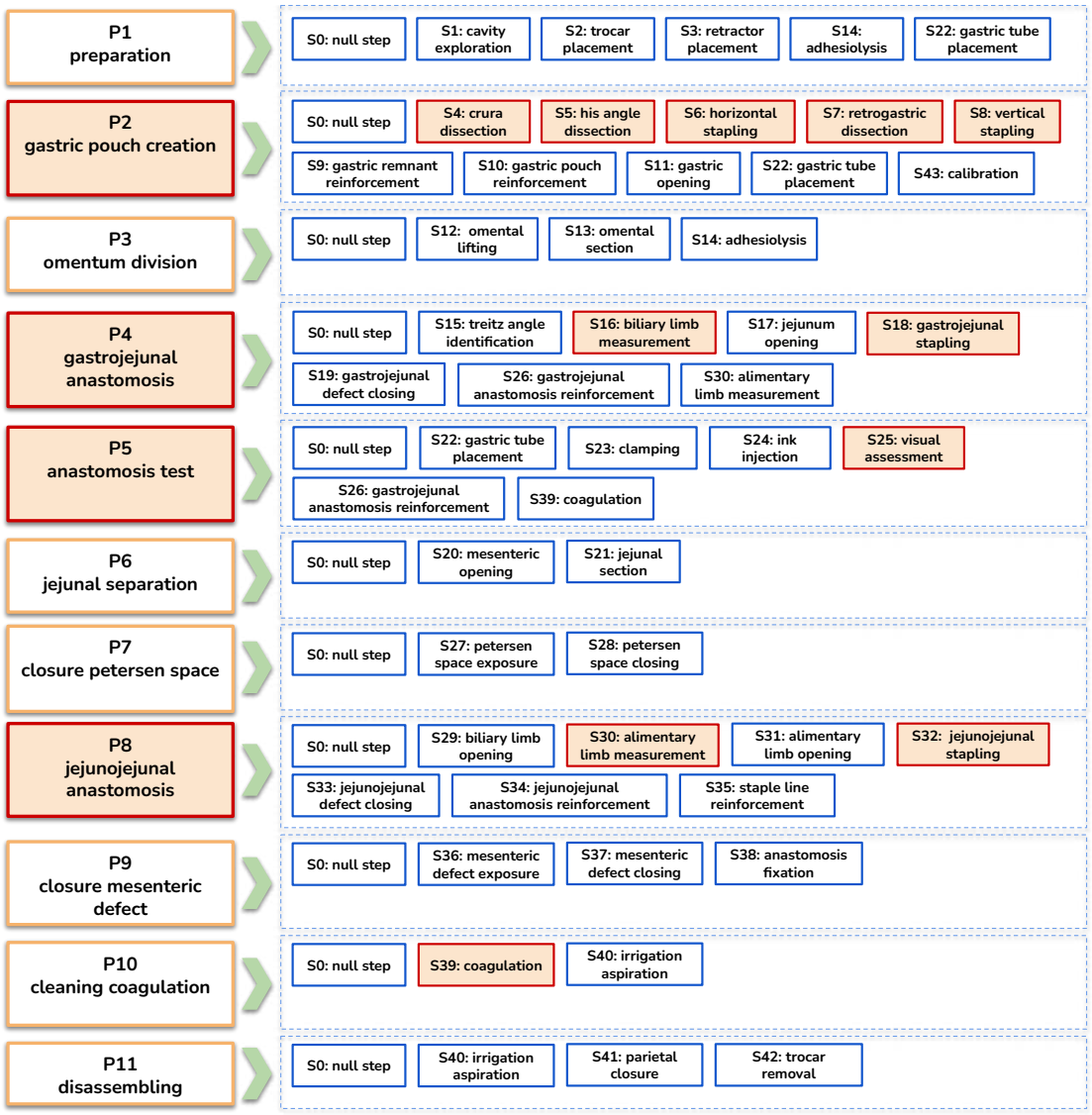


Figure A.3: Liste de toutes les phases et étapes définies dans l'ensemble de données avec leur relation hiérarchique. Les activités chirurgicales critiques sont surlignées en rouge.

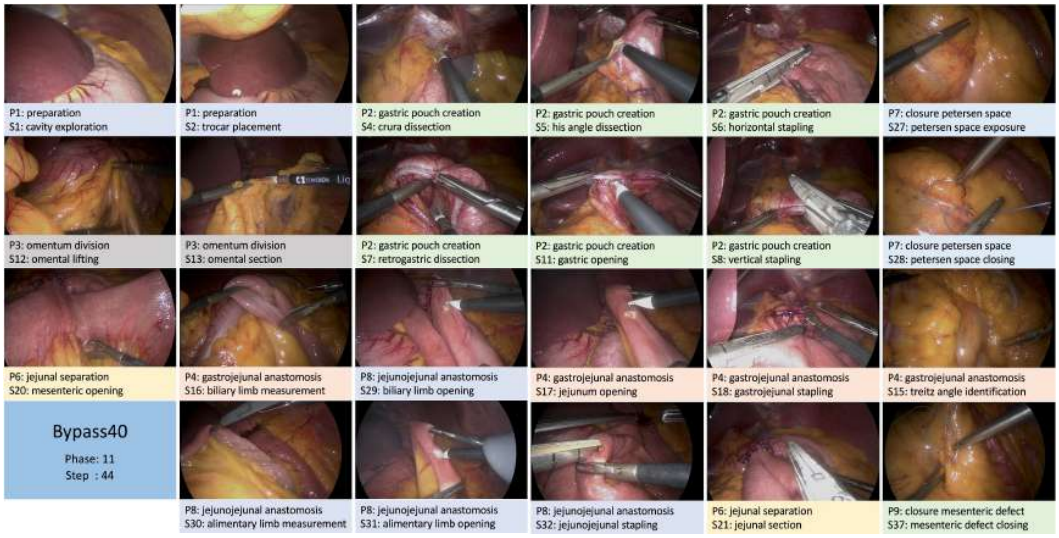


Figure A.4: Exemples d'images du jeu de données Bypass40.

chirurgie à un niveau plus grossier que les étapes. Les phases décrivent un ensemble d'objectifs chirurgicaux fondamentaux à atteindre afin de mener à bien la procédure chirurgicale, tandis que les étapes décrivent un ensemble d'actions chirurgicales à réaliser afin d'accomplir une phase chirurgicale. La procédure chirurgicale est segmentée en 44 étapes à grain fin, ainsi qu'en 11 phases plus grossières. Toutes les phases et étapes sont présentées dans la Figure A.3 et quelques exemples d'images tirées de l'ensemble de données sont présentés dans la Figure A.4. Ces deux types d'activités sont intéressants pour leur relation hiérarchique inhérente, qui est illustrée dans la figure. En outre, la figure met en évidence toutes les phases critiques, et les étapes critiques correspondantes, qui sont cliniquement connues pour être importantes pour les résultats chirurgicaux [Birkmeyer 2013].

Bypass40 est un jeu de données composé de 40 vidéos capturées par des caméras endoscopiques pendant des procédures de dérivation gastrique. Le jeu de données est entièrement annoté avec des étiquettes de segmentation d'activité pour les phases et les étapes. La reconnaissance de ces deux ensembles d'activités est essentielle pour la prise de décision et la navigation autonome des robots chirurgicaux. L'analyse du flux de travail de cette procédure est assez difficile en raison de multiples facteurs: longue durée des vidéos, fumée, sang et autres anomalies. En outre, les similitudes entre les phases et les fortes similitudes entre les étapes augmentent la complexité du problème de la reconnaissance des activités, ce qui entraîne une diminution des performances et une généralisation limitée des méthodes existantes. Il est donc nécessaire d'élaborer de nouvelles méthodes qui nous permettent de saisir les dépendances temporelles à long terme et de résoudre les ambiguïtés entre les phases et les étapes.

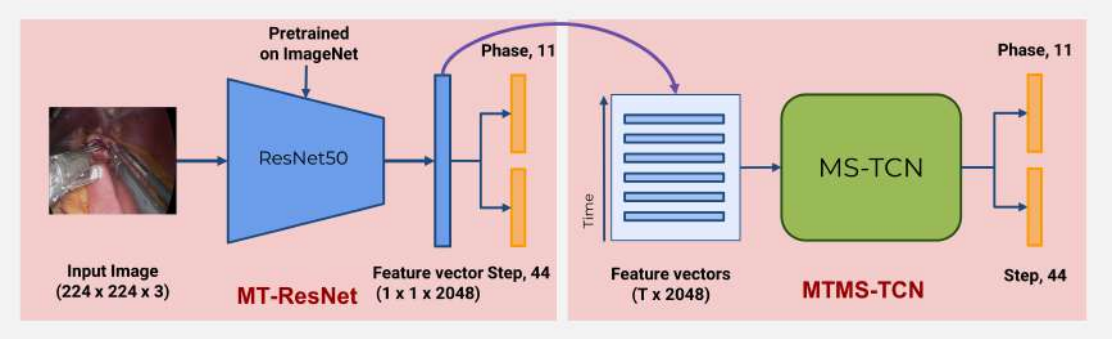


Figure A.5: Vue d'ensemble de la configuration de notre modèle. L'architecture multi-tâches de l'épine dorsale de l'extracteur de caractéristiques ResNet-50 à gauche et la configuration multi-tâches du modèle temporel TCN à droite.

Table A.1: Comparaison de base sur l'ensemble de données pour la reconnaissance conjointe des phases et des pas. Accuracy (ACC) est indiqué après une validation croisée 4 fois.

	Models	Phase ACC	Step ACC	Phase-Step ACC
No TCN	ResNet	82.1 ± 3.3	65.5 ± 2.0	54.9 ± 2.6
	MT-ResNet	81.7 ± 2.7	66.6 ± 2.4	64.8 ± 2.0
	ResNetLSTM	**89.1 ± 2.8**	71.3 ± 2.3	68.5 ± 2.3
	MT-ResNetLSTM	88.6 ± 2.7	**72.2 ± 2.0**	**70.7 ± 1.9**
Stage I	TeCNO	89.8 ± 3.5	75.1 ± 2.4	72.3 ± 3.0
	MTMS-TCN	**91.2 ± 2.9**	**76.1 ± 2.7**	**75.1 ± 2.8**
Stage II	TeCNO	89.9 ± 3.3	74.8 ± 2.5	71.9 ± 2.7
	MTMS-TCN	**90.9 ± 3.2**	**75.5 ± 3.1**	**75.1 ± 2.8**

A.2.2 Phase chirurgicale conjointe et reconnaissance des étapes

L'objectif de ce travail est la reconnaissance conjointe en ligne des phases et étapes chirurgicales. Nous proposons un pipeline de reconnaissance en ligne de l'activité chirurgicale comprenant les étapes suivantes: 1) Un ResNet-50 multi-tâches est utilisé comme extracteur de caractéristiques visuelles. 2) Un modèle de réseau convolutif temporel multi-tâches et multi-étapes (MTMS-TCN) affine la caractéristique extraite de l'image actuelle en encodant les informations temporelles déduites de l'analyse de l'historique. L'aperçu de la configuration du modèle est représenté sur la Figure A.5.

Comme la durée moyenne d'une opération chirurgicale peut varier de moins d'une demi-heure à plusieurs heures, il est difficile pour les modèles basés sur les LSTM d'exploiter les informations temporelles pour la reconnaissance de l'activité chirurgicale. Cela nous motive à explorer l'utilisation de réseaux de convolutions temporelles en raison de leur grand champ réceptif pour une résolution temporelle plus élevée. D'autre part, puisque la phase et l'étape capturent la même information à différents niveaux de granularité, l'étape multi-tâche exploite l'aspect complémentaire de ces activités pour mieux reconnaître les deux activités.

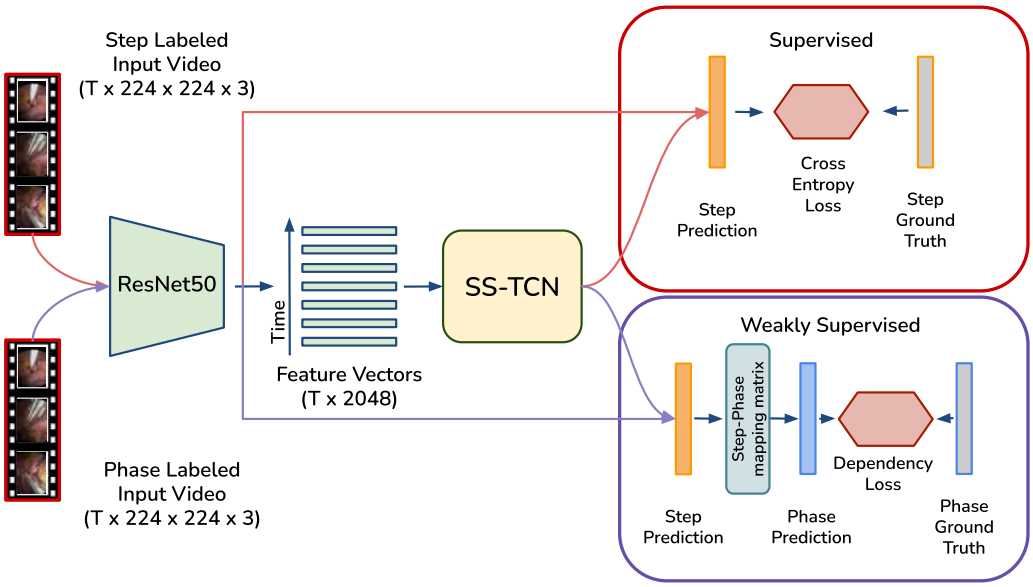


Figure A.6: Vue d'ensemble de notre configuration de modèle spatio-temporel de bout en bout: ResNet50 + SS-TCN (Single-Stage Temporal Convolutional Networks). Lorsque des étiquettes de phase sont disponibles, le modèle est entraîné par la voie supervisée (rouge) et la voie faiblement supervisée (violet) en utilisant les étiquettes de phase. Le modèle est entraîné de bout en bout en une seule étape d'apprentissage.

Nous évaluons efficacement notre méthode en comparant différents modèles (CNN, CNN + LSTM, CNN + TCN) dans des configurations mono-tâche et multi-tâches. Nos résultats (Tableau A.1) confirment notre hypothèse: 1) les TCN sont capables de capturer plus d'informations temporelles par rapport aux LSTM ; 2) une configuration multi-tâches permet de capturer des informations mutuelles et est en outre capable de reconnaître plus précisément les phases et les étapes.

A.2.3 Reconnaissance de l'activité chirurgicale à grain fin faiblement supervisée

Dans ce travail, nous proposons un apprentissage faiblement supervisé pour la tâche d'activité chirurgicale à grain fin, c'est-à-dire la reconnaissance. La principale motivation de la supervision faible est que la construction d'ensembles de données à grande échelle avec des annotations fines est extrêmement fastidieuse et nécessite un effort important pour valider la qualité des annotations. Nous utilisons des étiquettes de phase plus faciles à annoter comme supervision faible pour la reconnaissance des étapes. Puisque la phase et l'étape sont des activités définies à différents niveaux de granularité, notre approche exploite la relation hiérarchique étape-phase pour une supervision faible avec une fraction du jeu de données contenant des annotations d'étape. La Figure A.6 donne un aperçu de la configuration du modèle.

Nous introduisons une perte de dépendance pour modéliser la supervision faible en utilisant une matrice de correspondance étape-phase qui modélise la relation hiérarchique

Table A.2: Bypass40: Effet d'une supervision faible sur une quantité variable de vidéos étiquetées par étapes. Accuracy (ACC), Precision (PR), Recall (RE), et F1-score (F1) (%) sont rapportés. 'FSA' désigne le modèle entraîné pour la reconnaissance des étapes sans aucune annotation de phase. 'DEP' désigne la perte de dépendance ajoutée pour la supervision faible en utilisant les étiquettes de phase sur les vidéos restantes.

Model	# Videos Step	# Videos Phase	ACC	PR	RE	F1
FSA	3 (12%)	-	45.02 ± 9.96	26.62 ± 5.32	21.87 ± 4.70	19.44 ± 5.31
[Yu 2019]	3 (12%)	-	43.27 ± 11.8	23.63 ± 4.41	23.91 ± 5.71	19.77 ± 4.89
DEP	3 (12%)	21	$\mathbf{57.20 \pm 8.31}$	$\mathbf{33.44 \pm 6.04}$	$\mathbf{33.16 \pm 6.37}$	$\mathbf{29.38 \pm 6.11}$
FSA	6 (25%)	-	59.80 ± 10.17	37.19 ± 8.52	35.93 ± 7.31	32.15 ± 8.03
[Yu 2019]	6 (25%)	-	62.55 ± 10.09	40.63 ± 7.85	43.71 ± 8.35	37.68 ± 8.54
DEP	6 (25%)	18	$\mathbf{68.03 \pm 9.04}$	$\mathbf{50.05 \pm 6.82}$	$\mathbf{45.86 \pm 6.46}$	$\mathbf{42.05 \pm 7.44}$
FSA	12 (50%)	-	68.26 ± 8.31	47.57 ± 7.84	44.74 ± 7.59	41.30 ± 8.44
[Yu 2019]	12 (50%)	-	67.89 ± 11.04	46.26 ± 9.97	50.11 ± 8.20	43.41 ± 10.33
DEP	12 (50%)	12	$\mathbf{73.43 \pm 8.43}$	$\mathbf{53.40 \pm 7.43}$	$\mathbf{51.19 \pm 8.20}$	$\mathbf{48.34 \pm 8.85}$
FSA	18 (75%)	-	72.82 ± 6.76	50.60 ± 7.90	48.98 ± 8.33	46.08 ± 8.61
[Yu 2019]	18 (75%)	-	73.33 ± 10.15	$\mathbf{54.78 \pm 11.05}$	$\mathbf{57.21 \pm 8.51}$	$\mathbf{51.72 \pm 10.59}$
DEP	18 (75%)	6	$\mathbf{73.88 \pm 8.11}$	54.33 ± 6.38	51.79 ± 7.10	48.62 ± 7.49
FSA	24 (100%)	-	76.12 ± 7.39	54.23 ± 8.24	50.94 ± 7.53	48.17 ± 8.02

Table A.3: Bypass40: Effet du nombre de vidéos annotées par phase pour la reconnaissance des pas en utilisant la perte 'DEP' pour une supervision faible. Accuracy (ACC), Precision (PR), Recall (RE), et F1-score (F1) (%) sont rapportés pour des configurations avec 6, 12, et 24 vidéos entièrement annotées avec des étapes.

Model	# Videos Step	# Videos Phase	ACC	PR	RE	F1
FSA	6	-	59.80	37.19	35.93	32.15
DEP	6	3	62.15	40.48	37.15	33.48
DEP	6	6	67.94	46.17	42.61	39.67
DEP	6	12	68.07	47.18	43.18	40.42
DEP	6	18	68.03	50.05	45.86	42.05
FSA	12	-	68.26	47.57	44.74	41.30
DEP	12	3	72.79	50.10	48.39	45.06
DEP	12	6	72.43	53.02	51.20	47.26
DEP	12	12	73.43	53.40	51.19	48.34
FSA	24	-	76.12	54.23	50.94	48.17

Table A.4: CATARACTS: Effet d'une supervision faible sur une quantité variable de vidéos étiquetées par étapes. Accuracy (ACC), Precision (PR), Recall (RE), et F1-score (F1) (%) sont rapportés. 'FSA' désigne le modèle entraîné pour la reconnaissance des étapes sans aucune annotation de phase. 'DEP' désigne la perte de dépendance ajoutée pour la supervision faible en utilisant les étiquettes de phase sur les vidéos restantes.

Model	# Videos Step	# Videos Phase	ACC	PR	RE	F1
FSA	3 (12%)	-	48.47 ± 10.62	51.32 ± 11.91	37.44 ± 9.85	37.12 ± 10.15
[Yu 2019]	3 (12%)	-	59.61 ± 10.67	56.02 ± 14.31	**61.82 ± 14.45**	53.26 ± 13.61
DEP	3 (12%)	22	**66.78 ± 12.21**	**64.29 ± 12.50**	59.73 ± 11.93	**58.31 ± 12.73**
FSA	6 (25%)	-	69.51 ± 11.16	71.05 ± 14.13	56.70 ± 12.67	59.28 ± 13.50
[Yu 2019]	6 (25%)	-	74.62 ± 8.22	67.71 ± 11.48	**75.93 ± 12.48**	67.67 ± 12.46
DEP	6 (25%)	19	**75.28 ± 11.50**	**71.84 ± 14.30**	69.19 ± 12.72	**68.09 ± 13.97**
FSA	12 (50%)	-	78.02 ± 9.05	79.02 ± 13.20	69.55 ± 12.04	71.18 ± 13.04
[Yu 2019]	12 (50%)	-	77.84 ± 12.55	71.48 ± 13.41	**79.92 ± 15.28**	72.96 ± 14.46
DEP	12 (50%)	13	**79.94 ± 9.17**	**80.52 ± 12.93**	72.62 ± 11.91	**73.52 ± 13.29**
FSA	18 (75%)	-	82.5 ± 8.07	**82.58 ± 11.91**	76.05 ± 11.62	77.39 ± 12.12
[Yu 2019]	18 (75%)	-	78.59 ± 10.71	74.55 ± 14.17	**78.16 ± 12.64**	73.55 ± 13.67
DEP	18 (75%)	7	**82.64 ± 9.72**	82.20 ± 13.70	77.32 ± 12.70	**77.67 ± 13.56**
FSA	25 (100%)	-	83.37 ± 9.50	85.29 ± 12.05	78.96 ± 11.93	80.09 ± 13.34

entre eux. Conformément à nos travaux précédents, notre modèle se compose d'un réseau convolutif temporel à un seul étage (SS-TCN) pour la modélisation temporelle et de ResNet-50 comme colonne vertébrale. Nous formons le modèle de bout en bout en utilisant la perte de dépendance que nous proposons. Pour démontrer l'efficacité de notre méthode, nous réalisons des études d'ablation avec une configuration différente du modèle et de la formation. Avec n vidéos dans le jeu de données, dont k sont annotées avec des étapes et $n-k$ sont faiblement annotées avec des phases, d'abord, nous entraînons le modèle de base sur k vidéos dans une approche entièrement supervisée pour la tâche de reconnaissance des étapes et le comparons avec notre modèle proposé avec la perte de dépendance. De plus, nous analysons l'influence du nombre de vidéos supplémentaires avec des étiquettes de phase sur la performance du modèle en fixant k vidéos avec des annotations de pas et en variant le nombre de vidéos avec des annotations de phase de 0 à $n-k$. Les différentes expériences menées sur l'ensemble de données Bypass40 sont présentées dans les Tableaux A.2 & A.3.

Nos résultats montrent une amélioration de 5 à 13% de notre méthode par rapport à la ligne de base avec moins de 50% de l'ensemble de données annotées avec des étapes. Il est intéressant de noter que notre méthode, entraînée sur un ensemble de données comportant 50% de vidéos annotées de pas et 50% de vidéos annotées de phases, atteint des performances proches de celles du modèle de base supérieur entraîné sur l'ensemble de l'ensemble de données entièrement étiqueté avec des annotations de pas. Une expérimentation supplémentaire sur l'ensemble de données CATARACTS (Tableaux A.4) permet d'obtenir des performances similaires à celles de Bypass40, ce qui confirme notre hypothèse.

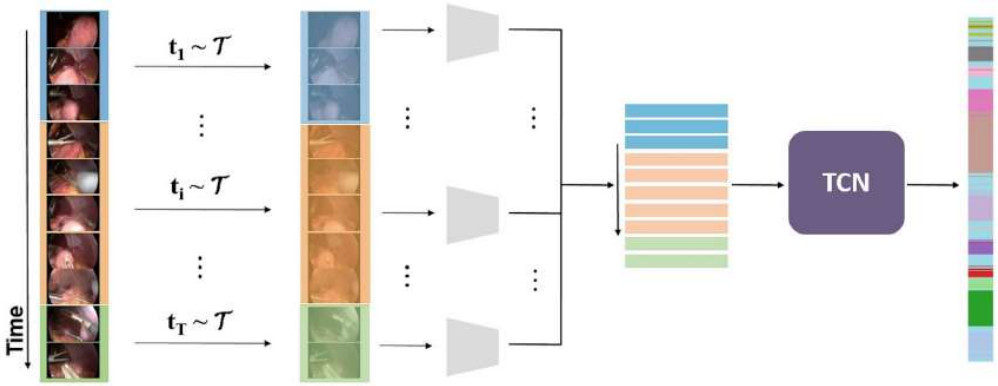


Figure A.7: Représentation pictographique de TRandAugment. Une vidéo est segmentée en T clips et une augmentation aléatoire t_i, échantillonnée à partir d'une liste de transformations τ, est appliquée au clip i. Les clips augmentés sont fusionnés pour former une nouvelle vidéo qui est transmise comme entrée lors de l'entraînement d'un réseau CNN+TCN de bout en bout qui prédit les phases ou les étapes.

.

A.2.4 Augmentations aléatoires temporelles pour la reconnaissance de l'activité chirurgicale

Ce travail introduit une nouvelle méthode d'augmentation des données pour les vidéos, appelée TRandAugment, qui est une technique simplifiée et automatisée inexplorée dans la littérature. TRandAugment s'appuie sur la méthode précédente (RandAugment) mais incorpore la dimension temporelle supplémentaire pour la tâche de reconnaissance de l'activité chirurgicale à deux niveaux de granularité, c'est-à-dire la phase et l'étape. Une représentation graphique de la méthode d'augmentation est présentée dans la Figure A.7.

TRandAugment, est conçu avec trois paramètres (M, N, T) incluant le paramètre supplémentaire T pour caractériser la dimension temporelle des vidéos. Une liste de $\|\tau\|= 9$ transformations est utilisée et appliquée avec une probabilité uniforme de $\frac{1}{\|\tau\|}$. L'idée de TRandAugment est d'appliquer différentes transformations à différents segments vidéo temporels. Ainsi, le paramètre T est introduit pour contrôler le nombre de segments temporels. Chaque vidéo est divisée en segments aléatoires $T' \in [1, T]$ et pour chaque segment i ($i \in [1, T']$), une transformation aléatoire $t_i \sim \tau$ est appliquée uniformément sur toutes les images de ce segment. M et N contrôlent la magnitude et le nombre de transformations appliquées à chaque segment.

Les résultats du Tableau A.5 montrent que nous pouvons trouver une bonne stratégie d'augmentation par une simple recherche de grille sur les paramètres de TRandAugment. De plus, nous montrons que les augmentations cohérentes dans le temps sont très bénéfiques lors de l'entraînement de modèles spatio-temporels sur de longs ensembles de données vidéo abordés en vision artificielle chirurgicale. TRandAugment, avec les meilleures valeurs (M, N, T), permet d'obtenir une amélioration de 1-10% par rapport aux méthodes existantes sur deux tâches (reconnaissance de phases et de pas) et sur deux jeux de données (Bypass40 et CATARACTS).

Table A.5: Comparaison de différentes méthodes sur les ensembles de données Bypass40 (BY40) et CATARACTS (CA50). * indique les modèles formés dans une configuration multitâche nécessitant des étiquettes de phase/étape supplémentaires.

Dataset Task	Method	$\|\tau'\|$	M, N, T	ACC	PR	RE	F1
CA50 Step	Custom	-	-, -, -	81.79±12.30	77.82±13.61	82.25±14.69	78.21±14.90
	RA	3	30, 1, -	80.45±10.33	76.48±13.00	81.34±13.56	76.87±14.01
	URA (ours)	10	30, 1, -	83.24±10.64	77.04±14.20	82.33±14.68	78.02±14.98
	TRA (ours)	10	30, 1, 5	**83.64±10.67**	**78.38±14.11**	**84.06±14.18**	**79.43±15.09**
BY40 Phase	Custom*	-	-, -, -	90.26 ± 6.44	84.74 ± 7.71	81.75 ± 9.12	81.31 ± 9.07
	URA (ours)	10	30, 3, -	**93.55 ± 3.24**	83.25 ± 7.80	86.07 ± 7.61	83.51 ± 7.93
	TRA (ours)	10	30, 2, 5	93.17 ± 4.27	**86.42 ± 8.50**	**86.70 ± 6.72**	**85.20 ± 8.40**
BY40 Step	Custom*	-	-, -, -	75.46 ± 9.34	55.58 ± 9.88	52.78 ± 9.22	50.35 ± 9.75
	URA (ours)	10	30, 2, -	80.55 ± 6.61	61.32 ± 8.11	62.13 ± 7.74	58.52 ± 8.46
	TRA (ours)	10	30, 2, 5	**80.80 ± 7.90**	**63.66 ± 9.08**	**63.94 ± 8.31**	**60.06 ± 9.22**

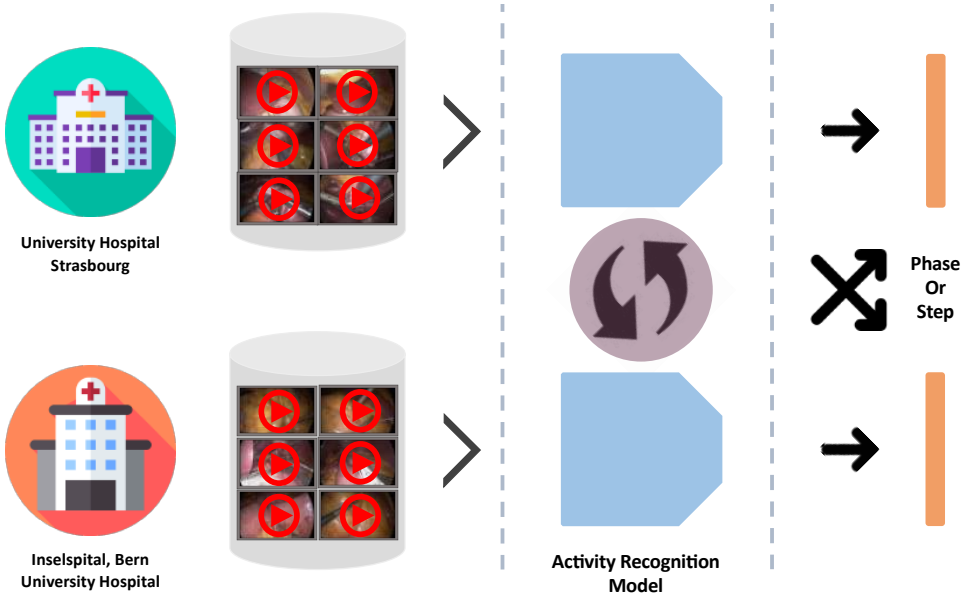


Figure A.8: Mise en place d'une étude inter-centres sur les modèles de reconnaissance de l'activité.

A.2.5 Etude de généralisation inter-centres

Pour l'objectif de cette étude, deux ensembles de données provenant de deux hôpitaux universitaires ont été créés. 1) StraBypass70 est un jeu de données qui étend Bypass40, composé de 70 vidéos LRYGB de l'hôpital universitaire de Strasbourg, France. 2) BernBypass70 est un jeu de données composé de 70 vidéos LRYGB réalisées à l'Inselspital, hôpital universitaire de Berne, Suisse. d. Dans cette étude, s'ils sont utilisés en combinaison, les jeux de données BernBypass70 et StraBypass70 sont appelés MultiBypass140.

Un réseau convolutif temporel multi-tâches et multi-étapes (MTMS-TCN) (Section A.2.2), un modèle d'apprentissage profond de pointe pour la reconnaissance des activ-

Table A.6: Performance of MTMS-TCN on different datasets on phase recognition.

Train	Test	ACC	PR	RE	F1
StrasBypass70	StrasBypass70	90.70 ± 6.92	82.32 ± 8.69	85.86 ± 7.70	82.31 ± 8.83
	BernBypass70	71.95 ± 13.98	38.38 ± 8.10	43.26 ± 10.44	35.69 ± 9.88
BernBypass70	StrasBypass70	63.63 ± 9.43	36.67 ± 5.00	38.44 ± 8.15	33.12 ± 5.52
	BernBypass70	85.01 ± 13.22	62.79 ± 10.99	66.41 ± 12.20	61.21 ± 11.73
MultiBypass140	StrasBypass70	90.14 ± 6.80	81.68 ± 7.93	83.79 ± 7.85	81.17 ± 8.09
	BernBypass70	85.97 ± 12.92	61.81 ± 10.92	67.04 ± 11.6	60.58 ± 11.32
	MultiBypass140	88.05 ± 10.53	71.75 ± 13.78	75.41 ± 12.97	70.88 ± 14.24

Table A.7: Performance of MTMS-TCN on different datasets on step recognition.

Train	Test	ACC	PR	RE	F1
StrasBypass70	StrasBypass70	78.79 ± 10.28	62.12 ± 7.14	64.79 ± 8.68	60.53 ± 8.23
	BernBypass70	49.57 ± 14.39	24.50 ± 6.57	29.50 ± 6.95	23.00 ± 6.47
BernBypass70	StrasBypass70	46.04 ± 11.00	30.36 ± 4.82	29.67 ± 6.19	24.69 ± 4.91
	BernBypass70	67.61 ± 13.51	52.75 ± 9.50	55.81 ± 11.41	50.08 ± 10.67
MultiBypass140	StrasBypass70	78.16 ± 10.07	62.12 ± 6.79	63.54 ± 8.15	59.87 ± 7.71
	BernBypass70	68.60 ± 13.35	52.38 ± 8.17	55.01 ± 9.59	49.69 ± 9.43
	MultiBypass140	73.38 ± 12.75	57.25 ± 8.95	59.28 ± 9.87	54.78 ± 10.00

ités chirurgicales, a été utilisé pour les différentes expériences présentées dans cet article. Sept configurations expérimentales ont été utilisées pour entraîner et évaluer le modèle d'apprentissage profond: 1) Formation et évaluation sur StraBypass70, 2) Formation et évaluation sur BernBypass70, 3) Formation sur StraBypass70 et évaluation sur BernBypass70, 4) Formation sur BernBypass70 et évaluation sur StraBypass70, 5) Formation et évaluation sur le jeu de données commun MultiBypass140, 6) Formation sur MultiBypass140 et évaluation sur StraBypass70, 7) Formation sur MultiBypass140 et évaluation sur BernBypass70.

Les résultats (Tableaux A.6 & A.7) démontrent la nécessité de présenter la variation des techniques chirurgicales et du flux de travail aux modèles d'apprentissage profond pour éviter le déficit de généralisation décrit dans la littérature. Il a été démontré que la distribution et la taille du jeu de données dues aux différentes techniques et flux de travail LRYGB entre les centres ont un impact majeur sur la performance du modèle. Ce travail souligne l'importance des jeux de données multicentriques pour l'entraînement et l'évaluation des modèles d'IA dans l'analyse de vidéos chirurgicales.

A.3 Conclusion

A.3.1 Résumé et contribution

L'objectif fondamental de cette thèse est de développer des méthodes pour la reconnaissance automatique d'activités chirurgicales à plusieurs niveaux de détails. Dans cette thèse, nous avons présenté quatre études abordant différents défis dans ce domaine: la reconnaissance d'activités à plusieurs niveaux, la dépendance à de grands ensembles de données étiquetées, l'entraînement optimal et la généralisation à d'autres centres. Tout d'abord, nous avons construit un grand ensemble de données de procédures LRYGB, appelé Bypass40, avec des étiquettes de phase et d'étape et nous avons proposé un modèle temporel multi-tâches. Ensuite, nous avons présenté une méthode d'apprentissage faiblement supervisée pour résoudre le problème de la rareté des étiquettes pour la reconnaissance des étapes en utilisant les étiquettes de phase comme signaux faibles. Pour optimiser l'entraînement des modèles spatio-temporels pour la reconnaissance de l'activité chirurgicale, nous avons proposé une méthode d'augmentation des données simplifiée et automatisée appelée augmentations aléatoires temporelles (TRandAugment). Enfin, nous étudions la propriété de généralisation de la méthode de reconnaissance d'activité de l'état de l'art sur des données provenant de deux centres cliniques différents. Pour ce faire, nous avons introduit deux ensembles de données, appelés StraBypass70 et BernBypass70, composés de 70 vidéos de procédures LRYGB provenant des centres cliniques de Strasbourg et de Berne, qui ont été entièrement annotées avec les phases et les étapes. La reconnaissance des activités chirurgicales à plusieurs niveaux est essentielle à la mise en œuvre de la CAS dans OR afin d'améliorer les communications homme-machine, d'accélérer l'exécution des manœuvres chirurgicales, de réduire la charge de travail et la tension chirurgicales, de réduire les erreurs chirurgicales, d'augmenter le nombre de patients et d'améliorer la qualité des soins.

A.3.2 Applications cliniques

La capacité de reconnaître automatiquement les activités chirurgicales à partir de vidéos endoscopiques pourrait permettre de déployer avec succès les systèmes d'aide à la chirurgie dans les salles d'opération. En particulier, ces systèmes seraient efficaces dans de nombreuses applications pré-, intra- et post-opératoires ciblant la formation chirurgicale, le contrôle de la sécurité, l'aide à la décision clinique, l'indexation des données et les rapports, l'exécution autonome, et d'autres. Une vue d'ensemble des applications des CAS est présentée dans la Figure A.9.

Figure A.9: Illustration de l'assistance contextuelle qui pourrait être fournie dans la salle d'opération et en dehors de celle-ci à l'aide de certains des systèmes développés dans le cadre de cette thèse.

References

[Agha 2015] R. A. Agha and A. J. Fowler. *The Role and Validity of Surgical Simulation.* International Surgery, vol. 100, no. 2, pages 350–357, February 2015. (Cited on page 10)

[Ahmadi 2006] S.-A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner and N. Navab. *Recovery of Surgical Workflow Without Explicit Models.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, pages 420–428. Springer Berlin Heidelberg, 2006. (Cited on pages 8, 17, and 18)

[Ahmidi 2017] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal and G. D. Hager. *A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery.* IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pages 2025–2041, September 2017. (Cited on pages 10 and 11)

[Alapatt 2021] D. Alapatt, P. Mascagni, A. Vardazaryan, A. Garcia, N. Okamoto, D. Mutter, J. Marescaux, G. Costamagna, B. Dallemagne and N. Padoy. *Temporally Constrained Neural Networks (TCNN): A framework for semi-supervised video semantic segmentation.* arXiv preprint arXiv:2112.13815, 2021. (Cited on page 78)

[Allan 2019] M. Allan, A. A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, L. C. García-Peraza, W. Li, V. I. Iglovikov, H. Luo, J. Yang, D. Stoyanov, L. Maier-Hein, S. Speidel and M. Azizian. *2017 Robotic Instrument Segmentation Challenge.* ArXiv, vol. abs/1902.06426, 2019. (Cited on page 36)

[Allan 2020] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes-Hurtado, E. Flouty, A. K. Mohammed, M. Pedersen, A. Kori, A. Varghese, G. Krishnamurthi, D. Rauber, R. Mendel, C. Palm, S. Bano, G. Saibro, C.-S. Shih, H.-A. Chiang, J. Zhuang, J. Yang, V. I. Iglovikov,

A. Dobrenkii, M. Reddiboina, A. Reddy, X. Liu, C. Gao, M. Unberath, M. Azizian, D. Stoyanov, L. Maier-Hein and S. Speidel. *2018 Robotic Scene Segmentation Challenge*. ArXiv, vol. abs/2001.11190, 2020. (Cited on page 36)

[Angrisani 2015] L. Angrisani, A. Santonicola, P. Iovino, G. Formisano, H. Buchwald and N. Scopinaro. *Bariatric Surgery Worldwide 2013*. Obesity Surgery, vol. 25, no. 10, pages 1822–1832, April 2015. (Cited on pages 39 and 114)

[Bar 2020] O. Bar, D. Neimark, M. Zohar, G. D. Hager, R. Girshick, G. M. Fried, T. Wolf and D. Asselmann. *Impact of data on generalization of AI for surgical intelligence applications*. Scientific Reports, vol. 10, no. 1, December 2020. (Cited on page 91)

[Battaglia 2021] E. Battaglia, J. Boehm, Y. Zheng, A. R. Jamieson, J. Gahan and A. M. Fey. *Rethinking Autonomous Surgery: Focusing on Enhancement over Autonomy*. European Urology Focus, vol. 7, no. 4, pages 696–705, July 2021. (Cited on page 102)

[Bawa 2021] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, K. Wang, L. Li, B. Wang, S. Zhao, L. Li, A. Stabile, F. Setti, R. Muradore and F. Cuzzolin. *The SARAS Endoscopic Surgeon Action Detection (ESAD) dataset: Challenges and methods*. ArXiv, vol. abs/2104.03178, 2021. (Cited on page 36)

[Bergstra 2011] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl. *Algorithms for Hyper-Parameter Optimization*. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 24. Curran Associates, Inc., 2011. (Cited on page 13)

[Bergstra 2013] J. Bergstra, D. Yamins and D. Cox. *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*. In S. Dasgupta and D. McAllester, editors, Proceedings of the 30th International Conference on Machine Learning, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. (Cited on page 13)

[Berlet 2022] M. Berlet, T. Vogel, D. Ostler, T. Czempiel, M. Kähler, S. Brunner, H. Feussner, D. Wilhelm and M. Kranzfelder. *Surgical reporting for laparoscopic cholecystectomy based on phase annotation by a convolutional neural network (CNN) and the phenomenon of phase flickering: a proof of concept*. International Journal of Computer Assisted Radiology and Surgery, vol. 17, no. 11, pages 1991–1999, May 2022. (Cited on pages xv, 100, and 101)

[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, October 2013. (Cited on page 117)

[Blum 2008a] T. Blum, N. Padoy, H. Feußner and N. Navab. *Modeling and Online Recognition of Surgical Phases Using Hidden Markov Models.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008, pages 627–635. Springer Berlin Heidelberg, 2008. (Cited on page 18)

[Blum 2008b] T. Blum, N. Padoy, H. Feußner and N. Navab. *Workflow mining for visualization and analysis of surgeries.* International Journal of Computer Assisted Radiology and Surgery, vol. 3, no. 5, pages 379–386, July 2008. (Cited on page 8)

[Blum 2010] T. Blum, H. Feußner and N. Navab. *Modeling and Segmentation of Surgical Workflow from Laparoscopic Video.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, pages 400–407. Springer Berlin Heidelberg, 2010. (Cited on pages 18 and 19)

[Bodenstedt 2017] S. Bodenstedt, M. Wagner, D. Katic, P. Mietkowski, B. F. B. Mayer, H. Kenngott, B. P. Müller-Stich, R. Dillmann and S. Speidel. *Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis.* CoRR, vol. abs/1702.03684, 2017. (Cited on pages 25 and 26)

[Bricon-Souf 2007] N. Bricon-Souf and C. R. Newman. *Context awareness in health care: A review.* International Journal of Medical Informatics, vol. 76, no. 1, pages 2–12, January 2007. (Cited on pages 6 and 112)

[Bukchin 2020] G. Bukchin, E. Schwartz, K. Saenko, O. Shahar, R. S. Feris, R. Giryes and L. Karlinsky. *Fine-grained Angular Contrastive Learning with Coarse Labels.* 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8726–8736, 2020. (Cited on pages xiii, 24, 25, and 30)

[Carreira 2019] J. Carreira, E. Noland, C. Hillier and A. Zisserman. *A short note on the kinetics-700 human action dataset.* arXiv preprint arXiv:1907.06987, 2019. (Cited on page 35)

[Charrière 2017] K. Charrière, G. Quellec, M. Lamard, D. Martiano, G. Cazuguel, G. Coatrieux and B. Cochener. *Real-time analysis of cataract surgery videos using statistical models.* Multimedia Tools and Applications, vol. 76, no. 21, pages 22473–22491, May 2017. (Cited on pages 9, 14, 21, 30, 60, 73, and 113)

[Charrière 2014] K. Charrière, 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, 2014. (Cited on pages 9, 14, 21, 30, and 113)

[Chu 2020] P. Chu, X. Bian, S. Liu and H. Ling. *Feature Space Augmentation for Long-Tailed Data.* In Computer Vision – ECCV 2020, pages 694–710. 2020. (Cited on page 77)

[Cobianchi 2023] L. Cobianchi, D. Piccolo, F. D. Mas, V. Agnoletti, L. Ansaloni, J. Balch, W. Biffl, G. Butturini, F. Catena, F. Coccolini, S. Denicolai, B. D. Simone, I. Frigerio, P. Fugazzola, G. Marseglia, G. R. Marseglia, J. Martellucci, M. Modenese, P. Previtali, F. Ruta, A. Venturi, H. M. Kaafarani, T. J. Loftus, K. L. Abbott, A. Abdelmalik, N. S. Abebe, F. Abu-Zidan, Y. A. Y. Adam, H. Adamou, D. M. Adamovich, F. Agresta, antonino Agrusa, E. Akin, M. Alessiani, H. Alexandrino, S. M. Ali, V. A. Mihai, P. M. Almeida, M. M. Al-Shehari, M. Altomare, F. Amico, M. Ammendola, J. Andreuccetti, E. Anestiadou, P. Angelos, A. Annicchiarico, A. Antonelli, D. Aparicio-Sanchez, antonella Ardito, G. Argenio, C. C. Arvieux, I. H. Askevold, B. T. Atanasov, G. Augustin, S. S. Awad, G. Bacchiocchi, C. Bagnoli, H. Bahouth, E. Baili, L. Bains, G. L. Baiocchi, M. Bala, C. Balagué, D. Balalis, E. Baldini, oussama Baraket, S. Baral, M. Barone, A. G. Barranquero, J. A. Barreras, G. A. Bass, Z. Bayhan, G. Bellanova, O. Ben-Ishay, F. Bert, V. Bianchi, H. Biancuzzi, C. Bidoli, R. B. Radulescu, M. B. Bignell, A. Biloslavo, R. Bini, P. Boati, G. Boddaert, B. Bogdanic, C. Bombardini, L. Bonavina, L. Bonomo, A. Bottari, K. Bouliaris, G. Brachini, A. Brillantino, G. Brisinda, M. M. Bulanauca, L. A. Buonomo, J. Burcharth, S. Buscemi, F. Calabretto, G. Calini, V. Calu, F. C. Campanile, R. C. Dall/Orto, A. Campos-Serra, S. Campostrini, R. Capoglu, J. M. Carvas, M. Cascella, G. C. Pattacini, V. Celentano, D. C. Centonze, M. Ceresoli, D. Chatzipetris, A. Chessa, M. M. Chiarello, M. Chirica, S. Chooklin, C. Chouliaras, S. Chowdhury, P. Cianci, N. Cillara, S. Cimbanassi, S. P. B. Cioffi, E. Colak, E. C. Ruiz, L. Conti, A. Coppola, T. C. D. Sa, S. D. Costa, V. Cozza, G. Curro', K. F. A.-S. A. Dabekaussen, F. D'Acapito, D. Damaskos, G. D'Ambrosio, K. Das, R. J. Davies, A. C. D. Beaux, S. P. D. L. Fernandez, A. D. Luca, F. D. Stefano, L. Degrate, Z. Demetrashvili, A. K. Demetriades, D. S. Detanac, A. Dezi, G. D. Buono, I. D. Carlo, P. D. Lascio, M. D. Martino, S. D. Saverio, B. Diaconescu, J. J. Diaz, R. Dibra, E. N. Dimitrov, V. P. Dinuzzi, S. Dios-Barbeito, J. F. A. Diyani, A. Dogjani, M. Domanin, M. D'Oria, V. D. Munoz-Cruzado, B. East, M. Ekelund, G. T. Ekwen, A. H. Elbaih, M. Elhadi, N. Enninghorst, M. Ernisova, J. P. Escalera-Antezana, S. Esposito, G. Esposito, M. Estaire, C. N. Farè, R. Farre, F. Favi, L. Ferrario, A. F. di Tor Vajana, C. Filisetti, F. Fleres, V. C. Fonseca, A. Forero-Torres, F. Forfori, L. Fortuna, E. Fradelos, G. P. Fraga, P. Fransvea, S. Frassini, G. Frazzetta, E. Pizzocaro, M. Frountzas, M. Gachabayov, R. Galeiras, A. A. G. Vazquez, S. Gargarella, I. U. Garzali, W. M. Ghannam, F. N. Ghazi, L. M. Gillman, R. Gioco, A. Giordano, L. Giordano, C. Giove, G. Giraudo, M. Giuffrida, M. G. Capponi, E. Gois, C. A. Gomes, F. C. Gomes, R. A. T. Gonsaga, E. Gonullu, J. Goosen, T. Goranovic, R. Gracia-Roman, G. M. P. Graziano, E. A. Griffiths, T. Guagni, D. B. Hadzhiev, M. G.

Haidar, H. K. S. Hamid, T. C. Hardcastle, F. Hayati, A. J. Healey, A. Hecker, M. Hecker, E. F. H. Garcia, A. M. Hodonou, E. C. Huaman, M. Huerta, A. F. Ibrahim, B. M. S. Ibrahim, G. Ietto, M. Inama, O. Ioannidis, A. Isik, N. Ismail, A. M. H. Ismail, R. F. Jailani, J. Y. Jang, C. Kalfountzos, S. N. R. Kalipershad, E. Kaouras, L. J. Kaplan, Y. Kara, E. Karamagioli, A. Karamarkovia, I. Katsaros, A. J. Kavalakat, A. Kechagias, J. Kenig, B. J. Kessel, J. S. Khan, V. Khokha, J. I. Kim, A. W. Kirkpatrick, R. Klappenbach, Y. Kluger, Y. Kobe, E. K. Lymperis, K. Y. Y. Kok, V. Kong, D. P. Korkolis, G. Koukoulis, B. Kovacevic, V. F. Kruger, I. A. Kryvoruchko, H. Kurihara, A. Kuriyama, A. Landaluce-Olavarria, P. Lapolla, A. Leppäniemi, L. Licari, G. Lisi, A. Litvin, A. Lizarazu, H. L. Bayo, V. Lohsiriwat, C. C. L. Moreira, E. Lostoridis, A. T. Luna, D. Luppi, G. M. M. V., M. Maegele, D. Maggiore, S. Magnone, R. V. Maier, P. Major, M. Manangi, andrea manetti, B. Mantoglu, C. Marafante, F. Mariani, A. Marinis, E. A. S. Mariot, G. Martines, A. M. Perez, C. Martino, P. Mascagni, D. Massalou, M. Massaro, B. Matías-García, G. Mazzarella, G. Mazzarolo, R. B. Melo, F. Mendoza-Moreno, S. Meric, J. Meyer, L. Miceli, N. V. Michalopoulos, F. Milana, A. Mingoli, T. S. Mishra, M. Mohamed, M. I. E. A. Mohamed, A. Y. Mohamedahmed, M. J. S. Mohammed, R. Mohan, E. E. Moore, D. Morales-Garcia, M. Muhrbeck, F. Mulita, S. M. S. Mustafa, E. M. Muttillo, M. D. Naimzada, P. H. Navsaria, I. Negoi, L. Nespoli, C. Nguyen, M. K. Nidaw, G. Nigri, I. Nikolopoulos, D. B. O'Connor, H. D. Ogundipe, C. Oliveri, S. Olmi, E. C. W. Ong, L. Orecchia, A. V. Osipov, M. F. Othman, M. Pace, M. Pacilli, L. Pagani, G. Palomba, D. Pantalone, A. Panyko, C. Paolillo, M. V. Papa, D. Papaconstantinou, M. Papadoliopoulou, A. Papadopoulos, D. Papis, N. Pararas, J. G. Parreira, N. G. Parry, F. Pata, T. Patel, S. Paterson-Brown, G. Pavone, F. Pecchini, V. Pegoraro, G. Pellino, M. Pelloni, A. Peloso, E. P. D. Pozo, R. G. Pereira, B. M. Pereira, A. L. Perez, S. Pérez, T. Perra, G. Perrone, A. Pesce, L. Petagna, G. Petracca, V. Phupong, B. Picardi, A. Picciariello, M. Piccoli, E. Picetti, E. P. Pikoulis, T. Pintar, G. Pirozzolo, F. Piscioneri, M. Podda, A. Porcu, F. Privitera, C. Punzo, S. Quaresima, M. A. Quiodettis, N. Qvist, R. Rahim, F. R. de Almeida, R. B. Ramely, H. K. Rasa, M. Reichert, A. Reinisch-Liese, A. Renne, C. Riccetti, M. R. Rodriguez-Luna, D. Roizblatt, A. Romanzi, L. Romeo, F. P. M. Roscio, R. B. Rosnelifaizur, S. Rossi, A. M. Rubiano, E. Ruiz-Ucar, B. E. Sakakushev, J. C. Salamea, I. Sall, L. B. Samarakoon, F. Sammartano, A. S. Arteaga, S. Sanchez-Cordero, D. P. M. Santoanastaso, M. Sartelli, D. Sasia, N. SATO, A. Savchuk, R. G. Sawyer, G. Scaioli, D. SCHIZAS, S. Sebastiani, B. Seeliger, H. A. S. Lohse, C. Seretis, G. Sermonesi, M. Serradilla-Martin, V. G. Shelat, S. Shlyapnikov, T. Sidiropoulos, R. L. Simoes, L. Siragusa, B. Siribumrungwong, M. Slavchev, L. Solaini, gabriele soldini, A. Sopuev, K. Soreide, A. SOVATZIDIS, P. F. Stahel, M. Strickland, M. A. H. Sultan, R. Sydorchuk, L. Sydorchuk, S. M. A. M. Syed, L. Tallon-Aguilar, A. M. Tamburini, N. Tamini, E. C. T. H. Tan, J. H. Tan, A. Tarasconi, N. Tartaglia, G. Tartaglia, D. Tartaglia, J. V. Tay-

lor, G. D. Tebala, R. A. T. Gonsaga, M. Teuben, A. Theodorou, M. Tolonen, G. Tomasicchio, A. Toro, B. Torre, T. Triantafyllou, G. T. Trigiante, M. Tripepi, J. Trostchansky, K. Tsekouras, V. Turrado-Rodriguez, R. Tutino, M. Uccelli, P. A. Uchikov, B. Ugarte-Sierra, M. T. Ukkonen, M. Vailas, P. G. Vassiliu, A. G. Vazquez, R. G. Vazquez, G. Velmahos, J. E. Verde, J. M. Verde, M. Veroux, J. Viganò, R. Vilallonga, D. Visconti, A. Vittori, M. Waledziak, T. Wannatoop, L. W. Widmer, M. S. J. Wilson, S. Woltz, T. H. Wong, S. Xenaki, B. Yu, S. Yule, S. K. Zachariah, G. Zacharis, C. Zaghi, A. D. Zakaria, D. A. Zambrano, N. Zampitis, B. Zampogna, S. Zanghì, M. Zantedeschi, K. Zapsalis, F. Zattoni and M. Z. and. *Surgeons' perspectives on artificial intelligence to support clinical decision-making in trauma and emergency contexts: results from an international survey.* World Journal of Emergency Surgery, vol. 18, no. 1, January 2023. (Cited on page 7)

[Cordts 2015] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele. *The Cityscapes Dataset.* In CVPR Workshop on The Future of Datasets in Vision, 2015. (Cited on page 35)

[Cubuk 2019] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le. *Autoaugment: Learning augmentation strategies from data.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 113–123, 2019. (Cited on pages 13, 27, 29, 30, 71, 114, and 115)

[Cubuk 2020] E. D. Cubuk, B. Zoph, J. Shlens and Q. V. Le. *Randaugment: Practical automated data augmentation with a reduced search space.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020. (Cited on pages 13, 29, 30, 31, 71, 72, 73, and 115)

[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 MICCAI, 2020. (Cited on pages xiii, 20, 21, 29, 30, 48, 50, 51, 58, 62, 72, and 74)

[Czempiel 2021] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam and N. Navab. *OperA: Attention-regularized transformers for surgical phase recognition.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 604–614. Springer, 2021. (Cited on pages 20, 72, and 74)

[Darzi 2004] A. Darzi and Y. Munz. *The Impact of Minimally Invasive Surgical Techniques.* Annual Review of Medicine, vol. 55, no. 1, page 223–237, February 2004. (Cited on page 5)

[Demir 2022] K. C. Demir, H. Schieber, D. Roth, A. Maier and S. H. Yang. *Surgical Phase Recognition: A Review and Evaluation of Current Approaches.* May 2022. (Cited on pages 8, 9, 14, 30, 113, and 114)

[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 CVPR09, 2009. (Cited on page 35)

[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, March 2016. (Cited on page 18)

[DeVries 2017a] T. DeVries and G. W. Taylor. *Dataset augmentation in feature space*. arXiv preprint arXiv:1702.05538, 2017. (Cited on page 114)

[DeVries 2017b] T. DeVries and G. W. Taylor. *Improved Regularization of Convolutional Neural Networks with Cutout*. arXiv preprint arXiv:1708.04552, 2017. (Cited on pages 27, 29, and 71)

[DiPietro 2016] R. S. DiPietro, C. S. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee and G. Hager. *Recognizing Surgical Activities with Recurrent Neural Networks*. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016. (Cited on page 23)

[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, April 2019. (Cited on page 23)

[Doughty 2018] H. Doughty, D. Damen and W. Mayol-Cuevas. *Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6057–6066, 2018. (Cited on page 100)

[Dowdy 2021] R. A. E. Dowdy, S. T. Mansour, J. H. Cottle, H. R. Mabe, H. B. Weprin, L. E. Yarborough, G. M. Ness, T. M. Jacobs and B. W. Cornelius. *Cardiac Arrest Upon Induction of General Anesthesia*. Anesthesia Progress, vol. 68, no. 1, pages 38–44, March 2021. (Cited on page 102)

[Eigen 2015] D. Eigen and R. Fergus. *Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture*. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 2650–2658, 2015. (Cited on pages 48 and 58)

[Fang 2019] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li and C. Lu. *Instaboost: Boosting instance segmentation via probability map guided copy-pasting*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 682–691, 2019. (Cited on pages 27 and 114)

[Farha 2019] Y. A. Farha and J. Gall. *MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation.* In CVPR, 2019. (Cited on pages xiii, 20, 21, and 30)

[Feng 2019] R. Feng, J. Gu, Y. Qiao and C. Dong. *Suppressing Model Overfitting for Image Super-Resolution Networks.* 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1964–1973, 2019. (Cited on pages 13 and 115)

[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, 2019. (Cited on pages xiii, 24, 25, and 30)

[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, Cham, 2018. Springer International Publishing. (Cited on pages 18, 25, and 26)

[Funke 2019] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz and S. Speidel. *Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video.* In MICCAI, 2019. (Cited on page 23)

[Gao 2014] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh*et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling.* In MICCAI workshop: M2cai, volume 3, 2014. (Cited on pages 10, 11, and 23)

[Gao 2020] X. Gao, Y. Jin, Q. Dou and P.-A. Heng. *Automatic Gesture Recognition in Robot-assisted Surgery with Reinforcement Learning and Tree Search.* In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 8440–8446, 2020. (Cited on page 23)

[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.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 593–603. Springer, 2021. (Cited on pages 20, 29, and 72)

[Garrow 2020] 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, B. P. Müller-Stich and F. Nickel. *Machine Learning for Surgical Phase Recognition.*

Annals of Surgery, vol. 273, no. 4, pages 684–693, November 2020. (Cited on pages 8, 9, 13, 14, 30, 113, and 114)

[Geirhos 2018] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann and W. Brendel. *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.* ArXiv, vol. abs/1811.12231, 2018. (Cited on pages 13 and 115)

[Gibaud 2018] B. Gibaud, G. Forestier, C. Feldmann, G. Ferrigno, P. Gonçalves, T. Haidegger, C. Julliard, D. Katić, H. Kenngott, L. Maier-Hein, K. März, E. de Momi, D. Á. Nagy, H. Nakawala, J. Neumann, T. Neumuth, J. R. Balderrama, S. Speidel, M. Wagner and P. Jannin. *Toward a standard ontology of surgical process models.* International Journal of Computer Assisted Radiology and Surgery, vol. 13, no. 9, pages 1397–1408, July 2018. (Cited on pages 8 and 9)

[Girshick 2018] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár and K. He. *Detectron*, 2018. (Cited on page 114)

[Gowda 2022] S. N. Gowda, M. Rohrbach, F. Keller and L. Sevilla-Lara. *Learn2Augment: Learning to Composite Videos for Data Augmentation in Action Recognition.* In European Conference on Computer Vision, pages 242–259. Springer, 2022. (Cited on pages 28, 29, and 30)

[Gurcan 2019] I. Gurcan and H. V. Nguyen. *Surgical Activities Recognition Using Multiscale Recurrent Networks.* In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2887–2891, 2019. (Cited on page 23)

[Hajj 2018] H. A. 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, July 2018. (Cited on page 113)

[Hajj 2019] H. A. Hajj, M. Lamard, P.-H. Conze, S. Roychowdhury, X. Hu, G. Maršalkaitė, O. Zisimopoulos, M. A. Dedmari, F. Zhao, J. Prellberg, M. Sahu, A. Galdran, T. Araújo, D. M. Vo, C. Panda, N. Dahiya, S. Kondo, Z. Bian, A. Vahdat, J. Bialopetravičius, E. Flouty, C. Qiu, S. Dill, A. Mukhopadhyay, P. Costa, G. Aresta, S. Ramamurthy, S.-W. Lee, A. Campilho, S. Zachow, S. Xia, S. Conjeti, D. Stoyanov, J. Armaitis, P.-A. Heng, W. G. Macready, B. Cochener and G. Quellec. *CATARACTS: Challenge on automatic tool annotation for cataRACT surgery.* Medical Image Analysis, vol. 52, pages 24–41, February 2019. (Cited on pages 11, 14, 36, 60, and 73)

[Han 2022] J. Han, P. Fang, W. Li, J. Hong, M. A. Armin, I. Reid, L. Petersson and H. Li. *You Only Cut Once: Boosting Data Augmentation with a Single Cut.* In

Proceedings of the 39th International Conference on Machine Learning, volume 162, pages 8196–8212, Jul 2022. (Cited on pages 27, 28, and 71)

[Haththotuwa 2020] R. N. Haththotuwa, C. N. Wijeyaratne and U. Senarath. *Worldwide epidemic of obesity.* In Obesity and Obstetrics, pages 3–8. Elsevier, 2020. (Cited on page 38)

[He 2016a] K. He, X. Zhang, S. Ren and J. Sun. *Deep Residual Learning for Image Recognition.* In CVPR, 2016. (Cited on page 20)

[He 2016b] K. He, X. Zhang, S. Ren and J. Sun. *Identity Mappings in Deep Residual Networks.* In Computer Vision – ECCV 2016. Springer International Publishing, 2016. (Cited on pages 48, 58, and 72)

[He 2017] K. He, G. Gkioxari, P. Dollar and R. Girshick. *Mask R-CNN.* In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017. (Cited on page 27)

[Hendrycks 2020] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer and B. Lakshminarayanan. *AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty.* Proceedings of the International Conference on Learning Representations (ICLR), 2020. (Cited on pages 29 and 71)

[Hepatology 2021] T. L. G. &. Hepatology. *Obesity: another ongoing pandemic.* The Lancet Gastroenterology & Hepatology, vol. 6, no. 6, page 411, June 2021. (Cited on page 38)

[Ho 2019] D. Ho, E. Liang, X. Chen, I. Stoica and P. Abbeel. *Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules.* In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 2731–2741, 2019. (Cited on page 27)

[Hong 2020] W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W. Chang and C.-S. Shih. *CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80.* ArXiv, vol. abs/2012.12453, 2020. (Cited on pages 14 and 36)

[Hwang 2016] S. Hwang and H.-E. Kim. *Self-Transfer Learning for Weakly Supervised Lesion Localization.* In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal and W. Wells, editors, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, pages 239–246, Cham, 2016. Springer International Publishing. (Cited on page 25)

[Jannin 2001] P. Jannin, M. Raimbault, X. Morandi and B. Gibaud. *Modeling Surgical Procedures for Multimodal Image-Guided Neurosurgery.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001, pages 565–572. Springer Berlin Heidelberg, 2001. (Cited on pages 8, 9, and 17)

[Jia 2017] Z. Jia, X. Huang, E. I. 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 page 25)

[Jin 2018] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu and P.-A. Heng. *SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network.* IEEE Transactions on Medical Imaging, vol. 37, no. 5, pages 1114–1126, May 2018. (Cited on pages 20, 48, 50, 51, 58, and 62)

[Jin 2020] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C. Fu and P. Heng. *Multi-Task Recurrent Convolutional Network with Correlation Loss for Surgical Video Analysis.* Medical image analysis, vol. 59, page 101572, 2020. (Cited on pages 20, 29, 48, 50, 58, 62, 72, and 113)

[Jin 2021] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou and P.-A. Heng. *Temporal Memory Relation Network for Workflow Recognition From Surgical Video.* IEEE Transactions on Medical Imaging, vol. 40, no. 7, pages 1911–1923, 2021. (Cited on page 20)

[Joerger 2017] G. Joerger, A. Y. Huang, B. L. Bass, B. Dunkin and M. Garbey. *Global Laparoscopy Positioning System with a Smart Trocar.* 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pages 359–366, 2017. (Cited on page 8)

[Kaijser 2018] M. A. Kaijser, G. H. van Ramshorst, M. Emous, N. J. G. M. Veeger, B. A. van Wagensveld and J.-P. E. N. Pierie. *A Delphi Consensus of the Crucial Steps in Gastric Bypass and Sleeve Gastrectomy Procedures in the Netherlands.* Obesity Surgery, vol. 28, no. 9, pages 2634–2643, April 2018. (Cited on page 114)

[Kannan 2020] 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, March 2020. (Cited on pages 10 and 22)

[Kassem 2022] H. Kassem, D. Alapatt, P. Mascagni, C. AI4SafeChole, A. Karargyris and N. Padoy. *Federated Cycling (FedCy): Semi-supervised Federated Learning of Surgical Phases.* IEEE Transactions on Medical Imaging, pages 1–1, 2022. (Cited on pages 27 and 78)

[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 Information Processing in Computer-Assisted Interventions, pages 158–167. Springer International Publishing, 2014. (Cited on pages 10 and 22)

[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, June 2015. (Cited on pages 8, 9, 11, 22, 58, and 113)

[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 page 35)

[Kim 2020] T. Kim, H. Lee, M. Cho, H. S. Lee, D. H. Cho and S. Lee. *Learning Temporally Invariant and Localizable Features via Data Augmentation for Video Recognition.* In Computer Vision – ECCV 2020 Workshops, pages 386–403. 2020. (Cited on pages 28 and 29)

[Kim 2022] T. Kim, J. Kim, M. Shim, S. Yun, M. Kang, D. Wee and S. Lee. *Exploring Temporally Dynamic Data Augmentation for Video Recognition.* arXiv preprint arXiv:2206.15015, 2022. (Cited on page 29)

[Kimata 2022] J. Kimata, T. Nitta and T. Tamaki. *ObjectMix: Data Augmentation by Copy-Pasting Objects in Videos for Action Recognition.* arXiv preprint arXiv:2204.00239, 2022. (Cited on pages 27 and 29)

[Kitaguchi 2022] D. Kitaguchi, T. Fujino, N. Takeshita, H. Hasegawa, K. Mori and M. Ito. *Limited generalizability of single deep neural network for surgical instrument segmentation in different surgical environments.* Scientific Reports, vol. 12, no. 1, July 2022. (Cited on page 91)

[Klank 2008] U. Klank, N. Padoy, H. Feussner and N. Navab. *Automatic feature generation in endoscopic images.* International Journal of Computer Assisted Radiology and Surgery, vol. 3, no. 3-4, pages 331–339, June 2008. (Cited on pages 17 and 18)

[Kranzfelder 2009] M. Kranzfelder, A. Schneider, G. Blahusch, H. Schaaf and H. Feussner. *Feasibility of opto-electronic surgical instrument identification.* Minimally Invasive Therapy & Allied Technologies, vol. 18, no. 5, pages 253–258, January 2009. (Cited on page 8)

[Kranzfelder 2012] 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, 2012. (Cited on pages 6, 7, and 112)

[Krizhevsky 2017] A. Krizhevsky, I. Sutskever and G. E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks.* Commun. ACM, vol. 60, no. 6, page 84–90, may 2017. (Cited on page 19)

[Laplante 2022] S. Laplante, B. Namazi, P. Kiani, D. A. Hashimoto, A. Alseidi, M. Pasten, L. M. Brunt, S. Gill, B. Davis, M. Bloom, L. Pernar, A. Okrainec and A. Madani. *Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy.* Surgical Endoscopy, August 2022. (Cited on page 9)

[Lavanchy 2022] J. L. Lavanchy, C. Gonzalez, H. Kassem, P. C. Nett, D. Mutter and N. Padoy. *Proposal and multicentric validation of a laparoscopic Roux-en-Y gastric bypass surgery ontology.* Surgical Endoscopy, October 2022. (Cited on pages 12 and 86)

[Lea 2015] C. Lea, G. D. Hager and R. Vidal. *An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks.* In 2015 IEEE Winter Conference on Applications of Computer Vision, pages 1123–1129, 2015. (Cited on page 23)

[Lea 2016a] C. Lea, A. Reiter, R. Vidal and G. D. Hager. *Segmental Spatiotemporal CNNs for Fine-Grained Action Segmentation.* In Computer Vision – ECCV 2016, pages 36–52. Springer International Publishing, 2016. (Cited on page 23)

[Lea 2016b] C. Lea, R. Vidal, A. Reiter and G. D. Hager. *Temporal Convolutional Networks: A Unified Approach to Action Segmentation.* In Lecture Notes in Computer Science, pages 47–54. Springer International Publishing, 2016. (Cited on pages xiii, 20, and 21)

[Lea 2016c] C. Lea, R. Vidal and G. D. Hager. *Learning convolutional action primitives for fine-grained action recognition.* In 2016 IEEE International Conference on Robotics and Automation (ICRA), pages 1642–1649, 2016. (Cited on page 23)

[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.* International Congress Series, vol. 1281, pages 862–867, May 2005. (Cited on page 6)

[Li 2020] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov and A. Zisserman. *The ava-kinetics localized human actions video dataset.* arXiv preprint arXiv:2005.00214, 2020. (Cited on page 35)

[Li 2021] J. Li, G. Zhu, C. Hua, M. Feng, B. Bennamoun, P. Li, X. Lu, J. Song, P. Shen, X. Xu, L. Mei, L. Zhang, S. A. A. Shah and Bennamoun. *A Systematic Collection of Medical Image Datasets for Deep Learning.* ArXiv, vol. abs/2106.12864, 2021. (Cited on page 36)

[Lim 2019] S. Lim, I. Kim, T. Kim, C. Kim and S. Kim. *Fast autoaugment.* Advances in Neural Information Processing Systems, vol. 32, 2019. (Cited on pages 13, 27, 29, 30, 71, 114, and 115)

[Lin 2006] H. C. Lin, I. Shafran, D. Yuh and G. D. Hager. *Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions.* Computer Aided Surgery, vol. 11, no. 5, pages 220–230, January 2006. (Cited on pages 17 and 23)

[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 page 35)

[Liu 2018a] B. Liu, X. Wang, M. Dixit, R. Kwitt and N. Vasconcelos. *Feature Space Transfer for Data Augmentation.* In CVPR, 2018. (Cited on page 77)

[Liu 2018b] D. Liu and T. Jiang. *Deep Reinforcement Learning for Surgical Gesture Segmentation and Classification.* In Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pages 247–255. Springer International Publishing, 2018. (Cited on page 23)

[Lopes 2019] R. G. Lopes, D. Yin, B. Poole, J. Gilmer and E. D. Cubuk. *Improving robustness without sacrificing accuracy with patch gaussian augmentation.* arXiv preprint arXiv:1906.02611, 2019. (Cited on pages 13, 15, 78, and 114)

[Maier-Hein 2017] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager and P. Jannin. *Surgical data science for next-generation interventions.* Nature Biomedical Engineering, vol. 1, no. 9, pages 691–696, September 2017. (Cited on pages xiii, 5, 6, 7, and 112)

[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, May 2016. (Cited on pages 8 and 17)

[Mascagni 2020] P. Mascagni, A. Vardazaryan, D. Alapatt, T. Urade, T. Emre, C. Fiorillo, P. Pessaux, D. Mutter, J. Marescaux, G. Costamagna, B. Dallemagne and N. Padoy. *Artificial Intelligence for Surgical Safety.* Annals of Surgery, vol. 275, no. 5, pages 955–961, November 2020. (Cited on pages 10 and 102)

[Mavroudi 2018] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali and R. Vidal. *End-to-End Fine-Grained Action Segmentation and Recognition Using Conditional Random Field Models and Discriminative Sparse Coding.* In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1558–1567, 2018. (Cited on page 23)

[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 and D. A. H. and. *SAGES consensus recommendations on an annotation framework for surgical video.* Surgical Endoscopy, vol. 35, no. 9, pages 4918–4929, jul 2021. (Cited on pages 8, 9, 11, and 113)

[Meling 2020] T. R. Meling and T. R. Meling. *The impact of surgical simulation on patient outcomes: a systematic review and meta-analysis.* Neurosurgical Review, vol. 44, no. 2, pages 843–854, May 2020. (Cited on page 10)

[Meljnikov 2009] I. Meljnikov, B. Radojcić, S. Grebeldinger and N. Radojcić. *History of surgical treatment of appendicitis.* Med. Pregl., vol. 62, no. 9-10, pages 489–492, September 2009. (Cited on page 4)

[Menegozzo 2019] G. Menegozzo, D. Dall'Alba, C. Zandonà and P. Fiorini. *Surgical gesture recognition with time delay neural network based on kinematic data.* In 2019 International Symposium on Medical Robotics (ISMR), pages 1–7, 2019. (Cited on page 23)

[Mohiuddin 2013] K. Mohiuddin and S. J. Swanson. *Maximizing the benefit of minimally invasive surgery.* Journal of Surgical Oncology, vol. 108, no. 5, page 315–319, August 2013. (Cited on page 5)

[Mumuni 2022] A. Mumuni and F. Mumuni. *Data augmentation: A comprehensive survey of modern approaches.* Array, vol. 16, page 100258, December 2022. (Cited on page 13)

[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, pages 1059–1067, 2019. (Cited on pages xiii, 12, 24, 25, 30, 78, and 113)

[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 Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 364–374. Springer International Publishing, 2020. (Cited on pages 10, 11, 14, 22, 23, 78, and 113)

[Nwoye 2022a] C. I. Nwoye, D. Alapatt, T. Yu, A. Vardazaryan, F. Xia, Z. Zhao, T. Xia, F. Jia, Y. Yang, H. Wang, D.-S. Yu, G. Zheng, X. Duan, N. Getty, R. Sánchez-Matilla, M. R. Robu, L. Zhang, H. Chen, J. Wang, L. Wang, B. Zhang, B. G. A. Gerats, S. Raviteja, R. Sathish, R. Tao, S. Kondo, W. Pang, H. Ren, J. R. Abbing, M. H. Sarhan, S. Bodenstedt, N. Bhasker, B. Oliveira, H. R. Torres, L. Ling, F. Gaida, T. Czempiel, J. L. Vilacca, P. Morais, J. C. Fonseca, R. M. Egging, I. N. Wijma, C. Qian, G. bin Bian, Z. Li, V. Balasubramanian, D. Sheet,

I. Luengo, Y. Zhu, S. Ding, J.-A. Aschenbrenner, N. E. van der Kar, M. Xu, M. Islam, L. Seenivasan, A. Jenke, D. Stoyanov, D. Mutter, P. Mascagni, B. Seeliger, C. Gonzalez and N. Padoy. *CholecTriplet2021: A benchmark challenge for surgical action triplet recognition.* ArXiv, vol. abs/2204.04746, 2022. (Cited on pages 10, 22, and 23)

[Nwoye 2022b] 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.* Medical Image Analysis, vol. 78, page 102433, May 2022. (Cited on pages xiii, 10, 11, 14, 22, 23, 30, 36, 78, and 113)

[Nwoye 2023] C. I. Nwoye, T. Yu, S. Sharma, A. Murali, D. Alapatt, A. Vardazaryan, K. Yuan, J. Hajek, W. Reiter, A. Yamlahi, F.-H. Smidt, X. Zou, G. Zheng, B. Oliveira, H. R. Torres, S. Kondo, S. Kasai, F. Holm, E. Özsoy, S. Gui, H. Li, S. Raviteja, R. Sathish, P. Poudel, B. Bhattarai, Z. Wang, G. Rui, M. Schellenberg, J. L. Vilaça, T. Czempiel, Z. Wang, D. Sheet, S. K. Thapa, M. Berniker, P. Godau, P. Morais, S. Regmi, T. N. Tran, J. Fonseca, J.-H. Nölke, E. Lima, E. Vazquez, L. Maier-Hein, N. Navab, P. Mascagni, B. Seeliger, C. Gonzalez, D. Mutter and N. Padoy. *CholecTriplet2022: Show me a tool and tell me the triplet – an endoscopic vision challenge for surgical action triplet detection*, 2023. (Cited on pages 10, 22, and 23)

[on Obesity 2000] W. C. on Obesity. *Obesity: preventing and managing the global epidemic. Report of a WHO consultation.* World Health Organization technical report series, vol. 894, pages i–xii, 1–253, 2000. (Cited on pages 38 and 113)

[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 Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, pages 102–109. Springer Berlin Heidelberg, 2007. (Cited on page 8)

[Padoy 2008] N. Padoy, T. Blum, H. Feußner, M. Berger and N. Navab. *On-line Recognition of Surgical Activity for Monitoring in the Operating Room.* In AAAI, 2008. (Cited on page 18)

[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, April 2012. (Cited on pages 8 and 18)

[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, March 2019. (Cited on pages xiii and 19)

[Pan 2021] T. Pan, Y. Song, T. Yang, W. Jiang and W. Liu. *Videomoco: Contrastive video representation learning with temporally adversarial examples*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11205–11214, 2021. (Cited on pages 13, 15, 78, and 114)

[Qian 2021] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie and Y. Cui. *Spatiotemporal Contrastive Video Representation Learning*. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6960–6970, 2021. (Cited on pages 13, 15, and 114)

[Quellec 2014] 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, April 2014. (Cited on pages 9 and 14)

[Rabiner 1989] L. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, vol. 77, no. 2, pages 257–286, 1989. (Cited on page 18)

[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*. International Journal of Computer Assisted Radiology and Surgery, May 2021. (Cited on pages 15, 29, 48, 57, 58, 60, 72, 73, and 74)

[Ramesh 2022] S. Ramesh, V. Srivastav, D. Alapatt, T. Yu, A. Murali, L. Sestini, C. I. Nwoye, I. Hamoud, S. Sharma, A. Fleurentin, G. Exarchakis, A. Karargyris and N. Padoy. *Dissecting Self-Supervised Learning Methods for Surgical Computer Vision*, 2022. (Cited on pages 15 and 78)

[Ramesh 2023a] S. Ramesh, D. Dall'Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorini and N. Padoy. *TRandAugment: temporal random augmentation strategy for surgical activity recognition from videos*. International Journal of Computer Assisted Radiology and Surgery, March 2023. (Cited on pages 15 and 70)

[Ramesh 2023b] S. Ramesh, D. Dall'Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorini and N. Padoy. *Weakly Supervised Temporal Convolutional Networks for Fine-grained Surgical Activity Recognition*. IEEE Transactions on Medical Imaging, pages 1–1, 2023. (Cited on pages 15, 29, 56, and 78)

[Reiley 2008] C. E. Reiley, H. C. Lin, B. Varadarajan, B. P. Vágvölgyi, S. Khudanpur, D. D. Yuh and G. Hager. *Automatic Recognition of Surgical Motions Using Statistical Modeling for Capturing Variability*. Studies in health technology and informatics, vol. 132, pages 396–401, 2008. (Cited on page 23)

[Reiley 2009] C. E. Reiley and G. D. Hager. *Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, pages 435–442. Springer Berlin Heidelberg, 2009. (Cited on page 10)

[Reynolds 2001] W. Reynolds Jr. *The first laparoscopic cholecystectomy.* JSLS, vol. 5, no. 1, pages 89–94, January 2001. (Cited on page 4)

[Rosen 2001] J. Rosen, B. Hannaford, C. Richards and M. Sinanan. *Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills.* IEEE Transactions on Biomedical Engineering, vol. 48, no. 5, pages 579–591, May 2001. (Cited on page 10)

[Roß 2021] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran, P. Bruno, P. Arbeláez, G.-B. Bian, S. Bodenstedt, J. L. Bolmgren, L. Bravo-Sánchez, H.-B. Chen, C. González, D. Guo, P. Halvorsen, P.-A. Heng, E. Hosgor, Z.-G. Hou, F. Isensee, D. Jha, T. Jiang, Y. Jin, K. Kirtac, S. Kletz, S. Leger, Z. Li, K. H. Maier-Hein, Z.-L. Ni, M. A. Riegler, K. Schoeffmann, R. Shi, S. Speidel, M. Stenzel, I. Twick, G. Wang, J. Wang, L. Wang, L. Wang, Y. Zhang, Y.-J. Zhou, L. Zhu, M. Wiesenfarth, A. Kopp-Schneider, B. P. Müller-Stich and L. Maier-Hein. *Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge.* Medical Image Analysis, vol. 70, page 101920, May 2021. (Cited on page 36)

[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, 2016. (Cited on page 23)

[Ryan 2021] D. Ryan, S. Barquera, O. B. Cavalcanti and J. Ralston. *The Global Pandemic of Overweight and Obesity.* In Handbook of Global Health, pages 739–773. Springer International Publishing, 2021. (Cited on page 38)

[Sakoe 1978] H. Sakoe and S. Chiba. *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pages 43–49, February 1978. (Cited on page 18)

[Schoeffmann 2018] K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus and D. Putzgruber. *Cataract-101: video dataset of 101 cataract surgeries.* In P. César, M. Zink and N. Murray, editors, Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018, pages 421–425. ACM, 2018. (Cited on pages 11 and 14)

[Sefati 2015] S. Sefati, N. Cowan and R. Vidal. *Learning Shared, Discriminative Dictionaries for Surgical Gesture Segmentation and Classification.* 10 2015. (Cited on page 23)

[Sharma 2022] S. Sharma, C. I. Nwoye, D. Mutter and N. Padoy. *Rendezvous in Time: An Attention-based Temporal Fusion approach for Surgical Triplet Recognition*, 2022. (Cited on pages 22, 23, and 113)

[Shi 2020] X. Shi, Y. Jin, Q. Dou and P.-A. Heng. *LRTD: long-range temporal dependency based active learning for surgical workflow recognition.* International Journal of Computer Assisted Radiology and Surgery, vol. 15, no. 9, pages 1573–1584, June 2020. (Cited on page 20)

[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, October 2021. (Cited on pages xiv, 13, 15, 25, 26, 27, 74, 78, and 114)

[Shorten 2019] C. Shorten and T. M. Khoshgoftaar. *A survey on Image Data Augmentation for Deep Learning.* Journal of Big Data, vol. 6, no. 1, July 2019. (Cited on page 13)

[Smaira 2020] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu and A. Zisserman. *A short note on the kinetics-700-2020 human action dataset.* arXiv preprint arXiv:2010.10864, 2020. (Cited on page 35)

[Speidel 2008] S. Speidel, G. Sudra, J. Senemaud, M. Drentschew, B. P. Müller-Stich, C. N. Gutt and R. Dillmann. *Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling.* In SPIE Medical Imaging, 2008. (Cited on page 9)

[Stauder 2016] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feußner and N. Navab. *The TUM LapChole dataset for the M2CAI 2016 workflow challenge.* ArXiv, vol. abs/1610.09278, 2016. (Cited on page 11)

[Su 2021] J.-C. Su and S. Maji. *Semi-Supervised Learning with Taxonomic Labels.* In British Machine Vision Conference, 2021. (Cited on page 25)

[Sznitman 2012] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager and P. Fua. *Data-Driven Visual Tracking in Retinal Microsurgery.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012, pages 568–575. Springer Berlin Heidelberg, 2012. (Cited on page 36)

[Taherkhani 2019] F. Taherkhani, H. Kazemi, A. Dabouei, J. Dawson and N. Nasrabadi. *A Weakly Supervised Fine Label Classifier Enhanced by Coarse Supervision.* In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6458–6467, 2019. (Cited on pages 25 and 30)

[Tao 2012] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager and R. Vidal. *Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation.* In Information Processing in Computer-Assisted Interventions, pages 167–177. Springer Berlin Heidelberg, 2012. (Cited on page 23)

[Tao 2013] L. Tao, L. Zappella, G. D. Hager and R. Vidal. *Surgical Gesture Segmentation and Recognition.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, pages 339–346. Springer Berlin Heidelberg, 2013. (Cited on page 23)

[Toti 2014] G. Toti, M. Garbey, V. Sherman, B. L. Bass and B. J. Dunkin. *A Smart Trocar for Automatic Tool Recognition in Laparoscopic Surgery.* Surgical Innovation, vol. 22, no. 1, pages 77–82, May 2014. (Cited on page 8)

[Touvron 2021] H. Touvron, A. Sablayrolles, M. Douze, M. Cord and H. Jegou. *Grafit: Learning fine-grained image representations with coarse labels.* In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 854–864, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. (Cited on pages 25 and 30)

[Twinanda 2014] A. P. Twinanda, J. Marescaux, M. D. Mathelin and N. Padoy. *Towards Better Laparoscopic Video Database Organization by Automatic Surgery Classification.* In Information Processing in Computer-Assisted Interventions, pages 186–195. Springer International Publishing, 2014. (Cited on page 22)

[Twinanda 2016] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin and N. Padoy. *Single- and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016.* ArXiv, vol. abs/1610.08844, 2016. (Cited on page 18)

[Twinanda 2017a] 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, January 2017. (Cited on pages xiii, 8, 11, 14, 19, 22, and 36)

[Twinanda 2017b] A. P. Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos.* Ph.d. theses, Université de Strasbourg, January 2017. (Cited on page 20)

[van Amsterdam 2019] B. van Amsterdam, H. Nakawala, E. Momi and D. Stoyanov. *Weakly Supervised Recognition of Surgical Gestures.* 2019 International Conference on Robotics and Automation (ICRA), pages 9565–9571, 2019. (Cited on page 25)

[van Amsterdam 2020] B. van Amsterdam, M. Clarkson and D. Stoyanov. *Multi-Task Recurrent Neural Network for Surgical Gesture Recognition and Progress Prediction.* 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 1380–1386, 2020. (Cited on page 23)

[van Amsterdam 2021] B. van Amsterdam, M. J. Clarkson and D. Stoyanov. *Gesture Recognition in Robotic Surgery: A Review.* IEEE Transactions on Biomedical Engineering, vol. 68, pages 2021–2035, 2021. (Cited on pages 8, 10, 14, 23, 30, and 113)

[van Amsterdam 2022] B. van Amsterdam, I. Funke, E. Edwards, S. Speidel, J. Collins, A. Sridhar, J. Kelly, M. J. Clarkson and D. Stoyanov. *Gesture Recognition in Robotic Surgery With Multimodal Attention.* IEEE Transactions on Medical Imaging, vol. 41, no. 7, pages 1677–1687, 2022. (Cited on page 23)

[van den Oord 2016] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu. *WaveNet: A Generative Model for Raw Audio.* In Arxiv, 2016. (Cited on pages 20, 48, and 58)

[Varadarajan 2009a] B. Varadarajan, C. Reiley, H. Lin, S. Khudanpur and G. Hager. *Data-Derived Models for Segmentation with Application to Surgical Assessment and Training.* In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, pages 426–434. Springer Berlin Heidelberg, 2009. (Cited on page 23)

[Varadarajan 2009b] B. Varadarajan, C. Reiley, H. Lin, S. Khudanpur and G. Hager. *Data-Derived Models for Segmentation with Application to Surgical Assessment and Training.* In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble and C. Taylor, editors, MICCAI, pages 426–434, 2009. (Cited on page 100)

[Varadarajan 2011a] B. Varadarajan. *Learning and inference algorithms for dynamical system models of dextrous motion.* PhD thesis, 2011. (Cited on page 23)

[Varadarajan 2011b] B. Varadarajan and S. Khudanpur. *Learning and inference algorithms for partially observed structured switching vector autoregressive models.* In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1281–1284, 2011. (Cited on page 23)

[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 International Publishing, 2018. (Cited on page 25)

[Vaughan 2016] N. Vaughan, B. Gabrys and V. N. Dubey. *An overview of self-adaptive technologies within virtual reality training.* Computer Science Review, vol. 22, pages 65–87, November 2016. (Cited on page 10)

[Vercauteren 2020] 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, January 2020. (Cited on pages 6, 7, 22, and 112)

[Wagner 2021] M. Wagner, B. P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, A. Reinke, T. Yu, A. Vardazaryan, C. I. Nwoye, N. Padoy, X. Liu, E.-J. Lee, C. Disch, H. Meine, T. Xia, F. Jia, S. Kondo, W. Reiter, Y. Jin, Y. Long, M. Jiang, Q. Dou, P.-A. Heng, I. Twick, K. Kirtaç, E. Hosgor, J. L. Bolmgren, M. Stenzel, B. von Siemens, H. G. Kenngott, F. Nickel, M. von Frankenberg, F. Mathis-Ullrich, L. Maier-Hein, S. Speidel and S. Bodenstedt. *Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark.* ArXiv, vol. abs/2109.14956, 2021. (Cited on page 36)

[Wang 2020] T. Wang, Y. Wang and M. Li. *Towards Accurate and Interpretable Surgical Skill Assessment: A Video-Based Method Incorporating Recognized Surgical Gestures and Skill Levels.* In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 668–678. Springer International Publishing, 2020. (Cited on page 23)

[Ward 2021] T. M. Ward, D. M. Fer, Y. Ban, G. Rosman, O. R. Meireles and D. A. Hashimoto. *Challenges in surgical video annotation.* Computer Assisted Surgery, vol. 26, no. 1, pages 58–68, January 2021. (Cited on page 12)

[Wu 2019] L. Wu, H. Zhao, H. Weng and D. Ma. *Lasting effects of general anesthetics on the brain in the young and elderly: "mixed picture" of neurotoxicity, neuroprotection and cognitive impairment.* Journal of Anesthesia, March 2019. (Cited on page 102)

[Xia 2021] T. Xia and F. Jia. *Against spatial–temporal discrepancy: contrastive learning-based network for surgical workflow recognition.* International Journal of Computer Assisted Radiology and Surgery, vol. 16, no. 5, pages 839–848, May 2021. (Cited on page 21)

[Yang 2017] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos and R. H. Taylor. *Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy.* Science Robotics, vol. 2, no. 4, March 2017. (Cited on pages xv and 103)

[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, vol. abs/1805.08569, 2018. (Cited on pages 12, 25, and 26)

[Yu 2015] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser and J. Xiao. *Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop.* arXiv preprint arXiv:1506.03365, 2015. (Cited on page 35)

[Yu 2019] 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*. International Conference on Information Processing in Computer-Assisted Interventions (IPCAI), 2019. (Cited on pages xiv, 12, 15, 25, 26, 48, 58, 61, 62, 63, 64, 65, 72, 78, 120, and 121)

[Yu 2020] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan and T. Darrell. *Bdd100k: A diverse driving dataset for heterogeneous multitask learning*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2636–2645, 2020. (Cited on page 35)

[Yun 2019] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe and Y. J. Yoo. *CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6022–6031, 2019. (Cited on pages 29 and 71)

[Zhang 2017] H. Zhang, M. Cissé, Y. Dauphin and D. Lopez-Paz. *mixup: Beyond Empirical Risk Minimization*. ArXiv, vol. abs/1710.09412, 2017. (Cited on page 29)

[Zhang 2020] J. Zhang, Y. Nie, Y. Lyu, H. Li, J. Chang, X. Yang and J. J. Zhang. *Symmetric Dilated Convolution for Surgical Gesture Recognition*. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pages 409–418. Springer International Publishing, 2020. (Cited on page 23)

[Zhong 2020] Z. Zhong, L. Zheng, G. Kang, S. Li and Y. Yang. *Random Erasing Data Augmentation*. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020. (Cited on pages 27, 28, and 71)

[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 MICCAI, 2018. (Cited on pages 18 and 19)

Université de Strasbourg

Sanat RAMESH

Multi-level Surgical Activity Recognition

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269 Université de Strasbourg

Résumé

Les innovations en matière de chirurgie mini-invasive ont amélioré les résultats pour les patients, mais ont accru la complexité des flux de travail chirurgicaux. L'optimisation du flux de travail en reconnaissant les activités chirurgicales est essentielle pour fournir une assistance contextuelle. Des recherches importantes ont été effectuées sur la reconnaissance des activités à granularité grossière (phases). Les méthodes de reconnaissance détaillée des activités sont essentielles pour mieux modéliser les flux de travail et faire progresser les capacités des systèmes contextuels. Cette thèse vise à développer des méthodes de reconnaissance d'activité multi-niveaux (phase et étape) à partir de vidéos de bypass gastrique laparoscopique Roux-en-Y (LRYGB). Tout d'abord, nous introduisons un vaste ensemble de données entièrement annoté avec des phases et des étapes et ciblons la reconnaissance conjointe. Ensuite, nous proposons une méthode d'apprentissage faiblement supervisé utilisant les phases comme signaux faibles pour la reconnaissance des pas. Par la suite, nous étudions l'augmentation des données pour un entraînement optimal de ces modèles, en concluant par une étude de généralisation sur un grand ensemble de données multicentriques.

Mots clés: Reconnaissance d'activité chirurgicale, bypass gastrique, reconnaissance de phases et d'étapes, apprentissage faiblement supervisé, augmentation vidéo temporelle, ensemble de données multicentriques

Résumé en anglais

Innovations in Minimally Invasive Surgery have improved patient outcomes but have increased the complexity of surgical workflows. Optimizing workflow by recognizing surgical activities is essential to provide context-aware assistance. Significant research has been done on recognizing coarse grained activities (phases). Methods for detailed activity recognition are essential to better model workflows and advance the capabilities of Context-Aware Systems. This thesis aims to develop multi-level (phase and step) activity recognition methods from Laparoscopic Roux-en-Y Gastric Bypass (LRYGB) videos. First, we introduce a large dataset fully annotated with phases and steps and target joint recognition. Next, we propose a weakly supervised learning method using phases as weak signals for step recognition. Subsequently, we investigate data augmentation for optimal training of these models concluding with a generalization study on a large multi-centric dataset.

Keywords: Surgical activity recognition, gastric bypass, phase and step recognition, weakly supervised learning, temporal video augmentation, multi-centric dataset