Computer Aided Medical Procedures
Prof. Dr. Nassir Navab

Dissertation

Symphony of Time: Temporal Deep Learning for Surgical Activity Recognition

Tobias M. Czempiel

TUM School of Computation, Information and Technology
Technische Universität München

Technische Universität München
TUM School of Computation, Information and Technology

Symphony of Time: Temporal Deep Learning for Surgical Activity Recognition

Tobias M. Czempiel

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Julia Schnabel

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab
2. Prof. Dr. Nicolas Padoy

Die Dissertation wurde am 22.02.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 13.07.2023 angenommen.

Tobias M. Czempiel
Symphony of Time:
Temporal Deep Learning for
Surgical Activity Recognition
Dissertation, Version 1.0

Technische Universität München
TUM School of Computation, Information and Technology
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 and Garching bei München

Abstract

Accurate identification of different activities of surgical procedures is a fundamental aspect of surgical workflow understanding and cognition in the Operating Room (OR). In order to better comprehend and build systems for the complex medical setting of the OR, it is necessary to analyze the activities that take place there, building systems that can understand activities in different hierarchies. One particularly promising area of research is the automatic recognition of surgical phases, which builds the foundation for the development of intra-operative decision support systems and has the potential to improve education and patient safety.

Researchers have investigated different methods to achieve these goals, including deep learning-based methods suited to processing temporal information. In particular, using temporal convolutions is an efficient way of processing sequential data and has been shown to have advantages over other temporal methods, such as recurrent neural networks. Additionally, attention mechanisms, specifically transformers, have been investigated to utilize the temporal reasoning power of deep learning models effectively. Through our research, we have proposed a pipeline for using both temporal convolutions and attention mechanisms to guide the temporal analysis of surgical activities. We utilize temporal convolutions to generate large temporal receptive fields, making them suitable for online activity analysis. We found that the attention weights in our transformer-based method can provide insights into the decision-making processes of the model and can further guide the attention toward descriptive frames for further improvements. We applied both approaches to predicting the phases of laparoscopic cholecystectomy procedures, demonstrating the prospect of these methods in terms of metrics. Our methods, TeCNO and OperA, can be viewed as different members of a symphony in the context of surgical activity recognition, where temporal analysis plays a crucial role. Like in a symphony, each member contributes a unique aspect to create a harmonious whole. The temporal convolutions in TeCNO and the attention-based transformer mechanism in OperA provide insights into the complex temporal patterns of surgical procedures.

Going forward, we are confident that our methods can be applied to a wide range of other clinical concepts, paving the way for a paradigm shift in surgical environments. In the future, signals and information will be automatically analyzed by models capable of understanding the intricate steps involved in complex surgical procedures, providing valuable insights that can aid decision-making and ultimately improve patient outcomes.

Zusammenfassung

Die genaue Identifizierung unterschiedlicher Aktivitäten bei chirurgischen Eingriffen ist ein grundlegender Aspekt des Verständnisses und der kognitiven Erfassung des Arbeitsablaufs im Operationssaal (OP). Um die komplexe medizinische Umgebung des OPs besser zu verstehen, ist es notwendig, die dort stattfindenden Aktivitäten zu analysieren und Systeme aufzubauen, die Aktivitäten in verschiedenen Hierarchien verstehen können. Ein besonders vielversprechender Forschungsbereich ist die automatische Erkennung von chirurgischen Phasen, die das Potenzial hat, die Patientensicherheit und Ausbildung von medizinischem Personal zu verbessern und den Weg für die Entwicklung von intraoperativen Entscheidungsunterstützungssystemen zu ebnen.

Forscher haben tiefe Lernmethoden (DL) untersucht, die zur Verarbeitung von Zeitinformationen geeignet sind, um diese Ziele zu erreichen. Insbesondere ist die Verwendung von temporalen Faltungen eine effiziente Möglichkeit zur Verarbeitung sequentieller Daten und hat Vorteile gegenüber anderen temporalen Methoden wie rekurrenten neuronalen Netzen gezeigt. Zusätzlich wurden Aufmerksamkeitsmechanismen, insbesondere Transformer, untersucht, um die zeitliche Argumentation von DL Modellen effektiv zu nutzen. In unsere Forschung haben wir eine Pipeline vorgeschlagen, um sowohl temporale Faltungen als auch Aufmerksamkeitsmechanismen zur Steuerung der temporalen Analyse von chirurgischen Aktivitäten zu nutzen. Temporale Faltungen sind ein effektiver Ansatz zur Erzeugung umfangreicher zeitlicher Rezeptionsfelder, die für die Online-Analyse von Aktivitäten geeignet sind. Wir haben festgestellt, dass die Aufmerksamkeitsgewichte von Transformer-basierten Methoden Einblicke in die Entscheidungsprozesse des Modells liefern und die Aufmerksamkeit auf beschreibende Frames lenken können, um weitere Verbesserungen zu erzielen. Wir haben beide Ansätze zur Vorhersage der Phasen von laparoskopischen Cholezystektomie Eingriffen angewendet und die Aussichten dieser Methoden in Bezug auf Metriken demonstriert.

Wir sind zuversichtlich, dass unsere Methoden in Zukunft auf eine Vielzahl anderer klinischer Konzepte angewendet werden können und den Weg für einen Paradigmenwechsel in chirurgischen Umgebungen ebnen können. In der Zukunft werden Signale und Informationen automatisch von Modellen analysiert, die in der Lage sind, die komplexen Schritte bei chirurgischen Eingriffen zu verstehen um wertvolle Einblicke zu liefern und bei der Entscheidungsfindung zu helfen, um letztendlich die Patientenergebnisse zu verbessern.

Acknowledgments

I am deeply grateful to my advisor Nassir Navab, whose unwavering support, guidance, and inspiration have been instrumental in shaping my academic journey.

I would also like to extend my thanks to the entire CAMP family, including Martina Hilla, Ulrich Eck, Thomas Wendler, and Ahmad Ahmadi, for their invaluable support.

My time at the chair was nothing short of amazing thanks to the IFL team. I am thankful for Matthias Keicher, Walter Simson, Farid Azampur, Maria Tirindelli, Christine Eilers, Lennart Bastian, Felix Holm, and Ege Özsoy, who not only proved to be great colleagues but also became close friends.

Special thanks to Magda Paschali and Benjamin Busam for their exceptional supervision and continuous support throughout my academic journey. Their inspiring guidance and mentoring, in both academic and personal spheres, have been invaluable from the very beginning of my PhD journey until today.

Thanks to the MITI research group, particularly Dirk Wilhelm and Daniel Ostler, for their exceptional support and expert guidance throughout my academic experience. Additionally, I am immensely grateful to Prof. Nicolas Padoy and all the members of the CAMMA group in Strasbourg for welcoming me into their exceptional group.

My friends have been an unwavering pillar of support and comfort during challenging times, providing invaluable encouragement and companionship throughout my journey.

I am also grateful to my parents, Matthias and Matthilde, and my sister, Mareike, for their love, acceptance, patience, and relaxed attitude, which provided me with stability and grounding. Finally, I am grateful to Coco Hannemann for her support, love, and belief in me, which have enabled me to embrace and enjoy every challenge that lies ahead.

Contents

Part I

Introduction

1 Introduction

Surgery is a field of medicine involving surgical procedures to treat or investigate various pathological conditions. The practice of surgery dates back thousands of years to pre-historic times before written historical records were created [127]. In the classical era Hippocrates (460-371 BC), one of the most noteworthy characters in the development of medicine, was not a promoter of surgical treatment and instead based his medical theory on *vis medicatrix naturae - the healing power of nature*. During the early days of surgery, Hippocrates and many others viewed it as a dangerous final option only to be considered after all conservative treatments had been exhausted [157].

The practice of surgery finally had its breakthrough in the 19th century. Before that, only insufficient anesthesia was achieved with alcohol and opium. In 1846 the dentist William Thomas Green Morton used an ether inhaler to narcotize a patient and successfully removed a neck tumor. Ether was later replaced by chloroform due to its destructive effects on the lungs. Next to the innovation in anesthesia, the advancements in antisepsis improved the survival rate of the patients immensely. In 1867 Joseph Lister found that the death rates for limb amputations dropped from 45% to 15% using linseed oil and carbolic acid as antiseptics. It is estimated that driven by the inventions of anesthesia and antisepsis surgery developed more in this short period than in the two thousand years before [157].

At the end of the 19th century, the first open Gastrectomy was performed, which had previously been impossible due to the high mortality rates associated with the procedure. The progress continued with the development of new inventions such as endoscopy (1868) and X-rays (1896), which led to the emergence of novel surgical directions. These imaging techniques facilitated the introduction of minimally invasive surgery (MIS), an alternative to open surgery, as well [166]. Compared to traditional open surgery, MIS, such as abdominal laparoscopy, can result in less pain for the patient, as well as a faster recovery and shorter hospital stays. In MIS, small ports are used to insert innovative rod-like instruments into the patient, rather than fully opening them up. As a result, the incisions for these ports are much smaller [166]. In addition to the surgical tools, a telescopic rod lens system or digital laparsocope is inserted, which serves as the eye of the surgeon. In 1989 the first laparoscopic cholecystectomy, the removal of the gallbladder, was performed and swiftly adapted to other intervention types. Even though the complexity of the surgery is increased with laparoscopic surgery, the aforementioned advantages are considerable. One way of reducing this new complexity, such as the challenging hand-eye coordination, which is inverted in laparoscopic surgery, is the use of robotic systems holding and guiding the tools while the surgeon controls the robot and does not interact directly with the patient. The combination of laparoscopic techniques and robotic surgery became popular in 2001 using a da Vinci robot for cholecystectomy surgery. Nowadays, some operation types are performed more frequently in the US using minimally invasive surgery compared to the open approaches [34].

The next era for surgery is fueled by the vast progress in Information Technology and computer assistance introduced to the health care system [83]. Currently, surgeons process specific domain knowledge of their medical specialty, and the surgeons use their experience to couple this knowledge with patient-specific individual data and case knowledge. For the future of surgery, systems capable of analyzing this information objectively and evidence-based manner are desired. A cognitive system capable of understanding different parts of an intervention and guiding and supporting the surgeon has many advantages. This information could be integrated as a surgical cockpit [123] that combines all of the information in one centralized system. In this direction, Surgical Data Science (SDS) research includes a broad range of applications and tasks with the goal of introducing cognitive systems in the OR.

The Surgical Environment 2

Contents

2.1 The Surgeon

To understand how SDS can support the surgeon, we first need to examine the role of a surgeon and the abilities required to fulfill this role.

Nowadays, the role of a surgeon is not just limited to performing surgical procedures, but also encompasses the ability to make critical decisions, work effectively as a member of a medical team and communicate with patients and their families. A surgeon must think quickly and make decisions under pressure while maintaining a high focus and attention to detail.

To become a surgeon, one must go through a rigorous educational process that can vary in different countries but always includes many years of school followed by medical school and surgical residency. After completing their residency, a surgeon may choose to specialize in a particular area of surgery, such as general surgery, orthopedic surgery, or neurosurgery, by completing an additional fellowship.

A surgeon should also possess specific physical abilities such as fine motor skills, manual dexterity, and the ability to stand for long periods of time. Good vision and the ability to stay focues for long periods of time are further necessary attribtues. In addition, a surgeon must be able to handle the physical demands of surgery, such as lifting and moving patients, as well as be able to tolerate the stress of the operating room environment. In addition to the physical attributes, specific mental and emotional capabilities are required, including the ability to manage stress and the pressure of making critical decisions.

Open surgery is the most traditional surgery, and the surgeon performs the procedure by making a large incision in the patient's body to access the area in question. The surgeon must have a good understanding of the anatomy and the surgical procedure and the ability to work with their hands in a precise and controlled manner.

During **Laparoscopic surgery**, the surgeon utilizes a laparoscope (figure 2.3), which is a slender, illuminated tube equipped with a camera, and small incisions to access the affected area. Due to the limited visibility, the surgeon must have a comprehensive understanding of laparoscopic techniques and be highly skilled in using specialized instruments. Additionally, having a good mental model of the anatomy is crucial for performing this type of surgery effectively.

Robotic master-slave surgery involves the surgeon sitting at a console and directing the robotic arms that carry out the procedure. A good understanding of the robotic system and the ability to work with specialized instruments are necessary for performing such surgery. The da Vinci surgical system is an example of such a system, offering improved precision, control, and dexterity. Proper training and certification are required for the specific robotic system to be operated by the surgeon.

In summary, a surgeon's role already faces numerous demands, including a wide range of medical knowledge, patient-specific data analysis, motoric skills, practical communication skills with medical teams and patients, stress-coping skills, fast decision-making in life-threatening situations, and understanding and proficiency with the latest technologies.

To enhance the safety, efficiency, and effectiveness of surgeries, there is a growing need to adopt more quantitative and objective-based decision-making processes. Achieving this requires analyzing vast amounts of patient and surgery data. However, processing such large volumes of data in real-time or within a limited time frame without introducing biases is a monumental task that surpasses human capabilities. Surgeons, therefore, require advanced computing solutions to assist them in these tasks [98].

2.2 Surgical Data Science

Surgical Data Science (SDS) is an interdisciplinary field that combines surgery and data science to support the decision-making process in surgery. The definition of SDS can be traced back to 2017 [98], and its primary objective is to optimize the quality of interventional healthcare by capturing, organizing, analyzing, and modeling patient and surgical data [96]. SDS provides surgeons with a quantitative approach to decision-making, thereby allowing for personalized patient treatment and improved surgical outcomes. This is achieved through analyzing structured and unstructured patient and surgical data, leveraging advanced computational techniques to extract meaningful insights.

Structured data sources in the operating room (OR) are well-organized data that can be easily analyzed [54], such as vital signs collected in a consistent format. On the other hand, unstructured data sources in the OR, such as video or audio recordings of surgical procedures, pose a challenge in analysis and require sophisticated techniques to extract relevant information.

The systematic collection and analysis of structured and unstructured data from the OR have the potential to enhance the understanding of surgical procedures, leading to the development of more effective decision support systems.

2.3 The Operating Room

A typical operating room (OR) is a sterile environment where surgical procedures are performed. It is generally equipped with various tools and equipment to ensure the procedure can be performed safely and effectively. In the earlier days of surgery, common rooms without special features were used and the patient was laid on a simple table without additional functionalities. This has changed over time and sepcialized OR equipment such as surgical tables and room layouts were developed [142].

To understand what data can be collected in the OR, it is mandatory to examine the OR team, set-up, and tools.

First, we analyze the OR team. In addition to the surgeon, there are several other roles in an OR team with dedicated responsibilities and tasks [18]:

- Anesthesiologist: responsible for administering anesthesia to patients and monitoring their vital signs during the procedure.
- Surgical nurse: preparing the patient for the procedure, setting up the OR, and helping the surgeon during the procedure.
- Surgical technologist: preparing and sterilizing the instruments, setting up the OR, and helping the surgeon during the procedure.
- Circulating nurse: maintaining the sterile environment in the OR, preparing and handling instruments and suture materials, and helping the surgeon during the procedure.
- Scrub tech: This person is responsible for passing instruments and other supplies to the surgeon during the procedure.

Each OR team member performs different action steps during surgery and plays a critical role in the procedure's success. Collecting data points on the single action steps of the OR team members can provide valuable insights into the surgery procedure [1].

Second, we examine the equipment and tools in the OR. Some of the tools and equipment that may be found in an OR include:

- Surgical tables: These tables are used to position the patient during the procedure. They are typically adjustable and can be tilted or raised to provide the best position for the surgeon.

- Surgical lights: These lights provide bright and even lighting to the surgical area, allowing the surgeon to see the area clearly.

- Surgical instruments: These include a variety of instruments such as scalpels, scissors, forceps, and retractors. The instruments are used to make incisions, remove tissue, and repair or suture blood vessels.

- Anesthesia equipment: This equipment is used to administer anesthesia to the patient, including anesthetic gases, intravenous (IV) lines, and monitors to measure the patient's vital signs.

- Imaging equipment: This equipment provides visual images of the surgical area, such as x-ray machines or ultrasound machines.

The tools used in surgery depend on the type of surgery: open, laparoscopic, and robotic, as shown in figure 2.1.

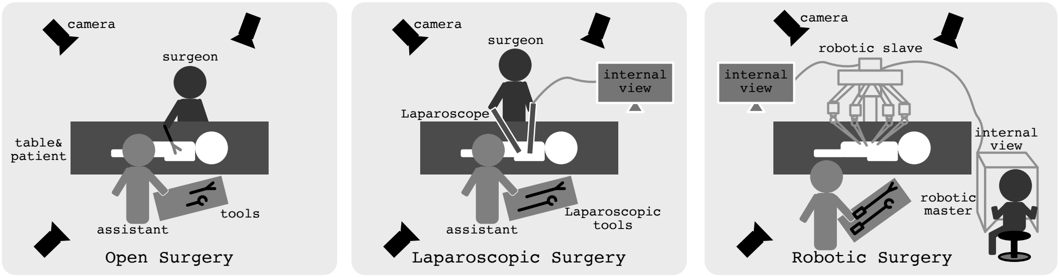


Fig. 2.1. Different surgery types left to right. Open surgery, minimally invasive Laparoscopic, and Robotic surgery. The cameras observing the external OR scene might be available in the future

Open surgery, also known as conventional surgery, is a traditional surgical procedure with a large incision to access the required body part. The tools used in open surgery typically include a variety of scalpels, scissors, forceps, and retractors. Scalpels are used to make incisions, scissors are used to cut through tissue, forceps are used to grasp and hold tissue, and retractors are used to hold open the surgical area [18].

Minimally invasive laparoscopic surgery is a surgical technique with small incisions, and instruments and a camera are inserted to visualize and operate on the internal organs. The tools in laparoscopic surgery are typically specialized instruments designed to be used through small incisions. These include a laparoscope, which is a thin tube with a camera and light attached, described in figure 2.3, as well as long, thin instruments such as graspers, scissors, and dissectors. Additionally, specialized instruments such as staplers are used to perform specific tasks.

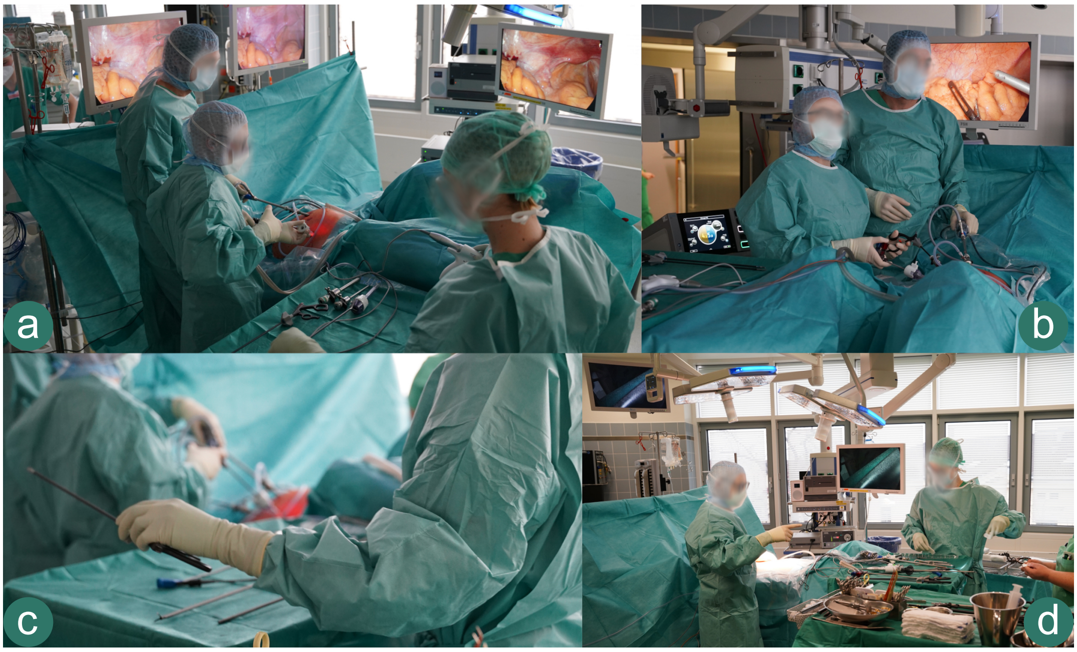


Fig. 2.2. Example of a laparoscopic procedure and surgery setup. a) The surgeons and assistant work together when changing tools. b) The assistant surgeon is holding the camera for the head surgeon. c) The assistant is preparing a laparoscopic instrument. d) OR setup with many different tools has to be organized. *Photo credit goes to MITI research group lead by Prof. Wilhelm*

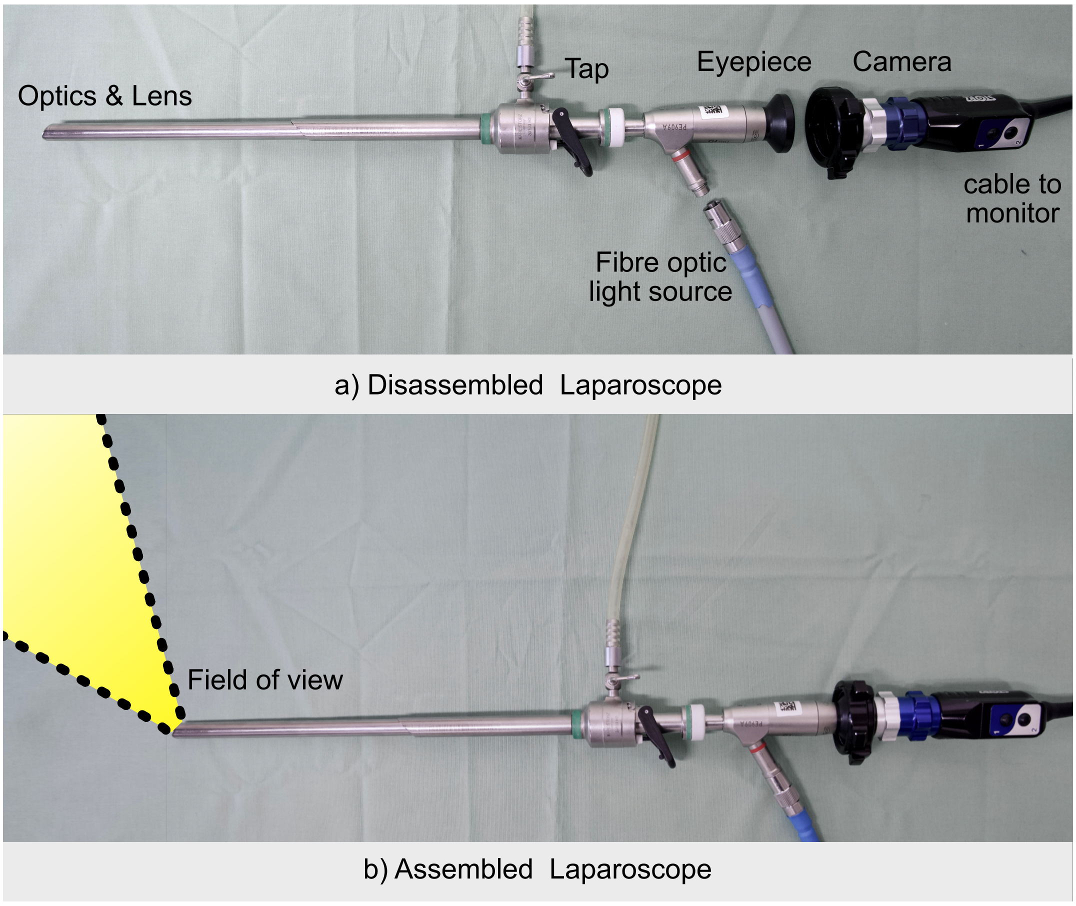


Fig. 2.3. In a) a disassembled laparoscope is shown and the different parts are named. In b) the field of view of the assembled laparoscope is visualized, which is not straightforward following the line of the rod but is sideways. The sideways field of view can be an advantage in certain procedures, such as gallbladder surgery, where the gallbladder is located in a deep and difficult-to-access area. *Photo credit goes to MITI research group lead by Prof. Wilhelm*

Robotic surgery [155] is a minimally invasive surgery that utilizes robotic technology to enhance the surgeon's ability to perform precise movements. The tools used in robotic master-slave surgery, such as the da Vinci surgical system [10], include robotic arms, which are controlled by the surgeon from a console as seen in figure 2.4, and a variety of instruments that can be attached to the arms, such as scissors, graspers, and dissectors. The da Vinci system also includes a high-definition 3D camera, which provides a magnified and detailed view of the surgical area. Using robotic systems the surgeon does not physically touch the patient if no emergencies or anomalies appear that force the surgeon to transition to open surgery.

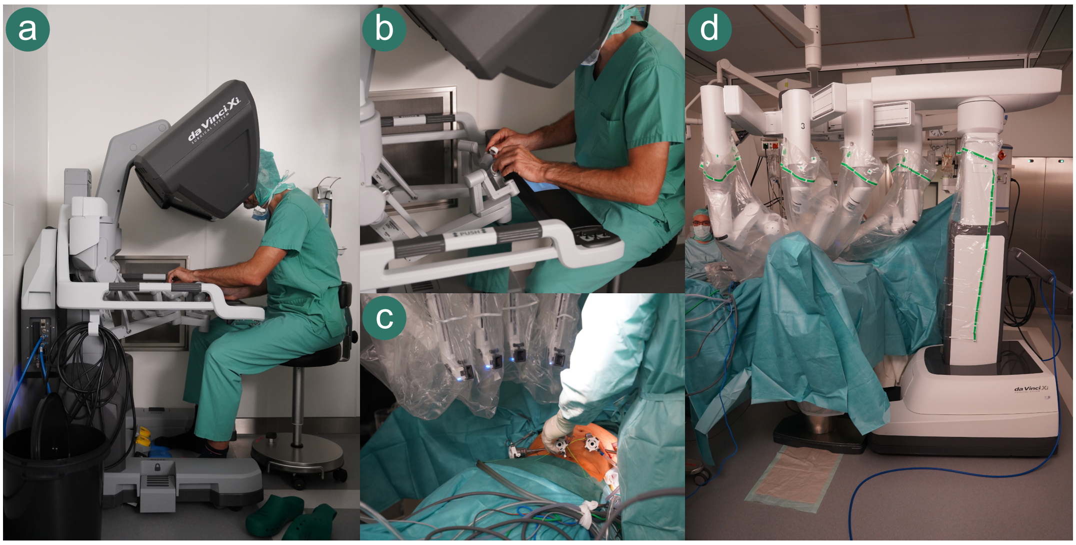


Fig. 2.4. Robotic Surgery (daVinci) a) shows the surgical console of the system (Master), b) detailed view of the human-machine interface c) ports for insertion of the tools into the abdominal cavity d) robotic system (Slave) with draping. *Photo credit goes to MITI research group lead by Prof. Wilhelm*

The overview of the roles, equipment, and tools in the OR highlights the importance of collecting data points on the OR environment and team to gain valuable insights into the surgical procedure. In the next chapter, we will discuss the signals in the OR that can be measured and recorded.

2.4 Signals in the OR

Measuring, displaying, and analyzing signals is crucial in modern operating rooms as they provide precise and accurate information to medical personnel during surgeries. They enable real-time monitoring of patients' vital signs, such as heart rate, blood pressure, and oxygen levels, and can be used for controlling medical equipment and devices, such as infusion pumps and ventilators. Signals allow for more efficient and effective medical treatments and reduce the risk of human error.

The digital signals in the operating room [156] can stem from many data sources. Some data sources are more structured, like patient information [54], anesthesia data, and tool data, while other data sources, like medical images and audio or video recordings, are more unstructured data sources. Each type of signal has unique advantages in terms of its ease of measurement and the feasibility of extracting meaning through analysis. In figure 2.5 an overview of common surgical signals is pictured.

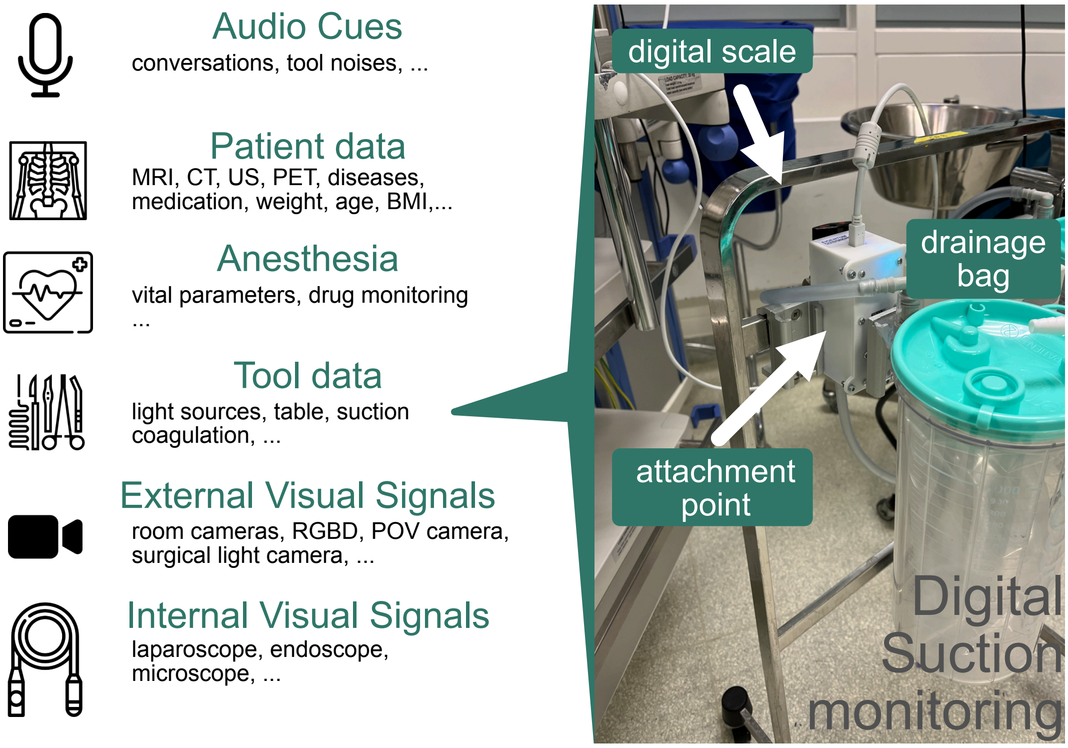

Audio Cues
conversations, tool noises, ...
Patient data
MRI, CT, US, PET, diseases, medication, weight, age, BMI,...
Anesthesia
vital parameters, drug monitoring ...
Tool data
light sources, table, suction coagulation, ...
External Visual Signals
room cameras, RGBD, POV camera, surgical light camera, ...
Internal Visual Signals
laparoscope, endoscope, microscope, ...
digital scale
drainage bag
attachment point
Digital Suction monitoring

Fig. 2.5. Signals in the OR by categories. To digitally measure the suction information, an example digital suction monitoring system is shown on the right using a scale to measure the volume and flow of liquids. *This figure has been designed using images from Flaticon.com. Photo credit goes to MITI research group lead by Prof. Wilhelm*

Patient records

Patient records, including preoperative medical imaging from sources such as MRI [168], CT [55], Ultrasound [28], or PET [128], are valuable information for medical personnel in the operating room. This information can be obtained from hospital information systems and provide insight into pathologies' location, size, and severity. Medical imaging data provides a more comprehensive picture of the patient's condition and is obtained before the surgery, unlike visual information obtained during the surgery. Patient information such as prior diseases and medications should also be stored in centralized patient records. Other factors such as patient weight, age, and BMI can also be helpful in certain situations for automatic methods.

Anesthesia

The information obtained from anesthesia is a valuable data source that can contribute to the standardization and interoperability of OR processes. The vital signs of patients, such as heart rate and narcotic state, can be used as a reliable indicator of OR occupancy [173]. Furthermore, detailed observation of these vital signs can facilitate promptly identifying adverse events, such as bleeding, by detecting rapid changes. Continuously monitoring anesthesia information over time provides a valuable resource, particularly in life-threatening situations, without additional hardware.

Tool data

Under the category of tool data, we encompass the activation states and levels from fixed OR equipment and specialized surgical instruments. The state of the light source or OR table [116] can provide valuable information, particularly in regard to the start and end of the surgical procedure. The surgical tool state from a coagulation device can be monitored through the digital output of the machine, which is a rich source of information as it reflects the interaction with tissue for coagulation or cutting. Using radiofrequency identification (RFID) instruments can be detected in real-time [80]. Other equipment, such as the suction device, can be monitored using specially designed hardware tools. As shown in Figure 2.5, we illustrate a digital suction monitoring system that consists of a digital scale attached to the drainage bag of the suction system. By using the density of blood, this system allows for the direct translation of weight increase into total blood volume. Furthermore, the change in blood volume over time can be calculated to assess the amount of bleeding.

Audio Cues

The audio cues within the operating room hold valuable information about the procedure. Recording conversations between the head and assistant surgeon makes distinguishing between stressful situations and routine steps easier. Commands given by the surgeon to the scrub tech can also be used to predict the following tool needed. Additionally, ambient tool noise, such as those from drills or saws, can provide valuable information on the timing of specific events [117].

However, the signal-to-noise ratio of audio cues in the OR is often not optimal, with irrelevant conversations between medical staff and background noises, such as fans, coughing, or the opening and closing of doors, having limited value. Using audio cues in the operating room also raises privacy concerns for the patient and medical staff, making their acquisition a rare occurrence.

2.4.1 Visual Signals

Visual information is critical in the modern operating room, providing real-time insight into the surgical environment. These signals can be broadly categorized into two groups - internal and external [38] as shown in figure 2.6. The internal group encompasses video signals that are generated within the OR itself, such as images captured by laparoscopes, endoscopes, and other similar instruments. These images give the surgeon a direct view of the internal anatomy and allow for precise surgical interventions. On the other hand, the external group encompasses visual signals generated from cameras positioned outside the OR, such as room cameras, point-of-view cameras, or cameras attached to robots. These cameras capture a broader view of the OR and provide important information about the overall surgical environment.

One advantage of visual signals is that they can be easily obtained in real-time during the surgical procedure using cameras. Another advantage is the level of detail and situational awareness provided by cameras, as they capture a visual representation of a known process. Problems or errors in the recordings can be more easily identified and discovered. Cameras can be used to monitor multiple aspects of the procedure, such as instrument positioning and patient anatomy, at the same time.

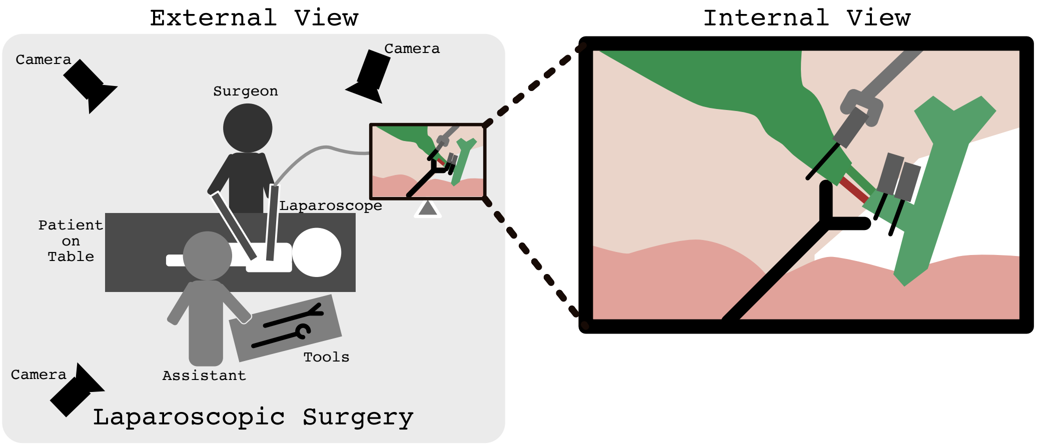


Fig. 2.6. External view of the OR and internal view inside the patient using minimally invasive surgery techniques. The external OR cameras observe the room and might be available in the future.

2.4.2 External Visual Signals

External visual signals in the OR can come from various types of cameras, camera-angles and puroses.

POV cameras

Point of view (POV) cameras worn by the surgeon can provide valuable information about the surgical perspective and allow for remote observation of the procedure. Currently, POV cameras are often integrated into head-mounted displays, such as the Microsoft HoloLens[1], providing an immersive experience for remote observers. Although POV cameras have the potential to provide benefits, their widespread adoption has been hindered by challenges. The use of POV cameras during surgery can interfere with the surgeon's line of sight by obstructing their view or causing distractions, which can potentially compromise the safety and accuracy of the procedure. With the advent of Augmented Reality (AR) technology [106], the use of POV cameras in the OR may become more prevalent in the future. POV cameras can provide a unique perspective and allow for iris tracking, which can provide insights into the surgeon's focus and decision-making process during the procedure. This information can be helpful in training and evaluation purposes, and may ultimately lead to improved surgical outcomes.

[1] https://www.microsoft.com/en-us/hololens

Robotic cameras

The robotic knee replacement system, MAKO surgical system from Stryker[2], is an example of a medical device that includes an external camera. However, it may not be possible to access the raw camera stream directly due to lack of standardization and concerns regarding the admission as a medical product. This limits the ability to use the visual information from existing external cameras for additional purposes. These cameras could provide useful information about the procedure, including the positioning of instruments and the patient's anatomy, as well as a more complete and accurate record of the surgical process.

Room Cameras

External room cameras in the operating room can capture the overall OR environment during procedures. These cameras are typically mounted on the walls or ceiling of the operating room and provide a broader perspective of the surgical scene. They are used for various purposes, such as monitoring the activity of the entire surgical team, providing a reference for patient position and instrument placement, and recording the procedure for education and review purposes.

Multi-view RGB room cameras provide a complete view of the surgical field [119], reducing the effect of single-view occlusions. These cameras use multiple viewpoints to capture images from different angles, which are combined to create a panoramic view of the OR. This allows for a more accurate representation of the surgical scene, particularly when tracking the movement of medical instruments and observing the patient's anatomy. A multi-view setup makes it possible to maintain visibility even when an object or a person moves into the view of one camera, as the other cameras can continue to capture the essential details.

In addition, RBGD cameras capture depth information and can be used to provide 3D scene representations [11]. This data can provide a more comprehensive view of the operating room, including the spatial arrangement of objects and the relative positions of surgical tools and the patient. Viewing the scene in 3D can help reduce occlusions and clarify the procedures being performed.

2.4.3 Internal Visual Signals

Internal cameras in the operating room are placed inside the body to observe the internal anatomy visually. These cameras are typically used in minimally invasive procedures, such as laparoscopy and endoscopy, to provide the surgeon with a view of the target area [104]. Internal cameras have the advantage of directly observing the surgical site delivering an accurate and detailed picture of the anatomy. Due to their small size, they can be inserted into tight spaces and used in procedures that would otherwise be difficult or impossible to perform using traditional open surgery. The images captured by internal cameras are usually displayed on a monitor in the operating room, providing the surgeon with real-time information to guide the procedure.

[2] https://patients.stryker.com/knee-replacement/options/mako-robotic-arm-assisted-total-knee

In recent years, stereo laparoscopy [104] has become a widespread technique in robotic surgery. Stereo endoscopes use two cameras to provide a stereo image, allowing the surgeon to perceive depth information and better control the robotic instruments. In addition to other advancements, Near-infrared laparoscopy produces views that are beyond what the human eye can perceive. Near-infrared laparoscopy is a surgical visualization technique that utilizes a near-infrared light source to enhance the visibility of tissues [5], particularly blood vessels, and nerves, during minimally invasive procedures. Near-infrared light penetrates the tissues deeper than visible light, allowing the surgical team to visualize the targeted structures clearly. This can be particularly useful during complex surgeries, where preserving blood vessels and nerves is crucial to ensuring positive patient outcomes.

2.4.4 Challenges

Interoperability

One of the significant challenges in the field of signals in the OR is the lack of interoperability between medical devices. Many medical devices operate as closed systems, making it difficult or impossible to extract and exchange data from them [169]. Even when data export is possible, the data formats are often inconsistent and do not conform to a common standard. Although protocol standards such as HL7 [16] exist, they come in many different variations, making it challenging to transfer information between devices. Efforts to promote common standards and interoperability in the OR are underway, with consortiums such as *OR.NET* [79] and *Smart Cyber Operating Theater* [115] working towards the goal of creating a more interconnected surgical environment. However, device manufacturers may not see the need to support these efforts, as they can benefit from the current lack of standardization by selling additional devices for integration with their equipment.

Data regulations

Surgical data collection, management, and usage are challenging tasks that require strict adherence to established standards of security and fidelity. The data collected in SDS often includes patient-specific information, meaning that data management in healthcare, particularly in surgery, must comply with various rules and regulations. Unfortunately, the healthcare data governance field is less grown than in other domains, making collecting and managing SDS data even more complicated. The General Data Protection Regulation (GDPR) was introduced in 2018, providing strict guidelines for collecting and handling personal data in the European Union (EU), including the entry and exit of data to or from the EU. Similarly, strict regulations exist in the USA and many other countries. Overall, it is essential to understand the regulatory framework and standards in place before engaging in SDS data collection and management to protect the privacy and security of patient data [97].

2.5 Operating Room of the Future

While the surgery and operating room of the future hold great promises, it is essential to note that many of these technologies are still in the development phase and have yet to become widely available. The OR of the future was already defined in previous works. In

2005 the OR2020 was defined [33] and in 2017 [98] the OR2030. The goals and predictions for the year 2020 could mainly not be achieved, and also the predictions for the year 2030 seem ambitious. Advancements in the field of surgery, including robotic surgery, and artificial intelligence, are continually evolving and improving. Still, it may be some time before they are fully integrated into surgical practice. Nonetheless, the potential benefits of these technologies make the future of surgery an exciting prospect.

The Operating Room of the future promises to be a highly advanced and integrated system, with all the information seamlessly flowing into a centralized system. In figure 2.7 the different components of the OR of the future are visualized. Different information sources are observed from the surgery but also pre-surgical data such as medical imaging from CT, MRI, and US. Additionally, the genetics and prior diseases of the patient are analyzed. During the procedure, cameras and OR room sensors observe all the activities, and the information is forwarded to machine intelligence to extract activity information and also creates a notification, e.g., for adverse events or wrong instrument use. This data, in combination with the other information sources, accumulated in a surgical cockpit [99]. This physical appearance of the surgical cockpit has yet to be defined, but one possible scenario would be that the cockpit is outside of the actual OR, and an expert is observing the parameters similar to a control tower in aviation. One cockpit for multiple surgeries at the same time could be a possibility when the critical steps of the different ORs are scheduled at different times with advanced resource management [17]. Other possible scenarios would be that with the use of AR, the surgical cockpit is integrated into head-mounted displays that the medical staff in the OR is wearing. This integration of AR would also allow remote experts to virtually join the intervention to guide novice surgeons in critical steps [106]. From the surgical cockpit, information can then be filtered and refined and brought directly into the OR. Here the medical staff is informed about and warned if the machine intelligence system did recognize any anomalies. This cognitive system is a key step for the OR of the future, as the visualization of all the data would overload the surgeon with information. A unified surgical display [150] could be used to identify the crucial information that the surgeon needs at different steps of the surgical workflow. The goal of the OR of the future is to provide real-time support to surgeons, enabling them to make informed decisions and improve the outcomes of surgical procedures. Thus, developing a cognitive surgical system that utilizes machine intelligence for a deeper understanding of the activities in the OR is crucial for the future of surgery.

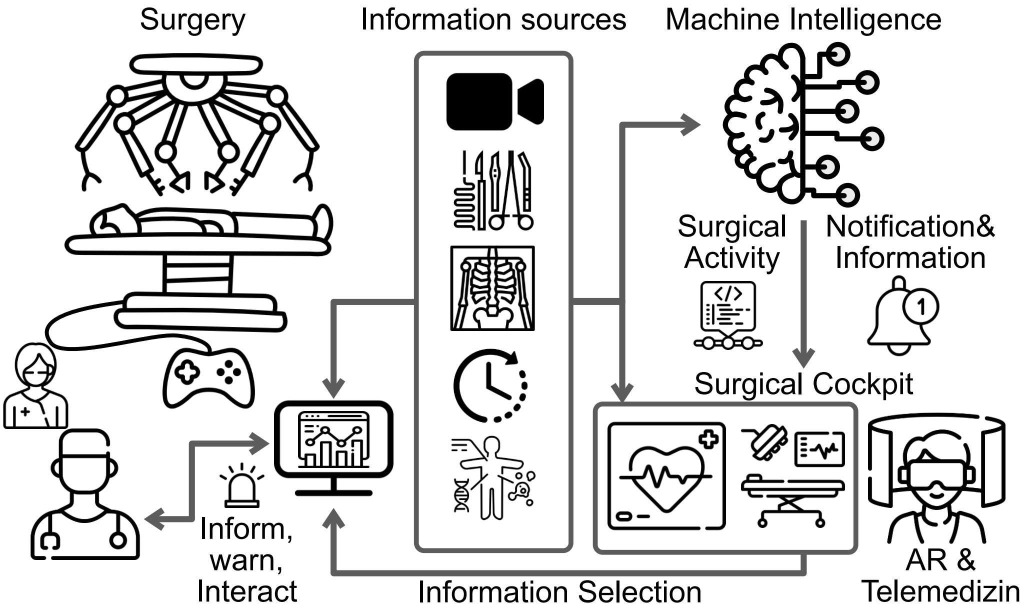


Fig. 2.7. The OR of the future consists of multiple parts besides Surgery. The information will be collected in a centralized system, and machine intelligence will be used to generate surgical activities and notifications. The surgical cockpit accumulates the information and has interfaces for telesurgery and AR and also towards the surgical room where the surgeon and medical staff are informed and warned and can interact with the system to see the information that is needed. *This figure has been designed using images from Flaticon.com*

3 Activity Recognition in the Operating Room

Contents

3.1 Introduction

Activity recognition in the operating room refers to identifying, analyzing, and categorizing the various actions, movements, tasks, and semantic phases performed by healthcare professionals during surgical procedures. The surgical workflow can be summarized with a sequence of events on various abstraction levels and details. Using surgical activity recognition systems, the goal is to automatically identify the events or activities using different signals, including external and internal cameras and digital tool signals. Developments in other areas further influence the progress of surgical activity recognition until today.

The recognition of the **surgical phase** was one of the first problems approached by utilizing binary surgical signals and tool states in the OR, such as scissors, graspers, and the suction device. The underlying surgical phase was recognized by comparing the sequence to a known sequence [2].

There exists a strong link between the analysis of surgical activities and the **surgical instrument**, and the joint recognition of both can be beneficial for both tasks. Automatic surgical tool recognition, initially performed with classical methods [146] by recognizing the tips of the instruments, is, therefore, a very relevant step for understanding the actions in the OR. Instruments could be tracked later with template matching [136] and more advanced learning-based methods. Different challenges [1][2] for tool detection [7] and also segmentation [6] brought further novelties into the field.

By tracking the instruments, the task of **surgical skill analysis** was presented [147] combining the idea of a surgical action with surgical skill such as tying a knot. More recent studies performed in vivo skill assessment on real surgeons assisted by machine intelligence meth-

[1] https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org

[2] https://robustmis2019.grand-challenge.org

ods [85]. The knowledge about the surgical skill level of a trainee can be an essential cue to understand further and classify actions and find anomalies in the workflow of the OR.

A more detailed and fine-grained analysis of laparoscopic surgeries was proposed using **Surgical action triplets** [72, 109] especially useful for complex scenarios. A triplet consists of an instrument, an action verb, and affected anatomy defined as <instrument, verb, target>. The direct prediction of action triplets from images is challenging but allows us to understand and investigate the processes in the OR in more detail [112, 114]

To increase safety in the OR, it is essential to analyze the activities and workflow and to monitor critical steps that can if performed poorly, negatively impact the patient. To address this, **critical view of safety** (CVS) for laparoscopic procedures was proposed to reduce the risk of bile duct injuries during cholecystectomy procedures. To achieve the CVS, essential anatomy must be visible on the laparoscopic video frame before the surgeon cuts the tissue [102]. To check whether or not CVS was achieved, computer vision methods can be used [101, 103].

On a very high abstraction level, recognizing the **procedure type**[69] can be challenging due to the high visual similarities between different procedure types, especially in MIS. The recognition of the **remaining surgery duration** (RSD) was also addressed, motivated by the need for smarter OR scheduling systems [3, 162].

Most of the works focused on recognizing laparoscopic videos or tool signals. But also, recognizing the **external workflow** in the OR is challenging. Using multiple ceiling-mounted cameras situated in different positions and analyzing the scene, the problem of occlusions in the complex OR environment can be addressed [149]. The complexity is further increased by introducing robotics to surgery, and the analysis of the **activities in robotic surgery** can help to compare robotic and minimally invasive robotic workflows [141, 145]. Different sensors have been used to analyze the external surgical workflow phases to approach the complexity and visual challenges [11].

The analysis of phases is very high-level information and does not explicitly encode the small activities in the OR between humans and objects. With the concept of **scene graphs** for the operating room [118], visual activities and interactions and semantic connections between tools can be encoded. Semantig surgical scene graphs can be used to understand the details of the surgery encoded as triplets defined as <subject, predicate, object> [119]. This representation can further be used for downstream tasks such as surgical phase or role prediction. Additionally, the scene graphs are a human-readable low-dimensional representation of the OR and can be used as a ritch knowledge base.

For the analysis of **surgical phases**, many innovations brought improvements in performance or showed ways to address the problem with a limited amount of data [176]. With the rise of more robust temporal methods [49], refinement of the predictions could be achieved. Despite advances in these methods, there are still significant challenges remaining for the task of surgical activity recognition, such as variability in patient anatomy, surgeon style, and limited availability and quality of training data.

3.2 Activity Granularity

In surgical process modeling, the notion of **Activity Granularity** was presented by Lalys et al. in their 2014 review of surgical procedures [83]. Although the aim of representing the surgery as a series of events is apparent, the level of detail at which this is accomplished is highly subjective.

The definition of the granularity level can be subjective but previous works proposed to differentiate between six primary granularity levels in the OR. **Low-level** information is defined on a frame or image level and describes the presence or absence of a person, structure, or tool. This is followed by the **Motion** granularity level, which describes a single movement, e.g. of a tool. Multiple motions represent an **Action**. The Action is the next higher granularity level and depicts a simple task over a short duration of time. For a real-world scenario, this could be as simple as e.g. opening a bottle, and for the laparoscopic case, it could be cutting tissue. The next higher hierarchy is the **Step** which is also often referred to as a task in the literature. One Step consists of a couple of short actions. Multiple Steps can further be summarized into Phases, and the succession of all phases describes one surgical **Procedure**.

In the literature, the terms action and activity are used interchangeably. In this thesis, activity is used to describe either of the four granularity levels Motion, Action, Step, or Phase [161]. Low-level information, such as the simple presence of an object, cannot be used to describe the surgery workflow and is therefore not considered an activity. The procedural level is already the highest granularity level which is why it is also not considered an Activity.

In this thesis, we want to focus on methods for analyzing surgical activities capable of encoding long temporal sequences. These methods consider many of the preceding information of an intervention to predict the current activity. For very fine granularity levels, the necessity of a temporal model might not be efficient. For the recognition of a motion of a tool, the knowledge about a particular step at the beginning of the procedure will be irrelevant. The higher we go in the granularity axis from motions toward actions, steps to phases, the more critical it becomes to use models with a long-term temporal understanding for the prediction task. The use of temporality can be compared to the mental model of a typical surgery that the medical staff uses to anticipate the next steps. This mental model is trained with practice and experience of the team [150], and we aim to translate this concept into automatic approaches.

In figure 3.1, the different granularity levels [83, 161] are depicted on the x-axis, and the potential benefit of a large temporal window is on the y-axis. Additionally, we summarized the different surgical activity applications from section 3.1 on the right and tried to align them towards a granularity level. The brackets visualize a range as a clear definition for the task towards one granularity level can be difficult.

While surgical tool detection and segmentation applications can profit only to a limited extent from a sizeable temporal window and temporal information, it is more beneficial for high-level activities. For the recognition of the **Surgical Phase**, larger temporal window is desired.

Likewise, Procedure Prediction and Remaining Surgery Duration prediction can be vastly improved by the addition of a sizeable temporal window.

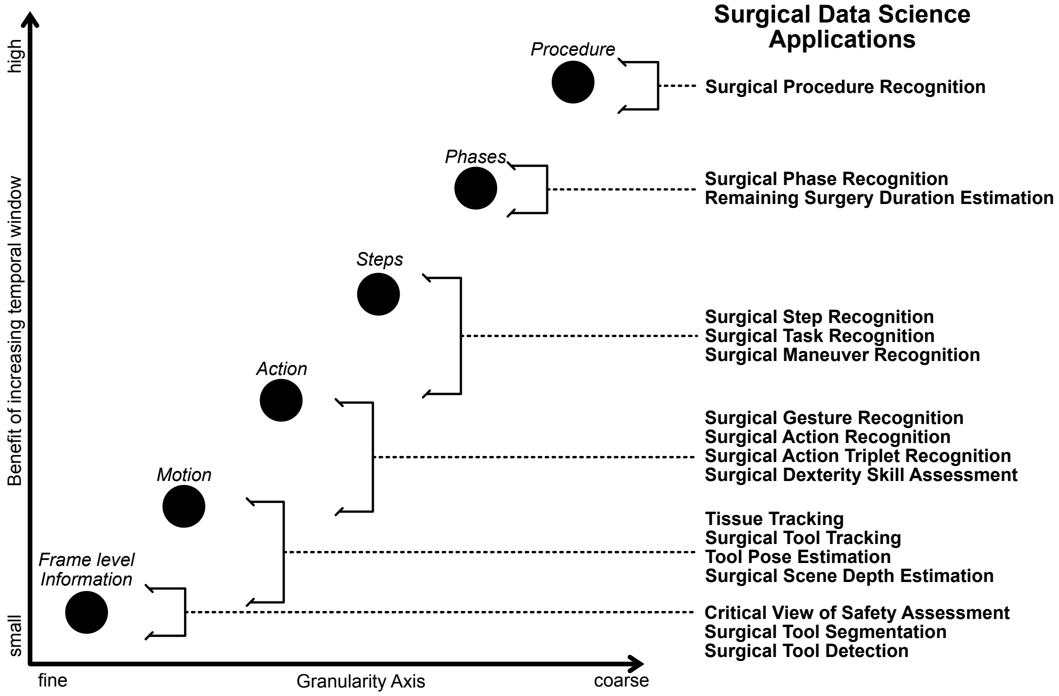

Surgical Data Science Applications
Procedure
Surgical Procedure Recognition
Phases
Surgical Phase Recognition
Remaining Surgery Duration Estimation
Steps
Surgical Step Recognition
Surgical Task Recognition
Surgical Maneuver Recognition
Action
Surgical Gesture Recognition
Surgical Action Recognition
Surgical Action Triplet Recognition
Surgical Dexterity Skill Assessment
Motion
Tissue Tracking
Surgical Tool Tracking
Tool Pose Estimation
Surgical Scene Depth Estimation
Frame level Information
Critical View of Safety Assessment
Surgical Tool Segmentation
Surgical Tool Detection
high
Benefit of increasing temporal window
small
fine
Granularity Axis
coarse

Fig. 3.1. The x-axis depicts the degree of refinement of activities, ranging from fine to coarse. The y-axis portrays the advantages of an increasing temporal window, from small to high. On the right-hand side, diverse applications of Surgical Data Science are presented and categorized according to granularity and temporal utilization.

3.3 Applications of Surgical Activity Recognition

The task of surgical activity recognition has been approached through various algorithms, each designed to address specific use cases. This section will provide an overview of the different algorithm classes, their applications for surgical activity recognition, and the potential for clinical values [100]. These algorithms can be divided into two main categories: online and offline (section 6.1). **Online** algorithms are employed in time-sensitive applications and in vivo and real-time predictions during in vivo interventions, while **Offline** algorithms are utilized in tasks that do not require immediate results and can be processed later when the data is recorded.

The applications and goals for these algorithms can be classified based on the level of granularity (3.1), to improve patient safety, reduce surgical errors, and facilitate communication in the operating room. As such, the analysis of surgical activities is a crucial step toward realizing the vision of the next-generation operating room [96, 98].

3.3.1 Motion, Steps & Action

This section discusses the fine granularity activity recognition cases, Motion, Step, or Action, and their use cases and applications.

The **offline recognition of motion**, such as surgical instrument tracking, serves as input to a variety of downstream tasks and could be used for Instrument motion tracking has been used to estimate the dexterity of surgical skill [65, 147] with the assumption that the path length of the tracked surgical instrument is influenced by the skill of the surgeon. This information could also be valuable for medical device manufacturers and could help design tools that are more fitting for particular surgical **steps** or **actions**. Particular motion and action patterns can also be used to label or tag particular intervention sections, allowing for fast querying of similar steps in other procedures for educational purposes.

The **online recognition of motion** can be used to identify the concentration or exhaustion level of the surgeon. The online recognition of tools and anatomy also builds the foundation for evaluating the critical view of safety, a task with direct clinical impact. Additionally, the tracking of anatomy and tools and even motion estimation [129] can serve as a way to notify the surgeon before approaching no zone area, a critical structure that should not be manipulated due to high risk of bleeding or permanent damage [148].

Surgical triplets are one way to encode surgical action at a more granular level. The detailed description of the subject, predicate, and object can further help understand the different parts of the surgical workflow. This fine level of understanding between the instrument (subject), tissue (object), and type of manipulation (predicate) helps to create increasing depth in the understanding task. The low dimensional representation of surgeries in these action triples helps to identify patterns and gives further insides into tool usage patterns for specific tissue interactions or ways to interpret and analyze the surgery data-driven better to understand the cause for anomalies and surgical errors.

Motion features can be useful for medical device manufacturers to design surgical tools that are better suited to specific steps or actions. This feedback can be iteratively applied to design new and innovative instruments which could be especially useful for MIS.

3.3.2 Phase & Procedural level

The recognition of **offline surgical procedures** can provide a means for categorizing recorded procedures in instances where cameras in the operating room (OR) continuously capture footage over an extended period, such as an entire day. In such a scenario, a large video file is produced that encompasses all procedures conducted during that day. Without manual triggers that signal the start of a new procedure, it can become challenging to assign videos accurately to the correct patients. However, through the implementation of procedure-level classification, it is possible to verify that each patient has been adequately paired with their corresponding video, including information on where the video file should be segmented and assigned to a new patient.

In regards to an automatic recognition system for an online surgical procedure, can be linked to a patient management system. Such a system could alert medical staff if an incorrect procedure is performed. Although it is rare for the wrong type of surgery to be performed on a patient, the potential impact on the patient is substantial. A tool that verifies that the planned procedure matches the actual procedure being performed can aid in reducing the risk of mistakes. Different surgical procedures require different tools to be used in the OR. Typically, the tools are prepared by clinical staff before the surgeon enters the OR. During this process, mistakes can result in delays for other surgeries as additional equipment needs to be procured. By analyzing the surgical procedure and comparing it to the tools that have been prepared in the room, a system could be established to verify that all necessary items required for a successful surgery are available.

The recognition of **offline surgical phases** presents several potential applications. One of these is the clustering and alignment of surgical cases based on the lengths of their phases, which can reveal correlations with patient information such as age, sex, gender, and BMI, or causal relationships between phase duration or order and patient outcomes. By highlighting problems in the surgical workflow, such information could be used to create targeted training systems, addressing specific areas of the OR that result in phase delays or errors.

The automatic extraction of a surgical protocol, which is highly desired for purposes such as archiving, education, and postoperative patient monitoring, could also benefit from offline phase recognition. Surgeons spend significant time documenting each intervention, even though the reports from different patients can be very similar if no anomalies appear. To address this issue, automated reporting systems have been proposed to standardize reports and establish a consistent nomenclature across healthcare providers, hospitals, and regions. Structured reporting, where reports are based on a predefined questionnaire rather than free text, has successfully promoted efficient and standardized reporting, as demonstrated by recent studies in reporting diagnostic findings [73]. Thus, the integration of phase recognition with structured reporting and algorithmic report generation has the potential to reduce the manual burden of report generation significantly.

The recognition of the **online surgical phases** is widely studied in surgical data science. Implementing a reliable and robust system for recognizing surgical phases can significantly impact patient safety and improve the overall efficiency of surgical procedures, as noted in several studies [49, 124].

One valuable application of automatic surgical phase recognition is improving the accuracy of remaining surgery duration (RSD) predictions [20], which is a crucial component of smart OR scheduling systems. These systems can help to optimize the use of OR resources, reduce patient waiting times, and increase hospital profits by maximizing the number of surgeries performed.

The recognition of surgical phases can play an essential role in context-aware decision support systems [61, 124]. During a surgical procedure, surgeons are tasked with making decisions based on limited information, leading to the initiation of different possible workflows. However, the automatic analysis of surgical phases provides objective data to support these decisions and helps to establish procedural standards. This objective data can be used to make informed decisions, reducing the risk of errors or deviations from the standard procedure, and ultimately improving patient outcomes.

Finally, real-time analysis of ongoing surgical interventions can provide valuable feedback to surgical teams and trigger alarm mechanisms in case of deviations or adverse events [61]. By analyzing statistics on individual surgical phases, the system can identify when a phase takes an unusual amount of time, which could indicate a deviation or anomaly.

In this way, surgical phase recognition serves as a valuable input for context-aware decision support systems with many different and diverse use cases and improvements potential for both medical staff and patients.

4 Contribution

In this thesis, we have made two significant contributions to the field of surgical activity and phase recognition. Firstly, we proposed a pipeline that utilizes temporal convolutional networks to efficiently and effectively model temporal relationships over long periods of time. Our work with TeCNO has opened up new avenues for exploring temporal methods in the field of surgical phase recognition. This breakthrough has inspired and motivated other researchers to explore innovative approaches, which is pushing the field towards novel solutions.

- **T. Czempiel**, M. Paschali, M. Keicher, W. Simson, H. Feussner, S.T. Kim, N. Navab. "*TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks.*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Lima, 2020

Next, we demonstrate a new technique for predicting surgical phases using an attention-based transformer method. Our approach involves directing attention toward relevant image features and includes analysis of low and high attention frames. Additionally, we evaluated the attention for each frame to gain deeper insights into the decision-making process of our model, identifying the frames that received the most and least attention for each phase. This analysis is an important first step toward model interpretation and trustworthy predictions.

- **T. Czempiel**, M. Paschali, D. Ostler, S.T. Kim, B. Busam, N. Navab. "*OperA: Attention-Regularized Transformers for Surgical Phase Recognition.*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

5 Outline

This thesis explores the use of machine learning techniques for surgical phase recognition, which can be used to optimize surgical workflow and facilitate surgical training.

In **Part I**, the work begins with an Introduction to surgical data science and the operating room of the future. It then delves into activity recognition in the operating room, exploring the granularity of surgical activity and the different applications of surgical activity recognition, including motion, steps, actions, phase, and procedural levels.

In **Part II**, the fundamentals of algorithmic classes and classical methods are covered, followed by a comprehensive review of machine learning techniques such as Random Forests, Deep Feedforward Networks, Convolutional Neural Networks, Hidden Markov Models, Recurrent Neural Networks (LSTM and GRU), Temporal Convolutions, and Attention. The work then focuses on surgical phase recognition approaches including approaches beyond RNNs. Different surgical phase datasets are described and metrics for evaluating surgical phase recognition are described in detail.

In **Part III**, the work proposes two different methodologies for temporal learning of surgical phase recognition. The first model, TeCNO, is using temporal convolutions and the proposed OperA model is using attention and transformers. The TeCNO method uses multi-stage temporal convolutions to learn discriminative temporal features from the input data. In contrast, the Opera method uses attention-regularized transformers to focus on the most informative parts of the surgical video. Both methods are compared with baseline approaches, and their effectiveness is evaluated using accuracy, recall, precision, F1 score, and phase recognition consistency. Finally, we identify several challenges and future research directions in the field of surgical data science, including the need to address information ambiguities, hierarchical action recognition, and the development of new datasets for surgical phase recognition.

Part II

Related Work

Fundamentals

6

Contents

6.1 Algorithmic Classes

In this section, we want to introduce different types of algorithm classes. We mainly differentiate between online and offline algorithms, as seen in figure 6.1.

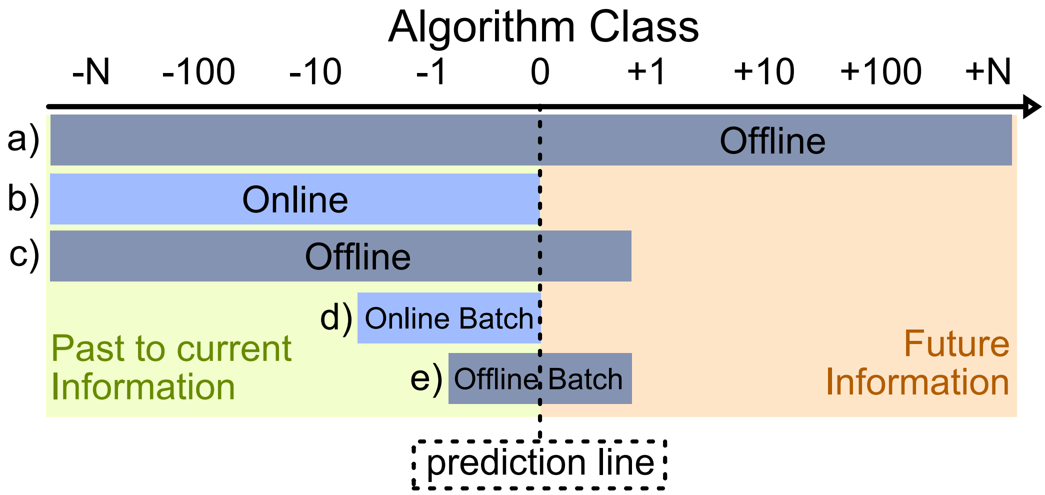


Fig. 6.1. Offline (a), Online(b), Offline-offset (c) and Batch Algorithm types (d,e) are visualized regarding their use of past and future information. The prediction line defines the time step for the prediction.

6.1.1 Online Algorithms

Online algorithms (figure 6.1, a) are characterized by automatic recognition parallel to the activity or surgical intervention [70]. In this setting, the detection cannot rely on any information or input from future time steps. This also means that algorithms might return a sub-optimal choice of prediction because they did not have all the information at the time. Additionally, algorithms have to be efficient enough to perform their inference in a reasonable time, as this can introduce additional delays, which can be problematic for specific online applications. Algorithms that can perform the inference in a reasonable time which is dependent on the sampling rate on the input, are also called real-time methods. Inference time can alternatively be reduced by upgrading the inference hardware, and algorithms that are now considered too computationally expensive (not real-time) for online applications in the future could be executed quickly if computational hardware is keeping up with Moore's law [140].

6.1.2 Offline Algorithms

In contrast to online algorithms, offline algorithms (figure 6.1, b) of activity recognition use the entirety of the past and future information accessible for prediction [70]. For the prediction of the first frame of a video or intervention, the algorithm can therefore look at all of the other future frames. This additional information also results in overall better activity recognition performance but can only be utilized for applications after the actual intervention.

Additionally, a small temporal offset k of a couple of seconds e.g. $k = 10s$ can be introduced to shift the online methods towards offline (figure 6.1, c). This offset allows the algorithm to look into the near future of the target time and the additional information of the entire past for smoother perceptions, especially around transitions. This setup would still be considered offline as it is using future information but could still be used in parallel to the intervention in a similar way as an online algorithm with a temporal offset k, by increasing k also the available information for the prediction of a frame increases and can further improve results. Once k is equal to the length of the sequence, the method is equivalent to the offline case (figure 6.1, a). For applications in the OR that require correct predictions and a slight delay is acceptable, k should be set as high as possible, optimizing performance with acceptable prediction delay. For easier comparison of the different methods in the literature, no offset is used with $k = 0$. The temporal shift of the prediction is a concept that is algorithm-agnostic, improving the results in most cases, and should be defined specifically for an application.

6.1.3 Batched Algorithms

In addition to the online and offline algorithms, the batched algorithms operate on parts of the sequence to reduce memory constraints or to achieve further efficiency optimizations. For the batched algorithms, there still exists the option for online (figure 6.1, d) and offline (figure 6.1, e) execution.

6.2 Classical Methods

Classical computer vision methods were developed before the advent of deep learning and neural networks. These methods are based on mathematical and statistical techniques and have been widely used in various applications such as object recognition, image processing, and activity recognition. In the following, some examples of relevant classical methods are summarized.

Template matching [25] is a classical method that has been used in computer vision for a long time. The basic idea of template matching is to slide the template over the image and compute a similarity measure at each position. The position with the highest similarity is considered the best match. This method can be used for object recognition, image registration, and tracking. The most widely used method for template matching is the Correlation Coefficient Method.

Optical flow [24] is a method in computer vision used to estimate the motion of objects in a video. Optical flow is based on analyzing the brightness patterns in consecutive video frames. If two adjacent frames are identical, the optical flow would be zero, as there is no flow in the scene. In optical flow, dynamic objects are highlighted, and static objects remain in the background. Optical flow has been widely used in video analysis, object tracking, and activity recognition.

One of the most famous classical methods is Scale-Invariant Feature Transform (SIFT) [94] introduced in 2004. SIFT is a feature descriptor invariant to image scale and rotation, used in object recognition and image alignment. The SIFT descriptor is based on the detection and description of local features in an image with distinct properties that are robust to changes in viewpoint, lighting, and noise. The key point detection process involves detecting local extrema in the scale-space of the image, which is then refined to precise locations using the Difference of Gaussian (DoG) function. Once key points are detected, a descriptor is constructed by sampling the gradient orientation of pixels in a small region around the key point and then building a histogram of these orientations. The descriptors from two images can be matched to find correspondences between both images to register, recognize or track objects or humans.

Random Sample Consensus (RANSAC) [45] is an iterative method for estimating the parameters of a mathematical model from a set of observed data that contains outliers. It can be used to fit models to image feature correspondences, which is beneficial when dealing with noisy or incomplete data. RANSAC can be combined with SIFT for image registration and object recognition tasks. The method starts by randomly selecting a subset of feature correspondences and then using these correspondences to estimate the model's parameters. These parameters are then used to classify all correspondences as inliers or outliers. The process is repeated many times, and the best set of inliers is selected as the final model.

Another classical method is Histograms of Oriented Gradients (HOG) [39]. HOG is a feature descriptor based on the distribution of gradient orientations in an image, and it has been used for human detection. HOG is a computationally efficient method, and it is robust to changes in lighting, viewpoint, and image resolution.

Speeded Up Robust Features (SURF) was introduced by Herbert Bay, Tinne Tuytelaars, and Luc Van Gool introduced in 2006 [14]. SURF is an extension of SIFT based on the detection and description of scale- and rotation-invariant interest points in an image used for object recognition and image registration.

In recent years, machine learning, deep learning, and neural networks have become increasingly popular in computer vision. Machine learning methods can learn from data and improve their performance over time, unlike classical methods, which typically rely on hand-crafted rules and assumptions about the data. Machine learning methods can automatically learn complex, non-linear relationships in the data, which classical methods often struggle to do. Additionally, machine learning methods can handle large and high-dimensional datasets. One such example is Support Vector Machines (SVMs) [35], a simple yet powerful algorithm that can be used for classification and regression tasks. SVMs create a decision boundary separating the different data classes. The goal of an SVM is to find the hyperplane that maximally separates the different classes while being as close as possible to as many of the data points as possible. Once the hyperplane is determined, new data points can be classified by observing which side of the decision boundary they fall.

Also, the combination of classical methods and machine learning is possible. SIFT features can encode important information about an image, such as texture and shape, and can be used as input to an SVM classifier which can be trained to classify the image.

6.3 Tree Based Approaches

In this section, random forests are introduced. Decision trees are the building blocks of random forests, an ensemble of decision trees.

6.3.1 Decision Tree

A decision tree is a tree-like graphical model that can be applied to classification and regression problems introduced by Breiman et al. in 1984 [22]. The tree is created by recursively partitioning the input data into subsets based on the values of the input features. This portioning is visualized in figure 6.2 where first the weight, then taste, and finally the shape of the input fruit is validated in the internal nodes. This validation can be performed for categorical, e.g., sour in taste, and numerical variables, e.g., weight. Following the path for the input, the leaf node represents a distinct prediction.

One of the main benefits of decision trees is that they are easy to interpret and comprehend. The tree structure allows the user to easily trace the path from the root of the tree to a leaf node and understand the decision-making process. This makes decision trees useful for feature selection and understanding the underlying relationships between the input features and the output.

One disadvantage of decision trees is that they can be sensitive to small changes in the data, which can result in significant changes in the tree's structure. This can lead to overfitting, where the tree becomes too complex and performs well on the training data but poorly on new unseen data. To avoid this, techniques such as pruning, bagging, and boosting [46] can be used, which improves the generalization performance.

6.3.2 Random Forest

Random forests were introduced in 2001 by Breiman et al. [21] and integrated the concept of decision trees with ensemble learning. A random forest combines multiple decision trees, where each decision tree is built using a different subset of the training data. The idea behind random forest is to create multiple decision trees and then combine their predictions, improving the overall performance of the model.

The decision trees in a random forest are constructed using a technique called bagging (bootstrap aggregating), which generates multiple training sets by randomly sampling the original data with replacement. This means that each decision tree is trained on a different subset of the data, and therefore, each tree will have a different set of splits, resulting in different predictions.

When a new input is presented to a random forest, each decision tree in the forest makes a prediction, and the final output is determined by a majority vote or by averaging the predictions as visualized in figure 6.2. Even if tree 2 wrongly classified the mango as an apple, most of the trees still predicted mango correctly. This ensemble of predictions is less prone to overfitting and performs better than a single decision tree.

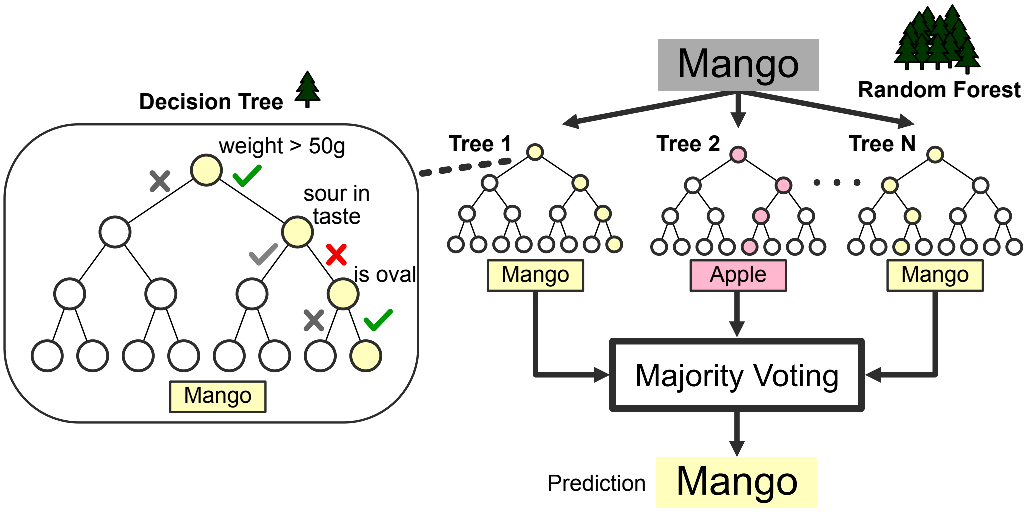


Fig. 6.2. The classification of a Mango using a Decision Tree and Random Forest is visualized. Even though Tree 2 incorrectly classified the Mango as an Apple the majority voting of many trees in the RF leads to the correct prediction.

The training of random forest includes multiple steps. The dataset is first divided into different parts (bootstrapping) to create multiple subsets of the data. Next, a decision tree is built for each subset, and the splitting conditions for each node are iteratively optimized to maximize the information gain [21]. Once all the decision trees are trained they are combined into one random forest. Additionally, pruning any unnecessary splits and feature selection can help to stabilize the training, reduce complexity, and help to generalize.

A popular alternative to random forests is XGBoost (Extreme Gradient Boosting) which performs exceptionally well on tabular data [29] and has been used widely for medical applications. While random forests can have advantages in handling missing, categorial, or imbalanced data, XGBoost compensates for these shortcomings with increased efficiency, accuracy, or less overfitting.

6.4 Deep Feedforward Networks

Deep feedforward networks, also known as feedforward neural networks or multi-layer perceptrons (MLPs), are artificial neural networks consisting of an input layer, one or more hidden layers, and an output layer. The input layer receives and passes the data through the hidden layers, where complex data representations are learned. The final output layer produces the network's prediction. This flow of information from the input layer through the intermediate computations until the output layer is why these methods are referred to as feedforward [50].

The hidden layers use activation functions, such as the rectified linear unit (ReLU) [105], to introduce non-linearity into the network and to allow the network to model complex relationships between the input data and the output predictions. The network weights and biases can be learned through a process called backpropagation [137], where the error between the network's predictions and the ground truth is used to update the weights and biases to minimize the error.

Building artificial neural networks with multiple layers is often called "deep" neural networks. Deep feedforward networks are used in various applications, such as image classification, speech recognition, and natural language processing. The deep structure allows the network to model complex relationships and representations in the data, achieving better performance than shallow machine learning algorithms.

6.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep learning neural network that is particularly effective for the image and video processing tasks. Even though CNNs were first invented and also patented by AT&T Bell Labs [1], Yann LeCun is still considered the most influential researcher to promote CNNs with his research in 1989 [89]. They are designed to process data with a grid-like topology like an image.

CNNs consist of a combination of convolutional layers, pooling layers, and classification layer, which was applied to the MNIST dataset to classify images of handwritten digits [90, 91]. Through additional research and components like the residual connections introduced by He et al. [57] the concept of CNNs is still one of the main building blocks of the latest image and video analysis tasks.

In figure 6.3, a basic example of the convolutional operation is shown. A 2-dimensional kernel is defined with kernel size $k = 2$. Each kernel element is multiplied with each corresponding element of the image and summed up. The resulting linear activation value is stored in an activation map, and a non-linear activation function, such as the ReLU, is used to introduce non-linearity. This operation covers the entire input space.

The pooling layer is then used to make the representation less sensitive to small input translations. For many tasks, it is not essential to know the pixel-perfect location e.g. the tip of a scalpel. It is sufficient to understand that the surgeon is holding it. Max pooling [181] is a simple choice for a pooling layer where in a rectangular neighborhood in the non-linear activation map, with pooling size ps, only the maximum value is kept. This way, the spatial size of the activation maps is also decreased by a factor of $1/ps$. The reduction of the size of the activation maps further helps with statistical efficiency and as the parameters of a CNN are directly related to the input size, the memory requirements for storing the learnable parameters can be reduced [50].

For a classification task, the pooled activation maps can then be fed into another set of layers for additional processing. This way, the spatial dimensions are further reduced while the feature dimensionality is increased. Finally, the activation maps are flattened, and the resulting

[1] `http://patentimages.storage.googleapis.com/df/7a/25/3d7d8123b5ed16/US5058179.pdf`

1-dimensional vector is used as input for a feedforward network to generate the final class predictions.

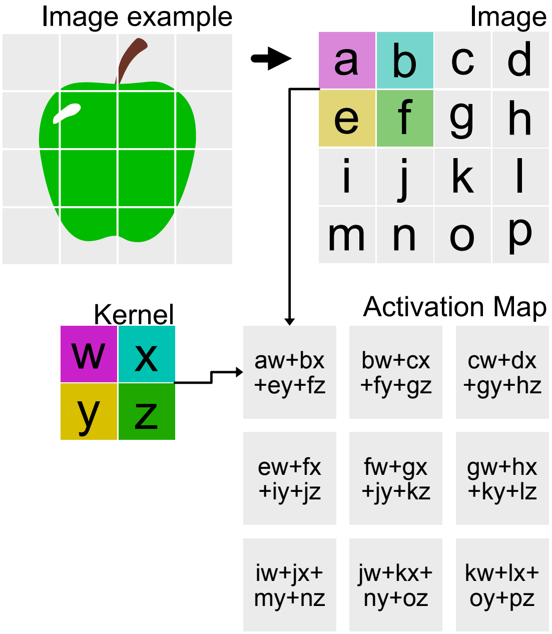


Fig. 6.3. Overview of the convolutional operation with a 2×2 Kernel on an image and the resulting computations for the parts of the Activation Map. Inspired byGoodfellow et al. [51]

In CNN training, the network learns to extract features from the input data that are applicable to the task and dataset it is trained on. The training process involves iteratively updating the weights and biases of the network's layers, which can be interpreted as the convolutional filters.

In the early layers of the CNN, the network can detect basic patterns, such as edges, corners, and blobs. Over time, the filters of the network evolve, as illustrated in figure 6.4. Initially, at Epoch 1, no discernible patterns are visible in the filters. However, as the network progresses in its training (e.g., at Epoch 50), the development of filters that capture the shape of blobs, edges, and corners becomes increasingly clear.

As the CNN progresses through the deeper layers, it develops an understanding of more intricate patterns and combinations of features, capturing more high-level features that are required for solving the problem at hand.

6.6 3D CNN

3D CNNs are an extension of 2D CNNs designed to process 3D data, such as 3D images from CT or MRI medical scans or 3D point clouds. 3D CNNs use 3D kernels to extract features from the input data. In figure 6.5 the comparison of a 2D input with a 2D kernel (top left) and 3D input with a 3D kernel is shown (bottom right). The kernel of the 3D CNN not only has a third dimension but also is slid in the third dimension additionally, H and W.

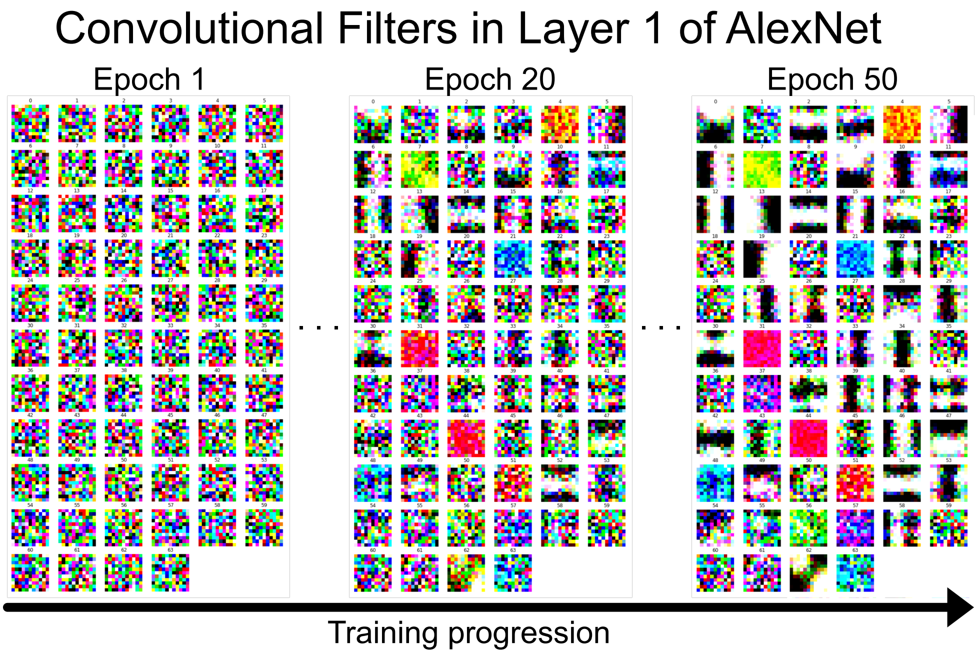


Fig. 6.4. The learning progression of the 64 filters in the AlexNet model on ImageNet [81]. From epoch 1 to epoch 50, the filters gradually attain more significance, revealing visible edge and blob patterns.

3D convolutions can also be used to capture the temporal information in a video or 2-dimensional time-series data by treating the third dimension as time (T). Instead of using 2D kernels to process multiple time points (2D + T) individually (top right). The 3D kernels observe the time dimension to capture the dependencies between frames, providing a way to model the dynamic changes over time (bottom left). This enables the network to learn spatio-temporal representations from the data, improving performance on tasks such as action recognition and video classification.

6.7 Dynamic Time-Warping

Dynamic Time-warping (DTW) is a method for time series analysis to measure and match two similar series with varying lengths with each other, initially proposed for matching spoken words [139]. A distance metric between both sequences is used to calculate an optimal path between them, by minimizing the cost of alignment using dynamic programming. Sequences, even if having varying lengths, can then be optimally matched using the DTW matrix. The algorithm can essentially stretch or skew parts of one sequence by matching one point of a sequence to many points of the target sequence (stretching) or matching many points of the target sequence to one point of the sequence (skewing).

Given two sequences $A = \{a_1, a_2, ..., a_i\}$ and $B = \{b_1, b_2, ..., b_j\}$ the matrix $DTW \in M_{i\times j}(\mathbb{R})$ is created. The DTW matrix is filled by first calculating the distance between the feature values of the corresponding points in the sequences $d(a_i, b_j)$. This distance measure can be freely selected and can be as simple as a euclidean distance as visualized in 6.6. The distance $d(a_i, b_j)$ is then added to the minimal value of three neighboring values in the DTW

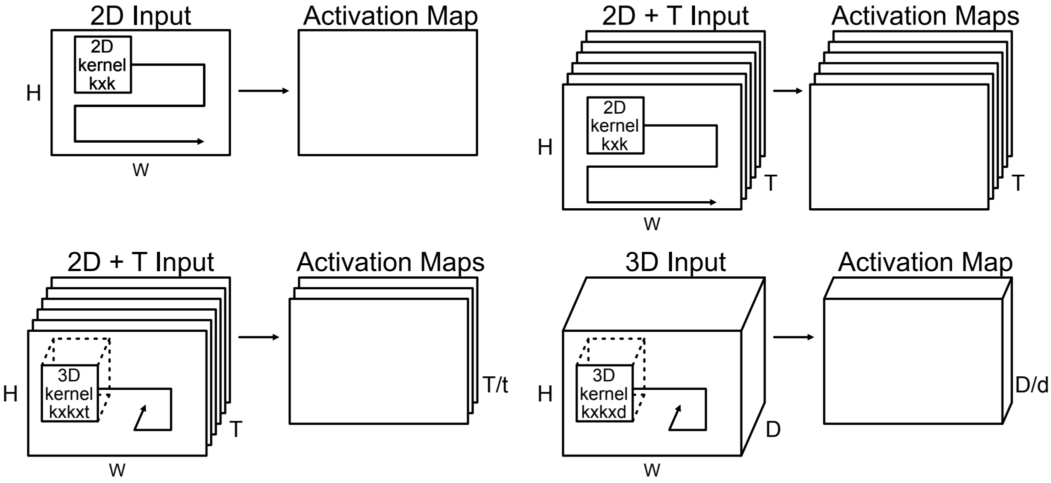
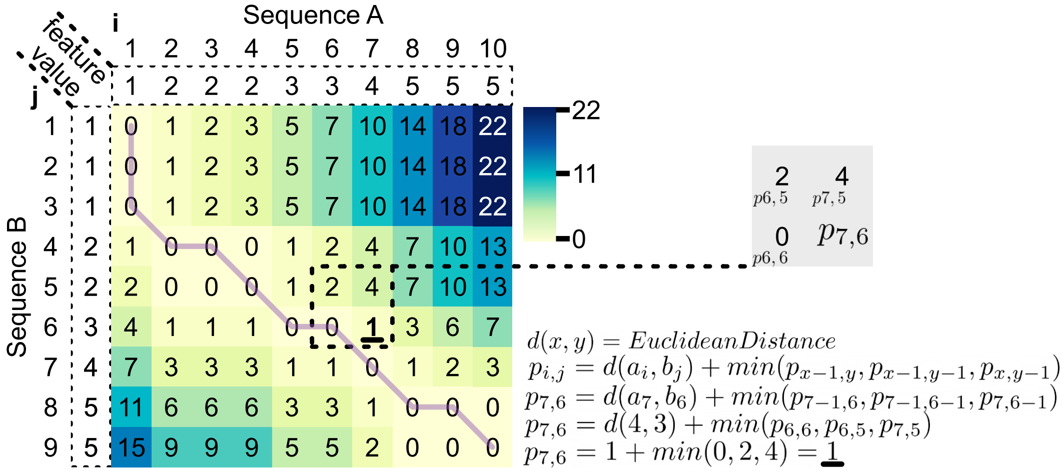


Fig. 6.5. Convolutional neural networks (CNNs) can be employed for 2D, 3D, and 2D+T inputs, which yield corresponding activation maps. 2D kernels only operate on the height (H) and width (W) of the input, whereas 3D kernels additionally move in the time (T) or depth (D) direction. The implementation of 3D convolutions for both the 2D+T and 3D cases is identical, but the differentiation is based on the type of input, with the temporal dimension (T) or the geometric dimensions..

matrix: $min(DTW_{i-1,j}, DTW_{i-1,j-1}, DTW_{i,j-1})$. This process is repeatedly continued to fill the entire DTW matrix, and negative indices in the calculations can be assumed to be zero.

Finally, the optimal path to match both sequences can be determined starting from $p_{i,j}$ and following the path of the smallest total weight through the DTW matrix as visualized with the purple line in figure 6.6. Every point in both sequences is matched with at least one of the other sequences and the synchronization of the sequences based on the distance function d has been completed.



j		1	2	3	4	5	6	7	8	9	10
		1	2	2	2	3	3	4	5	5	5
1	1	0	1	2	3	5	7	10	14	18	22
2	1	0	1	2	3	5	7	10	14	18	22
3	1	0	1	2	3	5	7	10	14	18	22
4	2	1	0	0	0	1	2	4	7	10	13
5	2	2	0	0	0	1	2	4	7	10	13
6	3	4	1	1	1	0	0	1	3	6	7
7	4	7	3	3	3	1	1	0	1	2	3
8	5	11	6	6	6	3	3	1	0	0	0
9	5	15	9	9	9	5	5	2	0	0	0

$$d(x,y) = EuclideanDistance$$
$$p_{i,j} = d(a_i, b_j) + min(p_{x-1,y}, p_{x-1,y-1}, p_{x,y-1})$$
$$p_{7,6} = d(a_7, b_6) + min(p_{7-1,6}, p_{7-1,6-1}, p_{7,6-1})$$
$$p_{7,6} = d(4,3) + min(p_{6,6}, p_{6,5}, p_{7,5})$$
$$p_{7,6} = 1 + min(0,2,4) = \underline{1}$$

Fig. 6.6. An example of Dynamic Time Warping (DTW), aligning two sequences A and B. The feature value for each sequence is employed to compute the DTW matrix values utilizing a predefined distance function, as exemplified in the illustration on the right. Subsequently, the path with the least total cost (optimal path) is selected to align the sequences.

6.8 Hidden Markov Models

The Hidden Markov Models (HMM) were first introduced by Baum et al. in the late 1960s and early 1970s [13, 130] and were implemented for speech processing. HMMs are probabilistic graphical methods to predict a sequence of hidden or unknown variables from a sequence of observable variables. HMMs are built upon the assumption that the probability of being in a particular state at a given time only depends on the state at the previous time step and not on any earlier states [53]. This assumption is known as the Markov property and it is what allows HMMs to make predictions about a system's future behavior based on its past observations. In other words, the Markov assumption is what makes HMM a powerful tool for modeling systems that change over time, such as speech recognition, weather forecasting, financial modeling, or medical applications.

A Hidden Markov Model is defined with $\lambda = (N, M, A, B, \pi)$

- N: The number of discrete states of the model with $states = (z_1, ..., z_N)$.

- M: The number of different observations that can be made with $observations = (x_1, ..., x_M)$.

- A: The state transition probability matrix of size $N \times N$, where the element $a_{i,j}$ represents the probability of transitioning from state i to state j.

- B: The observation probability matrix of size $N \times K$, where the element b_{ij} represents the probability of observing observation x_j when in state i.

- π: The initial state probability vector, where the element π_i represents the probability of starting in state i.

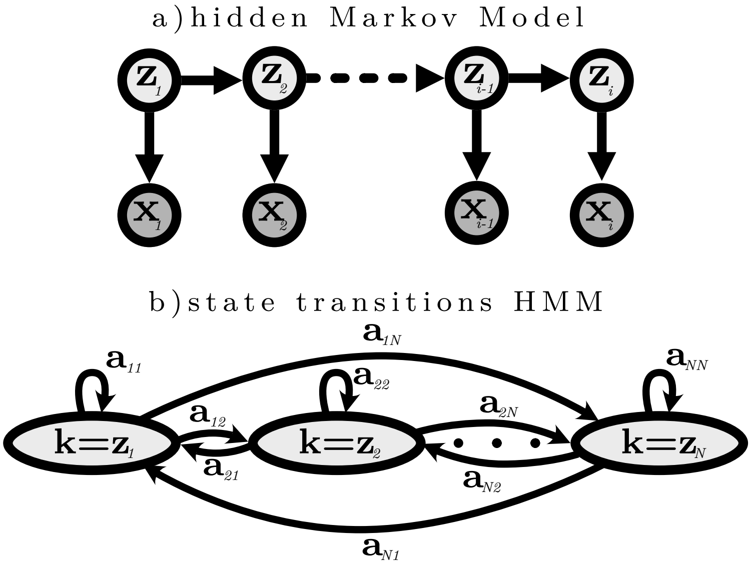


Fig. 6.7. a) A Hidden markov model (HMM) is visualized with the different states z and observations x. In b) the different state transition probabilities of an HMM are visualized. With the directed transition probabilities and states k.

The training of an HMM can be performed with different algorithms. The Baum-Welch algorithm [12] or Expectation-Maximization algorithm is an iterative method for training HMMs. It can be divided into two steps. In the **Expectation** step the current parameters of the model λ are used to calculate to predict the next observation probabilities x_t with t being the current temporal step. This is also called forward probability. Additionally, the backward probabilities are calculated where information about the future $t+1$ and onwards is considered. This forward-backward algorithm is used to calculate the likelihood of the observed data given the current estimates of the HMM parameters. In the **Maximization** step, the algorithm uses the calculated likelihood and expected sufficient statistics to update the estimates of the parameters to maximize the likelihood. This step usually involves updating the transition a and observation probabilities b of the HMM.

The training of HMMs can be done using other techniques, such as the Viterbi-training [130] that has been proposed to overcome some of the limitations of the original Baum-Welch algorithm. The training process of HMMs typically requires a large amount of data, and the quality of the estimated parameters will depend on the quality of the data.

6.9 Recurrent Neural Networks

Recurrent neural networks (RNNs) are adept at handling sequential data, such as time series and natural language. They are referred to as "recurrent" as they maintain an internal state that changes with each time step, allowing them to hold on to previous input information. RNNs were first introduced in the 80s [60], but it wasn't until the use of more advanced training algorithms and powerful GPUs that they gained popularity. Today, RNNs are utilized in a variety of applications, including image captioning, speech recognition, language translation, and activity recognition. An essential aspect of RNNs is the utilization of memory cells, which can retain information across multiple time steps. This capability allows RNNs to process sequences of varying lengths, which is particularly useful for natural language processing tasks where sentence length can vary greatly.

Given a single input $x \in \mathbb{R}^D$ with dimension D and a sequence $x = (x_1, x_2, ..., x_T)$ with timestep $t \in (1, 2, ..., T)$ the RNN generates its hidden states

$$h_t = \begin{cases} 0, & t = 0 \\ \phi(h_{t-1}, x_t), & otherwise \end{cases} \tag{6.1}$$

wit ϕ a nonlinear function. At time zero, no hidden state is available yet and is usually initialized as zero and continuously updated over the time steps. The update of the traditional tanh unit (tanh-RNN) [32] is then defined as

$$h_t = g(Wx_t + Uh_{t-1}) \tag{6.2}$$

with the sigmoid function g. The weights of the input and memory of the recurrent neuron are represented by W and U, respectively, which are learnable parameters that are adjusted during the training process.

It has been shown [15] that this formulation of the RNN is hard to train with large temporality T. Gradient-based optimization for traditional RNNs struggles with variations in gradient magnitudes and long-term dependencies being hidden by short-term dependencies. The gradients used to update the learnable parameters are prone to vanish or explode for the training of long sequences. Researchers have attempted to mitigate this by devising better learning algorithms or using methods such as clipped gradients or second-order methods.

An alternative solution to these challenges is the development of more advanced RNNs methods such as Long Short-Term Memory Networks (LSTM)[60] and Gated-Recurrent Units (GRU) [31], which have gained significant popularity in recent years. The sequential processing of RNNs is exemplified with an LSTM in figure 6.8. The output, and if applicable, memory from the previous timestep, is fed into the next step in a sequential way. The same concept is applicable when replacing the LSTM with GRU or tanh cells.

The sequential modeling of information can be extended so that information is not only modeled from past to future but additionally from future to past. This allows the network to capture both past and future context, giving it a more comprehensive understanding of the input sequence [143].

6.9.1 LSTM

Long short-term memory (LSTM) networks are designed to address the vanishing and exploding gradient problems of traditional RNNs (section 6.9). LSTMs were proposed in 1997 by

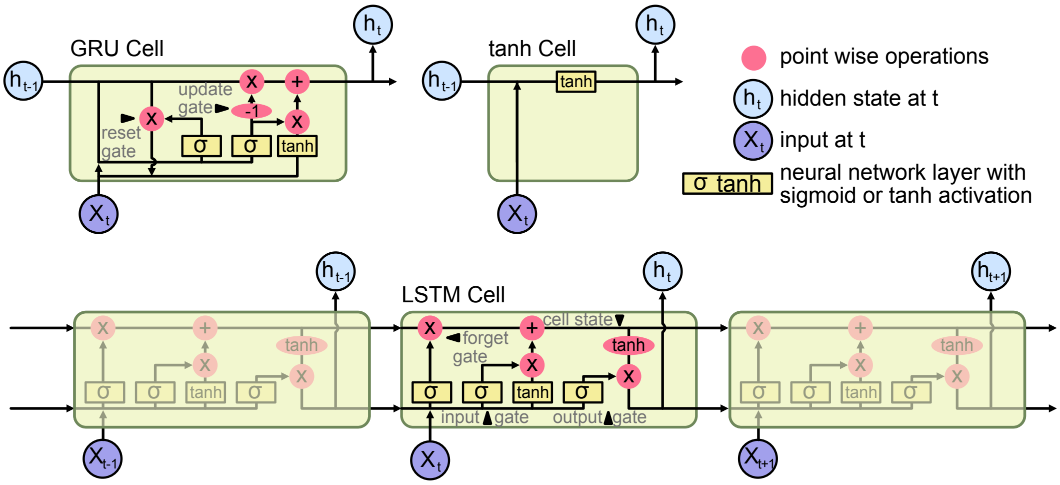


Fig. 6.8. Different RNN cells are visualized starting with the GRU cell including update and reset gate, tanh cell implemented using a simple tanh layer and the LSTM cell using input gate, forget gate, cell state and output gate. The information flow from x_{t-1} to x_{t+1} is exemplified for the LSTM case but functions similarly for the other cell types

Sepp Hochreiter and Jürgen Schmidhuber [60]. In figure 6.8 a LSTM cell is visualized and divided into different parts which are further described in more detail.

The forget gate is an essential part of an LSTM, designed to determine which information is no longer relevant or important for the current or future time steps and can therefore be ignored. If information from a single time step can be considered as noise the forget gate has the ability to largely disregard this noisy input, helping to maintain a valid memory. The forget gate is implemented as a sigmoid layer generating outputs from 0 to 1 and a value of 0 means that the information in the memory cell should be completely forgotten, while a value of 1 means that it should largely be remembered. The forget gate is one of the parts specifically designed to circumvent the vanishing and exploding gradient problem and helps to improve the ability to maintain long-term temporal dependencies of the input sequences.

Another important part of the LSTM network is the input gate. The input gate controls the flow of information into the memory cell, implemented as a sigmoid layer. Similar to the forget gate, a value of 0 means that no new information should be added, while a value of 1 means that new information should be added to the memory cell with high certainty. The input gate works in conjunction with the cell state and the activation function to control the flow of information into the memory cell. The input gate is updated based on the inputs, the previous hidden state, and the information stored in the memory cells. When the input gate is activated, new information is stored in the cell state, and this information is used to generate the output at the current time step.

The information accumulating in the cell state expresses the memory of the network, where information is stored and taken from one-time step to the next. The cell state is updated at each time step using the inputs, the previous hidden state, and the information in the cell state, intended to store information about long-term temporal dependencies in the input sequences. This is allowing the network to retain information over long periods of time and make predictions based on that information. The cell state is controlled by the input, forget, and output gates, controlling which information should be stored or forgotten.

Finally, the output gate is managing the flow of information out of the memory cell implemented as sigmoid layer. The output gate works with the information in the cell state and the activation function to determine the final output of the LSTM network. When the output gate is activated, the information stored in the cell state is passed through a tanh activation function and used as the output for the current time step.

6.9.2 GRU

The difference between GRUs and LSTMs is that LSTMs have three gates (input, forget, and output) whereas GRUs have only two gates (reset and update). The reset gate determines how much of the previous hidden state is forgotten and is comparable to the forget gate in LSTMs. The update gate defines how much of the new input is incorporated into the current hidden state. In GRUs the cell state and hidden state are not separated and therefore may struggle to store information for longer time periods. However, their simplicity with fewer learnable parameters and better regularization can overcome these limitations.

6.10 Temporal Convolutions

Temporal convolutions are a type of convolutional neural network (CNN) that are used to process sequential data [9] and are designed to take advantage of the temporal relationships between data points in a sequence by using convolutional layers that operate on a fixed-length window of the input data. They have been applied in various domains including audio [164], and video[88] and have shown to be effective in capturing long-term dependencies in sequential data, which is challenging for traditional recurrent neural networks (RNNs 6.9).

In a traditional CNN (section 6.5), the convolutional layers are applied to images, and each convolutional kernel operates on a small region of the image, the receptive field while scanning the input image. In a temporal convolution, the convolutional layers are applied to a sequence of data points, and the convolutional kernel scans the input sequence in a sliding window fashion. In figure 6.9 an example TCN network is shown consisting of three convolutional layers. The kernel size of the TCN is set to 3 which means that each TCN combines three features into one. The increase in dilation for each consecutive layer is essential for building up a longer temporal context. In figure 6.9 a dilated residual layer also implements a residual connection and 1x1 convolution in addition to the temporal convolutional operation Z. These additional design choices can help to improve temporal reasoning further. TCNs are able to extract features from the data that are invariant to small temporal shifts. This invariance to temporal shifts can be compared to the invariance to spacial shifts in traditional CNNs. The idea is to find if something happened rather than to precisely pinpoint the event to a particular step in the sequence.

In 2016, Lea et al. introduced temporal convolutions [88] as a means of hierarchically processing video data for action segmentation. Building on this work, in subsequent years,

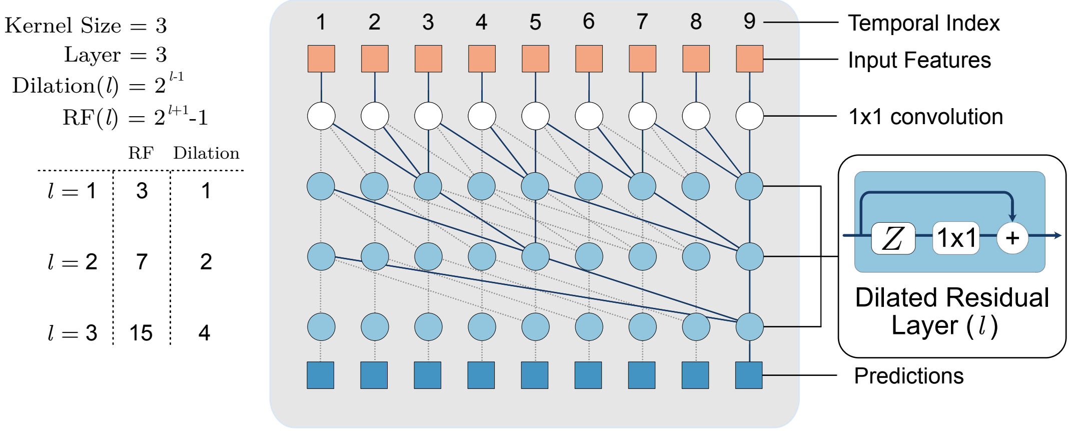


Fig. 6.9. Temporal Convolutional Network with three layers (l), kernel size 3 and temporal convolutian operation Z. The dilation of the temporal convolutions and resulting receptive field (RF) depends on the layer. In this example, the Dilated Residual Layer includes the 1x1 temporal convolution following the design in [44], but different arrangements and setups can be created. For each one-dimensional input one prediction is generated.

researchers adapted the use of temporal convolutions to other applications such as audio generation, resulting in improved performance due to a larger receptive field that allowed for higher temporal resolution. Specifically, these adaptations used dilated convolutions [164], which expanded the receptive field of the network. To further improve action segmentation, researchers developed Multi-Stage TCNs (MS-TCNs) [43, 93]. MS-TCNs consist of stacked predictor stages, each of which includes an individual multi-layer TCN. These stages incrementally refine the initial prediction of the previous stages, resulting in more accurate segmentation of actions in videos.

6.11 Attention

The attention mechanism in deep learning is a way to selectively focus on different parts of the input, typically in a sequence of inputs, for the purpose of generating a more expressive representation of the sequential data [165].

In figure 6.10, an overview of the attention mechanism is shown. The essential components of an attention mechanism include the query (Q), key (K), and value (V) vectors, where the query represents the current state of the model and the key and value represent the parts of the input that the model is attending to. The attention scores are computed by taking the dot product between the query and key (MatMul).

The purpose of scaling in the attention mechanism is to prevent attention scores from becoming too large, which would lead unoptimized training process. Scaling is usually performed by multiplying the attention scores with the square root of the dimension of the key vector. Masking in attention mechanisms prevents the model from attending to future steps, which would violate the causal constraint of sequences. This is typically done by zeroing elements in the attention weights corresponding to future tokens (Mask). Lastly, the attention weights are multiplied by the value vector. This weighted sum of the value vectors is used as the input to the next layer in the neural network, effectively allowing the model to attend to different parts of the input in a weighted manner.

The transformer architecture proposed by Vaswani et al. [165] further includes residual connections and layer normalization (Add&norm) for more efficient learning. In the original transformer architecture, the output of the transformer encoder is fed into a transformer decoder with a similar structure. The encoder-decoder structure is often used for sequence-to-sequence prediction where the input and output sequence can be different in length. When designing a network to predict for each timestep exactly one output (many-to-one), the decoder block can be skipped in favor of a simple classification head as depicted in figure 6.10.

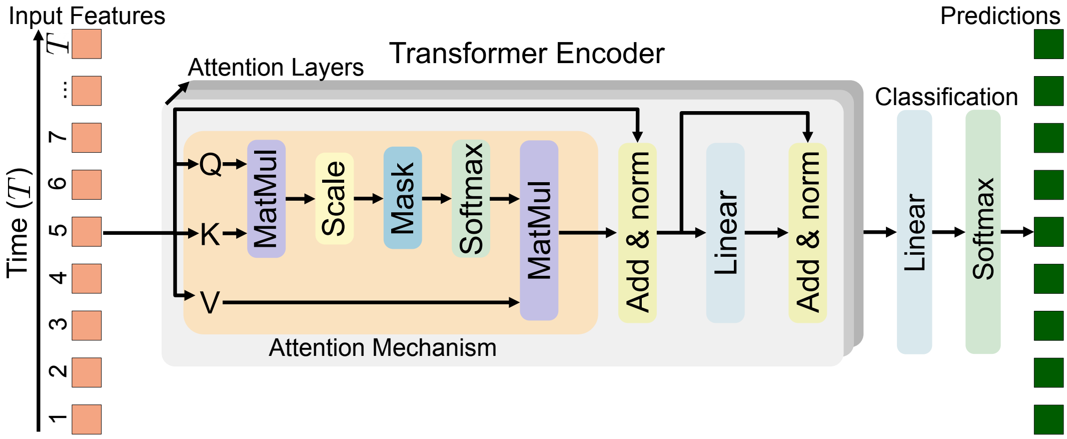


Fig. 6.10. On the left side of the figure the Input Features are visualized from 1 until T. A linear layer is converting the feature into Q,K and V the eseential part of the attention mechanism. Additional layers like Scaleing, Masking or Norm are helpful for further stability of the training and online masking can be used to create an online attention model.

Transformer models were initially introduced in the field of natural language processing (NLP) [165] and quickly became the state-of-the-art in numerous related tasks [23, 40]. Furthermore, the versatility of transformers has been showcased not only for vision tasks such as image classification [41] and text-to-image generation [132] but also in biology for the challenging protein folding problem with *AlphaFold* [59]. In surgical data sciences, transformers have also been explored only for surgical tools [78] classification.

Transformers and attention also provide a unique opportunity for model insights and explanations through the use of attention weights. While some researchers have expressed skepticism regarding the explanatory capabilities of attention mechanisms [62], this assumption has been challenged by others in the field [172]. The effectiveness of attention as an explanatory tool may be heavily dependent on the specific task being undertaken, highlighting the need for a task-specific approach to this issue.

Surgical Phase Recognition Approaches

7

Contents

7.1 Introduction

In chapter 6 we provided detailed explanations of classical, deep-learning, and temporal methods. Building on this foundation, we will now present works that have approached surgical phase recognition using a variety of the presented works. These studies demonstrate the diverse range of approaches that can be used for surgical phase recognition in the OR and provide valuable insights into the strengths and limitations of different methods.

The automatic understanding of the different workflow steps of a whole procedure was first proposed by Ahmadi et al. [2] in 2006. The authors used different tool usage signals from different surgical devices and combined them into one multidimensional state vector for each time step of surgery. To identify the phase transitions of an unseen surgery, the state vector of this unseen surgery was compared with a standard surgery with known surgical phase labels using a DTW model (section 6.7). DTW was further enhanced in other works by prioritizing signals with higher information gain of state vector [77, 120]. As the DTW analysis focuses on the analysis of the activities in an offline manner (section 6.1), HMMs (section 6.8) were used for online detection of the surgical activities [121] and temporal reasoning. Adding the visual information using low-level image features with the instrument state vectors showed improved performance [19]. Lalys et al. [84] combined the concept of visual information with statistics for microscopic videos to detect the surgical phases combining SVM (section 6.2) and HMMs. This concept was further extended through the manual definition of visual cues adapted to different surgery types [82]. DTW and HMMs were further used [125] for workflow recognition in laparoscopic cholecystectomy procedures from binary tool information. Stauder et al. [151] proposed to use Random Forests on tool signals (section 6.3). They predicted the surgical phase of laparoscopic cholecystectomies by feeding binary signals of the activities in the OR, such as OR light, table light, or suction device status, into a random forest model. The prediction of the phase is done per time step and does not consider any temporal information, which helps in predicting activities outside the typical workflow order but also leads to more noisy predictions. In their follow-up work [153], they added a Hidden Markov Model (section 6.8) to the output of the Random Forest and showed improved temporal reasoning of the combined setup.

Next to laparoscopic surgeries, the analysis of robotic surgeries was performed [170] using the kinematic information of the robot surgical gestures were identified, correlating with surgical steps or activities [178].

Further analysis and modeling of the workflow were conducted and a formal model of a procedure was built based on an in-depth analysis of surgery in terms of succession of possible steps and corresponding features [64]. Neumuth et al. [107] proposed ontologies to generate a structured representation of surgical workflows. This concept was further evolved, and validated [108] into Surgical Process Modeling [63] proposed for different procedures and types of interventions.

The recognition of phases in cataract surgeries from microscopic videos is another application area for surgical phase recognition [129]. Already published methods for recognizing laparoscopic phases could successfully be transferred to cataract surgeries. Some works also investigate the challenging environment of Intensive Care Units [87] combining action recognition and pixel-level segmentation of objects and persons. To recognize surgical actions such as a *patient-in* or *cleaning* action, ceiling-mounted surgical room cameras were used to capture the scene used as input for 3D Motion Features [122], which could be linked to the different actions in the room.

Most of the works presented so far relied on digital signals or low-level visual features extracted using classical methods. With the advent of the Convolutional neural network (section 6.5) and the introduction of more extensive and publicly available datasets [159], surgical phase recognition from raw videos of the interventions became feasible. In 2012, AlexNet [81] won the ImageNet Large Scale Visual Recognition Challenge. AlexNet is one of the first convolutional deep neural networks and inspired many follow-up studies in different domains, including surgical activity recognition.

Even though some works already used visual features created with image convolutions for surgical phase recognition and tool detection with random forests [138] on a small scale, one requirement for effectively using deep neural networks for image classification is large datasets with labeled annotations. In 2015 the first EndoVis Challenges (section 7.3) was introduced, and in 2016 the M2CAI dataset for the Surgical Workflow Recognition in the SensOR was released as part of the challenge. In the same year, the Cholec80 (section 7.3.2) dataset was introduced, which is, up to today, one of the largest and most used datasets, including surgical phase and tool labels. These datasets allowed the training of deeper networks with more learnable parameters and fueled the development of visual surgical phase recognition.

EndoNet [159, 160] was one of the first CNN methods for surgical phase recognition. EndoNet jointly conducted surgical tool and phase recognition with a shared feature extractor using multiple classification heads. Further, EndoNet was used to extract visual features of the frames, which were fed into HMM and LSTM (section 6.9.1) models. The combination of CNN and LSTM [161] showed strong performance, leading to more research in this direction.

In Endo2N2 [174], self-supervised pre-training for the visual feature extractor CNN was proposed by predicting the Remaining Surgery Duration (RSD). The pre-trained CNN was then fine-tuned using phase labels and a combination of CNN and LSTM in an end-to-end fashion. Jin et al. [65] combined the LSTM method with a ResNet [57] model and showed that the residual connections and the overall improvements of the ResNet architecture compared to its precessors could effectively boost the performance of phase recognition on the cholec80 dataset. The authors also built on previous works of surgical ontologies and used the prior knowledge of the surgical procedures to mask transitions between phases that never appear in the dataset showing additional performance gains. MTRCNet-CL [66] was proposed to address the correlation between the phase and tool information explicitly using a correlation loss, further improving the performance on Cholec80. Many other works built up on this success and iteratively improved the temporal reasoning and prediction of surgical phases with RNNs [49] for surgical phase recognition.

7.2 Advancing Beyond Recurrent Neural Networks

RNNs have become a popular choice for surgical phase analysis. However, training RNNs can be challenging due to the backpropagation through time (BPTT) algorithm, which involves gradient updates not only across layers but also across the temporal dimension of the sequence. This complexity can lead to the problem of vanishing or exploding gradients, which has been addressed through techniques such as Truncated BPTT, gradient clipping, and different activation functions. Moreover, architectural improvements like the introduction of forget gates in LSTM models have been proposed to alleviate this issue. While these techniques have their benefits, they also have their limitations, such as reduced temporal reasoning capabilities and potentially suboptimal performance due to not capturing long-term dependencies [52, 126, 135]. Therefore, alternative methods exploring novel techniques for encoding long temporal sequences of surgical procedures are explored.

One such direction is the use of multi-stage training [175], which has shown promise in enhancing the performance and efficiency of existing models. For instance, recent work has proposed a method that combines 3D convolutions (section 6.6) for short clip or batch-level temporality with LSTM for long temporal analysis of robotic interventions [145]. Lea et al. [86] have proposed the use of temporal convolutions (section 6.10) for surgical phase recognition, while avoiding the use of RNNs. In this approach, a CNN is first applied to extract spatial features for each frame, followed by temporal convolutions to refine the predictions. Although this method has limited complexity and a temporal window of only 60 seconds, it has demonstrated promising results and can be trained efficiently in parallel. Notably, the hierarchical modeling of the temporal convolutions resolves the vanishing or exploding gradient problem that can arise with LSTMs.

We, therefore, explored the use of TCNs further and proposed TeCNO, a multi-stage TCN with a ResNet-50 [57] feature extractor. Our approach aims to enhance modeling performance by increasing the modeling strength with multiple TCN stages and a two-stage training regime, allowing for the utilization of a large temporal receptive window. Detailed exposition of the development and results of our proposed method can be found in section 8.

TeCNO has served as a foundation for subsequent works, which have tackled more complex objectives, such as the anticipation of surgical phases. This challenging task involves predicting future phases, rather than the current ones, by utilizing the temporal relationships between surgical tools, segmentation, and phases through temporal convolutions. The model leverages past information leading up to the current time step, to forecast future phases with precision [177].

The revolutionary introduction of attention (section 6.11) for sequential modeling has made a significant impact in the medical field [58] as well as in surgical domains [111]. Our transformer-based method OperA uses a combination of CNNs to extract image features and attention for the temporal processing to predict surgical phases from long video sequences. We propose a novel attention regularized loss to focus our model on image features that lead to the correct prediction and, therefore, are more valuable. We further utilized attention weights to investigate high and low attention frames to find characteristic frames of the individual phases and to gain further insights into the decision-making process of our model. We show more details about our work in section 9.

Building up on OperA, intra-, and inter-video relations have been further explored to improve the accuracy of surgical semantic scene segmentation. By leveraging these relationships, the model could better understand the spatial and temporal dynamics of the surgical procedure, ultimately leading to more accurate predictions of surgical phases and segmentation [67]. Gao et al. [48] combined TCN and attention as a hybrid embedding aggregation to fuse spatial and temporal embeddings, allowing active queries based on spatial information and high speed inference. Valderrama et al. [163] presented a new transformer-based approach for studying surgical interventions using a dataset that includes annotations for long-term and short-term reasoning in robot-assisted surgeries. The approach provided a strong baseline for surgical scene understanding that leverages the multi-level annotations in the dataset toward holistic OR understanding.

In the field of surgical phase recognition, recent works have also explored other methods such as the use of graph neural networks and reinforcement learning in addition to attention and transformers. Graph neural networks offer a more generic and flexible approach for modeling temporal relationships by representing each frame in a video as a node in a graph with temporal connections defined by edges [68]. On the other hand, reinforcement learning has been utilized to identify surgical phase transitions without the need to examine all frames, which can reduce the amount of frames required for inference. However, this approach is limited to offline applications only [180]. To further improve the accuracy and speed of surgical phase recognition, auto-regressive inference strategies have been proposed, which have shown promising results in reducing inference bias while maintaining high speed [182]. Moreover, sequence-to-sequence modeling has been proposed for coarse-level phase segmentation in laparoscopic sacrocolpopexy, which can handle highly variable phase durations [179].

Given the limited availability of data in the surgical domain, self-supervised learning methods have garnered interest for enhancing surgical phase recognition [133]. Examples of such methods include MoCo [56], SimCLR [30], DINO [27], and SwaV [26] and demonstrated the potential of self-supervised learning to improve the performance of surgical phase recognition models without the need for additional labeled data.

Another promising approach to address the issue of limited data sharing in the surgical domain is federated learning. To this end, Kassem et al. proposed FedCy [71], a federated semi-supervised learning method that combines Federated Learning and self-supervised learning to improve surgical phase recognition. FedCy leverages a decentralized dataset of labeled and unlabeled videos, which enables the use of temporal patterns in the labeled data to guide unsupervised training on the unlabeled data. By doing so, FedCy has the potential to improve the accuracy of surgical phase recognition while addressing the challenge of data limitation and restricted data sharing. A closer examination of the practical application of phase recognition models exposed the challenge of accurately identifying underrepresented phases, underscoring the significance of comprehending the distribution of video data in real-world scenarios [76].

In a recent study, attention was used as a replacement for the commonly used CNN backbones in visual feature extraction for surgical phase recognition [38]. This approach was found to be more effective than the use of CNNs, as demonstrated in previous studies, suggesting also a shift beyond CNNs for the feature extraction step. In summary the recent research has demonstrated a trend of shifting away from RNNs towards different temporal approaches like TCNs and attention-based models. Our contributions have helped to clarify and advance this shift, inspiring other researchers to explore new temporal methods, ultimately driving the field towards innovative solutions.

7.3 Surgical Phase Datasets

7.3.1 Overview

In the field of SDS, various datasets have been proposed to facilitate research in surgical phase recognition and consist of recordings from different institutes, hospitals, and surgical approaches, such as laparoscopic, external, and open. As summarized in [49], some available datasets include the **M2CAI** dataset, one of the first datasets containing 41 laparoscopic cholecystectomy surgeries. The dataset was created by the Technical University of Munich [152] and the CAMMA research group [1] in Strasbourg [160] and is publicly available.

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

Another dataset that has gained attention is the **HeiCholec** dataset, which comprises 33 laparoscopic cholecystectomy videos from three surgical centers, totaling 22 hours of operation. It includes labels such as seven surgical phases, 250 phase transitions, 5514 occurrences of four surgical actions, 6980 occurrences of 21 surgical instruments from seven instrument categories, and 495 skill classifications in five skill dimensions. The dataset was used in the 2019 and 2021 Endoscopic Vision challenge [2], a sub-challenge for surgical workflow and skill analysis [171]. The **CATARACTS** dataset [3], which consists of 50 phacoemulsification cataract surgeries, has also been used to detect surgical tool presence and activity recognition. The dataset was divided into two sets (train, test) for surgical tool presence detection and three sets (train, dev, test) for activity recognition [4].

Additionally, the **cholecT50** dataset contains annotated labels for instrument-tissue interaction in laparoscopic cholecystectomies. It includes triplet annotations for every frame, indicating the instrument, verb, and target for surgical activity recognition. This dataset investigates the state-of-the-art in surgical fine-grained activity recognition [113].

The unavailability of many SDS and surgical phase recognition datasets [49, 145] due to regulatory concerns regarding patient confidentiality and data protection has been a major challenge (as discussed in section 2.4.4). However, it is expected that as these regulatory concerns are addressed, more datasets will become available to researchers. In this thesis, the focus is on three laparoscopic cholecystectomy workflow datasets, comprising of two proprietary in-house datasets and one publicly accessible dataset. To provide detailed information, the following sections comprehensively describe these datasets.

7.3.2 Cholec80

The publicly available Cholec80 dataset [159] is a collection of 80 intra-operative video recordings of laparoscopic cholecystectomy procedures performed by 13 surgeons at the University Hospital of Strasbourg. The videos have resolutions of 1920×1080 or 854×480 pixels and were recorded at 25 frames-per-second (4.6 million frames). Each frame of the videos has been manually assigned to one of seven classes representing different surgical phases summarized in table 7.1. The dataset also includes 7 tool annotations, summarized in figure 7.1. Tool annotations are sampled at 1fps and are provided for each video.

Split: There exists no official split for the dataset but many works [65, 66, 160] refer to the original split [159] proposed when the dataset was released. The original split uses the first forty videos (video ID 1-40) and the remaining forty (video ID 41-80) for testing. Eight of the forty training videos are used for validation (video ID 33-40). This splitting does not follow standard approaches for dataset splitting [113] considering video duration or type of intervention. Hence, other splitting approaches have been proposed[37] including techniques like cross-validation which is adequate to account for a limited amount of data.

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

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

Tab. 7.1. PhaseIDs and Phase Names of the different surgical phases in the Cholec80 dataset

Phase ID	**Phase Name**	**Number frames (@25fps)**	relative number
1	Preparation	214k	4.6%
2	Calot Triangle Dissection	1870k	40.6%
3	Clipping and Cutting	352k	7.6%
4	Gallbladder Dissection	1460k	31.7%
5	Gallbladder Packaging	190k	4.1%
6	Cleaning and Coagulation	358k	7.8%
7	Gallbladder Retraction	166k	3.6%

7.3.3 Cholec51

Cholec51 is an in-house dataset of 51 laparoscopic cholecystectomy videos with a resolution 1920×1080 pixels and a sampling rate of 1fps accounting for 85k frames. Cholec51 includes seven surgical phases that slightly differ from Cholec80 and have been annotated by expert physicians. There is no additional tool information provided. The annotation was done by medical experts. We defined a standard split for the dataset to compare our results objectively. We utilized 25 videos for training, 8 for validation, and 18 for testing. Our experiments for both datasets were repeated 5 times with random initialization to ensure the reproducibility of the results. In table 7.2, we show the different phase classes and their absolute and relative frequency.

Tab. 7.2. PhaseIDs and Phase Names of the different surgical phases in the Cholec51 dataset

Phase ID	**Phase Name**	**Number frames (@1fps)**	**relative numbers**
1	pre Preparation	0.2k	0.2%
2	Preparation	17.6k	20.7%
3	Clipping	16.6k	19.5%
4	Dissection	19.8k	23.3%
5	Haemostasis 1	9.1k	10.7%
6	Haemostasis 2	3.7k	4.5%
7	Retrieval	17.8k	20.9%

7.3.4 CSW

CSW (Cholecystectomy Surgical Workflow) is an in-house dataset of 85 laparoscopic cholecystectomy videos with a resolution 1920×1080 pixels and a sampling rate of 1fps. This accounts for 236k frames. CSW includes the 7 surgical phases of Cholec80, shown in Fig. 9.4, along with one additional phase Pre-preparation, used to describe frames before the Preparation phase which we gave the Phase ID 0. Even though the same phase names are used as in the

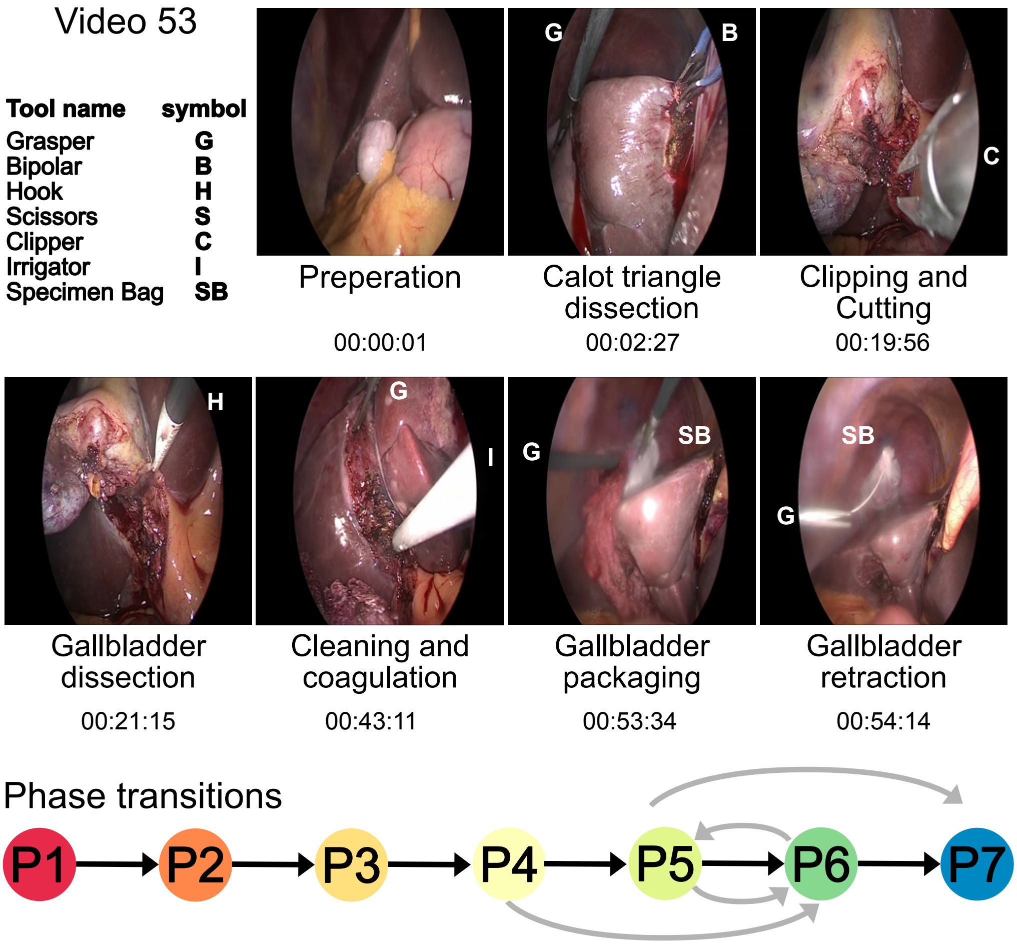


Fig. 7.1. Cholec80 video 53 example frames from each phase P1-7. Below each frame, the corresponding phase and time stamp in the video are shown. Tool information is visualized using the white Tool symbols. The Phase transition shown in black is the actual transition for the video and in grey is the other possible transition in the dataset.

Cholec80 dataset, the annotation protocol was adjusted. In table 7.3 we show the absolute and relative size of each class in terms of the number of frames next to the phase names. In Cholec80 the Calot Triangle Dissection has a relative frame number of 40.6% and only 18.7% for CSW respectively. Both adjacent phases (P1 - 8.9% and P3 - 15.4%) in the CSW dataset have higher relative frame counts than Cholec80 (P1 - 4.6% and P3 - 7.6%). Indicating that a slightly different annotation protocol was applied. The other phases have comparable frame counts, and the additional phase Pre-Preperation has only a limited amount of frames (2k or 0.9%). The phases have been annotated by expert physicians with no additional tool-presence information. 20 videos are utilized for testing and the remaining 65 videos are for training (52) and validation (13).

7.4 Challenges of Laparoscopic Image Data

Laparoscopic image data presents several challenges for analysis and interpretation even for advanced CAI systems [20, 92]. The images captured during laparoscopic procedures are often

Tab. 7.3. PhaseIDs and Phase Names of the different surgical phases and the number of frames absolute at 1fps and relative of the CSW dataset

Phase ID	Phase Name	Number frames (@1fps)	relative numbers
0	Pre-Preparation	2k	0.9%
1	Preparation	21k	8.9%
2	Calot Triangle Dissection	44k	18.7%
3	Clipping and Cutting	36k	15.4%
4	Gallbladder Dissection	42k	17.9%
6	Gallbladder Packaging	7k	3.0%
5	Cleaning and Coagulation	39k	16.6%
7	Gallbladder Retraction	44k	18.7%

of lower quality compared to those obtained through other imaging techniques. Even though the quality can be improved through novel lenses and specialized hardware, size constraints restrict the development of better vision. Factors such as variability of patient anatomy and surgeon style [47] can further be difficult for the analysis along with the limited availability and quality of video material [77]. Visual challenges for analysis and interpretation include occlusion, reflection artifacts, presence of smoke, blood occlusion, motion blur, out-of-body frames, restricted field-of-view, and dirtiness of lens and are visualized in figure 7.2. The use of instruments during laparoscopic procedures can obstruct the view of the camera leading to partial or complete occlusion of the lens. Additionally, the reflective surfaces of some organs and tissues can cause specular reflection, which can negatively impact the quality of the captured images. The presence of smoke, blood splatter, and other debris can also obscure the view and affect image quality. Motion blur can result from the movement of the camera or organs during the procedure, while out-of-body noise can come from electrical interference or other sources. The restricted field of view of laparoscopic cameras can make it challenging to get a comprehensive view of the surgical site, while dirtiness of the lens can impair image quality and clarity.

Further, the Global Ambiguities are visualized in figure 7.3. On the left side of the figure, the different phase names are shown together with a descriptive and ambiguous frame for each phase. It is interesting to see that the ambiguous frames for different phases look very similar to each other while the descriptive frames are very distinctive of their phase. These global ambiguities have to be addressed with models that use the global context of the video in predicting the phase.

Figure 7.4 illustrates a sequence of three frames ranging from $t-1$ to $t+1$, demonstrating the importance of considering the local context when analyzing a frame. Despite the challenges presented by factors such as occlusion (c), the dirtiness of the lens (a), or smoke (b), the inclusion of one step of temporal context can aid in resolving such issues. In the depicted example in figure 7.4 d, it is evident that local ambiguities may extend beyond a single step. In the case of predicting $t+1$, the method would require more than one additional step in

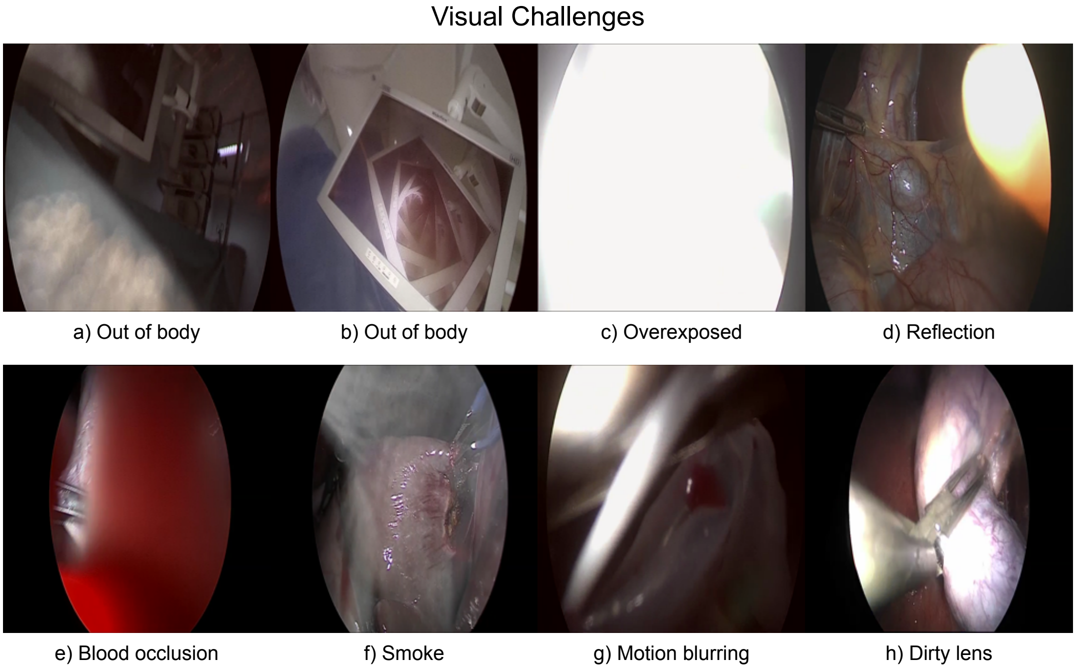


Fig. 7.2. Different visual example images are shown from the Cholec80 dataset [160] including "out of body", "overexposed", "reflection", "occlusion", "blurring"

the past and a greater amount of contextual information to overcome the persistent local ambiguity.

7.5 Metrics

To measure the quality of the surgical activity recognition task, multiple metrics have been proposed [125]. Incorrect detections within short phases may be hidden by overall accuracy but are revealed by precision and recall. In the following, the calculations are given with the true positives (TP), false positives (FP) and false negatives (FN).

7.5.1 Accuracy

Accuracy measures the correctness of the phase detection over a complete surgery in percent. The video-level Accuracy ($VidAcc$) is defined as the following:

$$VidAcc_v = \frac{TP_v}{T_v} \tag{7.1}$$

with T the number of frames in one vide v. The final overall Accuracy is calculated as the macro average over all videos with V the number of videos

$$Accuracy = \frac{1}{V}\sum_{v=0}^{V} VidAcc_v \tag{7.2}$$

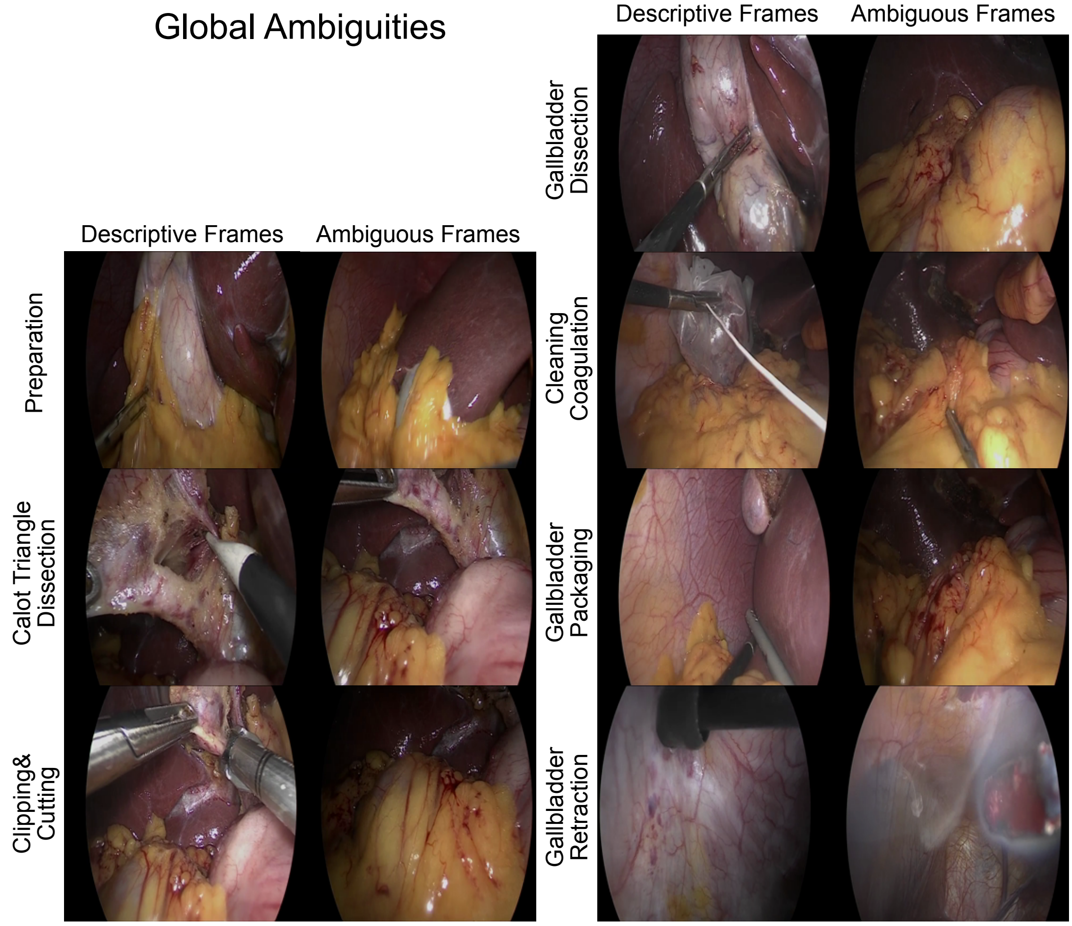


Fig. 7.3. Examples of Descriptive and Ambiguous frames for the same phase from the Cholec80 dataset [160].

7.5.2 Recall

Recall is calculated on a phase level by dividing the number of correct predictions by the length of the ground truth phase. For one video, the recall for one phase p can be calculated using the following equation.

$$VidRec_p = \frac{TP_p}{TP_p + FN_p} \tag{7.3}$$

To calculate the average recall for one phase for all videos with the total number of videos V:

$$Rec_p = \frac{1}{V}\sum_{v=0}^{V} VidRec_p^v \tag{7.4}$$

Finally, the average recall over all phases with P, the number of Phases, is calculated.

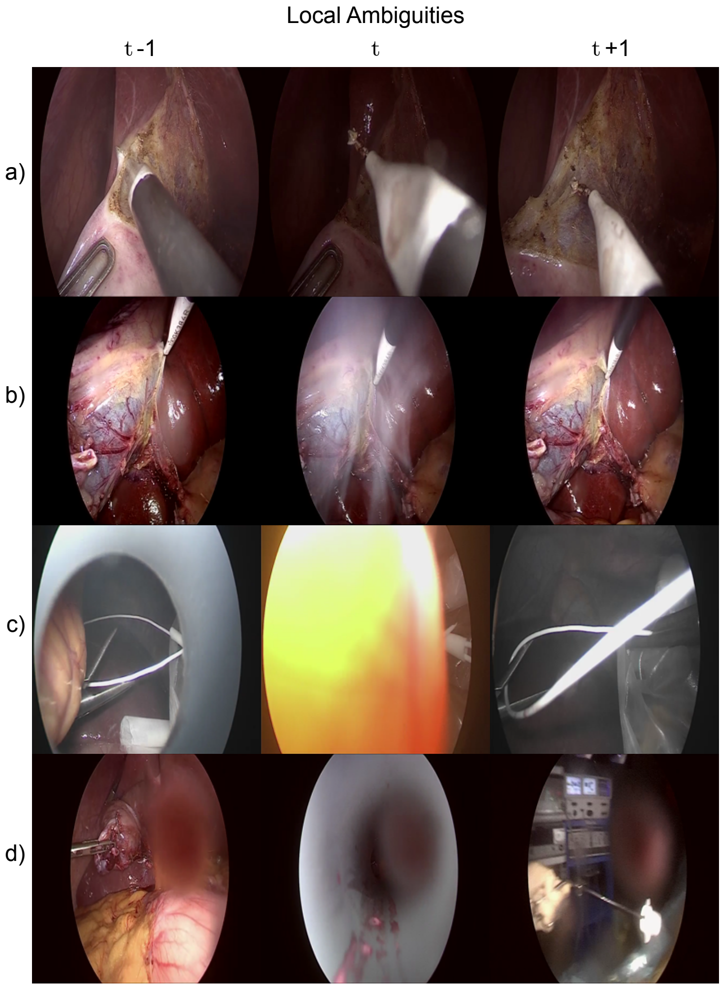


Fig. 7.4. Local ambiguities are visualized. At time t the frame information is not sufficient for a prediction. Including the surrounding time steps $t+1, t-1$ has to be considered to resolve the local ambiguities. The example images are from the Cholec80 [160]

$$Recall = \frac{1}{P}\sum_{p=0}^{P} Rec_p \tag{7.5}$$

In this average process the averaging is first done for each phase on all videos (V) and then averaging the class-level results of the phases (P). We refer to this averaging method as the VP average for the calculation of the recall.

7.5.3 Precision

Precision is calculated on a phase level by dividing the correct predictions of a phase with the sum of correct and incorrect predictions using the ground truth. For one video, the precision for one phase p can be calculated.

$$VidPrec_p = \frac{TP_p}{TP_p + FP_p} \tag{7.6}$$

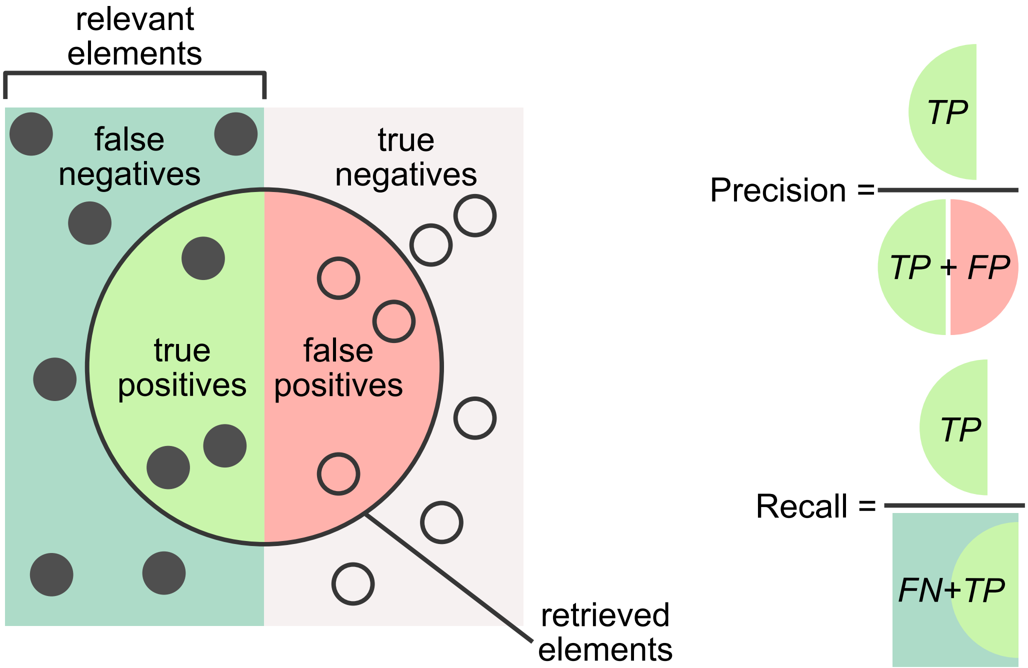


Fig. 7.5. Precision and Recall calculation. Abbreviations: True Positive (TP), False Positive (FP), False Negatives (FN)

To calculate the average precision for one phase p over all videos with the total number of videos V:

$$Prec_p = \frac{1}{V}\sum_{v=0}^{V} VidPrec_p^v \tag{7.7}$$

Finally, the average precision over all phases with P, the number of phases, is calculated

$$Precision = \frac{1}{P}\sum_{p=0}^{P} Prec_p \tag{7.8}$$

Similar to the recall calculation, we use video-phase averaging (VP) to calculate the precision. A graphical explanation for precision and recall are shown in figure 7.5.

7.5.4 F1

F1 is the harmonic mean of precision and recall. However, the precision and recall used for calculating the F-1 values can deviate from the ones presented earlier. While precision and recall are calculated with the video-phase (VP) averaging the F1 is calculated with the phase-video (PV) averaging. Initially, the recall and precision measures are computed for all phases (P) in each video, yielding the precision and recall values ($precision_{vid}$ and $recall_{vid}$) for each video. Subsequently, the results are averaged across all videos (V). To calculate the

F1 value with video-level (PV) averaging for a single video v, one can employ the following formula:

$$F1^v = 2 * \frac{precision_{vid}^v \cdot recall_{vid}^v}{recision_{vid}^v + recall_{vid}^v} \tag{7.9}$$

To calculate the average F1 over all videos V

$$F1 = \frac{1}{V} \sum_{v=0}^{V} F1^v \tag{7.10}$$

7.5.5 Handling *none* values

Class-level metrics such as precision and recall can take on undefined states in cases where a class is missing entirely from the label or the prediction.

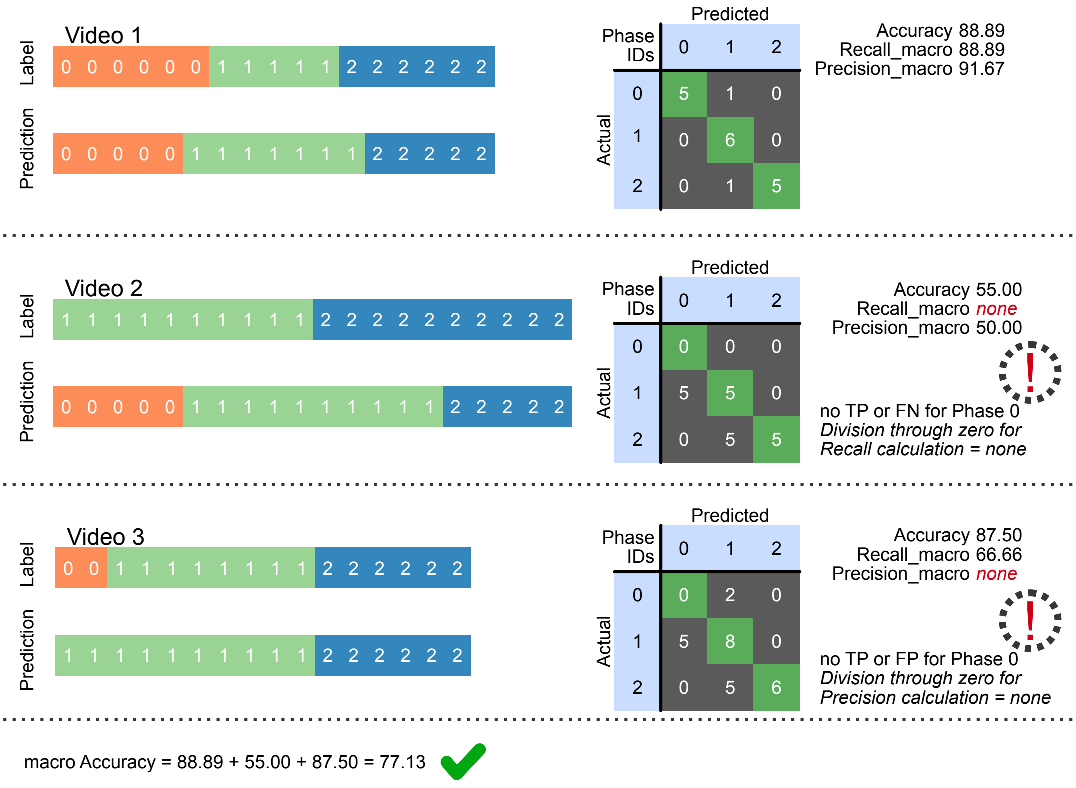


Fig. 7.6. Three example videos with predictions and the corresponding confusion matrices. In Video 2 the calculation for the calculation of Recall there is no TP or FN for Phase 0 leading to an undefined division with zero. For Video 3 for the calculation of Precision there is no TP or FP for Phase 0 leading to an undefined division with zero. The calculation of the Accuracy (macro) is not influenced by this problem.

These undefined states are expressed with *none* values, and for the recall calculation, this happens as described in 7.7 when one phase (e.g. Phase 0 in Video 2) does not appear in the label. In such scenarios, both true positives and false negatives are 0 for this phase leading to a division through 0, which is mathematically undefined. For the calculation of the precision, this case can also appear in cases where one phase is missing from the entire prediction (e.g. Phase 0 in Video 3). In this scenario, both true and false positives are zero, and the denominator of the precision calculation is, therefore, also zero - mathematically undefined. To deal with these undefined values there are two options.

1. treating *none* values as 0.

2. Ignoring *none* values and calculating results by ignoring them

Although treating *none* values as 0 is a simple approach to handle missing data, it can sometimes be misleading. If a particular phase does not appear in a video consistently, then even if the algorithm performs flawlessly, the evaluation metrics for that phase would be 0. This might create the impression that the algorithm is performing worse for that phase compared to other phases with potential errors, despite the fact that the algorithm is actually making correct predictions. Building the macro averages using zero is, however, easier. Both ways of accumulation through averaging: video-phase (VP) and phase-video (PV), lead to the same results, making it less prone to inconsistent metrics.

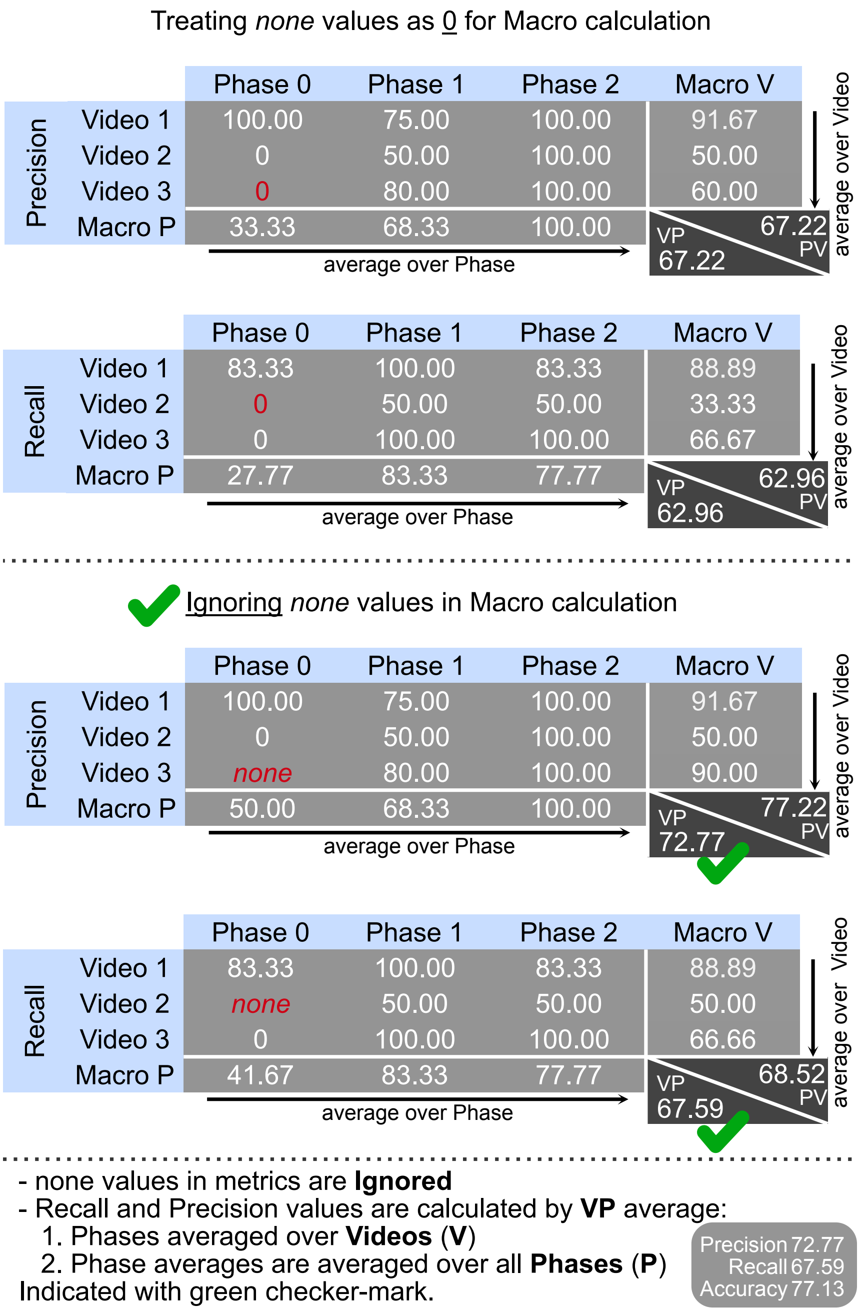
Ignoring *none* values seems more reasonable as a phase not always present in either prediction or label could still be evaluated by building the average values over the remaining results for one phase. Interestingly the accumulation (VP or PV) of the results can lead to different results in this case.

A video-phase average has the tendency to generate lower results as phases that are not always present in each video (minority class) can be more challenging to classify. If a minority class only appears in 2 of 3 possible events, the 2 potentially detrimental values of this class count equally as much as 3 values from a majority phase, which is always present, into the average. With the phase-video average, a minority class always has the same influence on the average for each video as any other class, and if missing, is ignored.

Therefore a phase-video (PV) average usually results in slightly improved results for datasets where phases are missing. However, the minority classes are not well represented this way, so for this thesis, we follow the video-phase average as highlighted in 7.7 and explained in 7.5.3 and 7.5.2.

In our work we decided to ignore *none* values and apply the VP accumulation for the precision and recall calculation.

Evaluation Metrics To comprehensively measure the results of the phase prediction, we deploy three different evaluation metrics suitable for surgical phase recognition [125], namely Accuracy (Acc), Precision (Prec), and Recall (Rec). Accuracy quantitatively evaluates the amount of correctly classified phases in the whole video, while precision, or positive predictive value, and recall, or true positive rate, evaluate the results for each individual phase [161].



Treating *none* values as 0 for Macro calculation

		Phase 0	Phase 1	Phase 2	Macro V
Precision	Video 1	100.00	75.00	100.00	91.67
	Video 2	0	50.00	100.00	50.00
	Video 3	0	80.00	100.00	60.00
	Macro P	33.33	68.33	100.00	VP 67.22 PV 67.22

average over Phase — average over Video

		Phase 0	Phase 1	Phase 2	Macro V
Recall	Video 1	83.33	100.00	83.33	88.89
	Video 2	0	50.00	50.00	33.33
	Video 3	0	100.00	100.00	66.67
	Macro P	27.77	83.33	77.77	VP 62.96 PV 62.96

average over Phase — average over Video

✓ Ignoring *none* values in Macro calculation

		Phase 0	Phase 1	Phase 2	Macro V
Precision	Video 1	100.00	75.00	100.00	91.67
	Video 2	0	50.00	100.00	50.00
	Video 3	*none*	80.00	100.00	90.00
	Macro P	50.00	68.33	100.00	VP 72.77 ✓ PV 77.22

average over Phase — average over Video

		Phase 0	Phase 1	Phase 2	Macro V
Recall	Video 1	83.33	100.00	83.33	88.89
	Video 2	*none*	50.00	50.00	50.00
	Video 3	0	100.00	100.00	66.66
	Macro P	41.67	83.33	77.77	VP 67.59 ✓ PV 68.52

average over Phase — average over Video

- none values in metrics are **Ignored**
- Recall and Precision values are calculated by **VP** average:
  1. Phases averaged over **Videos** (**V**)
  2. Phase averages are averaged over all **Phases** (**P**)

Indicated with green checker-mark.

Precision	72.77
Recall	67.59
Accuracy	77.13

Fig. 7.7. Metrics definition for Recall, Precision and Accuracy. All metrics are given in %. Treating the *none* values as zero or ignored is compared. Additionally the Phase-Video average (PV) and Video-Phase average (VP) is shown. Abbreviations: true positives (TP), false positive (FP), false negatives (FN)

Part III

Temporal Learning for Surgical Phase Recognition

8 Surgical Phase Recognition with Multi-Stage Temporal Convolutions

Contents

8.1 Introduction

In section 6.10, we introduced Temporal Convolutions and their effectiveness in analyzing long-temporal relationships. Their hierarchical process to capture activities in sequential data and videos is well suited for surgical phase recognition. By stacking multiple dilated convolutions, the size of the temporal receptive field can be increased efficiently. Additionally, Multi-Stage TCNs (MS-TCNs) can improve the results further as they can be used to distill the results incrementally.

We propose a pipeline employing multi-stage, multi-layer dilated TCNs for accurate and fast surgical phase recognition. Combining a large temporal receptive field with computational efficiency allows fast training and inference on long untrimmed surgical videos. We further propose the use of causal convolutions which allows the model to run online and name our approach TeCNO, derived from **Te**mporal **C**onvolutional **N**etworks for the **O**perating room.

8.2 Methodology

Our surgical workflow recognition pipeline TeCNO is made up of a few key steps. Firstly, we use a ResNet50 to gather visual information. Then, we enhance that information using a

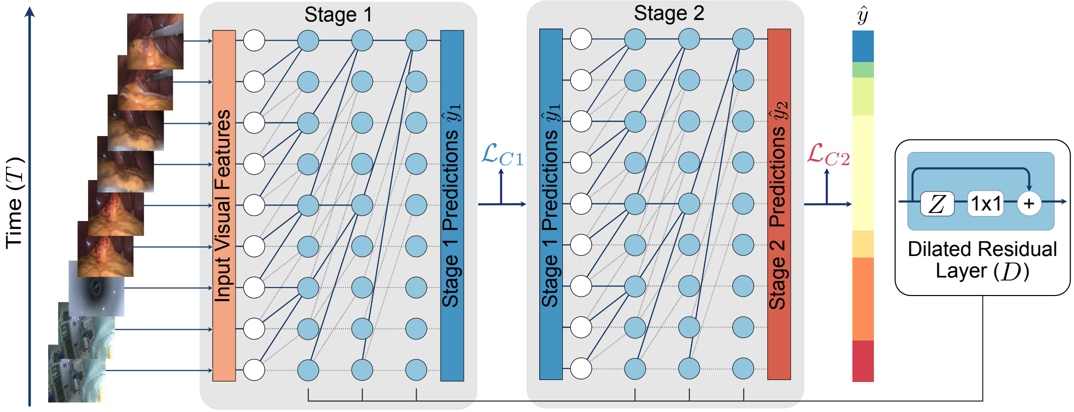

Time (T)
Stage 1
Stage 2
Input Visual Features
Stage 1 Predictions $\hat{y}_1$
$\mathcal{L}_{C1}$
Stage 1 Predictions $\hat{y}_1$
Stage 2 Predictions $\hat{y}_2$
$\mathcal{L}_{C2}$
$\hat{y}$
Z
1x1
+
Dilated Residual Layer (D)

Fig. 8.1. Overview of the proposed TeCNO multi-stage hierarchical refinement model. The extracted frame features are forwarded to Stage 1 of our TCN, which consists of 1D dilated convolutional and dilated residual layers D. Cross-Entropy Loss is calculated after each stage and aggregated for the joint training of the model. *Reproduced with permission from Springer Nature* [36].

2-stage causal TCN model, which basically takes into consideration what's happened before in the video to make a more informed decision about the current frame. This process is shown in figure 8.1.

8.2.1 Feature Extraction Backbone

We train a backbone ResNet50 [57] model to extract information from video frames. It can be used on its own to identify the phase of a surgical procedure using a single frame without further temporal information. Surgical datasets can include additional tool labels, which the backbone can utilize in two separate layers predicting phase label and tool identification. For phase recognition, an imbalanced multi-class problem, we use softmax activations and the weighted cross entropy loss to support the balancing of the classes. They use median frequency balancing [42] to calculate the class weights. We use a binary cross entropy loss after a sigmoid activation for the multi-label tool identification, as multiple tools can be present in a single frame.

We use a two-stage method setup of backbone and temporal method, which ensures that our temporal refinement pipeline is independent of the feature extractor and the available ground truth labels provided in the dataset. We tested different backbone architectures and configurations and show in Section 8.3 that TCNs can improve the predictions over various settings.

8.2.2 Temporal Convolutional Networks

We created TeCNO, a multi-stage temporal convolutional network for the temporal phase prediction task visualized in Fig. 8.1. The goal is to predict the class label $y_{1:t}$ for each frame in a given input video with $x_{1:t}, t \in [1, T]$ and T the total number of frames. To train and validate the method we use y_t the phase labels corresponding to the input videos.

Our model comprises only temporal convolutional layers and does not use pooling layers, decreasing the temporal resolution, or parameter-heavy fully connected layers following the design of MS-TCN and can analyze input sequences of varying lengths.

The first stage begins with a 1x1 convolutional layer to downscale the input feature dimension to the required dimension used in the subsequent layers. Following this downscaling, the rest of the stage is made up of dilated residual (D) layers detailed in Eq. 8.1 and Eq. 8.2. The central component of the dilated residual layer D is the dilated convolutional layer (Z) with

$$Z_l = ReLU(W_{1,l} * D_{l-1} + b_{1,l}) \tag{8.1}$$

$$D_l = D_{l-1} + W_{2,l} * Z_l + b_{2,l} \tag{8.2}$$

D_l is the output of D while Z_l is the result of the dilated convolution of kernel $W_{1,l}$ at layer l. To calculate Z_l the output of the previous layer D_{l-1} activated by a $ReLU$ is used(Eq. 8.1). $W_{2,l}$ is the kernel for the 1x1 convolutional layer with $*$ the convolutional operator and bias vectors $b_{1,l}, b_{2,l}$.

The original MS-TCN [43] was designed to not only consider information about the past for the prediction at time t but also include information of future time steps $\hat{y}_t(x_{t-n}, ..., x_{t+n})$ depending on both n past and n future frames. This acausal implementation is unsuitable for causal deployment, so we modified its implementation. The causal convolutions follow the design of acausal convolutions, including a kernel size of 3 with a dilation factor but the output of each convolution is shifted and the predicion $\hat{y}_t$ does no rely on future frames $\hat{y}_t(x_{t-n}, ..., x_t)$. This modification enables us to use our TeCNO model online.

We consecutively increase the dilation factor of the causal convolutions by two within the D layer for each consecutive layer. The increase in dilation directly influences the size of the temporal receptive field RF of the network (Eq.8.3) as visualized in Fig. 8.1.

While the first D layer with a dilation factor of one and kernel size of three can observe three time steps, an additional second D layer can already observe seven steps. Stacking more and more layers upon each other quickly increases the temporal receptive window further. The size of the temporal receptive field is depended on the number of D layers

$$RF(l) = (2)^{l+1} - 1 \tag{8.3}$$

It is also possible to skip, e.g., the first layer, and start with a high dilation factor. However, frames or information will be skipped, comparable to the downsampling of a video from higher to lower fps, and important frames with crucial information could be missed.

One main advantage of our method is that the resulting exponential increase of the receptive field comes with significantly reduced computational costs compared to models that construct larger temporal receptive fields by increasing the kernel size [164].

Multi-Stage TCN

The principal idea of adding multiple stages of the same network blocks is to further refine the output of the stages [110]. The extracted visual features serve as starting point of the sequential modeling in S_1. The output of stage S_1 is then used as the input for S_2. This process can be repeated M times to generate the final output visualized in figure 8.1. It is important to note that each output of $S_{1...M}$ has its own loss function, in our case, the weighted cross-entropy loss, as described in Eq. 8.4. We calculated the class weights w_c for the loss using median frequency balancing [42]. Our model TeCNO does not require tool information for the training and utilizes phase recognition labels exclusively.

$$\mathcal{L}_C = \frac{1}{M}\sum_{m}^{M}\mathcal{L}_{Cm} = -\frac{1}{M}\frac{1}{T}\sum_{m}^{M}\sum_{t}^{T} w_c y_{mt} \cdot log(\hat{y}_{mt}) \tag{8.4}$$

The final result is the prediction calculated after the last stage of the pipeline, where the conclusive refinement is performed.

8.2.3 Model Training

We used two laparoscopic cholecystectomy datasets with phase annotations for the training of TeCNO. The Cholec80 dataset introduced in Section 7.3.2 is one of the most frequently used datasets for surgical phase recognition. We downsampled the dataset from 25fps to 5fps resulting in a total of ~92000 frames. We followed separated the dataset into differents parts and used 40 videos for training, 8 for validation, and 32 for testing. Additionally we used an the in-house dataset Cholec51 (Section 7.3.3) with a sampling rate of 1fps.

For the surgical phase recognition task, we used TeCNO and trained it with the Adam optimizer and a learning rate of 5e-4 for 25 epoch. We present the test results from the model with the highest performance on the validation set. We utilized a batch size equal to the length of each video and using the PyTorch deep learning library. The training was run on an NVIDIA Titan V 12GB GPU using Polyaxon[1]. The source code for TeCNO is publicly available[2].

8.2.4 Ablative Testing

To determine an appropriate feature extractor for our MS-TCN model, we conducted experiments using two different CNN architectures: AlexNet [81] and ResNet50 [57]. We also experimented with varying numbers of TCN stages to determine which architecture most effectively captures the long temporal relationships in our surgical videos.

8.2.5 Baseline Comparison

We compared our TeCNO model with other surgical phase recognition models, such as PhaseLSTM [159], EndoLSTM [160], and MTRC-Net [66]. These models use LSTMs to capture the temporal information in surgical videos, as LSTMs have been shown to be more effective than HMMs in previous studies [174]. MTRCNet is trained end-to-end from surgical images to surgical phase prediction, while the other LSTM methods and TeCNO focus on

[1] https://polyaxon.com/

[2] https://github.com/tobiascz/TeCNO/

Tab. 8.1. Ablative testing results for AlexNet with no TCN () and increasing number of stages (#Stages) on Cholec80. Average metrics over multiple runs are reported (%) along with their respective standard deviation (±). *Reproduced with permission from Springer Nature* [36].

AlexNet			
#Stages	Accuracy	Precision	Recall
-	74.40 ± 4.30	63.06 ± 0.32	70.75 ± 0.05
I	84.04 ± 0.98	79.82 ± 0.31	79.03 ± 0.99
II	**85.31** ± 1.02	**81.54** ± 0.49	79.92 ± 1.16
III	84.41 ± 0.85	77.68 ± 0.90	**79.64** ± 1.6

Tab. 8.2. Ablative testing results for ResNet with no TCN () and increasing number of stages (#Stages) on Cholec80. Average metrics over multiple runs are reported (%) along with their respective standard deviation (±). *Reproduced with permission from Springer Nature* [36].

ResNet50			
#Stages	Accuracy	Precision	Recall
-	82.22 ± 0.60	70.65 ± 0.08	75.88 ± 1.35
I	88.35 ± 0.30	**82.44 ± 0.46**	84.71 ± 0.71
II	**88.56 ± 0.27**	81.64 ± 0.41	**85.24 ± 1.06**
III	86.49 ± 1.66	78.87 ± 1.52	83.69 ± 1.03

refining already extracted features. Since the Cholec51 dataset does not have surgical tool labels, EndoLSTM and MTRCNet could not be applied to the dataset. For the Cholec80 dataset, the feature extractors for all models were trained on phase and tool recognition, except for PhaseLSTM, which only used phase labels. For Cholec51, we trained the CNNs utilized to extract features only on the phase recognition task, as no further tool annotations were available for the dataset.

8.3 Results

8.3.1 Effect of Feature Extractor Architecture

ResNet50 outperforms AlexNet on all measured metrics with 2% to 8% accuracy improvements and even greater precision and recall, as summarized in Table 9.1. The advances of ResNet50 over AlexNet are coherent beyond all TCN stages, with improvements of up to 7% in precision and 6% in recall. We, therefore, selected ResNet50 as our feature extractor backbone network.

Tab. 8.3. Baseline Comparison for Cholec80 Dataset. The average metrics over multiple runs are reported (%) along with their respective standard deviation. *Reproduced with permission from Springer Nature* [36].

Cholec80			
Method	Accuracy	Precision	Recall
PhaseLSTM [160]	79.68± 0.07	72.85± 0.10	73.45± 0.12
EndoLSTM [161]	80.85± 0.17	76.81 ± 2.62	72.07± 0.64
MTRCNet [66]	82.76± 0.01	76.08± 0.01	78.02± 0.13
ResNetLSTM [65]	86.58± 1.01	80.53± 1.59	79.94± 1.79
TeCNO	**88.56**± 0.27	**81.64**± 0.41	**85.24**± 1.06

Tab. 8.4. Baseline Comparison for Cholec51 Dataset. EndoLSTM and MTRCNet require tool labels, therefore cannot be applied for Cholec51. The average metrics over multiple runs are reported (%) along with their respective standard deviation. *Reproduced with permission from Springer Nature* [36].

Cholec51			
Method	Accuracy	Precision	Recall
PhaseLSTM [160]	81.94± 0.20	68.84± 0.11	68.05± 0.79
EndoLSTM [161]	—	—	—
MTRCNet [66]	—	—	—
ResNetLSTM [65]	86.15± 0.60	70.45± 2.85	67.42± 1.43
TeCNO	**87.34**± 0.66	**75.87**± 0.58	**77.17**± 0.73

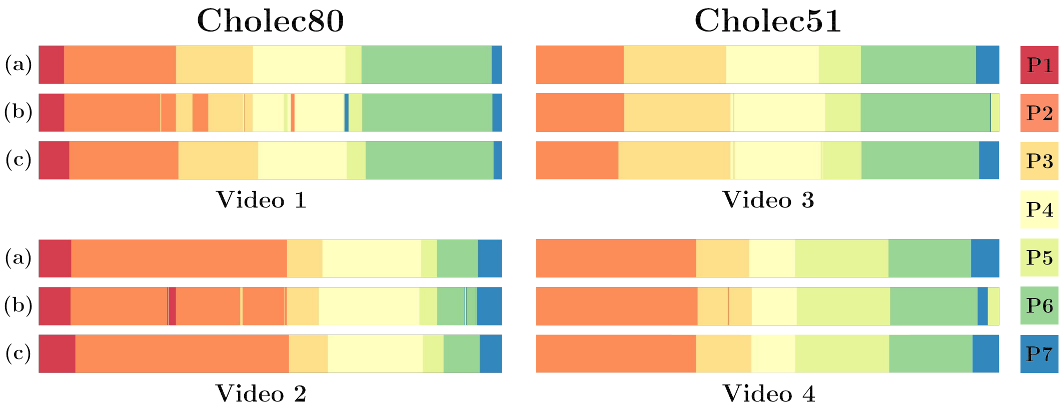


Fig. 8.2. Qualitative Results regarding quality of phase recognition for Cholec80 and Cholec51. (a) Ground Truth (b) ResNetLSTM Predictions (c) TeCNO Predictions. P1 to P7 indicate the phase label. *Reproduced with permission from Springer Nature* [36].

8.3.2 Effect of TCN and Number of Stages

We summarized the results of the temporal refinement over multiple stages using TCN in Table 9.1. We observe improvements in the accuracy of 10% and 6%, respectively, utilizing a single TCN Stage for both backbone architectures. These significant improvements in accuracy demonstrate the need for temporal modeling for the problem of surgical phase recognition. Additionally, it is noteworthy that the TCNs can improve the performance of different CNN feature extractor backbones. With the addition of a second TCN stage, the results are further enhanced on our metrics. By adding more complexity through an additional third Stage, the accuracy drops again by 1% and 2% for AlexNet and ResNet50, respectively. This decrease in performance suggests that a third TCN refinement stage leads to overfitting on the training set for our limited amount of data.

8.3.3 Comparative Methods

We also compared TeCNO to other surgical phase recognition methods which utilize LSTMs to encode the sequential information of the surgical phase recognition task. We summarize these results in Table 8.3 for Cholec80 and Table 8.4 for Cholec51. We can see that ResNetLSTM and TeCNO exceed PhaseLSTM [158] and EndoLSTM [158] substantially in accuracy by 6% and 8% for both datasets respectively. This difference in performance can be attributed to the use of the AlexNet backbone, which is a less capable backbone compared to ResNet50. The Comparison of the end-to-end trained MTRCNet against two-step ResNetLSTM and TeCNO is very insightful, and we can see that both two-step methods outperform the end-to-end model by 4% and 6% in accuracy. The difference between ResNetLSTM and TeCNO is limited, but TeCNO leads the accuracy metric by 1-2%. The difference between the two methods is more prominent in precision and recall, where TeCNO improves the results of ResNetLSTM by 6%-10%. The sizeable temporal resolution and receptive field of TeCNO enhance performance in under-represented and majority classes.

8.3.4 Phase Recognition Consistency

We show qualitative results on four different laparoscopic videos on both dataset in figure 8.2. The results showcase the ability of TeCNO for consistent and smooth results within and across different phases and phase transitions. We see that TeCNO can better analyze shorter phases like P5 or P7, which is an essential property as the phase durations are usually unbalanced. The robustness of TeCNO to missing phases can be seen in Video 3 and 4 where P1 is missing.

9 Attention-Regularized Transformers for Surgical Phase Recognition

Contents

9.1 Introduction

A recent advancement in machine learning that could help with the challenges of surgical workflow analysis is transformer networks [165] and their attention functionality. We gave an overview of attention methods in section 6.11 and summarized their capabilities for temporal and sequential modeling. Transformer networks have already shown success in the field of natural language processing [40] by creating relationships between current and earlier time steps using self-attention. Another benefit of transformers and self-attention is the ability to visualize the attention weights of a sequence, which provides insight into the model's decision-making process. We therefore introduced our **Oper**ation **A**ttention model OperA. OperA is a transformer-based method for online surgical phase prediction in laparoscopic interventions. The main contributions of OperA are that we used a transformer-based model for surgical phase recognition that surpasses other temporal refinement methods with a novel attention regularizer to extract the most relevant frames. We used the attention weights to visualize characteristic frames and evaluated OperA on two datasets.

9.2 Methodology

Our proposed model, OperA, combines a CNN and multiple self-attention layers. The CNN extracts visual features from each frame and the self-attention layers create relationships between the frame features. During training, we also use a novel regularizer to keep the attention focused on the most reliable image features. The full design of OperA can be seen in Fig. 9.1. For the visual feature extraction part of OperA, we trained a ResNet-50 [57]

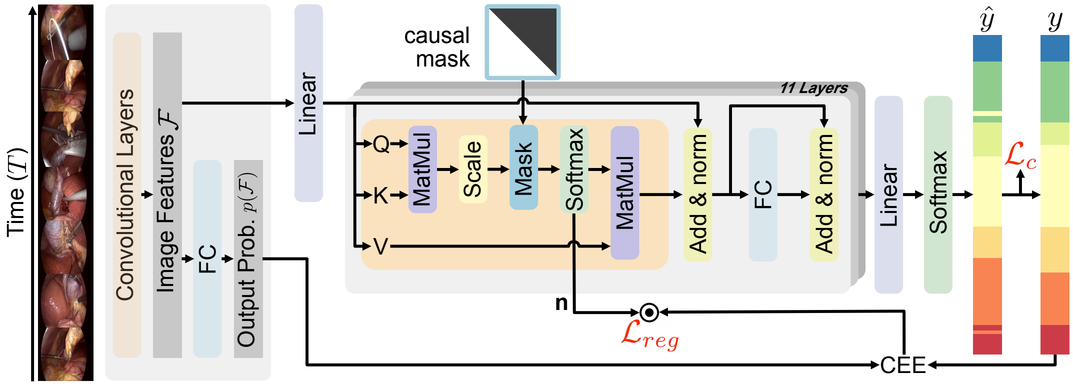


Fig. 9.1. Overview of the proposed OperA model. Image features $\mathcal{F}$ are used as input for the transformer. The output logits $p(\mathcal{F})$ of the feature extraction backbone are used in combination with the normalized frame-wise attention weights **n** to regularize the attention. *Reproduced with permission from Springer Nature* [37].

CNN as described in section 8.2.1. We trained this CNN on the phase recognition task and if tool informations was available additionally on the surgical tool detection task. The output of the CNN is a set of image features, $\mathcal{F}$, for each frame, represented as a vector of $\mathbb{R}^{2048}$ numbers. The CNN also provides class probabilities, $p(\mathcal{F}) \in [0,1]^c$, for each frame, indicating the likelihood that the frame belongs to each of the c classes.

9.2.1 Sequential Transformer Network

Our model, OperA, is based on the popular Transformer architecture [165] with a unique addition - our attention regularization. Transformers have the ability to model long sequences by relating each input feature to every other input feature, regardless of their position in the sequence, using self-attention [75]. Our method OperA is visualized in figure 9.1. The information flow from input to output follows the traditional attention mechanism. First wuery Q, key K and value V are computed using a linear layer and used as the inputs for the scaled dot product attention $(Q, K, V) = \text{Linear}(\mathcal{F}) \in \mathbb{R}^{3d}$ with $d = 64$.

$$\text{AttentionWeights}(Q, K) = \text{softmax}\left(\text{mask}\left(\frac{QK^T}{\sqrt{d}}\right)\right) \tag{9.1}$$

$$\text{Attention}(Q, K, V) = \text{AttentionWeights}(Q, K)V \tag{9.2}$$

We use 11 consecutive transformer encoder layers in our architecture, each following the vanilla transfomer architecture introduced in section 6.11. One layer consists of a scaled dot-product attention layer, normalization layer [8] and residual connection [57]. Following the design of the Vision Tranformer [41] architecture, we use a linear layer followed by a softmax activation to estimate the class-wise probabilities y time step in the sequence. We kept the loss function simple and used a median frequency balanced cross-entropy loss [42] $\mathcal{L}_c$. We prevent information flow from future time steps for predicting the current step through the causal masking [134] of the scaled attention weights. The causal mask is implemented as a binary mask $M \ni \{0, 1\}$ where 0 represents illegal future attention and one causal attention. This causal masking makes our method OperA suitable for online applications in the OR.

9.2.2 Normalized Frame-Wise Attention

The Attention weights $A = \text{AttentionWeights}(Q, K)$ for each frame express how much the other frames influenced the features from the current step. Each frame in the sequence accounts for one column in A. A_{ij} with $i = j = t$ encodes the influence of the frame at step t to itself or more general it quantifies how much attention is being paid by frame i (query) to frame j (key). M_{ij} will be zero if the key index j is larger than the query index i as we want to restrict our model to only respect previous events in the video. Each row in matrix A sum up to one due to the softmax activation step (Eq. 9.1). By summing up the values in A column-wise with $a_j := A[:, j]$ the total attention value a_j for each frame at time t can be observed. The causal masking of A ensures that any future frames cannot influence the encoding of the earlier time steps in the sequence. This also leads to the effect that the last frame at time $j = T$ of a sequence of length T can only contribute to itself. The first frame of the sequence can, however, attend to all frames of the sequence, contributing T times. We use a normalization step where we divide the total attention of each frame by the number of possible attentions m_t for the same frame calculated by $m_j := M[:, j]$ with $M_{ij} = 0$ if $j > i$ else 1. This normalization allows us to compare the total attention of the frames with each other regardless of their sequence position. We defined the frame-wise attention n using the $\mathbf{L}_1$ norm $||.||_1$ as:

$$\mathbf{n} = (n_1, \ldots, n_T) \text{ with } n_j = \frac{||a_j||_1}{||m_j||_1} \tag{9.3}$$

In figure 9.2 we visualized an example of the attention matrix A and details about for the calculation of the frame-wise attention n.

9.2.3 Attention Regularization

The visual feature embeddings extracted by a CNN backbone network form the input to OperA. In contrast to language tokens in NLM or pixel inputs for Visual Transformers the visual embeddings can be considered as noisy input. The resulting embedding will be inadequate for frames where the CNN prediction is incorrect. This problem could be circumvented by end-to-end training of the CNN backbone and OperA, which is unfortunately impossible due to the length of the input surgical videos. Therefore, we follow a different approach to this problem and try to make OperA focus on high-quality visual embeddings that led to correct classification results in the backbone. High-quality visual embeddings have higher softmax probabilities, output confidence, and lower cross-entropy loss values. This relationship between visual feature quality and attention can be modeled by comparing the normalized frame-wise attention weights with the prediction error of our CNN.

$$\mathcal{L}_{reg} = \langle \mathbf{n}, \text{CEE}\left(p(\mathcal{F}), y\right) \rangle \tag{9.4}$$

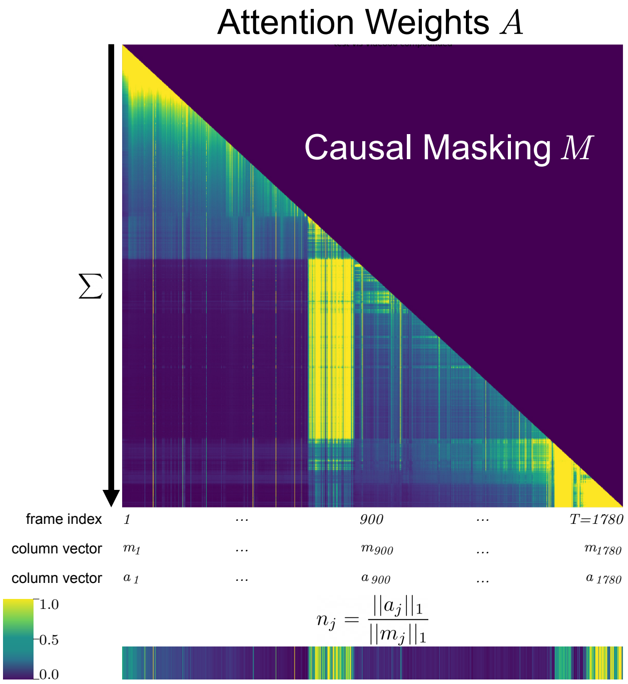


Fig. 9.2. The attention weights A are visualized with the upper triangular causal masking using M on an example with sequence length 1780. We use the column-wise summation of A and M to calculate the normalized frame-wise attention n using the $\mathbf{L}_1$ norm

We introduce the Cross-Entropy Evaluation error (CEE) describing the residual error of $p(\mathcal{F})$ and the label y. For the training of OperA, the weights of the backbone CNN are not further optimized due to the memory limitations mentioned above. We multiply the CEE values with the normalized frame-wise attention **n** to penalize the model if a high attention value was generated for a frame with a large Evaluation error (CEE). To focus only on the direct relationship with the input visual features we only apply the proposed regularization on the first attention layer.

The combined loss function is denoted as: $\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_{reg}$

We additionally used the normalized frame-wise attention to interpret if OperA is currently focussing on informative frames for each phase and extract the frames with the Highest (HA) and Lowest (LA) Attention for the qualitative results.

9.3 Experimental Setup

Datasets

For the evaluation of OperA we use two challenging surgical workflow intra-operative video datasets of laparoscopic cholecystectomy procedures. Cholec80 (section 7.3.2) and CSW (section 7.3.3). For all the experiments 5-fold cross validation is performed and the datasets were sub-sampled to 1fps. For the balancing of our loss function we set λ to 1.

Tab. 9.1. Ablative testing results for 6 and 11 transformer layers and with the addition of Attention Regularization (Reg). Average metrics over 5 folds are reported (%) with the corresponding standard deviation (±). *Reproduced with permission from Springer Nature* [37].

Layers	Reg	Accuracy	Precision	Recall
			Cholec80	
6	-	90.35 ± 0.71	80.64 ± 1.41	86.48 ± 0.61
6	✓	90.49 ± 0.70	81.38 ± 0.29	**86.98** ± **0.61**
11	-	90.37 ± 0.86	81.60 ± 0.40	86.23 ± 0.34
11	✓	**91.26** ± **0.64**	**82.19** ± **0.70**	86.92 ± 0.86
			CSW	
6	-	84.88 ± 1.43	82.76 ± 1.43	87.20 ± 1.02
6	✓	85.41 ± 0.95	83.00 ± 1.34	87.41 ± 1.66
11	-	85.02 ± 1.01	82.89 ± 1.20	**87.82** ± **0.75**
11	✓	**85.77** ± **0.95**	**83.32** ± **1.52**	87.68 ± 1.08

In previous works, the Cholec80 dataset was divided into 32 videos for training, 8 for validation, and 40 for testing. However, we decided to increase the number of validation videos from 8 to 12 because we felt that 8 videos was not a large enough sample size to determine the best model for the test set. Similarly, we also increased the number of training videos from 32 to 48. We used 5-fold cross-validation on the 48 training and 12 validation videos. Following previous studies [36, 65, 66, 161], we kept a separate test set of 20 videos that was not part of the cross-validation. During the test phase, we used the best model from each cross-validation split and took the average of the results from the 5 models on the unseen test set. All the results we report, including the baselines, were obtained using the same data split and cross-validation method.

For the in-house CSW dataset we utilized 20 videos for testing and the remaining 65 videos for training. From the training videos we used for each fold 52 videos for training and 13 for validation.

Model Training

Our model, OperA, was trained to recognize surgical phases using the Adam optimization algorithm with a starting learning rate of 1e-5 for 30 epochs. To evaluate its performance, we selected the best performing model from each of the five folds of our cross-validation based on its performance on the validation set. Each batch in our model was equivalent to the length of one video. We implemented OperA using the PyTorch framework and ran it on an NVIDIA Titan V GPU with 12GB of memory, using the Polyaxon platform [1].

[1] https://polyaxon.com/

Evaluation Metrics and Baselines

For the evaluation we used the metrics proposed in section 7.5. We use the video-level Accuracy (Acc) and the Precision and Recall values [125] averaged over the 5 splits. We identified the most suitable number of attention layers and tested the effect of our novel attention regularization (section 9.2.3) in an ablative testing step. We benchmark OperA with different surgical phase recognition baselines.

9.4 Results

Effect of Layers and Regularization In Table 9.1, we perform a comparison between models that were trained with 6 and 11 attention layers, and examine the effect of attention regularization. 11 attention layers represent the maximum number of layers that could be accommodated within the GPU memory. The results indicate a slight improvement of approximately ∼1% in terms of Accuracy and Precision for 11 attention layers, with similar Recall values for both datasets. Furthermore, the results show that the attention regularization results in a marginal improvement of around ∼1% for both datasets and the number of layers considered. Subsequently, we will demonstrate that the attention regularization not only results in a small increase in model performance, but also enhances the quality of the highest attention video frames.

Baseline Comparison In our evaluation of surgical phase recognition methods, we compare OperA to other models, including ResNet-50 as the feature extraction backbone, ResLSTM [65] and MTRCNet-CL [66] which incorporate LSTMs for temporal refinement. Unlike the other models, MTRCNet-CL is trained in an end-to-end manner, combining CNN feature extraction with LSTM training. However, due to memory constraints, only a limited portion of the video can be used per batch in this approach. In contrast, ResLSTM, TeCNO, and OperA utilize pre-trained image features, allowing for full video sequence analysis in a single pass.

We conducted a comparison of OperA with several methods for surgical phase recognition. In this comparison, we found that the models incorporating temporal refinement outperformed the ResNet-50 baseline by a substantial margin, ranging from 4-10% for Cholec80 and 9-12% for CSW. The MTRCNet-CL, which utilizes an end-to-end approach combining CNNs with LSTMs, was outperformed by the other temporal models by 2-6%, possibly due to the limited video sequence length that can be processed in a single batch. OperA, with or without positional encoding (PE) [165], outperformed the other temporal models in terms of accuracy for Cholec80 by 2-6%, showcasing the ability of transformers to model long temporal dependencies. For the CSW dataset, OperA without PE showed improved accuracy by 0.6-3%, as well as improved precision and recall over all baselines. It was observed that PE slightly decreased the performance for both datasets, which may be due to the longer sequence lengths in surgical videos compared to NLP tasks.

Predictions and Attention Values

In Figure 9.3, we present a visual analysis of the ground truth, predictions, and attention values for video 66 in the Cholec80 dataset. It is observed that the predictions made by OperA

Tab. 9.2. Baseline comparisons for Cholec80 and CSW. MTRCNet-CL requires tool information, thus cannot be used for CSW. We report the average metrics over 5-fold cross validation along with their respective standard deviation (±). *Reproduced with permission from Springer Nature* [37].

	Accuracy	Precision	Recall
	Cholec80		
ResNet-50	81.21 ± 1.16	68.35 ± 1.61	78.31 ± 1.14
ResLSTM	87.94 ± 0.80	80.26 ± 1.12	84.43 ± 0.85
MTRCNET-CL	85.64 ± 0.21	79.31 ± 0.97	82.67 ± 0.114
TeCNO	89.05 ± 0.79	80.90 ± 0.75	**87.44** ± **0.64**
OperA + PE	90.20 ± 1.45	80.78 ± 1.42	86.08 ± 0.89
OperA	**91.26** ± **0.64**	**82.19** ± **0.70**	86.92 ± 0.86
	CSW		
ResNet-50	73.90 ± 1.89	69.06 ± 1.38	74.20 ± 1.63
ResLSTM	82.97 ± 1.18	82.08 ± 1.57	86.15 ± 0.94
MTRCNET-CL	–	–	–
TeCNO	85.09 ± 1.67	82.03 ± 0.20	86.50 ± 0.43
OperA + PE	83.67 ± 1.54	81.34 ± 1.60	86.94 ± 0.98
OperA	**85.77** ± **0.95**	**83.32** ± **1.10**	**87.68** ± **0.71**

are more consistent and smoother than those made by the CNN model. Furthermore, the high attention (HA) frames, indicated by triangles (△), correspond to frames where both CNN and OperA predictions are accurate. On the other hand, low attention (LA) frames, indicated by downward triangles (▽), are situated where the CNN predictions are incorrect, demonstrating the effectiveness of the attention regularization in OperA.

Highest and Lowest Attention Frames Transformers and self-attention have the potential to offer model insights and explanation, which is an important aspect worth mentioning. While some studies have put forward that attention has limited ability to explain [62], this claim has been disputed [172], stressing that it is necessary for each research work to clarify their notion of explanation as it can be highly dependent on the specific task being addressed. Despite the ongoing discussions, we conducted an investigation into the attention of our model in order to gain further insights. In figure 9.4, we present a visual comparison of the highly attended (HA) and lowly attended (LA) frames for the models trained with and without attention regularization. Our examination reveals that LA frames are generally less informative for the corresponding surgical phase. However, as indicated by the blue boxes, the model without attention regularization exhibits minimal attention on frames that contain surgical tools that are distinctive of their respective phase. On the other hand, HA frames are more diverse and reflective of their surgical phase. As demonstrated by the blue boxes, for the model trained with attention regularization, surgical tools are present in all phases except "Preparation", emphasizing the strong connection between tools and surgical phases despite the attention

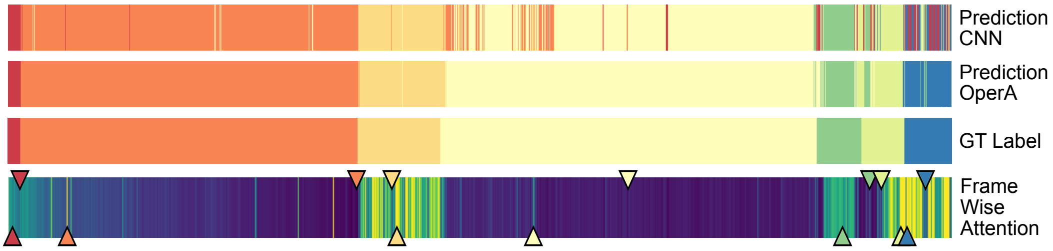


Fig. 9.3. Qualitative results of the predictions per phase for video 66 from Cholec80 using the feature extraction CNN and OperA compared to the ground truth labels. In the frame-wise attention, brighter (yellow) color corresponds to higher attention, darker (blue) color to lower attention. The position of the LA frames for each phase is denoted with ▽ and the HA frames with △. *Reproduced with permission from Springer Nature* [37].

model not being trained on tool information. The red boxes in figure 9.4 highlight the HA frames for the model without attention regularization. These frames are not indicative of their phase and appear quite similar to each other in the cases of "Cleaning Coagulation" and "Gallbladder Retraction".

In figure 9.5, we show two more example videos from the test set and their lowest and highest attention frame per phase with and without the regularization. In Video 77, interestingly, the LA frames for the "Calot Triangle Dissection" phase can both be seen as descriptive of their phase, and we see the same is observable for Video 80. One hypothesis is that most of the frames in this phase are descriptive, as the surgeon has to have direct sight of the coagulation tool during the dissection because otherwise, he could cut in the wrong structure or damage blood vessels. Therefore, the surgeon himself is very careful not to lose sight of the surgical scene, producing frames with overall higher value. For other phases, like the Gallbladder Packaging phase on both videos, the LA frames are darker and blurry, and there is not much information in the frames. For the HA of this phase for both videos, the Bag is visible. In this case, the results generated with and without regularization are very similar. This could mean that the model did not need further guidance using our proposed regularization, and the CNN image features were already descriptive enough. For other phases, "Clipping & Cutting" in Video 80, the use of regularization helped to identify more informative HA frames in which the clipper is clearly visible. These observations highlight the advantages of the proposed attention regularization and its potential for surgical video summarization or key frame detection. Further research into this direction is required using different analysis methods for model interpretation [74, 95, 154] and explaination [144, 167]. Further it would be interesting to design a quantitative metric to measure the information value of a frame to confirm the presented qualitative results. One intuitive direction would be using the tool or triplet annotations, assuming that informative frames include meaningful interactions between tool and tissue. For specific interventions and phases, this assumption could be misleading in cases were local occlusions of the lens with blood or heavy smoke are integral part of a surgical activity.

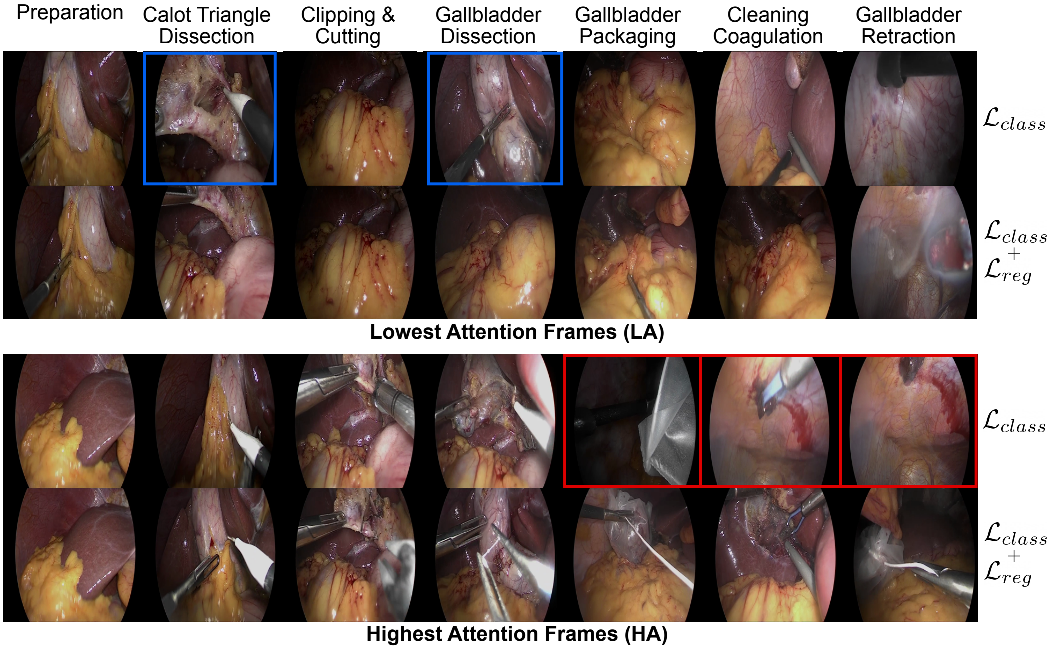

Preparation
Calot Triangle Dissection
Clipping & Cutting
Gallbladder Dissection
Gallbladder Packaging
Cleaning Coagulation
Gallbladder Retraction
$\mathcal{L}_{class}$
$\mathcal{L}_{class} + \mathcal{L}_{reg}$
Lowest Attention Frames (LA)
$\mathcal{L}_{class}$
$\mathcal{L}_{class} + \mathcal{L}_{reg}$
Highest Attention Frames (HA)

Fig. 9.4. Visualization of frames of video 66 of Cholec80 with highest (HA) and lowest (LA) attention per phase for the models with and without attention regularization. Blue and red boxes denote frames of the model without regularization that have low attention, while they are descriptive of their phase and high attention, while they are not. *Reproduced with permission from Springer Nature* [37].

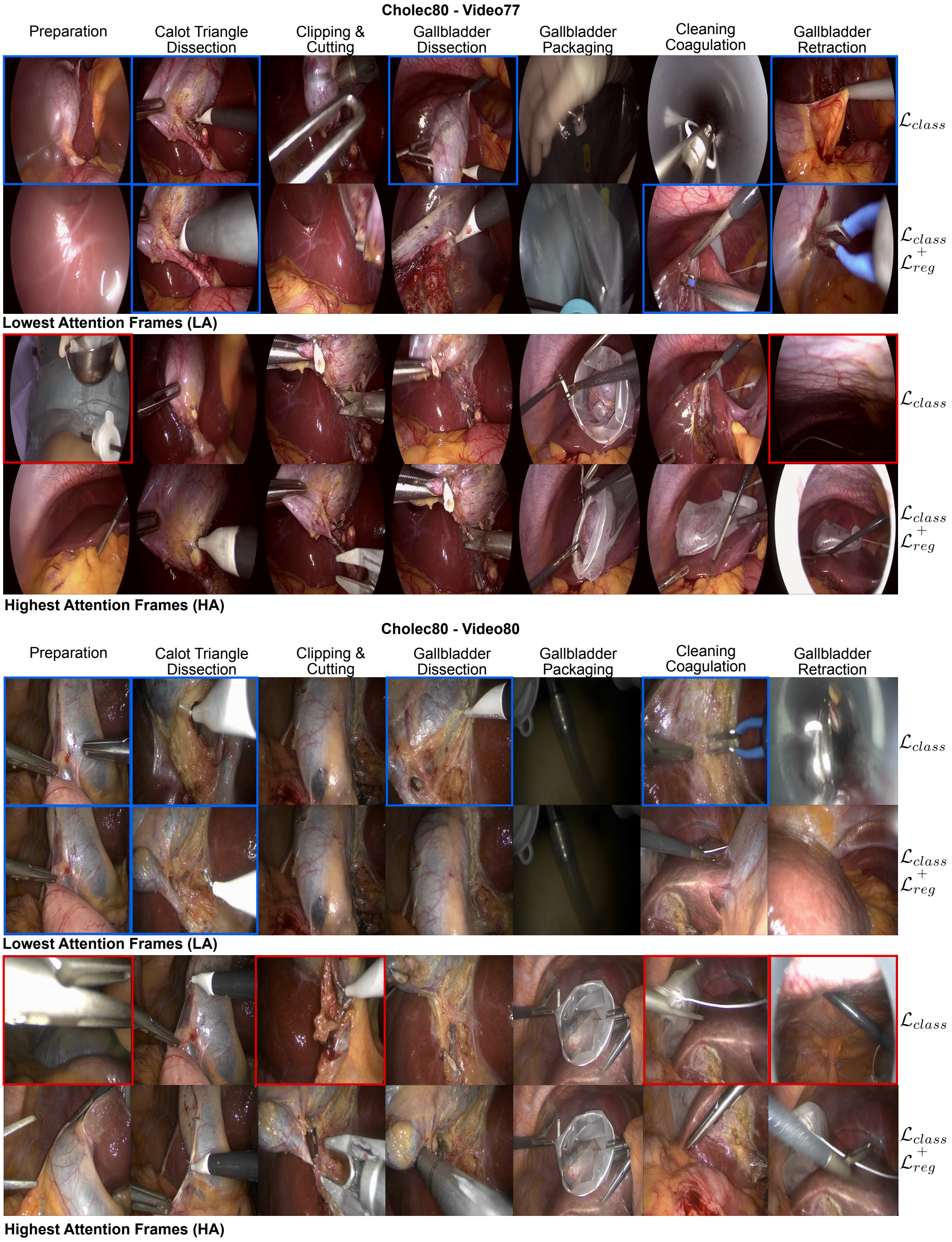


Fig. 9.5. Two more videos from Cholec80 are visualized with maximum and minimum attention per phase for the models with and without attention regularization. Blue and red boxes denote frames of the model without regularization that have low attention, while they are descriptive of their phase and high attention, while they are not. *Reproduced with permission from Springer Nature* [37].

10 Conclusion

In our method **TeCNO** (chapter 8) we present the a network for surgical phase recognition using Temporal Convolutions. By incooperating TCNs into a multi-stage network we showed an efficient approach to refine the predictions from stage to stage. For its online capabilities a shift in the temporal convolutions was implemented and TeCNO is capable of processing the visual feature of an entire video-sequence for an improved prediction compared to the frequently used LSTM baseline methods. One of the key features of temporal convolutions is the smoothness of their predictions one important aspect that are difficult to express in a single numerical value. Our work with TeCNO has shown that LSTMs can be effectively replaced with other temporal methods, such as TCNs, for the challenging task of surgical phase recognition. This breakthrough has inspired and encouraged other researchers to investigate new and innovative temporal approaches, pushing the field towards exciting new solutions.

In **OperA** (chapter 9) we challenge yet again the temporal architecture using transformers for the temporal refinement of the surgical phases. The attention mechanism of the transformer is build to model input specific weights and follows a more complex multi-layer architecture. The online capabilities of OperA are ensured with a causal masking of the attention weights and we used a novel attention regularization technique to guide our attention towards frames with descriptive visual cues. In order to gain deeper insights into the decision-making process of our model, we performed an evaluation of the attention for each frame. This analysis allowed us to identify the frames that received the most and least attention for each phase, which is an important first step towards model interpretation and trustworthy predictions.

Both works can be further improved and validated on larger and more diverse datasets with limited order in the phase succession. We think that both methods contribute to the scientific community and can serve as starting points for future directions. While TeCNO has advantages in efficiency and computational cost, OperA not only achieves improved accuracy, recall, and precision for surgical phase recognition but also provides a built-in mechanism to understand and investigate the decision-making process by observing the attention weights. This capability is crucial for interpreting the model's behavior and building trust in its predictions.

Future Directions

11

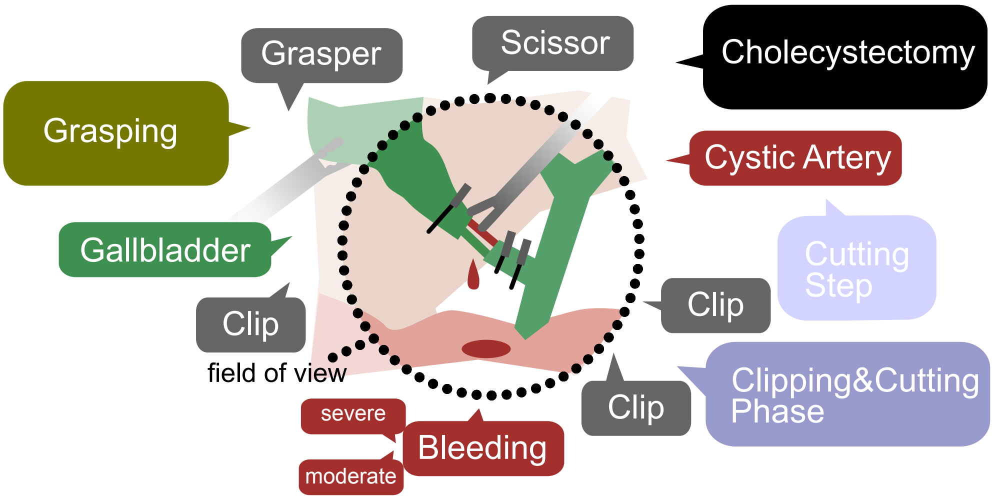


Fig. 11.1. The cutting step of a cholecystectomy surgery is depicted. The different colored boxes describe different kinds of information. Different design choices have to be made for a consistent annotation. The field of view can be restricted due to hardware limits(dotted circle).

11.1 Data Ambiguities

For this future work chapter, I first want to zoom out and again look at the task of visual laparoscopic phase recognition. In figure 11.1, the cutting step of a laparoscopic cholecystectomy surgery is visualized. When looking at this figure, several questions must be addressed before recording and annotating data or training algorithms. In this example, during the cutting, bleeding has occurred, but it is highly subjective to quantify the severity of the bleeding. Some surgeons might consider this *severe* or *moderate* depending on many factors, such as surgical training or prior experience. Additionally, the activity level has to be defined to summarize the event. The *Clipping Step*, *Clipping & Cutting Phase*, and *Cholecystectomy* all describe the setting but in different granularities, from fine to coarse. There is no definite answer to which granularity should be preferred, as it has to match the use case and possible applications. One approach would always be to select the finest granularity, but this also means more annotation effort. Additionally, three instances of the *Clip* are shown. For a binary tool or instrument detection, it is not trivial to deal with multiple instances. Still, this problem could be addressed by changing the binary label to a numeric one. The differentiation between different instances can be addressed using bounding box labels or semantic segmentation. Finally, the field of view of the laparoscope can be limited (dotted line in figure 11.1) and might not be large enough to envision all the interactions in a scene. Even if the *Grasper* is not visible in the camera scene, it can still grasp the Gallbladder, building up tension to safely cut the cystic artery. However, if the annotation of the data is done post-surgically and only based on the visual information, there is no easy way to be sure about the state of instruments outside the field of view. Here the combination of visual information with digital information observing the state of surgical instruments creates a valuable synergy.

There are various design considerations and challenges when creating the ground truth annotation for a single surgical frame. It is unrealistic to aim for perfect datasets, as we will always encounter obstacles, and addressing them would require an excessive amount of time and resources for annotation.

11.2 Addressing Data Challenges in Surgical Data Science

The future of surgical data science (SDS) and activity recognition in the operating room (OR) requires large amounts of labeled data for training, validation, and testing. However, due to the aforementioned challenges and data privacy concerns, it seems unrealistic to expect such large datasets to be available. Additionally, sharing information across institutions is still a significant obstacle that has yet to be resolved. Even if exhaustively annotated datasets for the OR exist, their sharing and exchange pose challenges.

Three possible approaches can be taken to address this problem. The first concept involves training algorithms in each hospital without sharing data and using semantic representations, such as surgical scene graphs, as an interface for secure and data-compliant transfer. Scene graphs do not contain identifying information about a person or medical staff but still encode the scene with high semantics on an excellent granularity.

Using the scene graph representation, downstream tasks such as phase recognition can be approached, but the algorithms would only consider the scene graph without the raw image. One challenge for this approach is that hospitals still need to acquire training data for image-to-scene graph generation. In other words, the hospital must develop a way to generate a scene graph representation of the OR environment from the raw image data. This can be done through a variety of computer vision techniques, such as object detection and semantic segmentation.

The second approach involves learning pipelines where a model is constantly refined and fine-tuned in the OR under an active learning regime. The surgeon could indicate if the AI is outputting wrong results, for example in the case of a new intervention or unknown tools, which would create an annotation that can be used for the next iteration of training. This approach could be coupled with federated learning, which has data privacy and efficiency advantages, as only the model weights have to be exchanged, which is far smaller. However, this approach requires that hospitals use compatible algorithms or setups to allow the flow of information, which seems unrealistic as even the flow within one hospital is unoptimized.

Lastly, the use of unstructured data from different sources could be utilized to build up contrastive learning approaches that aim to generate a common embedding from two unstructured data sources using models such as CLIP [131]. One direction is the use of surgical reports, educational instructions, or asking surgeons to explain the surgery during the interventions. After the translation from audio to text the text embedding and visual embedding can be aligned. This way large amounts of unlabeled and unstructured data sources can be used efficiently reducing the amount of annotated data necessary.

In summary, all three approaches are addressing the challenges of data privacy and sharing in the OR environment. The first approach relies on semantic representations and image-to-scene graph generation, while the second approach involves active learning and federated learning and the third approach utilizes large amounts of unstructured data in an unsupervised fashion. While each approach has its own challenges, they provide potential solutions for the problem of limited and difficult-to-share surgical data.

11.3 Advanced Approaches for Surgical Activity Understanding

As mentioned before scene graphs are an important aspect to consider when analyzing medical environments as they process the ability to express interactions on a very granular level for semantic understanding of a scene in a graphical and interpretable way. The integration of temporality has not been fully explored in current research on scene graphs. In the medical domain, integrating temporal consistency could be especially helpful for dealing with recurring activities of the same or similar type but with different intentions.

The definition of surgical activities can be ambiguous as it is of multi-level in nature. Instead of optimizing for just one arbitrary activity level, researchers should consider the joint prediction and analysis of different granularities. A promising research direction in this area is hierarchical activity recognition, which can provide fine-grained and high-level analysis of activities that build up on each other.

Another promising avenue for improving surgical activity analysis is to combine cameras for external analysis with a more detailed view, such as from a laparoscopic instrument, microscope, or bedside camera. This hierarchical camera setup allows for the acquisition of auxiliary information that can be used for self-supervised training, reducing the need for costly expert annotations. By combining information from different sensors or tools, it is possible to represent a patient or procedure holistically and with temporal refinement.

Lastly, in the context of surgical activity understanding, the issue of model reasoning and interpretation has not been given sufficient attention. The subjective nature of decision-making and the potential for disagreements between annotators highlight the importance of exploring new avenues to approach uncertainty. One possible strategy is to incorporate uncertainty modeling into the labeling process, which would enable the recognition of ambiguous or uncertain states rather than forcing them into binary labels. This approach would be in line with the practices of expert surgeons who themselves sometimes disagree on judgments regarding a scene or activity. The incorporation of uncertainty modeling would facilitate transparent decision-making and potentially enable the development of algorithms capable of providing trustworthy expert opinions, a prospect that could revolutionize the field.

In my opinion, the realization of the OR of the future (from 2020 [33] to 2030 [98] towards OR2040) can only be accomplished through sustained dedication and gradual advancements. While the research community in SDS is actively collaborating towards this objective, it is apparent that the combined efforts of academic research, industrial partners, and forward-thinking governments are necessary to achieve this vision. It is essential that these entities remain faithful and courageous and, most importantly, are willing to learn from their mistakes.

Part IV

Appendix

A

Authored and Co-authored Publications

Authored

1. **T. Czempiel**, M. Paschali, M. Keicher, W. Simson, H. Feussner, S.T. Kim, N. Navab. "*TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks.*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Lima, 2020

2. **T. Czempiel**, M. Paschali, D. Ostler, S.T. Kim, B. Busam, N. Navab. "*OperA: Attention-Regularized Transformers for Surgical Phase Recognition.*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

3. **T. Czempiel**, A. Sharghi, M. Paschali, N. Navab, and O. Mohareri. "*Surgical workflow recognition: From analysis of challenges to architectural study.*" Computer Vision – ECCV MCV, arXiv preprint arXiv: 2203.09230, 2022

4. **T. Czempiel**, C. Rogers, M. Keicher, M. Paschali, R. Braren, E. Burian, M. Makowski, N. Navab, T. Wendler, S. T. Kim. "*Longitudinal self-supervision for covid-19 pathology quantification*" arXiv preprint arXiv:2203.10804, 2022.

Co-authored

1. E. Özsoy, E. P. Örnek, U. Eck, **T. Czempiel**, F. Tombari, and N. Navab. "*4D-OR: Semantic Scene Graphs for OR Domain Modeling*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Singapore, 2022

2. L. Bastian, **T. Czempiel**, C. Heiliger, K. Karcz, U. Eck, B. Busam, N. Navab. "*Know your sensors — a modality study for surgical action classification.*", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2022

3. C. Nwoye, D. Alapatt, T. Yu, A.Vardazaryan, F.Xia, Z. Zhao, T. Xia, F. Jia, Y. Yang, H. Wang, D. Yu, G. Zheng, X. Duan, N. Getty, R. Sanchez-Matilla, M. Robu, L. Zhang, H. Chen, J. Wang, L. Wang, B. Zhang, B. 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. Vilaca, P. Morais, J. Fonseca, R. Mae Egging,

I. Nicole Wijma, C. Qian, G. 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, N. Padoy. "*Cholectriplet2021: A benchmark challenge for surgical action triplet recognition.*" arXiv preprint arXiv:2204.04746, 2022.

4. M. Keicher, K. Mullakaeva, **T. Czempiel**, K. Mach, A. Khakzar, N. Navab. "*Few-shot Structured Radiology Report Generation Using Natural Language Prompts.*" arXiv preprint arXiv:2203.15723, 2022.

5. M. Berlet, T. Vogel, D. Ostler, **T. Czempiel**, M. Kähler, S. Brunner, H. Feussner, D. Wilhelm, 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, 2022.

6. R. Hartwig, M. Berlet, **T. Czempiel**, J. Fuchtmann, T. Rückert, H. Feussner, D. Wilhelm. "*Image-based supportive measures for future application in surgery.*" Chirurgie (Heidelberg, Germany), 2022.

7. L. Bernhard, R. Krumpholz, Y. Krieger, **T. Czempiel**, A. Meining, N. Navab, T. Lüth, D. Wilhelm. "*Plafokon: A new concept for a patient-individual and intervention-specific flexible surgical platform.*" Surgical Endoscopy, 2022.

8. S.T. Kim, L. Goli, M. Paschali, A. Khakzar, M. Keicher, **T. Czempiel**, E. Burian, R. Braren, N. Navab, T. Wendler. "*Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs.*" International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021

9. D. M. Hedderich, M. Keicher, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, **T. Czempiel**, J. E. Spiro, D. P. dos Santos, D. Heim, C. Zimmer, D. Rückert, J. S. Kirschke, N. Navab. "*AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging.*" in Healthcare, Multidisciplinary Digital Publishing Institute, 2021

10. M. Keicher*, H. Burwinkel*, D. Bani-Harouni*, M. Paschali, **T. Czempiel**, E. Burian, M.R. Makowski, R. Braren, N. Navab, T. Wendler. "*U-GAT: Multimodal Graph Attention Network for COVID-19 Outcome Prediction.*" arXiv preprint arXiv:2108.00860, 2021 (Equal Contribution)

B Abstracts of Publications not Discussed in this Thesis

Surgical workflow recognition: From analysis of challenges to architectural study

T. Czempiel, A. Sharghi, M. Paschali, N. Navab, and O. Mohareri, Computer Vision – ECCV MCV, arXiv preprint arXiv: 2203.09230, 2022

Abstract. Algorithmic surgical workflow recognition is an ongoing research field and can be divided into laparoscopic (Internal) and operating room (External) analysis. So far, many different works for the internal analysis have been proposed with the combination of a frame-level and an additional temporal model to address the temporal ambiguities between different workflow phases. For the External recognition task, Clip-level methods are in the focus of researchers targeting the local ambiguities present in the operating room (OR) scene. In this work, we evaluate the performance of different combinations of common architectures for the task of surgical workflow recognition to provide a fair and comprehensive comparison of the methods for both settings, Internal and External. We show that the methods particularly designed for one setting can be transferred to the other mode and discuss the architecture effectiveness considering the main challenges for both Internal and External surgical workflow recognition.

Longitudinal self-supervision for covid-19 pathology quantification

T. Czempiel, C. Rogers, M. Keicher, M. Paschali, R. Braren, E. Burian, M. Makowski, N. Navab, T. Wendler, S. T. Kim. arXiv preprint arXiv:2203.10804, 2022

Quantifying COVID-19 infection over time is an important task to manage the hospitalization of patients during a global pandemic. Recently, deep learning-based approaches have been proposed to help radiologists automatically quantify COVID-19 pathologies on longitudinal CT scans. However, the learning process of deep learning methods demands extensive training data to learn the complex characteristics of infected regions over longitudinal scans. It is challenging to collect a large-scale dataset, especially for longitudinal training. In this study, we want to address this problem by proposing a new self-supervised learning method to effectively train longitudinal networks for the quantification of COVID-19 infections. For this purpose, longitudinal self-supervision schemes are explored on clinical longitudinal COVID-19 CT scans. Experimental results show that the proposed method is effective, helping the model better exploit the semantics of longitudinal data and improve two COVID-19 quantification tasks.

4D-OR: Semantic Scene Graphs for OR Domain Modeling

E. Özsoy, E. P. Örnek, U. Eck, **T. Czempiel**, F. Tombari, and N. Navab. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Singapore, 2022. *Reproduced with permission from Springer Nature*

Surgical procedures are conducted in highly complex operating rooms (OR), comprising different actors, devices, and interactions. To date, only medically trained human experts are capable of understanding all the links and interactions in such a demanding environment. This paper aims to bring the community one step closer to automated, holistic and semantic understanding and modeling of OR domain. Towards this goal, for the first time, we propose using semantic scene graphs (SSG) to describe and summarize the surgical scene. The nodes of the scene graphs represent different actors and objects in the room, such as medical staff, patients, and medical equipment, whereas edges are the relationships between them. To validate the possibilities of the proposed representation, we create the first publicly available 4D surgical SSG dataset, 4D-OR, containing ten simulated total knee replacement surgeries recorded with six RGB-D sensors in a realistic OR simulation center. 4D-OR includes 6734 frames and is richly annotated with SSGs, human and object poses, and clinical roles. We propose an end-to-end neural network-based SSG generation pipeline, with a rate of success of 0.75 macro F1, indeed being able to infer semantic reasoning in the OR. We further demonstrate the representation power of our scene graphs by using it for the problem of clinical role prediction, where we achieve 0.85 macro F1. The code and dataset will be made available upon acceptance.

Know your sensors — a modality study for surgical action classification

L. Bastian, **T. Czempiel**, C. Heiliger, K. Karcz, U. Eck, B. Busam, N. Nava. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2022.

The surgical operating room (OR) presents many opportunities for automation and optimization. Videos from various sources in the OR are becoming increasingly available. The medical community seeks to leverage this wealth of data to develop automated methods to advance interventional care, lower costs, and improve overall patient outcomes. Existing datasets from OR room cameras are thus far limited in size or modalities acquired, leaving it unclear which sensor modalities are best suited for tasks such as recognizing surgical action from videos. This study demonstrates that surgical action recognition performance can vary depending on the image modalities used. We perform a methodical analysis on several commonly available sensor modalities, presenting two fusion approaches that improve classification performance. The analyses are carried out on a set of multi-view RGB-D video recordings of 18 laparoscopic procedures.

Cholectriplet2021: A benchmark challenge for surgical action triplet recognition

C. Nwoye, D. Alapatt, T. Yu, A.Vardazaryan, F. Xia, Z. Zhao, T. Xia, F. Jia, Y. Yang, H. Wang, D. Yu, G. Zheng, X. Duan, N. Getty, R. Sanchez-Matilla, M. Robu, L. Zhang, H. Chen, J. Wang, L. Wang, B. Zhang, B. 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. Vilaca, P. Morais, J. Fonseca, R. M. Egging, I. Nicole Wijma, C. Qian, G. 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, N. Padoy. arXiv preprint arXiv:2204.04746, 2022

Context-aware decision support in the operating room can foster surgical safety and efficiency by leveraging real-time feedback from surgical workflow analysis. Most existing works recognize surgical activities at a coarse-grained level, such as phases, steps or events, leaving out fine-grained interaction details about the surgical activity; yet those are needed for more helpful AI assistance in the operating room. Recognizing surgical actions as triplets of <instrument, verb, target> combination delivers comprehensive details about the activities taking place in surgical videos. This paper presents CholecTriplet2021: an endoscopic vision challenge organized at MICCAI 2021 for the recognition of surgical action triplets in laparoscopic videos. The challenge granted private access to the large-scale CholecT50 dataset, which is annotated with action triplet information. In this paper, we present the challenge setup and assessment of the state-of-the-art deep learning methods proposed by the participants during the challenge. A total of 4 baseline methods from the challenge organizers and 19 new deep learning algorithms by competing teams are presented to recognize surgical action triplets

directly from surgical videos, achieving mean average precision (mAP) ranging from 4.2% to 38.1%. This study also analyzes the significance of the results obtained by the presented approaches, performs a thorough methodological comparison between them, in-depth result analysis, and proposes a novel ensemble method for enhanced recognition. Our analysis shows that surgical workflow analysis is not yet solved, and also highlights interesting directions for future research on fine-grained surgical activity recognition which is of utmost importance for the development of AI in surgery.

Few-shot Structured Radiology Report Generation Using Natural Language Prompts

M. Keicher, K. Mullakaeva, **T. Czempiel**, K. Mach, A. Khakzar, N. Navab. arXiv preprint arXiv:2203.15723, 2022

Chest radiograph reporting is time-consuming, and numerous solutions to automate this process have been proposed. Due to the complexity of medical information, the variety of writing styles, and free text being prone to typos and inconsistencies, the efficacy of quantifying the clinical accuracy of free-text reports using natural language processing measures is challenging. On the other hand, structured reports ensure consistency and can more easily be used as a quality assurance tool. To accomplish this, we present a strategy for predicting clinical observations and their anatomical location that is easily extensible to other structured findings. First, we train a contrastive language-image model using related chest radiographs and free-text radiological reports. Then, we create textual prompts for each structured finding and optimize a classifier for predicting clinical findings and their associations within the medical image. The results indicate that even when only a few image-level annotations are used for training, the method can localize pathologies in chest radiographs and generate structured reports.

Image-based supportive measures for future application in surgery

R. Hartwig, M. Berlet, **T. Czempiel**, J. Fuchtmann, T. Rückert, H. Feussner, D. Wilhelm. Chirurgie (Heidelberg, Germany), 2022 **Background**: The development of assistive technologies will become of increasing importance in the coming years and not only in surgery. The comprehensive perception of the actual situation is the basis of every autonomous action. Different sensor systems can be used for this purpose, of which video-based systems have a special potential. **Method**: Based on the available literature and on own research projects, central aspects of image-based support systems for surgery are presented. In this context, not only the potential but also the limitations of the methods are explained.

Results: An established application is the phase detection of surgical interventions, for which surgical videos are analyzed using neural networks. Through a time-based and transformative analysis the results of the prediction could only recently be significantly improved. Robotic camera guidance systems will also use image data to autonomously navigate laparoscopes in the near future. The reliability of the systems needs to be adapted to the high requirements in surgery by means of additional information. A comparable multimodal approach has already been implemented for navigation and localization during laparoscopic procedures. For this purpose, video data are analyzed using various methods and these data are fused with other sensor modalities. **Discussion**: Image-based supportive methods are already available for various tasks and will become an important aspect for the surgery of the future; however, in order to be able to be reliably implemented for autonomous functions, they must be embedded in multimodal approaches in the future in order to provide the necessary security.

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

M. Berlet, T. Vogel, D. Ostler, **T. Czempiel**, M. Kähler, S. Brunner, H. Feussner, D. Wilhelm, M. Kranzfelder. International Journal of Computer Assisted Radiology and Surgery, 2022

Purpose Surgical documentation is an important yet time-consuming necessity in clinical routine. Beside its core function to transmit information about a surgery to other medical professionals, the surgical report has gained even more significance in terms of information extraction for scientific, administrative and judicial application. A possible basis for computer aided reporting is phase detection by convolutional neural networks (CNN). In this article we propose a workflow to generate operative notes based on the output of the TeCNO CNN. **Methods** Video recordings of 15 cholecystectomies were used for inference. The annotation of TeCNO was compared to that of an expert surgeon (HE) and the algorithm based annotation of a scientist (HA). The CNN output then was used to identify aberrance from standard course as basis for the final report. Moreover, we assessed the phenomenon of 'phase flickering' as clusters of incorrectly labeled frames and evaluated its usability. **Results** The accordance of the HE and CNN was 79.7aberrant course with AUCs of 0.91 and 0.89 in ROC analysis regarding number and extend of concerned frames. Finally, we created operative notes based on a standard text, deviation alerts, and manual completion by the surgeon. **Conclusion** Computer-aided documentation is a noteworthy use case for phase recognition in standardized surgery. The analysis of phase flickering in a CNN's annotation has the potential of retrieving more information about the course of a particular procedure to complement an automated report.

Plafokon: A new concept for a patient-individual and intervention-specific flexible surgical platform

L. Bernhard, R. Krumpholz, Y. Krieger, **T. Czempiel**, A. Meining, N. Navab, T. Lüth, D. Wilhelm. Surgical Endoscopy, 2022

Background Research in the field of surgery is mainly driven by aiming for trauma reduction as well as for personalized treatment concepts. Beyond laparoscopy, other proposed approaches for further reduction of the therapeutic trauma have failed to achieve clinical translation, with few notable exceptions. We believe that this is mainly due to a lack of flexibility and high associated costs. We aimed at addressing these issues by developing a novel minimally invasive operating platform and a preoperative design workflow for patient-individual adaptation and cost-effective rapid manufacturing of surgical manipulators. In this article, we report on the first in-vitro cholecystectomy performed with our operating platform.

Methods The single-port overtube (SPOT) is a snake-like surgical manipulator for minimally invasive interventions. The system layout is highly flexible and can be adapted in design and dimensions for different kinds of surgery, based on patient- and disease-specific parameters.

For collecting and analyzing this data, we developed a graphical user interface, which assists clinicians during the preoperative planning phase. Other major components of our operating platform include an instrument management system and a non-sterile user interface. For the trial surgery, we used a validated phantom which was further equipped with a porcine liver including the gallbladder.

Results Following our envisioned preoperative design workflow, a suitable geometry of the surgical manipulator was determined for our trial surgery and rapidly manufactured by means of 3D printing. With this setup, we successfully performed a first in-vitro cholecystectomy, which was completed in 78 min.

Conclusions By conducting the trial surgery, we demonstrated the effectiveness of our PLAFOKON operating platform. While some aspects – especially regarding usability and ergonomics – can be further optimized, the overall performance of the system is highly promising, with sufficient flexibility and strength for conducting the necessary tissue manipulations.

Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs

S.T. Kim, L. Goli, M. Paschali, A. Khakzar, M. Keicher, **T. Czempiel**, E. Burian, R. Braren, N. Navab, T. Wendler. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Strasbourg, 2021. *Reproduced with permission from Springer Nature*

Chest computed tomography (CT) has played an essential diagnostic role in assessing patients with COVID-19 by showing disease-specific image features such as ground-glass opacity and consolidation. Image segmentation methods have proven to help quantify the disease and even help predict the outcome. The availability of longitudinal CT series may also result in an efficient and effective method to reliably assess the progression of COVID-19, monitor the healing process and the response to different therapeutic strategies. In this paper, we propose a new framework to identify infection at a voxel level (identification of healthy lung, consolidation, and ground-glass opacity) and visualize the progression of COVID-19 using sequential low-dose non-contrast CT scans. In particular, we devise a longitudinal segmentation network that utilizes the reference scan information to improve the performance of disease identification. Experimental results on a clinical longitudinal dataset collected in our institution show the effectiveness of the proposed method compared to the static deep neural networks for disease quantification.

AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging

D. M. Hedderich, M. Keicher, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, **T. Czempiel**, J. E. Spiro, D. P. dos Santos, D. Heim, C. Zimmer, D. Rückert, J. S. Kirschke, N. Navab.Healthcare, Multidisciplinary Digital Publishing Institute, 2021

Successful adoption of artificial intelligence (AI) in medical imaging requires medical professionals to understand underlying principles and techniques. However, educational offerings tailored to the need of medical professionals are scarce. To fill this gap, we created the course "AI for Doctors: Medical Imaging". An analysis of participants' opinions on AI and self-perceived skills rated on a five-point Likert scale was conducted before and after the course. The participants' attitude towards AI in medical imaging was very optimistic before and after the course. However, deeper knowledge of AI and the process for validating and deploying it resulted in significantly less overoptimism with respect to perceivable patient benefits through AI ($p = 0.020$). Self-assessed skill ratings significantly improved after the course, and the appreciation of the course content was very positive. However, we observed a substantial drop-out rate, mostly attributed to the lack of time of medical professionals. There is a high demand for educational offerings regarding AI in medical imaging among medical professionals, and better education may lead to a more realistic appreciation of clinical adoption. However, time constraints imposed by a busy clinical schedule need to be taken into account for successful education of medical professionals.

U-GAT: Multimodal Graph Attention Network for COVID-19 Outcome Prediction

M. Keicher*, H. Burwinkel*, D. Bani-Harouni*, M. Paschali, **T. Czempiel**, E. Burian, M.R. Makowski, R. Braren, N. Navab, T. Wendler. arXiv preprint arXiv:2108.00860, 2021 (Equal Contribution)

During the first wave of COVID-19, hospitals were overwhelmed with the high number of admitted patients. An accurate prediction of the most likely individual disease progression can improve the planning of limited resources and finding the optimal treatment for patients. However, when dealing with a newly emerging disease such as COVID-19, the impact of patient- and disease-specific factors (e.g. body weight or known co-morbidities) on the immediate course of disease is by and large unknown. In the case of COVID-19, the need for intensive care unit (ICU) admission of pneumonia patients is often determined only by acute indicators such as vital signs (e.g. breathing rate, blood oxygen levels), whereas statistical analysis and decision support systems that integrate all of the available data could enable an earlier prognosis. To this end, we propose a holistic graph-based approach combining both imaging and non-imaging information. Specifically, we introduce a multimodal similarity metric to build a population graph for clustering patients and an image-based end-to-end Graph Attention Network to process this graph and predict the COVID-19 patient outcomes: admission to ICU, need for ventilation and mortality. Additionally, the network segments chest CT images as an auxiliary task and extracts image features and radiomics for feature fusion with the available metadata. Results on a dataset collected in Klinikum rechts der Isar in Munich, Germany show that our approach outperforms single modality and non-graph baselines. Moreover, our clustering and graph attention allow for increased understanding of the patient relationships within the population graph and provide insight into the network's decision-making process.

Bibliography

[1] N. Ahmad, A. A. Hussein, L. Cavuoto, et al. “Ambulatory movements, team dynamics and interactions during robot-assisted surgery”. In: *BJU international* 118.1 (July 2016), pp. 132–139 (cit. on p. 7).

[2] S. A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner, and N. Navab. “Recovery of surgical workflow without explicit models”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4190 LNCS (2006), pp. 420–428 (cit. on pp. 19, 51).

[3] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy. “Deep neural networks predict remaining surgery duration from cholecystectomy videos”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10434 LNCS (2017), pp. 586–593 (cit. on p. 20).

[4] H. Al Hajj, M. Lamard, P. H. Conze, et al. “CATARACTS: Challenge on automatic tool annotation for cataRACT surgery”. In: *Medical Image Analysis* 52 (2019), pp. 24–41 (cit. on p. 56).

[5] E. M. Aleassa and K. M. El-Hayek. “Video atlas of intraoperative applications of near infrared fluorescence imaging”. In: (2020), p. 275 (cit. on p. 15).

[6] M. Allan, S. Kondo, S. Bodenstedt, et al. “2018 Robotic Scene Segmentation Challenge”. In: 14 (Jan. 2020), p. 16 (cit. on p. 19).

[7] M. Allan, A. Shvets, T. Kurmann, et al. “2017 Robotic Instrument Segmentation Challenge”. In: (Feb. 2019) (cit. on p. 19).

[8] J. L. Ba, J. R. Kiros, and G. E. Hinton. “Layer Normalization”. In: (2016) (cit. on p. 78).

[9] S. Bai, J. Z. Kolter, and V. Koltun. “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling”. In: *CoRR* abs/1803.01271 (2018) (cit. on p. 47).

[10] G. H. Ballantyne. “Robotic surgery, telerobotic surgery, telepresence, and telementoring: Review of early clinical results”. In: *Surgical Endoscopy and Other Interventional Techniques* 16.10 (Oct. 2002), pp. 1389–1402 (cit. on p. 10).

[11] L. Bastian, T. Czempiel, C. Heiliger, et al. “Know your sensors — a modality study for surgical action classification”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 0.0 (2022), pp. 1–9 (cit. on pp. 14, 20).

[12] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. “A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains”. In: *The Annals of Mathematical Statistics* 41.1 (1970), pp. 164–171 (cit. on p. 44).

[13] L. E. Baum and T. Petrie. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. In: *The Annals of Mathematical Statistics* 37.6 (Dec. 1966), pp. 1554–1563 (cit. on p. 43).

[14] H. Bay, T. Tuytelaars, L. Van Gool, A. Leonardis, H. Bischof, and A. Pinz. “SURF: Speeded up robust features”. In: *Lecture Notes in Computer Science* 3951 (Jan. 2006), pp. 404–417 (cit. on p. 36).

[15] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult”. In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166 (cit. on p. 45).

[16] T. Benson and G. Grieve. “HL7 Version 2”. In: *Principles of Health Interoperability: SNOMED CT, HL7 and FHIR*. Cham: Springer International Publishing, 2016, pp. 223–242 (cit. on p. 15).

[17] R. Bharathan, R. Aggarwal, and A. Darzi. “Operating room of the future”. In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 27.3 (June 2013), pp. 311–322 (cit. on p. 16).

[18] A. Bleakley, J. Bligh, and J. Browne. “Medical Education for the Future”. In: Advances in Medical Education 1 (2011) (cit. on pp. 7, 8).

[19] T. Blum, H. Feußner, and N. Navab. “Modeling and segmentation of surgical workflow from laparoscopic video”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6363 LNCS.PART 3 (2010), pp. 400–407 (cit. on p. 51).

[20] S. Bodenstedt, M. Wagner, L. Mündermann, et al. “Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data”. In: *International Journal of Computer Assisted Radiology and Surgery* 14.6 (2019), pp. 1089–1095 (cit. on pp. 24, 58).

[21] L. Breiman. “Random forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32 (cit. on pp. 37, 38).

[22] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984 (cit. on p. 37).

[23] T. B. Brown, B. Mann, N. Ryder, et al. “Language models are few-shot learners”. In: *arXiv* (2020) (cit. on p. 49).

[24] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. “High accuracy optical flow estimation based on a theory for warping”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3024 (2004), pp. 25–36 (cit. on p. 35).

[25] R. Brunelli. *Template matching techniques in computer vision: theory and practice*. 2009 (cit. on p. 35).

[26] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: (June 2020) (cit. on p. 55).

[27] M. Caron, H. Touvron, I. Misra, et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2021 (cit. on p. 55).

[28] V. Chan and A. Perlas. “Basics of ultrasound imaging”. In: *Atlas of Ultrasound-Guided Procedures in Interventional Pain Management* (2011), pp. 13–19 (cit. on p. 11).

[29] T. Chen and C. Guestrin. “XGBoost: A scalable tree boosting system”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu (2016), pp. 785–794 (cit. on p. 38).

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A simple framework for contrastive learning of visual representations”. In: *37th International Conference on Machine Learning, ICML 2020*. Vol. PartF168147-3. 2020 (cit. on p. 55).

[31] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014), pp. 103–111 (cit. on p. 45).

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *NIPS Deep Learning and Representation Learning Workshop* (Dec. 2014) (cit. on p. 45).

[33] K. Cleary, A. Kinsella, and S. K. Mun. "OR 2020 workshop report: Operating room of the future". In: *International Congress Series* 1281 (May 2005), pp. 832–838 (cit. on pp. 16, 92).

[34] K. Cleary and T. M. Peters. *Image-guided interventions: Technology review and clinical applications*. Aug. 2010 (cit. on p. 3).

[35] C. Cortes, V. Vapnik, and L. Saitta. "Support-vector networks". In: *Machine Learning 1995 20:3* 20.3 (Sept. 1995), pp. 273–297 (cit. on p. 36).

[36] T. Czempiel, M. Paschali, M. Keicher, et al. "TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by A. L. Martel, P. Abolmaesumi, D. Stoyanov, et al. Cham: Springer International Publishing, 2020, pp. 343–352 (cit. on pp. 70, 73, 74, 81).

[37] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab. "OperA: Attention-Regularized Transformers for Surgical Phase Recognition". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, et al. Cham: Springer International Publishing, 2021, pp. 604–614 (cit. on pp. 56, 78, 81, 83–86).

[38] T. Czempiel, A. Sharghi, M. Paschali, N. Navab, and O. Mohareri. "Surgical Workflow Recognition: from Analysis of Challenges to Architectural Study". In: *Computer Vision – ECCV 2022 – ECCV MCV* (2022) (cit. on pp. 12, 55).

[39] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005* I (2005), pp. 886–893 (cit. on p. 36).

[40] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1.Mlm (2019), pp. 4171–4186 (cit. on pp. 49, 77).

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (Oct. 2020) (cit. on pp. 49, 78).

[42] 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)*. IEEE, Dec. 2015 (cit. on pp. 70, 72, 78).

[43] Y. A. Farha and J. Gall. "MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2019-June. IEEE, June 2019, pp. 3570–3579 (cit. on pp. 48, 71).

[44] Y. A. Farha and J. Gall. "MS-TCN: Multi-stage temporal convolutional network for action segmentation". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (2019), pp. 3570–3579 (cit. on p. 48).

[45] M. A. Fischler and R. C. Bolles. "Random sample consensus". In: *Communications of the ACM* 24.6 (June 1981), pp. 381–395 (cit. on p. 36).

[46] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (Aug. 1997), pp. 119–139 (cit. on p. 37).

[47] I. Funke, S. T. Mees, J. Weitz, and S. Speidel. "Video-based surgical skill assessment using 3D convolutional neural networks". In: *International Journal of Computer Assisted Radiology and Surgery* 14.7 (2019), pp. 1217–1225 (cit. on p. 59).

[48] 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: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12904 LNCS (2021), pp. 593–603 (cit. on p. 54).

[49] C. R. Garrow, K.-F. Kowalewski, L. Li, et al. “Machine Learning for Surgical Phase Recognition: A Systematic Review.” In: *Annals of surgery* 273.4 (2021), pp. 684–693 (cit. on pp. 20, 24, 53, 55, 56).

[50] I. Goodfellow, Y. Bengio, and A. Courville. “Deep Learning”. In: *MIT Press* (2016) (cit. on pp. 38, 39).

[51] I. Goodfellow, Y. Bengio, and A. Courville. “Deep Learning”. In: *MIT Press* (2016) (cit. on p. 40).

[52] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. “LSTM: A Search Space Odyssey”. In: *IEEE Transactions on Neural Networks and Learning Systems* 28.10 (Oct. 2017), pp. 2222–2232 (cit. on p. 53).

[53] V. N. Gudivada, D. Rao, and V. V. Raghavan. “Chapter 9 - Big Data Driven Natural Language Processing Research and Applications”. In: *Big Data Analytics*. Ed. by V. Govindaraju, V. V. Raghavan, and C. R. Rao. Vol. 33. Handbook of Statistics. Elsevier, 2015, pp. 203–238 (cit. on p. 43).

[54] S. Hansen and A. James Baroody. “Electronic Health Records and the Logics of Care: Complementarity and Conflict in the U.S. Healthcare System”. In: *https://doi.org/10.1287/isre.2019.0875* 31.1 (Aug. 2019), pp. 57–75 (cit. on pp. 6, 10).

[55] J. T. Hathcock and R. L. Stickle. “Principles and concepts of computed tomography”. In: *The Veterinary clinics of North America. Small animal practice* 23.2 (1993), pp. 399–415 (cit. on p. 11).

[56] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Nov. 2019), pp. 9726–9735 (cit. on p. 55).

[57] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 770–778 (cit. on pp. 39, 53, 54, 70, 72, 77, 78).

[58] K. He, C. Gan, Z. Li, et al. “Transformers in Medical Image Analysis: A Review”. In: (Feb. 2022) (cit. on p. 54).

[59] L. Heo and M. Feig. “High-accuracy protein structures by combining machine-learning with physics-based refinement.” eng. In: *Proteins* 88.5 (May 2020), pp. 637–642 (cit. on p. 49).

[60] S. Hochreiter and J. Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 44–46).

[61] A. Huaulmé, P. Jannin, F. Reche, J. L. Faucheron, A. Moreau-Gaudry, and S. Voros. “Offline identification of surgical deviations in laparoscopic rectopexy”. In: *Artificial Intelligence in Medicine* 104.May 2019 (2020) (cit. on pp. 24, 25).

[62] S. Jain and B. C. Wallace. “Attention is not Explanation”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, {NAACL-HLT} 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*. Association for Computational Linguistics, 2019, pp. 3543–3556 (cit. on pp. 49, 83).

[63] P. Jannin and X. Morandi. “Surgical models for computer-assisted neurosurgery”. In: *NeuroImage* 37.3 (Sept. 2007), pp. 783–791 (cit. on p. 52).

[64] P. Jannin, M. Raimbault, X. Morandi, and B. Gibaud. “Modeling surgical procedures for multimodal image-guided neurosurgery”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 2208. 2001 (cit. on p. 52).

[65] Y. Jin, Q. Dou, H. Chen, et al. “SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network”. In: *IEEE Transactions on Medical Imaging* 37.5 (2018), pp. 1114–1126 (cit. on pp. 23, 53, 56, 74, 81, 82).

[66] Y. Jin, H. Li, Q. Dou, et al. “Multi-task recurrent convolutional network with correlation loss for surgical video analysis”. In: *Medical Image Analysis* 59 (2020) (cit. on pp. 53, 56, 72, 74, 81, 82).

[67] Y. Jin, Y. Yu, C. Chen, Z. Zhao, P. A. Heng, and D. Stoyanov. “Exploring Intra- and Inter-Video Relation for Surgical Semantic Scene Segmentation”. In: *IEEE Transactions on Medical Imaging* 41.11 (Nov. 2022), pp. 2991–3002 (cit. on p. 54).

[68] A. Kadkhodamohammadi, I. Luengo, and D. Stoyanov. “PATG: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos”. In: *International Journal of Computer Assisted Radiology and Surgery* 17.5 (May 2022), pp. 849–856 (cit. on p. 54).

[69] S. Kannan, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy. “Future-State Predicting LSTM for Early Surgery Type Recognition”. In: *IEEE Transactions on Medical Imaging* 39.3 (Mar. 2020), pp. 556–566 (cit. on p. 20).

[70] R. M. Karp. “On-line algorithms versus off-line algorithms: How much is it worth to know the future?” In: *Mathematical Social Sciences* 25.3 (1993) (cit. on p. 34).

[71] H. Kassem, D. Alapatt, P. Mascagni, A. Karargyris, and N. Padoy. “Federated Cycling (FedCy): Semi-supervised Federated Learning of Surgical Phases”. In: *IEEE transactions on medical imaging* PP (2022) (cit. on p. 55).

[72] D. Katić, A. L. Wekerle, F. Gärtner, et al. “Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8498 LNCS (2014), pp. 158–167 (cit. on p. 20).

[73] M. Keicher, K. Mullakaeva, T. Czempiel, K. Mach, A. Khakzar, and N. Navab. “Few-shot Structured Radiology Report Generation Using Natural Language Prompts”. In: (Mar. 2022) (cit. on p. 24).

[74] A. Khakzar, S. Musatian, J. Buchberger, et al. “Towards Semantic Interpretation of Thoracic Disease and COVID-19 Diagnosis Models”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12903 LNCS (2021), pp. 499–508 (cit. on p. 84).

[75] Y. Kim, C. Denton, L. Hoang, and A. M. Rush. “Structured attention networks”. In: *International Conference on Learning Representations* (2017), pp. 1–21 (cit. on p. 78).

[76] K. Kirtac, N. Aydin, J. L. Lavanchy, et al. “Surgical Phase Recognition: From Public Datasets to Real-World Data”. In: *Applied Sciences (Switzerland)* 12.17 (Sept. 2022), p. 8746 (cit. on p. 55).

[77] U. Klank, N. Padoy, H. Feussner, and N. Navab. “Automatic feature generation in endoscopic images”. In: *International Journal of Computer Assisted Radiology and Surgery* 3.3 (2008), pp. 331–339 (cit. on pp. 51, 59).

[78] S. Kondo. “LapFormer: surgical tool detection in laparoscopic surgical video using transformer architecture”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* (2020) (cit. on p. 49).

[79] M. Köny, J. Benzko, M. Czaplik, et al. “The Smart Operating Room: smartOR”. In: *Distributed Networks* (Dec. 2014), pp. 291–315 (cit. on p. 15).

[80] M. Kranzfelder, A. Schneider, A. Fiolka, et al. “Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology”. In: *The Journal of surgical research* 185.2 (2013), pp. 704–710 (cit. on p. 12).

[81] A. Krizhevsky, I. Sutskever, and G. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Neural Information Processing Systems* 25 (2012) (cit. on pp. 41, 52, 72).

[82] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin. “A framework for the recognition of high-level surgical tasks from video images for cataract surgeries”. In: *Ieee Transactions on Bio-Medical Engineering* 59.4 (Apr. 2012), p. 966 (cit. on p. 52).

[83] F. Lalys and P. Jannin. “Surgical process modelling: A review”. In: *International Journal of Computer Assisted Radiology and Surgery* 9.3 (Sept. 2014), pp. 495–511 (cit. on pp. 4, 21).

[84] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin. “Surgical phases detection from microscope videos by combining SVM and HMM”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6533 LNCS (2011), pp. 54–62 (cit. on p. 51).

[85] J. L. Lavanchy, J. Zindel, K. Kirtac, et al. “Automation of surgical skill assessment using a three-stage machine learning algorithm”. In: *Scientific Reports 2021 11:1* 11.1 (Mar. 2021), pp. 1–9 (cit. on p. 20).

[86] C. Lea, J. H. Choi, A. Reiter, and G. D. Hager. “Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World”. In: *M2CAI - Satellite Workshop of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2016) (cit. on p. 53).

[87] C. Lea, J. Facker, G. Hager, R. Taylor, and S. Saria. “3D Sensing Algorithms Towards Building an Intelligent Intensive Care Unit”. In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2013 (2013), pp. 136–40 (cit. on p. 52).

[88] 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 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9915 LNCS. 2016, pp. 47–54 (cit. on p. 47).

[89] Y. LeCun. *Generalization and network design strategies*. Tech. rep. University of Toronto, 1989 (cit. on p. 39).

[90] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2323 (cit. on p. 39).

[91] Y. LeCun and C. Cortes. “MNIST handwritten digit database”. In: *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* 7 (2010) (cit. on p. 39).

[92] G. Lecuyer, M. Ragot, N. Martin, L. Launay, and P. Jannin. “Assisted phase and step annotation for surgical videos”. In: *International Journal of Computer Assisted Radiology and Surgery* (2020) (cit. on p. 58).

[93] S. Li, Y. A. Farha, Y. Liu, M.-M. Cheng, and J. Gall. “MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (June 2020), pp. 3570–3579 (cit. on p. 48).

[94] D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110 (cit. on p. 35).

[95] S. M. Lundberg and S. I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems* 2017-December (May 2017), pp. 4766–4775 (cit. on p. 84).

[96] L. Maier-Hein, M. Eisenmann, C. Feldmann, et al. “Surgical Data Science: A Consensus Perspective”. In: (2018), pp. 1–29 (cit. on pp. 6, 22).

[97] L. Maier-Hein, M. Eisenmann, D. Sarikaya, et al. *Surgical data science – from concepts toward clinical translation*. Vol. 76. 2022, pp. 0–3 (cit. on p. 15).

[98] L. Maier-Hein, S. S. Vedula, S. Speidel, et al. “Surgical data science for next-generation interventions”. In: *Nature Biomedical Engineering* 1.9 (2017), pp. 691–696 (cit. on pp. 6, 16, 22, 92).

[99] P. Mascagni and N. Padoy. “OR black box and surgical control tower: Recording and streaming data and analytics to improve surgical care”. In: *Journal of Visceral Surgery* 158.3 (June 2021), S18–S25 (cit. on p. 16).

[100] P. Mascagni, D. Alapatt, L. Sestini, et al. “Computer vision in surgery: from potential to clinical value”. In: *npj Digital Medicine* 5.1 (Oct. 2022), p. 163 (cit. on p. 22).

[101] P. Mascagni, D. Alapatt, T. Urade, et al. “A Computer Vision Platform to Automatically Locate Critical Events in Surgical Videos: Documenting Safety in Laparoscopic Cholecystectomy”. In: *Annals of surgery* 274.1 (July 2021), e93–e95 (cit. on p. 20).

[102] P. Mascagni, C. Fiorillo, T. Urade, et al. “Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety”. In: *Surgical Endoscopy* 34.6 (June 2020), pp. 2709–2714 (cit. on p. 20).

[103] P. Mascagni, A. Vardazaryan, D. Alapatt, et al. “Artificial Intelligence for Surgical Safety: Automatic Assessment of the Critical View of Safety in Laparoscopic Cholecystectomy Using Deep Learning”. In: *Annals of surgery* 275.5 (2022) (cit. on p. 20).

[104] R. K. Mishra. “Textbook of laparoscopy for surgeons and gynecologists”. In: (2021) (cit. on pp. 14, 15).

[105] V. Nair and G. E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Internation Conference on Machine Learning* (2010) (cit. on p. 38).

[106] N. Navab, A. Martin-Gomez, M. Seibold, et al. “Medical Augmented Reality: Definition, Principle Components, Domain Modeling, and Design-Development-Validation Process”. In: *Journal of Imaging 2023, Vol. 9, Page 4* 9.1 (Dec. 2022), p. 4 (cit. on pp. 13, 16).

[107] T. Neumuth, N. Durstewitz, M. Fischer, et al. “Structured recording of intraoperative surgical workflows”. In: *Medical Imaging 2006: PACS and Imaging Informatics* 6145 (Mar. 2006), 61450A (cit. on p. 52).

[108] T. Neumuth, P. Jannin, G. Strauss, J. Meixensberger, and O. Burgert. “Validation of Knowledge Acquisition for Surgical Process Models”. In: *Journal of the American Medical Informatics Association : JAMIA* 16.1 (Jan. 2009), p. 72 (cit. on p. 52).

[109] T. Neumuth, G. Strauß, J. Meixensberger, H. U. Lemke, and O. Burgert. “Acquisition of process descriptions from surgical interventions”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4080 LNCS (2006), pp. 602–611 (cit. on p. 20).

[110] A. Newell, K. Yang, and J. Deng. *Stacked Hourglass Networks for Human Pose Estimation*. Tech. rep. (cit. on p. 71).

[111] M. Nguyen, T. T. Bui, Q. Van Nguyen, T. T. Nguyen, and T. Van Pham. “LAPFormer: A Light and Accurate Polyp Segmentation Transformer”. In: (Oct. 2022) (cit. on p. 54).

[112] C. I. Nwoye, C. Gonzalez, T. Yu, et al. “Recognition of Instrument-Tissue Interactions in Endoscopic Videos via Action Triplets”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12263 LNCS (2020), pp. 364–374 (cit. on p. 20).

[113] C. I. Nwoye and N. Padoy. “Data Splits and Metrics for Method Benchmarking on Surgical Action Triplet Datasets”. In: (2022), pp. 1–11 (cit. on p. 56).

[114] C. I. Nwoye, T. Yu, C. Gonzalez, et al. “Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos”. In: *Medical Image Analysis* 78 (May 2022), p. 102433 (cit. on p. 20).

[115] J. Okamoto, K. Masamune, H. Iseki, and Y. Muragaki. “Development concepts of a Smart Cyber Operating Theater (SCOT) using ORiN technology”. In: *Biomedizinische Technik* 63.1 (Feb. 2018), pp. 31–37 (cit. on p. 15).

[116] D. Ostler, M. Kranzfelder, R. Stauder, D. Wilhelm, H. Feussner, and A. Schneider. “A centralized data acquisition framework for operating theatres”. In: *2015 17th International Conference on E-Health Networking, Application and Services, HealthCom 2015*. Institute of Electrical and Electronics Engineers Inc., 2015, pp. 1–5 (cit. on p. 12).

[117] D. Ostler, M. Seibold, J. Fuchtmann, et al. “Acoustic signal analysis of instrument–tissue interaction for minimally invasive interventions”. In: *International Journal of Computer Assisted Radiology and Surgery* 15.5 (May 2020), pp. 771–779 (cit. on p. 12).

[118] E. Özsoy, E. P. Örnek, U. Eck, F. Tombari, and N. Navab. “Multimodal Semantic Scene Graphs for Holistic Modeling of Surgical Procedures”. In: *CoRR* (2021) (cit. on p. 20).

[119] E. Özsoy, E. P. Örnek, U. Eck, T. Czempiel, F. Tombari, and N. Navab. “4D-OR: Semantic Scene Graphs for OR Domain Modeling”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13437 LNCS (2022), pp. 475–485 (cit. on pp. 14, 20).

[120] 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 ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 10.Pt 1 (2007), pp. 102–109 (cit. on p. 51).

[121] N. Padoy, T. Blum, H. Feussner, M. O. Berger, and N. Navab. “On-line recognition of surgical activity for monitoring in the operating room”. In: *Proceedings of the National Conference on Artificial Intelligence* 3 (2008), pp. 1718–1724 (cit. on p. 51).

[122] N. Padoy, D. Mateus, D. Weinland, M. O. Berger, and N. Navab. “Workflow monitoring based on 3D motion features”. In: *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009* (2009), pp. 585–592 (cit. on p. 52).

[123] N. Padoy. “Machine and deep learning for workflow recognition during surgery”. In: *Minimally Invasive Therapy and Allied Technologies* 28.2 (2019), pp. 82–90 (cit. on p. 4).

[124] N. Padoy. “Machine and deep learning for workflow recognition during surgery”. In: *Minimally Invasive Therapy and Allied Technologies* 28.2 (2019), pp. 82–90 (cit. on p. 24).

[125] N. Padoy, T. Blum, S. A. Ahmadi, H. Feussner, M. O. Berger, and N. Navab. “Statistical modeling and recognition of surgical workflow”. In: *Medical Image Analysis* 16.3 (2012), pp. 632–641 (cit. on pp. 52, 60, 65, 82).

[126] R. Pascanu, T. Mikolov, and Y. Bengio. *On the difficulty of training recurrent neural networks*. May 2013 (cit. on p. 53).

[127] P. Petrone, M. Niola, P. Di Lorenzo, et al. “Early medical skull surgery for treatment of post-traumatic osteomyelitis 5,000 years ago”. In: *PLoS ONE* 10.5 (May 2015) (cit. on p. 3).

[128] R. A. Powsner, M. R. Palmer, and E. R. Powsner. “Essentials of Nuclear Medicine Physics and Instrumentation”. In: (2013) (cit. on p. 11).

[129] G. Quellec, K. Charrière, M. Lamard, et al. “Real-time recognition of surgical tasks in eye surgery videos”. In: *Medical image analysis* 18.3 (2014), pp. 579–590 (cit. on pp. 23, 52).

[130] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286 (cit. on pp. 43, 44).

[131] A. Radford, J. W. Kim, C. Hallacy, et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: (Feb. 2021) (cit. on p. 91).

[132] A. Ramesh, M. Pavlov, G. Goh, et al. *Zero-Shot Text-to-Image Generation*. 2021 (cit. on p. 49).

[133] S. Ramesh, V. Srivastav, D. Alapatt, et al. “Dissecting Self-Supervised Learning Methods for Surgical Computer Vision”. In: (July 2022), p. 1 (cit. on p. 55).

[134] R. Al-Rfou, D. Choe, N. Constant, M. Guo, and L. Jones. “Character-Level Language Modeling with Deeper Self-Attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019) (cit. on p. 79).

[135] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. B. Schön. “Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness”. In: (2019) (cit. on p. 53).

[136] N. Rieke, D. J. Tan, M. Alsheakhali, et al. “Surgical tool tracking and pose estimation in retinal microsurgery”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9349 (2015), pp. 266–273 (cit. on p. 19).

[137] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1986 (cit. on p. 39).

[138] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow. “Tool and Phase recognition using contextual CNN features”. In: (Oct. 2016) (cit. on p. 52).

[139] H. Sakoe and S. Chiba. “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (1978), pp. 43–49 (cit. on p. 41).

[140] R. R. Schaller. “Moore’s law: past, present, and future”. In: *IEEE Spectrum* 34.6 (June 1997), pp. 52–55 (cit. on p. 34).

[141] A. Schmidt, A. Sharghi, H. Haugerud, D. Oh, and O. Mohareri. *Multi-view Surgical Video Action Detection via Mixed Global View Attention*. Vol. 12904 LNCS. Springer International Publishing, 2021, pp. 626–635 (cit. on p. 20).

[142] A. Schneider and H. Feussner. “Chapter 13 - Training and Simulation”. In: *Biomedical Engineering in Gastrointestinal Surgery* (2017). Ed. by A. Schneider and H. Feussner, pp. 491–512 (cit. on p. 7).

[143] M. Schuster and K. K. Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681 (cit. on p. 45).

[144] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *International Journal of Computer Vision* 128.2 (Oct. 2016), pp. 336–359 (cit. on p. 84).

[145] A. Sharghi, H. Haugerud, D. Oh, and O. Mohareri. “Automatic Operating Room Surgical Activity Recognition for Robot-Assisted Surgery”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12263 LNCS (2020), pp. 385–395 (cit. on pp. 20, 53, 56).

[146] S. Speidel, J. Benzko, S. Krappe, et al. “Automatic classification of minimally invasive instruments based on endoscopic image sequences”. In: *Medical Imaging* 7261 (Mar. 2009), pp. 106–113 (cit. on p. 19).

[147] S. Speidel, M. Delles, C. Gutt, and R. Dillmann. “Tracking of instruments in minimally invasive surgery for surgical skill analysis”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4091 LNCS (2006), pp. 148–155 (cit. on pp. 19, 23).

[148] S. Speidel, G. Sudra, J. Senemaud, et al. “Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling”. In: *https://doi.org/10.1117/12.770385* 6918 (Mar. 2008), pp. 326–333 (cit. on p. 23).

[149] V. Srivastav, T. Issenhuth, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi, and N. Padoy. “MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation”. In: *MICCAI-LABELS* (Aug. 2018) (cit. on p. 20).

[150] R. Stauder. *Context Awareness for the Operating Room of the Future*. Tech. rep. 2018 (cit. on pp. 16, 21).

[151] R. Stauder, A. Okur, L. Peter, et al. “Random forests for phase detection in surgical workflow analysis”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8498 LNCS (2014), pp. 148–157 (cit. on p. 52).

[152] R. Stauder, D. Ostler, M. Kranzfelder, S. Koller, H. Feußner, and N. Navab. “The TUM LapChole dataset for the M2CAI 2016 workflow challenge”. In: (Oct. 2016) (cit. on p. 55).

[153] R. Stauder, D. Ostler, T. Vogel, et al. “Surgical data processing for smart intraoperative assistance systems”. In: *Innovative Surgical Sciences* 2.3 (Sept. 2017), pp. 145–152 (cit. on p. 52).

[154] M. Sundararajan, A. Taly, and Q. Yan. “Axiomatic Attribution for Deep Networks”. In: *34th International Conference on Machine Learning, ICML 2017* 7 (Mar. 2017), pp. 5109–5118 (cit. on p. 84).

[155] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario. “Medical Robotics and Computer-Integrated Surgery”. In: *Springer Handbooks* (2016), pp. 1657–1684 (cit. on p. 10).

[156] *The Integrated Operating Room - HCD Magazine* (cit. on p. 10).

[157] L. L. Thurston. “The History of Surgery”. In: *Introduction to Surgery for Students*. Ed. by F. R. A., K. Ahmed, and P. Dasgupta. Cham: Springer International Publishing, 2017, pp. 1–15 (cit. on p. 3).

[158] A. P. Twinanda, J. Marescaux, M. De Mathelin, and N. Padoy. *Single-and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016*. Tech. rep. (cit. on p. 75).

[159] 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”. In: (2016), pp. 1–7 (cit. on pp. 52, 53, 56, 72).

[160] 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”. In: *IEEE Transactions on Medical Imaging* 36.1 (2017), pp. 86–97 (cit. on pp. 53, 55, 56, 60–62, 72, 74).

[161] A. P. Twinanda, N. Padoy, M. J. Troccaz, and G. Hager. “Vision-based Approaches for Surgical Activity Recognition Using Laparoscopic and RBGD Videos”. In: *Thesis* Umr 7357 (2017) (cit. on pp. 21, 53, 65, 74, 81).

[162] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy. “RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations”. In: *IEEE transactions on medical imaging* 38.4 (Apr. 2019), pp. 1069–1078 (cit. on p. 20).

[163] N. Valderrama, P. Ruiz Puentes, I. Hernández, et al. “Towards Holistic Surgical Scene Understanding”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13437 LNCS (2022), pp. 442–452 (cit. on p. 54).

[164] A. Van Den Oord, S. Dieleman, H. Zen, et al. *WAVENET: A GENERATIVE MODEL FOR RAW AUDIO*. Tech. rep. (cit. on pp. 47, 48, 71).

[165] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* 2017-Decem.Nips (2017), pp. 5999–6009 (cit. on pp. 48, 49, 77, 78, 82).

[166] V. Velanovich. “Laparoscopic vs open surgery: A preliminary comparison of quality-of- life outcomes”. In: *Surgical Endoscopy* 14.1 (Jan. 2000), pp. 16–21 (cit. on p. 3).

[167] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Medical Image Analysis* 79 (July 2022), p. 102470 (cit. on p. 84).

[168] M. T. Vlaardingerbroek and J. A. den Boer. “Magnetic Resonance Imaging”. In: (1996) (cit. on p. 11).

[169] V. Voigt, R. Rossaint, and M. Czaplik. “Der vernetzte Operationssaal”. In: *Telemedizin: Grundlagen und praktische Anwendung in stationaeren und ambulanten Einrichtungen*. Ed. by G. Marx, R. Rossaint, and N. Marx. Berlin, Heidelberg: Springer Berlin Heidelberg, 2021, pp. 437–442 (cit. on p. 15).

[170] S. Voros and G. D. Hager. “Towards "real-time" tool-tissue interaction detection in robotically assisted laparoscopy”. In: *Proceedings of the 2nd Biennial IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, BioRob 2008* (2008), pp. 562–567 (cit. on p. 52).

[171] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, et al. “Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark”. In: (Sept. 2021) (cit. on p. 56).

[172] S. Wiegreffe and Y. Pinter. “Attention is not not explanation”. In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2020), pp. 11–20 (cit. on pp. 49, 83).

[173] Y. Xiao, P. Hu, H. Hu, et al. “An algorithm for processing vital sign monitoring data to remotely identify operating room occupancy in real-time”. In: *Anesthesia and Analgesia* 101.3 (2005), pp. 823–829 (cit. on p. 11).

[174] 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”. In: (2018) (cit. on pp. 53, 72).

[175] F. Yi, Y. Yang, and T. Jiang. *Not End-to-End: Explore Multi-Stage Architecture for Online Surgical Phase Recognition*. 2022 (cit. on p. 53).

[176] 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”. In: (2018) (cit. on p. 20).

[177] K. Yuan, M. Holden, S. Gao, and W. Lee. “Anticipation for surgical workflow through instrument interaction and recognized Signals”. In: *Medical Image Analysis* 82 (Nov. 2022), p. 102611 (cit. on p. 54).

[178] L. Zappella, B. Béjar, G. Hager, and R. Vidal. “Surgical gesture classification from video and kinematic data”. In: *Medical Image Analysis* 17.7 (Oct. 2013), pp. 732–745 (cit. on p. 52).

[179] Y. Zhang, S. Bano, A. S. Page, J. Deprest, D. Stoyanov, and F. Vasconcelos. “Large-scale surgical workflow segmentation for laparoscopic sacrocolpopexy”. In: *International Journal of Computer Assisted Radiology and Surgery* 17.3 (Mar. 2022), pp. 467–477 (cit. on p. 55).

[180] Y. Zhang, S. Bano, A. S. Page, J. Deprest, D. Stoyanov, and F. Vasconcelos. “Retrieval of Surgical Phase Transitions Using Reinforcement Learning”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13437 LNCS (2022), pp. 497–506 (cit. on p. 54).

[181] Y. T. Zhou and R. Chellappa. “Computation of optical flow using a neural network”. In: *IEEE 1988 International Conference on Neural Networks* (1988), pp. 71–78 (cit. on p. 39).

[182] X. Zou, W. Liu, J. Wang, R. Tao, and G. Zheng. “ARST: auto-regressive surgical transformer for phase recognition from laparoscopic videos”. In: *https://doi.org/10.1080/21681163.2022.2145238* (2022) (cit. on p. 55).

List of Figures

List of Tables