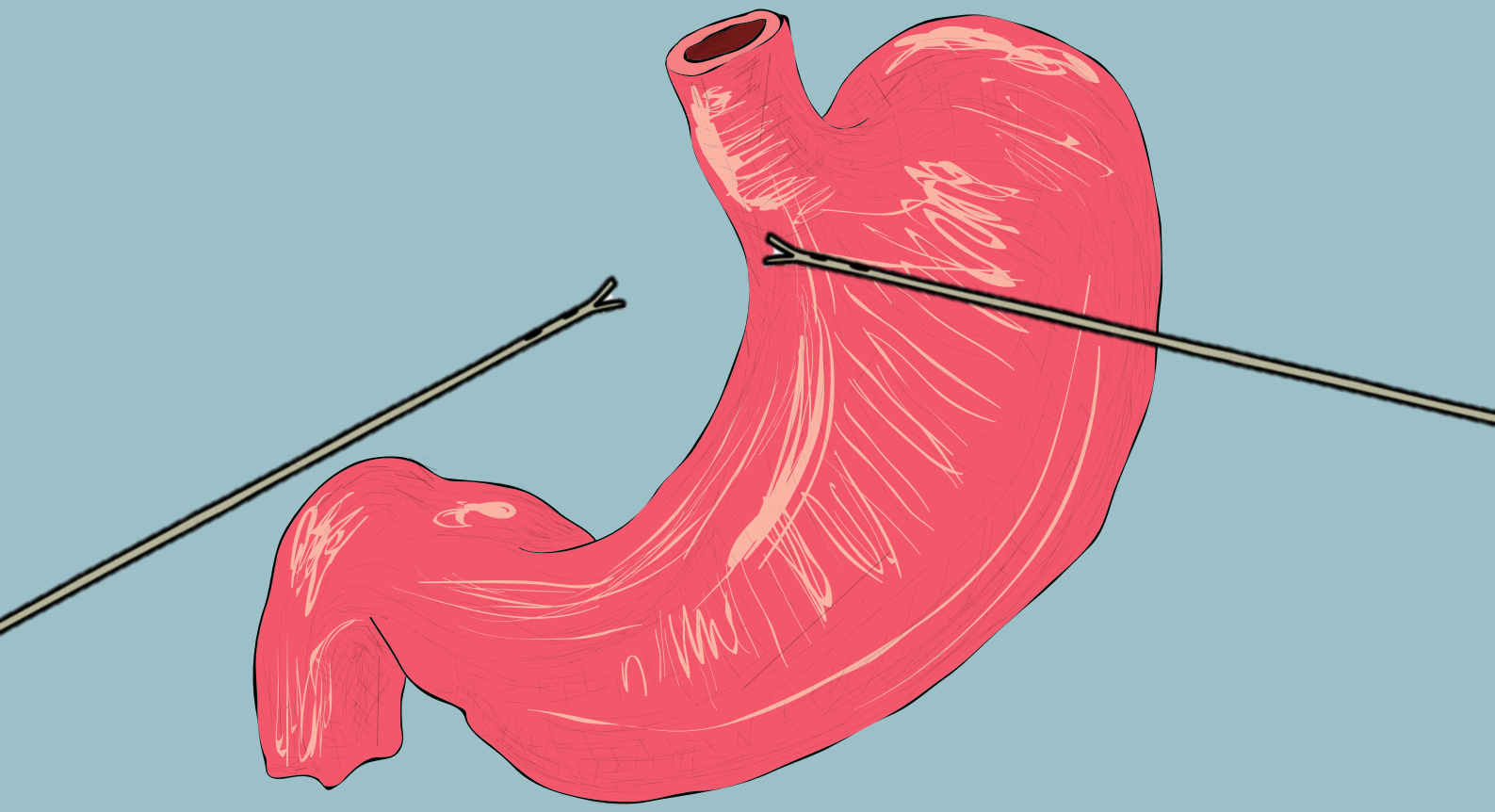


ARTIFICAL INTELLIGENCE

FOR TOOL AND ACTION DETECTION IN LAPAROSCOPIC FUNDOPLICATION SURGERY VIDEOS

MASTER THESIS OF ELKE CROONEN

MASTER THESIS

Artificial intelligence for tool and action detection in laparoscopic fundoplication surgery videos

Author
Elke Croonen, BSc

Chairman & Medical Supervisor Institution
Prof. dr. I.A.M.J. Broeders

Technical Supervisor University
Dr. J.M. Wolterink

Process Supervisors University
Drs. J. de Witte

External Member University
MSc. J.K. van Zandwijk

Date
25-07-2023

Meander medical center
Department of Surgery & Center for AI

University of twente
Faculty of Science and Technology

meander MEDISCH CENTRUM

UNIVERSITY OF TWENTE.

Abstract

Electrosurgical devices are widely used in modern surgical interventions, providing precise tissue cutting, coagulation, and sealing. However, ensuring optimal hemostasis quality while minimizing tissue damage remains a significant challenge. Excessive energy utilization can lead to adverse effects such as thermal injury, prolonged healing, and increased postoperative complications. Therefore, it is important to create more knowledge about the usage of electrosurgical tools during surgery. To address this, a research project at the Meander Medical Centre aims to develop an automatic objective assessment of energy usage during surgery. During this these two methods are developed. One for tool detection and one for tool activations detection.

The first method presents a YOLOv7 network developed for surgical tool detection in 65 laparoscopic fundoplication videos. Utilizing a semi-supervised approach with active learning techniques, with 36% of 13.518 frames manually annotated. The network achieves high precision of 0.90, recall of 0.81, mAP@0.5 of 0.81, and mAP@0.5:0.95 of 0.58. This study emphasizes the importance of a semi-supervised and active learning approach to enhance data labeling efficiency. Future research should consider incorporating temporal information and expanding the network's training on diverse surgical procedures and tools.

The second method presents a method to acquire the number and length of activations of an electrosurgical tool, the Enseal, in laparoscopic fundoplication surgery videos. The study involves acquiring video and audio data from 10 laparoscopic fundoplication surgeries to train an action recognition network for detecting the activation of the Enseal tool. The audio recordings of the Gen11 (the energy generator of the Enseal) are processed to generate a ground truth label for Enseal tool activations. The network architecture comprises an I3D feature extractor and a MSTCN++. The MSTCN++ generates frame-wise labels from the feature maps and processed audio, enabling action detection. The network achieves a frame-wise accuracy of 90.74, a segmental edit distance of 86.86, segmental F1@0.1 of 73.99, F1@0.25 of 70.17, and F1@0.5 of 54.90. The results of the action detection network show promise, considering the limited dataset of only 10 videos. However, there is room for improvements, which include increasing the dataset size, enhancing the feature extractor, incorporating data augmentation techniques, and exploring additional spatial information. These enhancements are expected to enhance the developed methodology's accuracy and robustness.

This master thesis shows the potential of utilizing deep learning networks for detecting energy devices and actions during laparoscopic surgery. The proposed methods hold promise for creating an automatic objective assessment of energy usage, enhancing surgical outcomes, and contributing to surgeons' self-improvement and benchmarking. Future research should focus on refining the current methods and combining the methods with active bleeding detection, evaluating the effectiveness of tool activations. Additionally, exploring alternative data acquisition methods, such as energy monitoring devices, could provide valuable insights into the amount of energy usage and its correlation with tool activation and effectiveness.

Contents

Abbreviations

Acc - accuracy
AP - average precision
CNN - Convolutional neural network
Edit - segmental edit distance
fps - frames per second
FUSE - Fundamental Use of Surgical Energy
GERD - Gastroesophageal Reflux Disease
HAR - Human activity recognition
I3D - Inflated 3D ConvNet
IoU - Intersection Over Union
LES - Lower Esophageal Sphincter
LSTM - Long-short term memory network
MSE - Mean Squared Error
mAP - mean Average Precision
MSTCN - Multi-stage temporal convolutional network
NSAIDs - non-steroidal anti-inflammatory drugs
P - precision
PAN - Path Aggregation Network
R - recall
ResNet - Residual neural network
R-CNN - Region Based Convolutional Neural Networks
RNN - Recurrent neural network
SAGES - Society of American Gastrointestinal and Endoscopic Surgeons
SGD - Stochastic Gradient Descent
SPP - Spatial Pyramid Pooling
std - standard deviation
YOLO - You Only Look Once

1. Introduction

Electrosurgical devices are widely used in modern surgical interventions, providing precise tissue cutting, coagulation, and sealing.[1] These procedures involve the utilization of various electrosurgical devices, including monopolar and bipolar devices, each employing multiple current types (for more detailed information, please refer to Section 1.1). However, ensuring optimal hemostasis quality while minimizing tissue damage remains a significant challenge.[2] Excessive energy utilization can lead to adverse effects such as thermal injury, prolonged healing, and increased postoperative complications.[1,3] Therefore, it is important to create awareness of energy usage during surgery.

In previous years, manufacturers have introduced tools that promise to deliver optimal hemostasis during electrosurgical procedures with the help of feedback control. These feedback-controlled sealing devices stop electrical flow when optimal hemostasis is realized.[4] However, these tools may sometimes cause more damage than necessary.[2] In certain situations, it may be preferable to provide less energy since minimizing thermal spread to neighboring tissues could be more important than ensuring perfect hemostasis. The potential damage caused by excessive energy can outweigh the consequences of minor bleeding. Therefore, surgeons can adopt a dynamic approach when using the sealing tool. The dynamic approach or dynamic usage of the electrosurgical tool refers to situations where the surgeon performs a complete activation of the tool in certain cases, while in others, the full cycle is not completed. A full sealing cycle is characterized by the feedback-controlled tool emitting an end sound, indicating that electrical flow has stopped.[5] An example of a situation where the surgeon may need to perform a full sealing cycle is when dealing with vessels. However, when operating near critical structures like nerves, surgeons should be cautious and sometimes stop sealing before the end of a full sealing cycle, even if optimal hemostasis has not been reached. By dynamically using the sealing tool, surgeons can optimize patient outcomes and ensure the safety and well-being of their patients.[2]

It is crucial to provide surgeons with understanding of their energy usage during surgery, as they currently lack insights into this aspect.[6] It is believed that higher energy usage during surgical procedures may lead to increased harm to the patient.[7] By gaining deeper insights into energy usage and dynamic use of electrosurgical tools, we can obtain more information regarding this belief and its implications. Therefore, the Meander Medical Centre started with research into energy usage during surgery. The aim is to develop a method to create an automatic objective assessment of energy usage during surgery.

Information needed to create an automatic objective assessment will be provided from an energy monitoring device and laparoscopic surgery videos. Figure 1 provides an overview of the topics covered by this research line. The energy monitoring device will capture energy data throughout surgical procedures, which will be subsequently utilized to generate an energy report. In parallel, laparoscopic videos will be utilized for tool detection, tool activation detection, and bleeding detection. Subsequently, this data can be utilized to evaluate the effectiveness of the activation of the tool, thereby providing valuable insights into achieving optimal hemostasis while minimizing tissue damage. Throughout this research, data will be collected from laparoscopic fundoplication surgeries, which are performed as a treatment for gastroesophageal reflux disease (GERD)(Section 1.2).

Fundoplication surgery (Section 1.2) is specifically chosen for this research for several reasons. Firstly, only one electrosurgical device, the Enseal tool from Ethicon[8], is used during these surgeries. This simplifies the data collection process as energy monitoring and analysis of laparoscopic videos only need to be performed for one device. Additionally, fundoplication surgeries have distinct phases, with the sealing tool being utilized in only one part of the procedure. This is beneficial as it requires less computational power to use only one part of the video for the algorithms that will be used during this research. Furthermore, the dynamic usage of the Enseal during this procedure holds significant importance. Firstly, avoiding causing harm to adjacent tissues, such as the vagus nerve, the stomach, and the parietal pleura, is crucial.[9,10] But also ensuring optimal hemostasis is essential when dissecting the

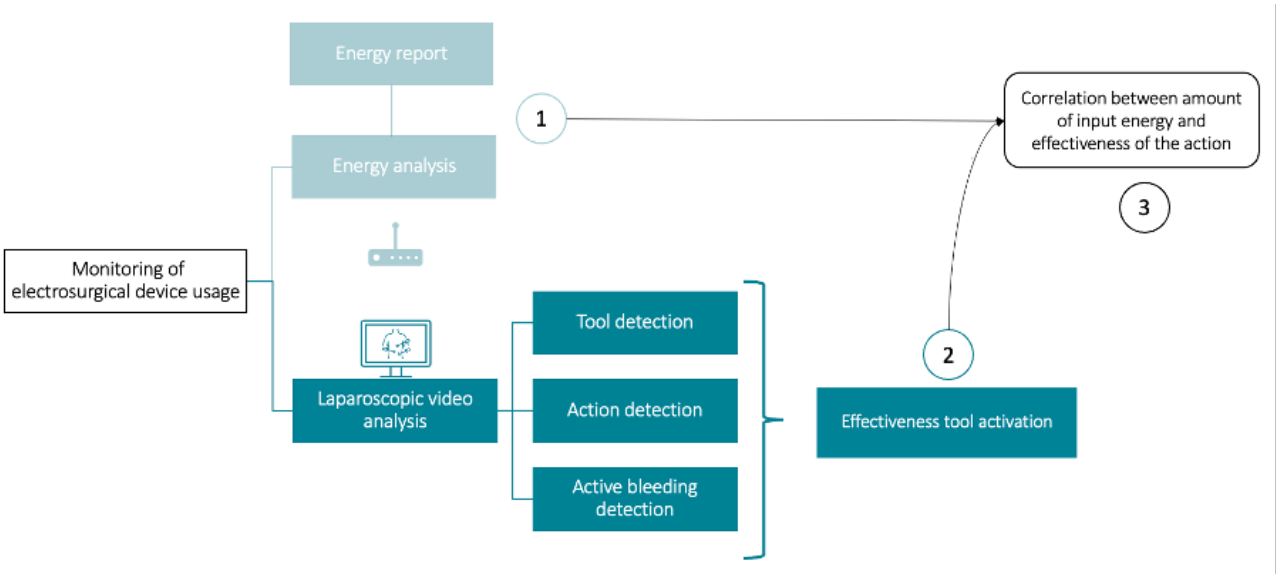


Figure 1: An overview of the energy monitoring research line of Meander Medical Center. Part one (light blue) presents an energy analysis method utilizing an energy monitoring device. This data can be used to generate an energy report. Part two (dark blue) utilizes laparoscopic videos to extract tool localization, tool action, and active bleeding information. Within this study, deep learning networks are employed for tool detection and tool activation detection. Future research aims to combine these techniques with active bleeding detection to evaluate the effectiveness of tool activation, which means optimal hemostasis while minimizing tissue damage. Furthermore, part one and part two can be combined to analyze the correlation between the amount of input energy and the effectiveness of the activation, which is indicated by the number 3 in the Figure.

gastric fundus, as the short gastric vessels retract into the tissue during dissection. Without effective hemostasis, bleeding may occur, and controlling the movement of these vessels can be challenging.[11]

This master thesis investigates the feasibility of extracting tool information from laparoscopic fundoplication videos obtained from Meander Medical Center (highlighted in dark blue in Figure 1). The specific focus is on tool and action detection. The research aims to address the following question: To what extent can the location and activation of surgical tools be recognized in laparoscopic fundoplication videos with the use of artificial intelligence?

To address this question, a YOLOv7[12] network is trained to detect surgical tools in laparoscopic fundoplication surgery videos from Meander Medical Center. A semi-supervised method with aspects of active learning is employed to label the video data efficiently. Detailed information regarding the tool detection method and results can be found in Chapter 2. Additionally, the activation of the Enseal tool is detected using a feature extractor along with a multi-stage temporal convolutional network (MSTCN). To efficiently label the data, labels were created from audio recordings during the surgical procedures. More information about the tool activation detection can be found in Chapter 3.

1.1. Technical background

More than 80% of surgical procedures performed today involve devices that apply energy to tissues, with significant advantages for tissue dissection, hemostasis, and ablation.[6] The following sections will provide information about the current types of electrosurgical devices and their respective applications. Furthermore, information about the tissue effects, the electrical circuit of electrosurgical devices, and feedback-controlled devices will be provided.

1.1.1. Current types

Electrosurgery refers to tissue cutting, fulguration, and coagulating using high-frequency electrical currents.[13] The waveforms of this current can be altered to produce different tissue effects.

Cutting current uses a pure, non-modulated sinusoidal waveform, see Figure 2. This waveform achieves a higher average power than other alternating waveforms with equal peak voltage. Due to the high current density produced by the high average power, a smooth cut without causing extensive thermal damage can be achieved.[14]

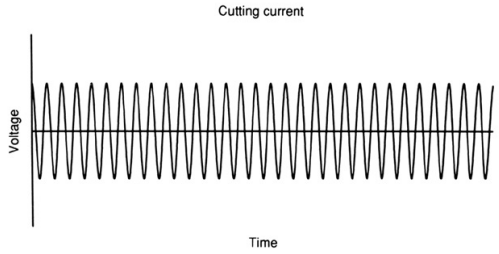


Figure 2: Cutting current[14]

Coagulation current consists of intermittent bursts of damped sine waves of high peak voltage, see Figure 3. High peak voltages result in high tissue temperatures and severe thermal destruction. This makes this type of current ideal for coagulating bleeding vessels.[14]

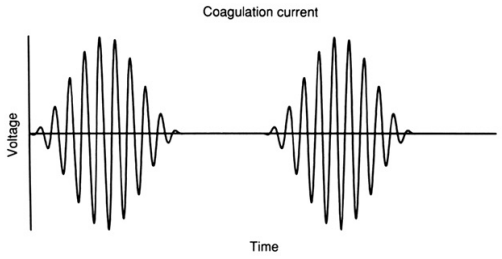


Figure 3: Coagulation current[14]

Blended current are created by alternating cutting and coagulation currents, producing a higher peak-to-peak voltage, as shown in Figure 4. This waveform is then delivered in intermittent bursts at a rate determined by the settings of the electrosurgical generator. Blended currents contain a lower average power than a pure sinusoidal waveform because of the duty cycle.[14]

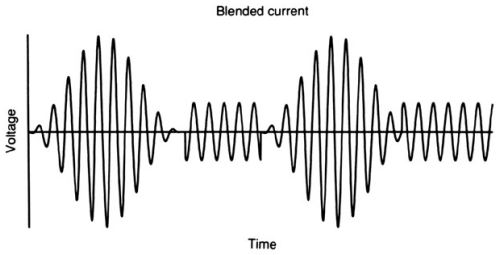


Figure 4: Blended current [14]

1.1.2. Tissue effects

The heat from electrosurgical instruments can activate tissues in a variety of ways. An illustration of the different tissue effects is shown in Figure 5.

Electrosurgical cutting involves the use of electric sparks to divide tissue, focusing intense heat at the surgical site. The surgeon achieves a maximum concentration of current by applying sparks to the tissue. The surgeon should keep the electrode slightly away from the tissue to generate this spark. This technique generates a significant amount of heat in a very short time, leading to tissue vaporization. [15]

Fulguration, sparking with the coagulation waveform, coagulates and chars tissue over a broad area. The widely dispersed arcs cause a rapid decrease in current density, which means that less heat will be generated in the deeper structures. Also, the thin carbon layer and desiccated tissue beneath it form an insulating barrier, decreasing the probability of subsequent arc strikes in the same location. [14] Less heat is produced due to the low duty cycle of approximately 6 percent. As a result, a coagulum is formed instead of cellular vaporization. Overcoming the high impedance of air requires the coagulation waveform to have a significantly higher voltage than the cutting current. [15]

Desiccation occurs when the electrode comes into direct contact with the tissue. Optimal desiccation is achieved using the "cutting" current. When the electrode touches the tissue, the current concentration decreases, leading to reduced heat generation and the absence of cutting action. Instead, the cells dry out and form a coagulum without vaporization or explosion. [15]

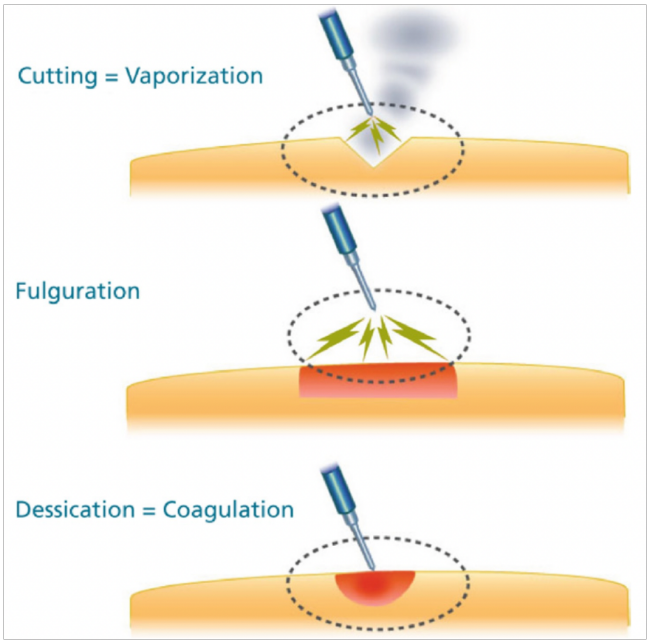


Figure 5: Illustration depicting the three electrosurgical tissue effects. (A) Electrosurgical cutting demonstrates the division of tissue through the generation of electric sparks, resulting in vaporization. (B) Electrosurgical fulguration shows the coagulation and charring of tissue over a wide area using the coagulation waveform. (C) Electrosurgical desiccation exhibits the drying out of tissue cells and the formation of a coagulum when the electrode is in direct contact, without vaporization or explosion. [15]

1.1.3. Electrical circuit

During electrosurgery, an optimal electrical circuit is crucial. This means that there is a closed current pathway along which electricity flows. Any currents flowing in a path or direction other than the intended pathway may lead to undesired outcomes, such as alternate site burns or shocks. [2–4,6,7]

Monopolar circuit consists of an electrosurgical generator, a small active electrode, and a large dispersive electrode pad on the patient, which is connected to the electrosurgical generator. A representation of a monopolar circuit is visualized in the right part of Figure 6. The large size of the dispersive electrode pad lowers the current density at its placement site, preventing unwanted burns. [15,16]

Bipolar circuit only consists of an electrosurgical generator and a tool with two electrodes of equal size. Both electrodes are located at the site of coagulation. See the left side of Figure 6 for an example. Because the path of least resistance is the shortest distance between the electrodes, there is a low probability that the current will travel via an alternate pathway. [15,16] Bipolar electrosurgery is safer than monopolar electrosurgery as it limits tissue penetration, reduces the risk of burns, and provides better visibility during procedures. [2,3]

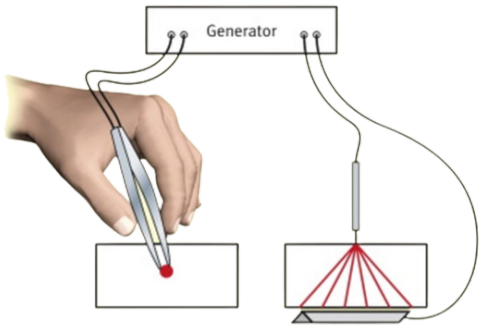


Figure 6: A representation of the monopolar and bipolar circuit. On the left, a bipolar electrosurgery tool with two electrodes of equal size. On the right, a monopolar electrosurgery tool with a small active electrode and a large dispersive electrode pad.[17]

1.1.4. Advanced bipolar electrosurgical tools

In previous years, advanced bipolar electrosurgical tools have been developed. These instruments utilize low-voltage electricity and high mechanical pressure to fuse vessel walls, resulting in the complete closure of vessel lumens. Incorporated within these devices are feedback sensors that regulate the delivery of electricity based on changes in tissue impedance, allowing for precise modulation of tissue desiccation and sealing.[2,18] Enseal (Ethicon, USA) and ligasure (Medtronic, USA), are the first and most commonly used devices. In their bench-top testing research, Chekan et al.[19] demonstrated that Enseal G2 sealers exhibit significant advantages over LigaSure. These advantages include superior uniform compression, stronger and more consistent vessel sealing, and reduced tissue sticking. During my master thesis, I focused on investigating the tool activation of the Enseal, shown in Figure 7. The associated energy generator of the Enseal is the Gen11 (Figure 8).[8] It receives signals from the Enseal device and adjusts the energy output accordingly.[20] When the tool is activated, the Gen11 emits a beeping sound. Once optimal hemostasis is achieved (determined by tissue impedance), the electrical flow is turned off, and the Gen11 emits an end sound to signal the surgeon. However, in practice, completing a full sealing cycle is not beneficial in specific cases. Chapter 3 will go into further detail on this.

Figure 7: An Enseal sealing tool, an adaptive feedback-controlled bipolar electrosurgical device from ETHICON.[8]

Figure 8: GEN11 from ETHICON, an energy generator for the Enseal sealing tool. It receives signals from the Enseal device and adjusts the energy output accordingly.[20] When the tool is activated, the Gen11 emits a beeping sound. Once optimal hemostasis is achieved (determined by tissue impedance), the electrical flow is turned off, and the Gen11 emits an end sound to signal the surgeon.[5]

1.2. Clinical background

The research described in Chapters 2 and 3 uses laparoscopic videos of fundoplication surgeries. Fundoplication surgery is performed as a treatment for gastroesophageal reflux disease (GERD). GERD is a medical condition characterized by the backflow of stomach acid into the esophagus. This section will provide an overview of GERD and the treatment of GERD using fundoplication surgery.

1.2.1. GERD

Symptoms GERD is characterized by the persistent movement of stomach content into the esophagus, causing symptoms and/or complications.[21–23] Symptoms include dental corrosion, dysphagia, heartburn, pain when swallowing, regurgitation, non-cardiac chest pain, and extraesophageal symptoms such as chronic cough, hoarseness, reflux-induced laryngitis, or asthma. Long-term problems include esophagitis, esophageal stricture, and Barrett's esophagus.[23]

Cause Frequent acid reflux is due to poor closure of the lower esophageal sphincter (LES), which connects the esophagus and stomach. Figure 9 depicts a representation of acid reflux.

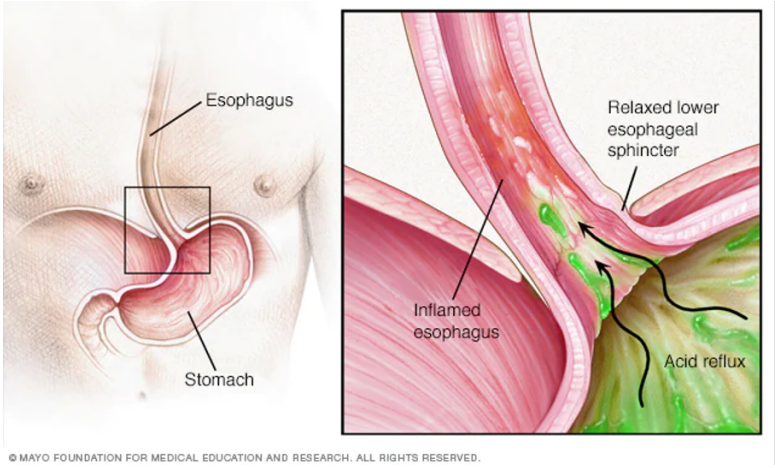


Figure 9: Acid reflux occurs when the sphincter muscle at the lower end of the esophagus relaxes at the wrong time, allowing stomach acid to flow back into your esophagus. This can cause heartburn and other signs and symptoms. Frequent or constant reflux can lead to GERD.[24]

Risk factors Risk factors include obesity, pregnancy, smoking, having a hiatal hernia, and using certain medications. Medications that could cause or aggravate the disease include benzodiazepines, calcium channel blockers, tricyclic antidepressants, non-steroidal anti-inflammatory drugs (NSAIDs), and asthma medications.[21]

Diagnosis The diagnosis is determined by the presence of symptoms and by additional clinical examinations. The clinical examination may include endoscopy, esophageal pH monitoring, esophageal manometry, and (swallow) X-ray.[25] Grades A through D are used to classify reflux esophagitis, see Figure 10.[26]

Treatment In addition to diet and lifestyle improvements, GERD may be treated with medication or surgery. A proton-pump inhibitor like omeprazole is frequently used as the first course of treatment. If medication does not achieve the intended result, a fundoplication surgery can be suggested.[25]

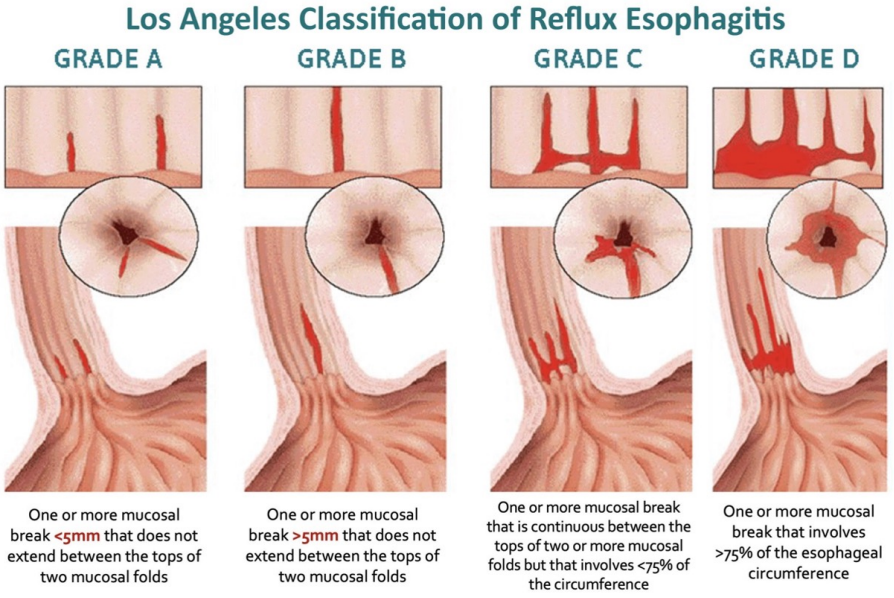


Figure 10: Los Angeles Classification of Reflux Esophagitis.[26]

1.2.2. Fundoplication surgery

Fundoplication surgery is a commonly performed surgical procedure used to treat GERD or hiatal hernia. The main objective of this surgical technique is to strengthen the lower esophageal sphincter (LES) and prevent the backflow of stomach acid into the esophagus.[21–23] The surgery can be carried out using laparoscopic methods or with the assistance of a surgical robot.[27]

During the fundoplication surgery, the surgeon wraps the upper part of the stomach, known as the gastric fundus, around the lower end of the esophagus. This creates a barrier that limits stomach acid reflux, effectively reinforcing the weakened LES, which is often the underlying cause of GERD. To address concurrent hiatal hernia, where the fundus moves up through an enlarged opening in the diaphragm, the surgeon may constrict the hiatus further using surgical sutures or a patch[28].

The most common type of fundoplication is the Nissen fundoplication, which involves a total (360°) wrap of the fundus around the esophagus. Figure 11 depicts a representation of this method. Other alternative fundoplication techniques include the Thal (270° anterior), Belsey (270° anterior transthoracic), Dor (anterior 180-200°), Lind (300° posterior), and Toupet fundoplications (posterior 270°)[29]. The choice of surgical technique is based on the surgeon's preference. Research comparing partial and total wrap procedures shows mixed results; however, the literature generally supports fewer complications and comparable symptom relief with partial wraps. However, recurrent symptoms are more likely with partial fundoplication. Anterior partial wraps are considered less durable than complete and posterior partial wraps.[30–33]

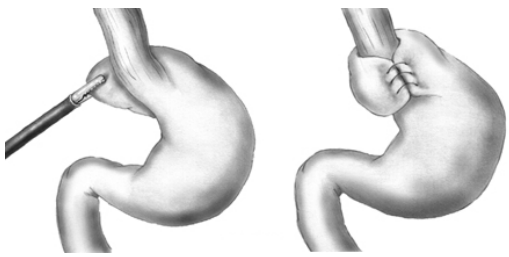


Figure 11: Illustration of a total (360°) wrap of the fundus around the esophagus, also know as the Nissen fundoplication.[34]

During laparoscopic or robot-assisted Nissen fundoplication, the surgeon opens the gastrohepatic ligament for right crural dissection. Subsequently, the surgeon exposes the phrenogastric ligament for left crural dissection and division of the short gastric vessels. Throughout this dissection, the surgeon must exercise caution with regard to surrounding structures, such as the vagus nerve, the stomach, and the parietal pleura.[35]

An optimal electrosurgical tool is required to successfully perform the dissection of these ligaments. This tool should be capable of dissecting the stomach vessels while preserving delicate tissues like the vagus nerve. Additionally, the surgeon must fully utilize the capabilities of the tool to ensure a successful procedure. Therefore, research into the effectiveness of tool activations is needed.

2. Surgical-tool detection

A Semi-Supervised YOLO Network approach for laparoscopic fundoplication surgery videos

Abstract

Introduction Laparoscopic surgery has revolutionized the field of surgery by providing minimally invasive procedures with benefits such as reduced trauma, faster recovery, and improved outcomes. Accurate tool detection during laparoscopic surgeries is crucial for understanding surgical techniques, standardizing best practices, and enhancing surgical education and patient safety.
Method A deep learning network is developed using the You Only Look Once version 7 (YOLOv7) architecture. The network is trained on 65 laparoscopic fundoplication surgery videos using a semi-supervised approach with active learning aspects to label the data efficiently. Only 36% of the data is labeled manually.
Results The network achieved an overall precision of 0.90, recall of 0.81, mAP@0.5 of 0.81, and mAP@0.5:0.95 of 0.58.
Discussion and conclusion The results show promising performance in terms of precision, recall, and mAP. This study emphasizes the importance of a semi-supervised and active learning approach to enhance data labeling efficiency. Future research should consider incorporating temporal information and expanding the network's training on diverse surgical procedures and tools.

2.1. Introduction

Laparoscopic surgery, a minimally invasive approach to surgical interventions, has transformed the field of surgery by offering advantages such as reduced patient trauma, faster recovery times, and improved surgical outcomes.[36] As laparoscopic surgeries become more popular, there is a growing need to improve surgical techniques and gain insights for both surgeons and trainees. One crucial aspect of improving laparoscopic surgery lies in understanding tool usage during the procedure. Accurate tool detection can provide information about the number and types of tools that are used, move patterns, standardize best practices, and aid in surgical education.[37]

By analyzing tool usage, variations in techniques among surgeons, patterns that lead to successful outcomes, and potential areas for improvement can be identified. This could be used for training or skill-level benchmarking.[38] Moreover, accurate tool detection could further enhance patient safety by identifying potential instrument misplacements or unintended interactions during the surgery[37]. Surgical education also stands to gain from tool detection systems. Real-time feedback on tool usage can assist trainees in understanding proper instrument handling techniques, improve hand-eye coordination, and develop critical decision-making skills. The collected data can be utilized to create educational resources such as annotated videos and interactive simulations, enabling more efficient and comprehensive surgical training programs.[39]

In addition to the mentioned advantages of surgical tool detections, it also holds significant value for research on energy device usage during surgeries. Currently, there is insufficient information available regarding energy usage in surgical procedures, and it is believed that higher energy usage may pose increased risks to the patients.[6,7] By gaining a deeper understanding of energy usage during surgery, we can gather more information about this belief and its implications. Therefore, this research aims to develop a method for tool detection in laparoscopic surgeries, with a specific emphasis on the Enseal tool, an electrosurgical device used during fundoplication surgeries. This tool detection network serves

as a foundational step toward further research into energy device usage and activation during surgery. The research question is as follows: To what extent can surgical tools be detected in laparoscopic fundoplication videos? The novelty of this research lies in the approach to labeling the data for the network, as only 36% of the data has been manually labeled.

2.2. Related work

Various deep-learning networks have been developed for object detection and object segmentation. When detecting objects in an image, using bounding boxes around the objects is a more efficient approach than pixel-wise classifications used in segmentation. Therefore, object detection is often preferred when working with large amounts of data.[40] Examples of object detection algorithms include the Mask R-CNN[41], which combines object detection with instance segmentation, allowing precise localization and segmentation of objects, and EfficientDet[42], a high-performance and efficient object detection architecture. However, one of the most widely recognized and utilized networks for object detection is the You Only Look Once (YOLO) network[43].

YOLO stands out for its real-time processing capabilities and has demonstrated promising results for tool detection in laparoscopic surgeries. Choi et al.[44] achieved a mean Average Precision (mAP) of 72.26% on the m2cai16-tool dataset[45] and Yang et al.[46] achieved a mAP of 91.58% on the Cholec80 dataset[47], highlighting the effectiveness of YOLO in accurately detecting surgical instruments within laparoscopic surgery videos.

Since the labeling of data for tool detection is a time-consuming task, several studies have been done on deep learning approaches such as semi-supervised training and using weakly labeled data. For example, Vardazaryan et al.[48] conducted a study where they trained a modified ResNet[49] to generate image-level annotations. Raw localization maps were used to identify the predicted positions of the tools. These maps were created by converting the 512 feature maps from the network through a convolution layer with 1x1 kernels. The predicted location of the tool was determined based on the maximum activation within the map. Their approach achieved a mAP of 88.8% on the Cholec80[47] dataset. Ali et al[50], created a method for a semi-supervised student-teacher network that was able to achieve a mAP50 of 90.25% and a mAP50:95 of 46.88% on the m2cai16-tool-locations[50] dataset with only 10% annotated (2300 frames).

Building upon previous research findings and advancements in object detection, this study will deploy a YOLO network for tool detection, incorporating a semi-supervised training approach with active learning elements to efficiently label the data.

2.3. Methodology

During this research, a YOLOv7 network is trained for tool detection of four surgical instruments commonly used during fundoplication surgeries. The network is trained with 65 laparoscopic fundoplication surgery videos from the Meander Medical Center, where 36% of the frames were manually labeled to train the initial model in a supervised manner. Subsequently, a semi-supervised approach was employed, leveraging the YOLO network to label a portion of the remaining data automatically. To further enhance accuracy, active learning techniques were integrated into the process, with manual re-labeling of frames with unreliable assigned labels. This chapter provides an overview of the network and dataset used in this study. First, the YOLO network architecture will be discussed, highlighting the improvements introduced in YOLOv7. It then discusses the loss functions employed during training, the dataset utilized for model development and testing, and the labeling process, which incorporates semi-supervised and active learning techniques. Finally, the chapter discusses the metrics

used to evaluate the network's performance.

2.3.1. YOLO architecture

In 2015 Redmond et al.[51] presented the YOLO network, a new deep-learning approach for object detection. Instead of approaching object detection as a classifier problem, they approached it as a regression problem. They separately predicted bounding boxes and corresponding class probabilities with a single neural network, as illustrated in Figure 12. This approach enabled end-to-end optimization of the entire detection pipeline.

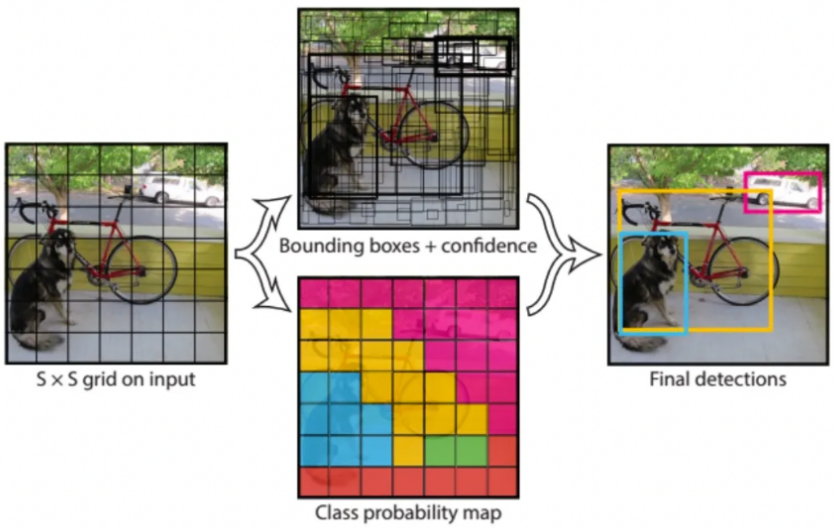


Figure 12: The YOLO network takes a 2D image or video as input and divides the image into an S × S grid. For each grid cell, the model predicts B bounding boxes, along with the confidence scores for those boxes, and C class probabilities.[52]

The network consists of three main components: the backbone, the neck, and the dense prediction, as visualized in Figure 13. 2D images or videos undergo feature extraction through the backbone, a deep neural network primarily composed of convolutional layers, with the objective of extracting essential features from the images or video frames.[43] The neck in the YOLO network refers to a series of additional convolutional layers that come after the backbone. These layers further process and refine the features obtained from the backbone. The neck is crucial for integrating multi-scale information, allowing the YOLO network to detect objects of various sizes within the input image.[43,53] The dense prediction aspect of YOLO refers to how the network generates object predictions across the entire image grid. Unlike traditional object detectors that use region proposal techniques, YOLO divides the input image into a grid and predicts bounding boxes, objectness scores, and class probabilities for each grid cell (Figure 12). Each grid cell is responsible for detecting objects that are centered within its boundaries. This dense prediction approach enables YOLO to efficiently detect multiple objects in a single forward pass.[43]

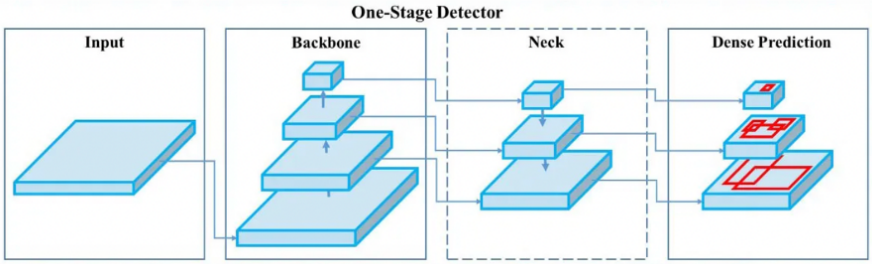


Figure 13: A YOLO network consists of a backbone, neck, and dense prediction/head. The neck of the network enables object detection at various scales, by giving feature maps from different spatial resolutions from the backbone as input for the dense prediction of the network. Modified from [43].

YOLOv7 [12] is an improved version of the previous YOLOv4 model with modifications aimed at increasing accuracy without additional training costs. YOLOv7 consist of a CSPDarknet53 [54] backbone, SPP [55] and PAN [56] Neck, and a YOLOv3 [53] dense prediction, similar to the YOLOv4 model. However, the following improvements have been implemented: it introduces architectural changes called E-ELAN, which improve learning ability. It also implements compound model scaling for concatenation-based models, maintaining the properties that the model had at the initial design and also the optimal structure. Trainable enhancements include planned re-parameterized convolution, merging computational modules, and coarse-to-fine hierarchical labels for deep supervision. Overall, YOLOv7 optimizes the model's learning, scalability, and accuracy while preserving its initial design. Please find Appendix 5.1 for more information regarding the improvements.

2.3.2. Loss functions

YOLOv7 network uses a combination of three different loss functions, just like previous models, to train the model: classification loss, localization loss, and confidence/objectness loss. [43,51]

1. Classification loss: This loss function is used to predict the class probabilities of the objects in an image. It is calculated using the cross-entropy loss between the predicted class probabilities and the true class labels.
2. Localization loss: This loss function is used to predict the bounding boxes of the objects in an image. It is calculated using the Mean Squared Error (MSE) between the predicted and true bounding box coordinates.
3. Confidence/Objectness loss: The confidence of object presence is the objectness loss. It is the probability the box contains an object multiplied by the intersection over union (IoU) between the predicted box and the ground truth. [57,58] IoU is the area of overlap between two bounding boxes divided by the area of union.

2.3.3. Dataset

The dataset comprises 65 videos of fundoplication surgery procedures recorded at the Meander Medical Centre between January 2020 and November 2022. The videos have a mean duration of 52.7 minutes and a frame rate of 25 frames per second. The frames used to train the network were extracted from the videos at a rate of 1 frame per 5 seconds. This resulted in 41,171 frames. The frame rate was chosen to create differences between the frames, so that the dataset would be diverse and not consist of multiple frames that are similar, which would happen if the original frame rate of 25 frames per second was used.

2.3.4. Manual labeling process and network training

A YOLOv7 network from Wang et al.[12] was initially trained using a publicly available dataset, the Cholec80 dataset[47] (training 1, Figure 15). This network was then employed to pre-label a single video, which was subsequently uploaded to Label Studio[59] for manual correction of the pre-labeled data. During this labeling, the four tools used during fundoplication surgery are labeled: the Enseal, the grasper, the irrigator, and the scissors. Figure 14 shows a screenshot of a labeled frame in Label Studio. The frames and labels are then used to further train the network from training 1, which you can see in Figure 15 as training 2. Ten additional videos were pre-labeled using the network trained in training 2, and then randomly selected frames were manually corrected in Label Studio. Finally, for training 3, the network from training 1 was further trained using the randomly selected frames from the 11 videos. The number of frames used during each training can be found in Figure 15. This is the total number of frames in the training, validation, and test set together.

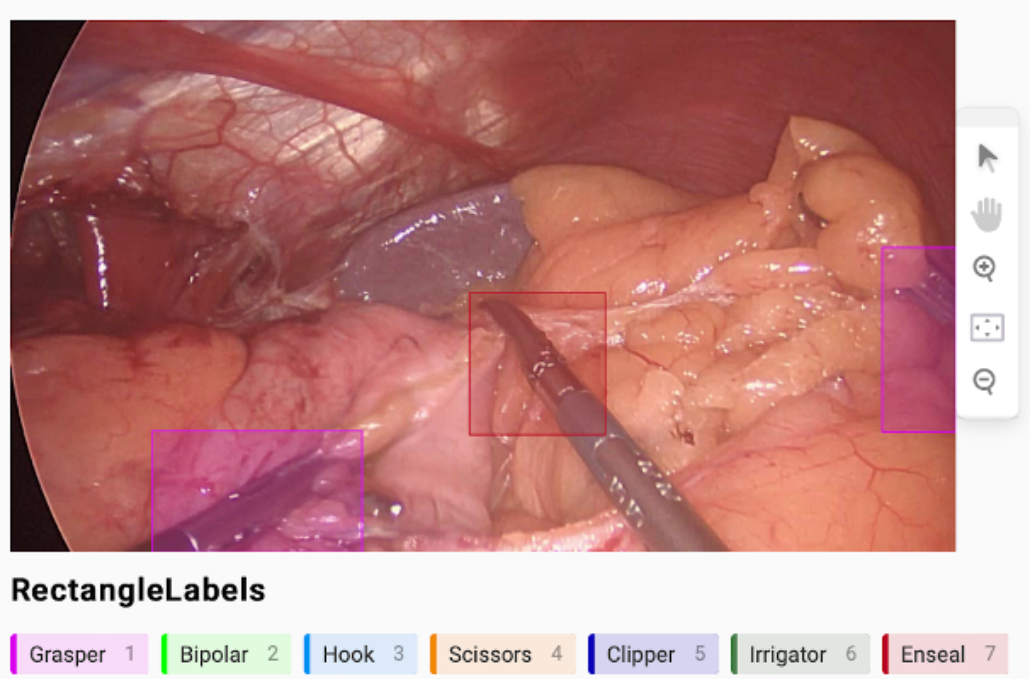


Figure 14: A screenshot of a labeled frame in Label Studio. With a red bounding box around the Enseal tool and two pink bounding boxes around the graspers. The frames are uploaded with labels created by one of the networks from training 1 to 4 in Figure 15. A bounding box can be moved to the right location to correct the labels, a different class can be selected when selecting the bounding box around a tool, or a new bounding box can be created. To create a new bounding box, one of the seven suggestions under the image can be selected and a bounding box can be created around the tool. This image includes seven classes instead of the four classes in the new dataset, as it was trained on the Cholec80 dataset containing six classes, with an additional class, Enseal, which I incorporated.

Not all frames in the videos were manually labeled due to the significant time and effort required. Instead, for training 3 a total of 4497 frames were manually labeled, of which 3416 were used to create the train set, 386 frames were assigned to the validation set, and 695 frames were assigned to the test set. Following this, the network was employed to pre-label all 65 videos in the dataset, including the 11 manually labeled videos.

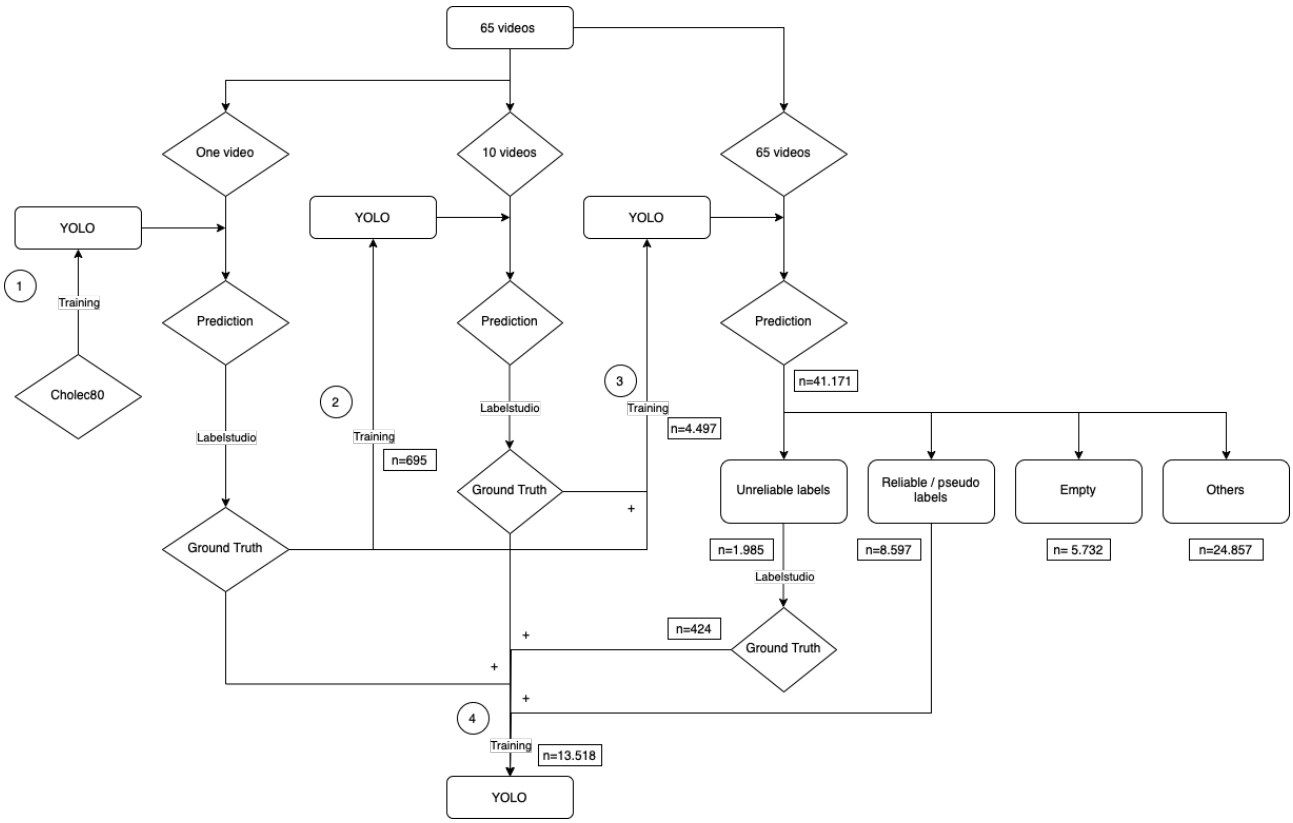


Figure 15: The flowchart illustrates the data flow and labeling process for 65 videos. The flowchart shows four stages of training, represented by numbered circles. Training 1 involves pretraining with the publicly available cholec80 dataset. Training 2 is performed using one video while Training 3 utilizes 11 videos. Subsequently, the network trained in Training 3 is employed to label all 65 videos. During the labeling process, a distinction is made based on the confidence scores of the predicted outcome. This process will be explained in Section 2.3.5. The four categories include Unreliable labels (which were subsequently relabeled in the label studio), Pseudolabels (labels with high reliability), Empty labels (where no tool was present), and Others. Finally, Training 4 is conducted using the 11 videos along with the pseudo labels and the relabeled unreliable labels

2.3.5. Semi-supervised and active learning

To efficiently label the remaining data, semi-supervised and active learning aspects are incorporated into the labeling process. The network, trained with 3416 frames, is employed to generate predictions for the entire dataset, which comprises 41171 frames. The frames are predicted with a confidence score threshold of 0.2, 0.3, and 0.7. The confidence score threshold ensures that only the boxes with a box probability higher than the threshold remain in the output. Subsequently, the predictions are classified into four distinct categories: reliable, unreliable, empty, and other.

Reliable labels are those that possess the same label when predicted with a confidence level threshold of 0.2 (pred_0.2) and 0.7 (pred_0.7). The predictions with a confidence level threshold of 0.2 are used because sometimes pred_0.7 shows only one tool, even though there are multiple tools in the frame. Sometimes the bounding boxes around these tools have a lower confidence score. Therefore, I stated that if pred_0.2 gives the same output as pred_0.7, it is probably a reliable label. These labels are then used as pseudo labels for training 4 (see Figure 15). This process is called semi-supervised learning; the pseudo-labels created by the network are used for further training.

Unreliable labels are those that do not possess the same label when predicted with a confidence level of 0.2(pred_0.2) and 0.3(pred_0.3). Both predictions are used as the filtering process now excludes boxes with a confidence score ranging between 0.2 and 0.3, which require correction. To improve the accuracy of these labels, a manual correction process was employed using Label Studio[59], resulting in more reliable annotations. These corrected labels were then utilized for training 4 (see Figure 15).

This approach incorporates elements of active learning, as only the labels with a low confidence level require manual intervention and correction.

Empty labels are frames that do not have a label when predicted with a confidence level of 0.2. The remaining frames that do not fall into any of the aforementioned categories are classified as others. This process is shown in Algorithm 1.

As shown in Figure 15, all 65 videos are predicted with the YOLO network of training 3. For training 4 the corrected unreliable labels, the reliable labels, and the 11 manually labeled videos are used. The labels from the manually labeled videos overwrite the labels of these frames in the corrected unreliable labels and reliable labels datasets because manually labeled frames are considered as ground truth.

Algorithm 1 Point selection

```
Predict images with confidence of 0.2 (predict_0.2), 0.3 (predict_0.3), and 0.7 (predict_0.7)
if predict_0.2 = predict_0.7 then
    Pseudo-label is reliable
else if predict_0.2 ≠ predict_0.3 then
    Psuedo-label is unreliable → manual correction
else if predict_0.2 is empty then
    Psuedo-label is empty
else
    Pseudo-label is classified as "other"
end if
```

2.3.6. Evaluation metrics

The results will be evaluated by calculating the precision, the recall, F1, the mean average precision (mAP) with an IOU threshold of 0.5 (mAP@0.5), and mAP with an IOU threshold between 0.5 and 0.95 (mAP@0.5:0.95). mAP is the mean of each class's average precision (AP), which is the area under the Precision-Recall curve.[60]

2.4. Experiment and Results

2.4.1. Dataset Split

The videos are divided into 49 training videos, 8 validation videos, and 8 test videos. The 11 manual annotated videos are divided over the dataset, eight in the training, one in the validation, and two in the test set. This resulted in 12486 frames in the training dataset, 1405 in the validation dataset, and 1682 in the testing dataset. The number of times the class occurs in the dataset can be found in Table 1. As you can see in Table 1, the test set contains more empty images than the train and val datasets. This is because the test set contains all frames, including the empty frames because this is the most representative of the original videos. However, the experiment in Chapter 2.4.3 reveals that training without the empty frames improved the network's results. Therefore, only the manually labeled empty frames remain in the training and validation set.

Class	Train set	Val set	Test set
Total labels	17360 (100%)	1541 (100%)	2528 (100%)
Enseal labels	3875 (22%)	388 (25%)	628 (25%)
Grapser labels	13278 (77%)	1148 (74%)	1883 (75%)
Irrigator labels	84 (0.5%)	2 (0.01%)	10 (0.04%)
Scissors labels	123 (0.7%)	3 (0.02%)	8 (0.03%)
Total number of images	10673	963	1882
Number of empty images	36	1	421

Table 1: Number of labels per class in the training, validation, and test dataset

2.4.2. Proposed method

For this application, a YOLOv7 network from Wang et al.[12] is trained on a single NVIDIA A4000 GPU. The parameters used during training can be found in Appendix 5.2. The YOLOv7 is pre-trained with the publicly available dataset Cholec80[47]. The batch size is 16, and freeze is set to 10. By setting freeze to 10, the first 10 layers are frozen, implying that the weights of the remaining layers are utilized to compute loss and are updated by the optimizer. This approach reduces the required computational power compared to regular training and enables faster training times.[61] The following values are the result of grid hyperparameter search: optimizer (SGD), IOU threshold (0.2), classification loss gain (cls=0.5), localization loss gain (box=0.05), objectness loss gain (obj=0.7), number of epochs (300), and augmentation (e.g. scale, flip, mixup, mosaic) as mentioned in Appendix 5.2. IoU threshold is a threshold during training. If a bounding box has an IoU less than the specified threshold, that bounding box is not taken into consideration.[57]

Figure 2 shows the prediction of the trained network of a frame during surgery with bounding boxes around the tools with a confidence score.

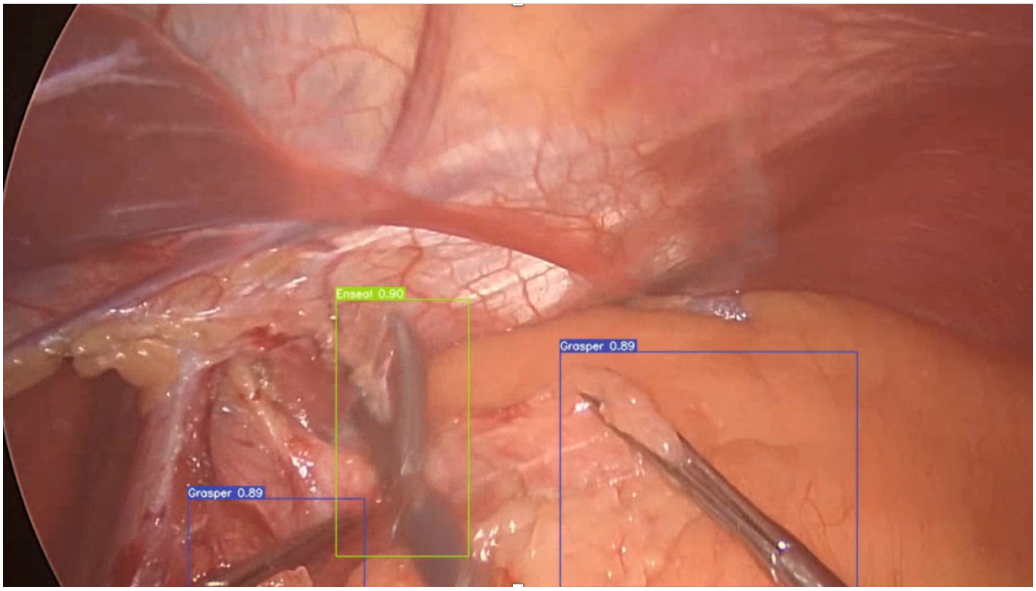


Table 2: A prediction generated by the network, with a bounding box around the Enseal with a confidence of 0.90, and two prediction bounding boxes around the graspers with both a confidence of 0.89.

The precision, recall, F1, mAP@0.5, and mAP@0.5:0.95 scores for the proposed method are displayed in Table 3. The results indicate an overall precision of 0.901, recall of 0.808, F1 of 0.852, mAP@0.5 of 0.81, and mAP@0.5:0.95 of 0.577. Remarkable is the difference in scores between the classes, which is the result of an imbalanced dataset.

Class	P	R	F1	mAP@0.5	mAP@0.5:0.95
All	0.901	0.808	0.852	0.810	0.577
Enseal	0.950	0.895	0.922	0.964	0.726
Grasper	0.935	0.914	0.924	0.966	0.724
Irrigator	0.727	0.800	0.761	0.669	0.446
Scissors	0.989	0.625	0.766	0.639	0.415

Table 3: The precision (P), recall (R), mAP@:0.5, and mAP@0.5:0.95 of all classes

The confusion matrix in Figure 16 indicates that in 95% of the ground truth bounding boxes around the Enseal were accurately classified as Enseal.

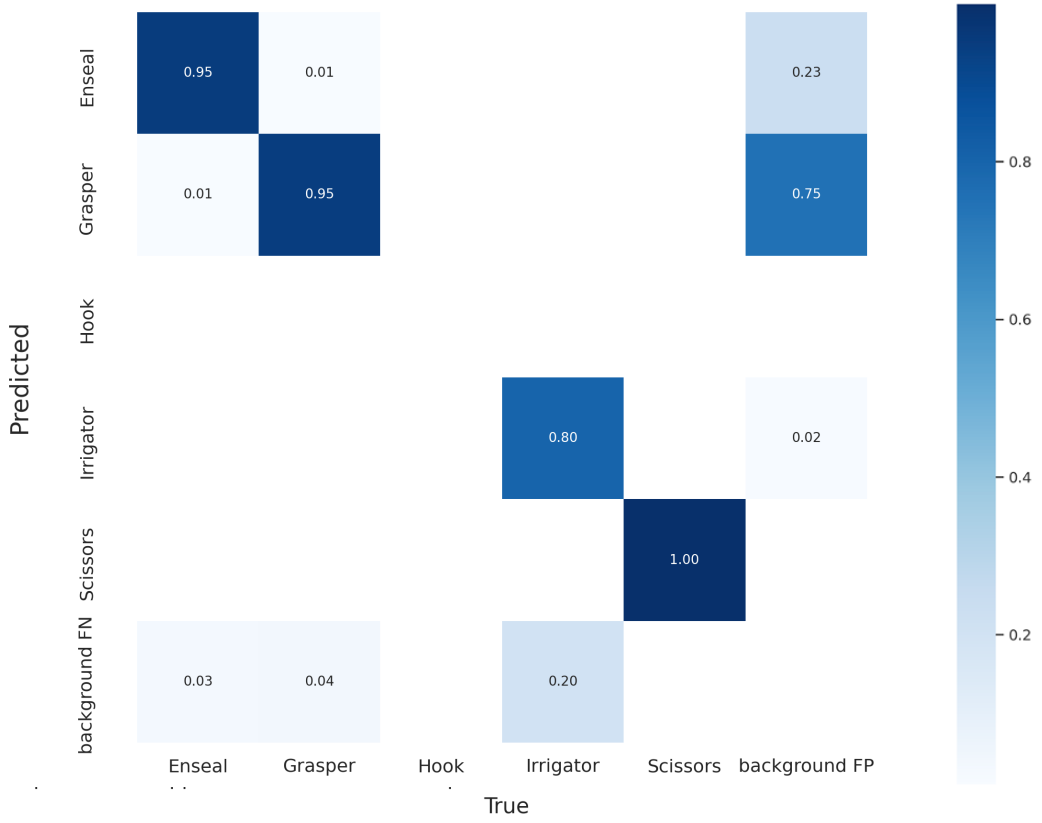


Figure 16: Confusion matrix of predicted vs. true classifications on the test dataset.

A graph of the F1 score and precision-recall of the different tools can be found in Appendix 5.3.

2.4.3. Experiments

Effectiveness of initial network To demonstrate the significance of selecting the optimal initial network, an experiment was conducted between two networks. One network had initial weights of YOLOv7 trained on the COCO data set (a large image recognition dataset for object detection, consisting of 330,000 images, each annotated with 80 object categories)[62] and one network had initial weights of YOLOv7 trained on the Cholec80 data set[47]. The dataset and the remaining parameters

were the same during both trainings. The results are presented in Table 4. This study shows that the performance of both initial networks is comparable, suggesting that pre-training with a similar data set is not a requirement for training a network with a dataset of this size.

Initial weights	P	R	F1	mAP@0.5	mAP@0.5:0.95
Cholec80	0.901	**0.808**	0.852	0.810	**0.577**
YOLOv7	**0.972**	0.765	**0.856**	**0.825**	0.573

Table 4: The results of the experiment conducted to demonstrate the significance of selecting the optimal initial network. The table compares the performance of two networks: one initialized with weights from YOLOv7 trained on the COCO dataset, and the other initialized with weights from YOLOv7 trained on the Cholec80 dataset. P= precision, R=recall, F1=F1, mAP@0.5 = the mean average precision (mAP) with an IOU threshold of 0.5, and mAP@0.5:0.95= mAP with an IOU threshold between 0.5 and 0.95.

Effectiveness of data set To show the effectiveness of using pseudo-labels predicted by the network, manually labeling unreliable data, and using empty labels as well, the network is trained with three data sets:

1. Manually annotated data;
2. Manually annotated data, reliable pseudo-labels, and manually corrected unreliable pseudo-labels (as shown in Figure 15);
3. Manually annotated data, reliable pseudo-labels, manually corrected unreliable pseudo-labels, and empty labels (as shown in Figure 15);

The training dataset and the remaining parameters were the same during all trainings. The results are shown in Table 5. The experiment demonstrates that leveraging predictions from the network enhances the network's performance. Also, it shows that adding empty labels to the dataset does not improve the performance. This experiment highlights the significance of using semi-supervised and active learning techniques to improve the network's overall performance.

Data set	Train	Val	P	R	F1	mAP@0.5	mAP@0.5:0.95
Annotated data	n=2805	n=266	0.852	0.793	0.821	0.789	0.564
Without empty labels	**n=10637**	**n=962**	**0.901**	**0.808**	**0.852**	**0.810**	**0.577**
With empty labels	n=15223	n=1555	0.891	0.796	0.841	0.795	0.570

Table 5: The results of a network trained with three datasets. 1) Manually annotated data. 2) Manually annotated data, reliable pseudo-labels, and manually corrected unreliable pseudo-labels (as depicted in Figure 15). 3) Manually annotated data, reliable pseudo-labels, manually corrected unreliable pseudo-labels, and empty labels (as depicted in Figure 15). Train = number of frames in training dataset, val= number of frames in validation dataset. P= precision, R=recall, F1=F1, mAP@0.5 = the mean average precision (mAP) with an IOU threshold of 0.5, and mAP@0.5:0.95= mAP with an IOU threshold between 0.5 and 0.95.

2.5. Discussion and conclusion

The research question was as follows: To what extent can the surgical tools be detected in laparoscopic fundoplication videos? With a special focus on efficiently labeling data and the Enseal sealing tool. The results of the network show promising performance based on several evaluation metrics, with an overall precision of 0.901, recall of 0.808, F1 of 0.852, mAP@0.5 of 0.810, and mAP@0.5:0.95 of 0.577. When comparing the results of this research to those of similar studies mentioned in the related work section, it becomes evident that the outcomes of this study are comparable. For example, Choi et al.[44] achieved a mean Average Precision (mAP) of 72.26% on the m2cai16-tool dataset[45], however, they only used annotated data. Yang et al.[46] achieved a higher mAP of 91.58% on the Cholec80 dataset, but again with only annotated data. When comparing my results to the results of Ali et al[50], that trained a semi-supervised student-teacher network on the m2cai16-tool-locations[50] dataset with only

10% annotated (2300 frames), I achieved comparable results. They achieved a higher mAP50 of 90.25% and a lower mAP50:95 of 46.88%.

It is important to note that the results on the whole dataset are relatively lower than those of Enseal and grasper alone. This is due to the fact that there were only a few irrigators (0.5%) and scissors (0.7%) available in the dataset, which was not enough data to learn from. However, the network's ability to accurately detect the Enseal tool remains the most important aspect for further research into electrosurgical devices. With a precision of 0.95 for Enseal detection, the network successfully identifies the Enseal tool. Moreover, this study underscores the significance of employing a semi-supervised and active learning approach. Since labeling data for tool detection is time-intensive, incorporating methods such as pre-labeling, generating pseudo labels, and manually correcting unreliable labels maximizes efficiency and reduces manual effort.

A limitation of this study is the focus on a single type of surgery. In future research concerning tool detection, it would be beneficial to train the network on various other surgical procedures and incorporate a wider range of surgical tools. Additionally, increasing the frames per second could significantly expand the dataset, up to 125 times larger. Also, if I would do this research again, I would train the network for creating pseudo-labels (training 3 in Figure 15) only with a training and validation dataset. I now used a training, validation, and test set, but for this use, a test set is not necessary. I could have added the test set to the training set, which could have improved the network. An additional limitation of the study is the absence of temporal information for the network. Therefore, integrating this temporal aspect into the network has the potential to enhance its performance. For further research, using models that incorporate temporal information could improve the network. An example of a temporal boosted YOLO model is a model created by Alqaysi et al.[63]. The model achieved a mean average precision (mAP) of 92% when evaluated on a Skagen bird dataset. This network outperformed the regular YOLOv4 model, which achieved a mAP of 60% on the same dataset. This model can be used as an example for further research.

Nevertheless, in the present case, the network has already demonstrated high accuracy and is considered adequate for further research purposes into electrosurgical devices.

3. Detection of energy device activation

A MSTCN++ based Enseal tool activation detection with audio labeling for laparoscopic fundoplication surgery videos

Abstract

Introduction Electrosurgical procedures are widely used in modern surgery, offering precise tissue cutting, coagulation, and sealing. However, achieving optimal hemostasis while minimizing tissue damage remains challenging, as excessive energy use can lead to adverse effects and complications. This study aims to get more information about energy usage of electrosurgical devices during surgery. A method is created to acquire the number and length of activations of the Enseal (an electrosurgical tool) in laparoscopic fundoplication surgery videos.
Method The study involves acquiring video and audio data from 10 laparoscopic fundoplication surgeries to train an action recognition network for detecting the activation of the Enseal tool. The audio recordings of the Gen11 (energy generator of the Enseal) are processed to generate a ground truth label for Enseal tool activations. The network architecture comprises an Inflated 3D ConvNet (I3D) feature extractor and a multi-stage temporal convolutional network ++ (MSTCN++). The MSTCN++ generates frame-wise labels from the feature maps and processed audio, enabling action detection.
Results The obtained results are as follows: a frame-wise accuracy of 90.74, a segmental edit distance of 86.86, segmental F1@0.1 of 73.99, F1@0.25 of 70.17, and F1@0.5 of 54.90.
Discussion and Conclusion The results of the action detection network show promise, considering the limited dataset of only 10 videos. However, there is room for improvements, which include increasing the dataset size, enhancing the feature extractor, incorporating data augmentation techniques, and exploring additional spatial information. These enhancements are expected to enhance the developed methodology's accuracy and robustness.

3.1. Introduction

Electrosurgical procedures are widely used in modern surgical interventions, providing precise tissue cutting, coagulation, and sealing.[1] However, ensuring optimal hemostasis quality while minimizing tissue damage remains a significant challenge.[2] Excessive energy utilization can lead to adverse effects such as thermal injury, prolonged healing, and increased postoperative complications.[1,3] Therefore, it is important to create awareness of energy usage during surgery.

It is believed that higher energy usage during surgical procedures may lead to increased harm to the patient.[7] By gaining deeper insights into energy usage with the help of automatic objective assessment, we can obtain more information regarding this belief and its implications. The first step into automatic objective assessment is acquiring the number and duration of tool activations. This can be done by monitoring the direct energy uses from the tool, but it could also be done by monitoring the number of activations and the length of the activations from the laparoscopic surgery videos. Focusing on the latter has the advantage that all surgeries that have already been performed can be used without additional information. Also, it can be used in other hospitals, without the need for additional intraoperative recordings. Deep learning networks could be used to calculate the number and duration of tool activations from laparoscopic surgery videos.

This research aims to address the following question: To what extent can the activation of Enseal (an electrosurgical tool) be recognized in laparoscopic fundoplication videos with the use of artificial intelligence? In order to address this question, a multi-stage temporal convolutional network (MSTCN)

is trained using 10 laparoscopic fundoplication surgery videos. The videos are labeled using audio recordings of the electrosurgical device's generator (the Gen11) acquired during the surgery.

3.2. Related work

Action recognition networks have already made significant advancements, with numerous deep learning models developed. Action detection involves both recognizing and temporally localizing actions or activities in videos. This is useful for identifying actions on a frame-by-frame basis. A few examples of deep learning networks used for action detection include convolutional neural network (CNN)[64], recurrent neural network (RNN), and long-short term memory network (LSTM)[65]. However, CNNs have limitations in capturing temporal features effectively, while RNNs suffer from short-term memory problems, which might result in overlooking crucial information in long input sequences. Although LSTM networks address some of the limitations of RNNs, they require significant computational and memory resources for training due to their multiple gate operations.[66] To tackle these limitations, Sekaran et al.[66] introduced a network called the multiscale temporal convolutional network (MSTCN).

MSTCN has already demonstrated good results on publicly available datasets such as the gtea dataset[67] (accuracy of 80.1 %), the salad50 dataset[68] (accuracy of 83.7%), and the breakfast dataset[69] (accuracy of 67.6 %). Moreover, MSTCN has also been applied specifically to laparoscopic surgery videos, for example on the Cholec80 dataset[47], a dataset of laparoscopic cholecystectomy surgery videos and related surgical phases. Czempiel et al.[70] showed promising results on the Cholec80 dataset, with an accuracy of 88.56, a precision of 81.64, and a recall of 85.24. However, the surgical phases dataset is different from the dataset that is used in this research. Surgical phases are longer actions and there are multiple classes. The dataset used in this research consists of short activations and has only one class. This research will show if the MSTCN++ is able to detect short activations as well.

3.3. Methodology

To address the task of action detection of the Enseal in laparoscopic surgery videos, a two-stage network architecture is employed, composed of a feature extractor and MSTCN++. Figure 17 shows an overview of this method. The MSTCN++ is trained in a supervised way. To create labels for the videos, audio of the energy generator of the Enseal is recorded during surgery. The audio is processed and ground truth labels are created (Section 3.3.1). The second part of the method consists of an inflated 3D ConvNet (I3D) feature extractor, which is responsible for extracting spatial and temporal information from input video frames and generating feature maps as an output (Section 3.3.2). Subsequently, the MSTCN++ network (Section 3.3.3) creates a frame-wise output from the feature maps and ground truth labels from the audio.

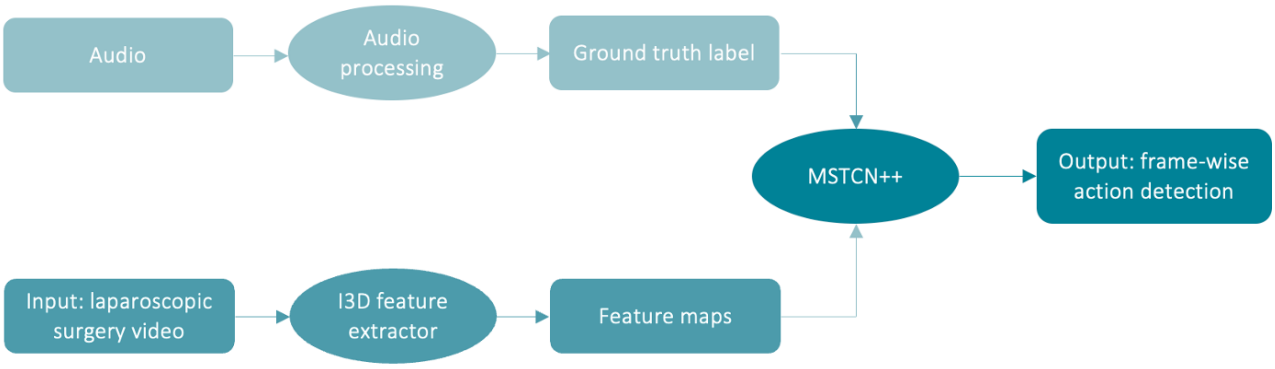
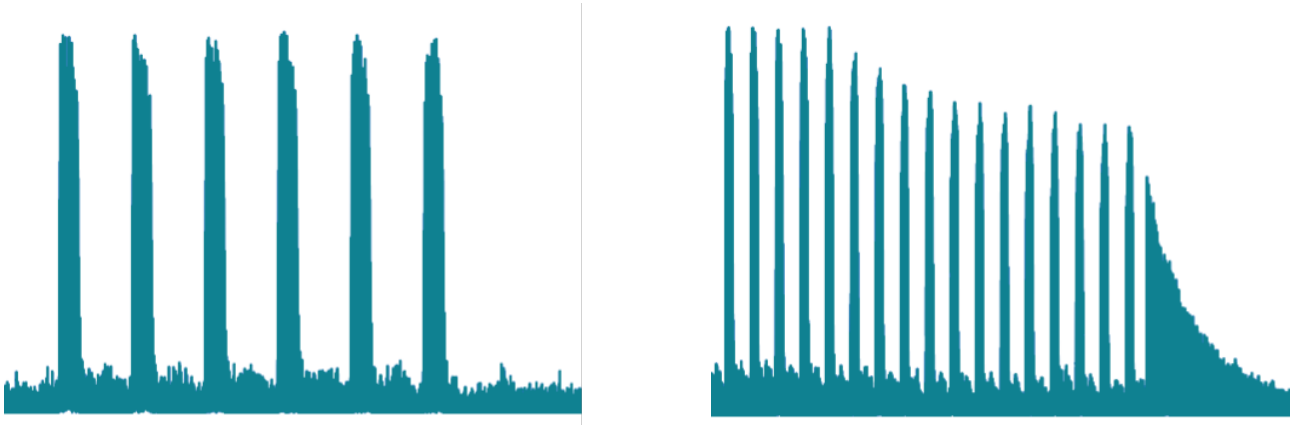


Figure 17: An overview of the method. Light blue: The recorded audio is processed to generate ground truth labels (Section 3.3.1). Dark blue: The method for action detection involves creating feature maps by converting input videos through an I3D feature extractor (Section 3.3.2). Subsequently, the feature maps and ground truth labels are provided as input to the MSTCN++ network (Section 3.3.3). The network is trained with the data and will create a frame-wise prediction of the action detection in the videos.

3.3.1. Generating action detection labels from audio recordings during surgery

During surgery, the Gen11, the energy generator of the Enseal tool, emits bleeping sounds during activation of the Enseal. This audio is recorded and will undergo processing to create a label for the laparoscopic fundoplication surgery videos. In order to use the audio recordings from the surgery as a label for the activation recognition network, the activation sounds need to be filtered from the audio file. There are two distinctive sounds: one is the activation sound, which consists of a series of bleeping sounds (Figure 18.a), and the other is the end sound, which occurs when the full cycle is finished (Figure 18.b).



(a) (b)

Figure 18: (a) An example of an activation signal in the audio, which consists of multiple bleeping sounds. (b) An example of a full-cycle activation signal in the audio consists of multiple bleeping sounds and one end sound.

To filter out the activation sound, the audio signal is first filtered with a band-pass filter with frequencies between 292 and 364, and then the peaks of the signal are identified. To account for potential noise in the signal that could be mistaken for an activation, a constraint is implemented, requiring a minimum of two consecutive bleeping sounds to occur. To filter out the end-cycle sound, the audio signal is filtered with a band-pass filter between 410 and 481. This signal is then convolved with a segment of the audio signal where the end sound is present. A representation of this method can be found in Figure 19.

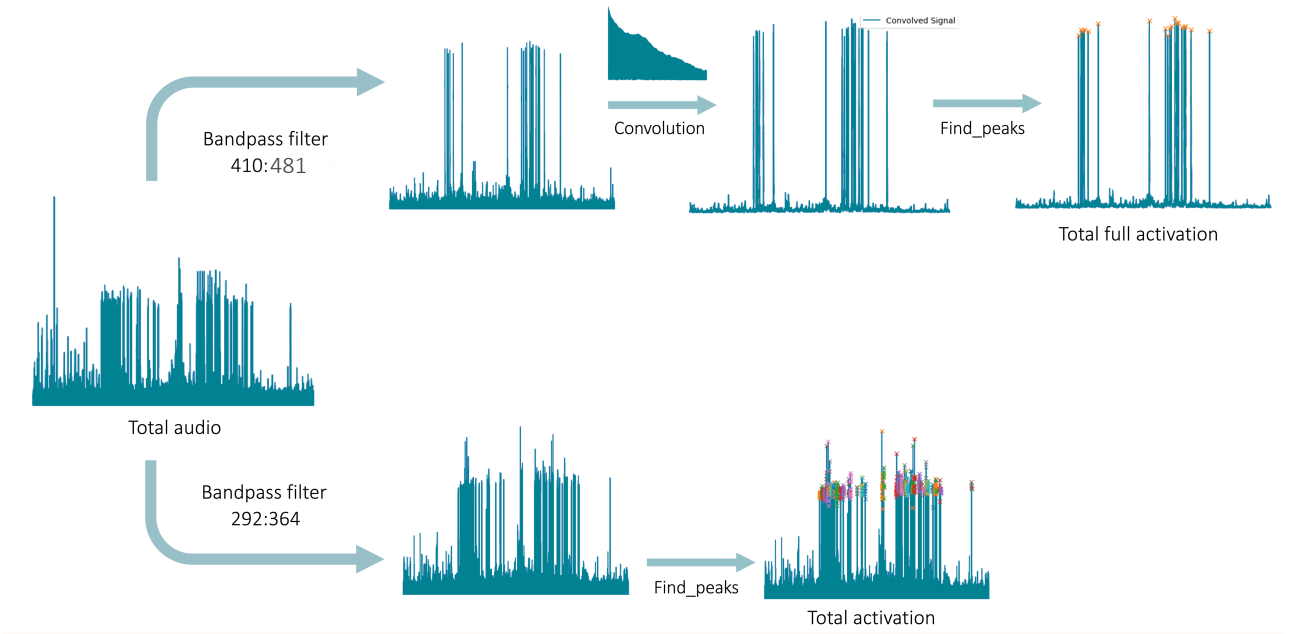


Figure 19: A overview of the audio processing method. The top line illustrates the filtering of the full cycle activations. First, the signal is filtered using a bandpass filter between 410 and 481. Next, a convolution is applied with a fragment of the signal to obtain the end signal. The full cycle activations are then detected using the *find_peaks* function. The bottom line demonstrates the audio being filtered with a bandpass filter between 292 and 364. Subsequently, the activations are identified using the *find_peaks* function, with the constraint that at least two consecutive bleeping sounds must occur. With this method, the number and temporal location of all activations and full activations can be found.

A numpy array will be generated, containing ones between the first and last peak of the activation, while all other values will be set to zeros. In cases where the activation concludes with a full-cycle end sound, an additional one second (or 48,000 samples) of ones will be appended after the last activation. This is done because the tool is still activated during the end activation, despite not being identified as an activation during the normal activation detection process. Figure 20 provides a visual representation of the resulting array.

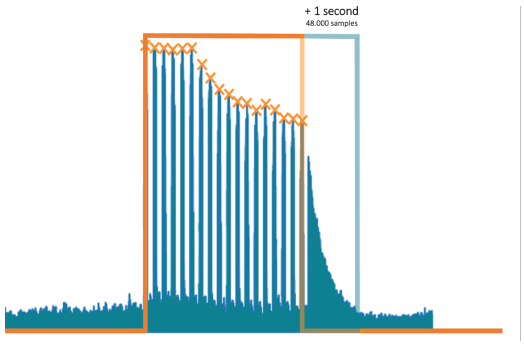


Figure 20: Audio signal of a full cycle activation shown in blue, with orange crosses representing each peak detected using the method to find total activations (described in Figure 19). The orange line indicates the resulting activation array, which spans from the first to the last peak. Because this audio signal ends with the full-cycle end sound, an additional one second (48.000 samples) of ones will be added to the activation array.

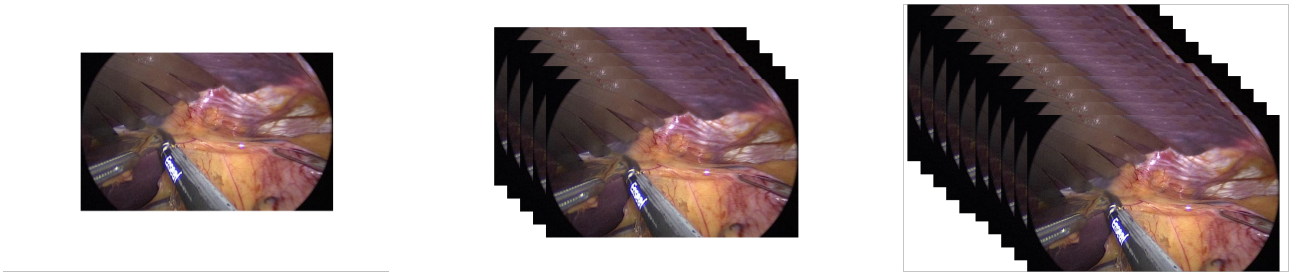
To use the pre-processed audio signal as a label for the corresponding video, the numpy array generated from the audio will be resampled to match the frame rate of the video. The first activation in the audio will be detected automatically and will then be aligned with the first manually selected activation in the video. After processing the audio recording, it can be used as a ground truth label for the feature maps for training the MSTCN++ network.

3.3.2. Video feature extraction with I3D feature extractor

The feature extraction stage plays a crucial role in capturing spatiotemporal information from the input video data. In this work, an I3D network from Github[71], pre-trained on Kinetics 400 (a large-scale publicly available video dataset used for action recognition in videos[72]), is employed as a feature extractor. The I3D network is designed to extract spatial and temporal information and create feature maps. The input for the I3D feature extractor is a video and two input values, step size, and stack size. The number of frames to step before extracting the next features is called the step size. The number of frames of which a feature map will be created is called the stack size (Figure 21). The output of the network is an array with n number of feature maps. The number of feature maps is the total number of frames in the video divided by the step size. Using feature maps instead of full videos as input during MSTCN++ training significantly reduces the required computational power because generating feature maps is now conducted once before training the network instead of each iteration.

For training the MSTCN++ in the following stage, each stack of frames must have a label that reflects whether or not the activation is present within the sequence of frames. The label of a stack of multiple frames needs to be chosen; this can be the label of each frame in the stack. During this research, the chosen frame of the stack will be called the "position of the ground truth". For example, in a stack of 10 frames, the ground truth label may correspond to the first frame, allowing the network to look nine frames in the future. Alternatively, the ground truth label might correspond to the last frame, allowing the network to look nine frames in the past. Finally, if the ground truth label matches the middle frame (i.e., frame 5), the network can look five frames into the past and five frames in the future (Figure 22).

After the videos and the audio recordings are processed, the MSTCN++ can be trained. Experiments will be conducted to determine the importance of stack size and the ground truth label position.



a) Stack size of 1 b) Stack size of 5 c) Stack size of 10

Figure 21: A representation of different stack sizes. The stack size represents the number of frames of the video as input for the feature extractor.

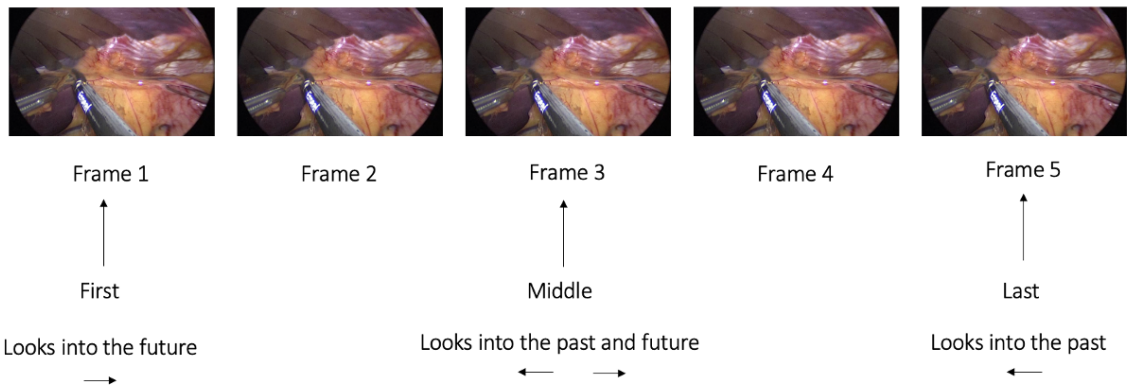


Figure 22: A representation of the position of the ground truth label within the frame sequences. When position 'first' is chosen, the label of the stack frames will correspond to the label of the first frame. This gives the network the ability to look four frames in the future. When position 'middle' is chosen, the label of the stack frames will correspond to the middle frame, frame three in this example. The network is able to look two frames in the past and two frames in the future. When position 'last' is chosen, the label of the stack frames will correspond to the last frame. This will give the network the ability to look four frames in the past.

3.3.3. MSTCN++ Architecture

The feature maps and processed audio recordings will be the input for the MSTCN++. MSTCN++ is an improved version of the MSTCN model, a deep-learning model for video action recognition. The model incorporates temporal convolutions and can process videos at their full temporal resolution, resulting in improved performance.[73] An overview of the MSTCN++ model is shown in Figure 23.

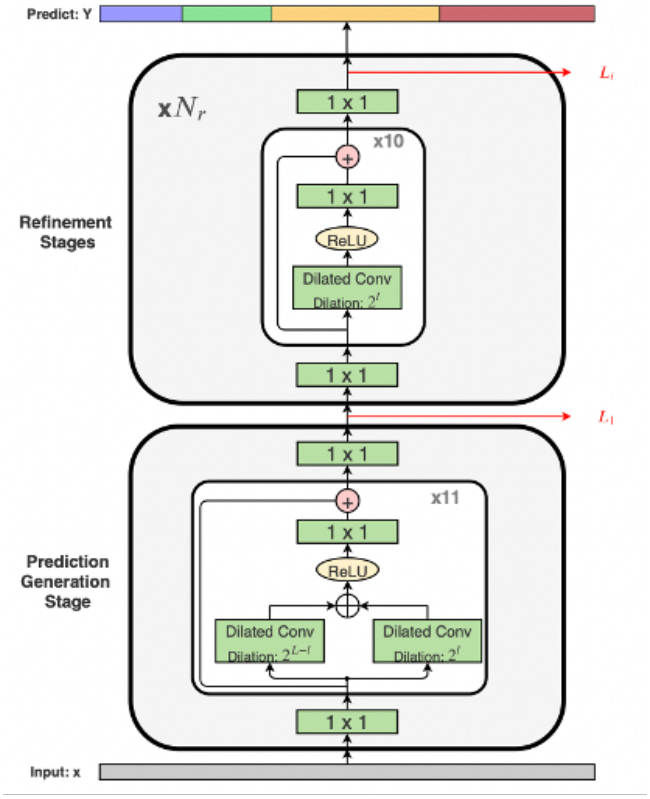


Figure 23: Overview of MS-TCN++. Initially, an SS-TCN model with dual dilated layers is employed to generate an initial prediction. This prediction is further refined in a step-by-step manner through a series of N_r refinement stages. Each refinement stage consists of an SS-TCN with dilated residual layers and a loss layer.[73]

The model is composed of multiple single-stage TCNs (SS-TCN). The first stage, the prediction generation stage, is responsible for generating the initial prediction, whereas the remaining stages refine this prediction. To avoid the limitations of pooling layers, which reduce temporal resolution, and fully connected layers, which force the model to operate on inputs of fixed size and massively increase the number of parameters, each stage only consists of temporal convolutional layers. One stage consists of a 1 x 1 convolutional layer, followed by multiple dilated 1D convolutional layers that increase the receptive field without adding more parameters through extra layers or kernel sizes. The final layers of the stage involve a 1x1 convolution applied to the output of the last dilated convolutional layer, followed by a SoftMax activation to generate the probabilities for the output classes.[73]

The MSTCN++ model introduces multiple enhancements, compared to the MSTCN model.[73]

- Dual dilated layers are added, that combine large and small receptive fields.
- Prediction phase and the refinement phase are decoupled. Therefore, it is possible to have a different number of layers in the prediction and refinement stages.
- And parameters are shared between the refinement stages, which results in a more compact model without compromising performance.

The input of the first stage, the prediction generation stage, are frame-wise features. The following stages only get the frame-wise probabilities as input, without any additional features.[73] The output of the model is a frame-wise prediction. In my case, this is an array of zeros and ones because there is only one class: Enseal tool is activated (1) or not (0).

3.3.4. Loss functions

During training, multiple loss functions will be used, following the approach described in Li et al.'s article[73]. The used loss function is a combination of a classification loss and a smoothing loss. The classification loss consists of a cross-entropy loss. I added a weight to correct for an unbalanced dataset[74]:

$$\mathcal{L}_{cls} = \frac{1}{T}\sum_{t} -w_c \log(y_t,c)$$

Where T is the length of the video, w_c the weight per class c, $y_{t,c}$ the predicted probability for the ground truth label c at time t.

A smoothing loss is added to reduce the over-segmentation errors. Which is a truncated MSE over the frame-wise log probabilities:

$$\mathcal{L}_{T-MSE} = \frac{1}{TC}\sum_{t,c} \tilde{\triangle}^2_{t,c}$$

with $\tilde{\triangle}_{t,c}$ being the modified version of $\triangle_{t,c}$:

$$\tilde{\triangle}_{t,c} = \begin{cases} \triangle_{t,c} & : \triangle_{t,c} \leq \tau \\ \tau & : \text{otherwise} \end{cases}$$

τ represents a threshold value. If $\triangle_{t,c}$ is less than or equal to the threshold, $\tilde{\triangle}_{t,c}$ takes the same value as $\triangle_{t,c}$. Otherwise, it is set to the threshold value, τ. By adding this threshold, this approach reduces the influence of large outliers.[75] $\triangle_{t,c}$ is the absolute difference between the logarithm of the class probability at time t and the logarithm of the class probability at the previous time step $t-1$:

$$\triangle_{t,c} = |\log y_{t,c} - \log y_{t-1,c}|$$

The final loss function for a single stage is a combination of the above-mentioned losses:

$$\mathcal{L}_s = \mathcal{L}_{cls} + \lambda \mathcal{L}_{T-MSE}$$

where λ is a model hyper-parameter to determine the contribution of the different losses. Finally, the complete model is trained by minimizing the sum of the losses over all stages:

$$\mathcal{L} = \sum_s \mathcal{L}_s$$

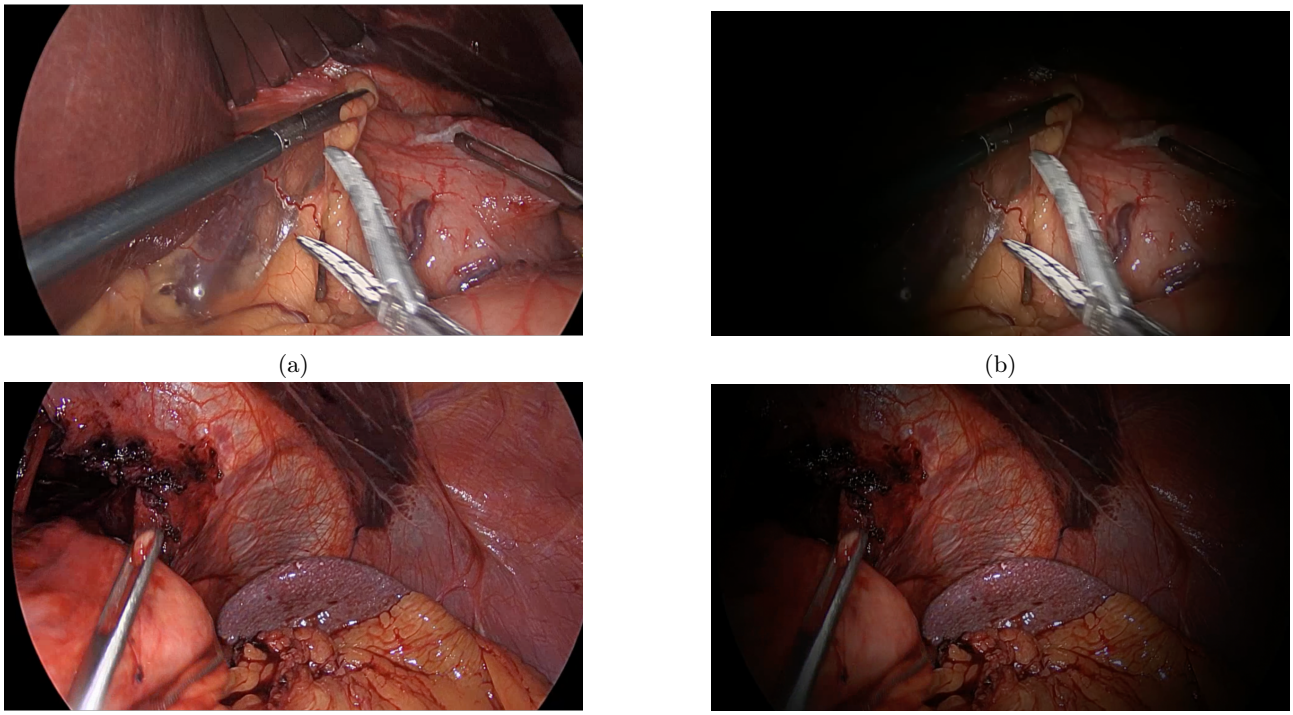
3.3.5. Dataset

The dataset consists of 10 laparoscopic fundoplication surgery videos and corresponding audio recordings, acquired between March 2023 and May 2023 at Meander Medical Center. The videos were acquired with a frame rate of 25 frames per second and have an average length of 1:05:11 hours. The part of the surgery where the Enseal is used will be cut out, as described in Section 3.3.6. This method resulted in ten videos with an average length of 0:25:35 hours. The Enseal tool was activated for approximately 13.8% of the total recording time. The audio during the surgery is recorded using the Voice Memos app[76] on an iPhone 13 positioned on the Gen11, which is the generator of the Enseal tool and emits the activation sound. The audio is recorded at a sample rate of 48000 and subsequently uploaded to Python to process it to labels, as described in Section 3.3.1.

3.3.6. Pre-processing videos

Before the videos of the laparoscopic fundoplication surgery are used as input for the feature extractor, the videos are pre-processed. Fundoplication surgery is comprised of multiple phases. The enseal is only used during the first phase of the surgery. In order to maintain balance within the dataset, only this first phase of the surgery was used for the dataset. This is done by manually cutting the part of the video where the Enseal is detected. Afterward, these videos will be given to the feature extractor as described in Section 3.3.2.

Also, an experiment will be conducted to determine the importance of providing the neural network with additional information about the location of the Enseal. The tool detection network introduced in Chapter 2 will be employed to detect the Enseal tool. The detected information will then be utilized to provide the network with direction regarding the specific areas it should focus on. A Gaussian sphere with a sigma of 150 pixels centered on the midpoint of the bounding box around the Enseal is applied to the frame, resulting in a spotlight effect around the Enseal. The result of this method is shown in Figure 24. In the frames where there is no Enseal detected, a Gaussian sphere with a sigma of 200 pixels will be centered around the midpoint of the whole frame. This is necessary because the tool detection network is not 100% accurate, and there are instances where it fails to detect the Enseal. In such cases, providing the network with the entire frame as input, rather than a black frame, could help. The network will also be trained with these videos as input, and this experiment's results can be found in Section 2.4.3.



(a) (b) (c) (d)

Figure 24: Results before and after Gaussian pre-processing (a) shows a frame of the video without Gaussian pre-processing. (b) shows a frame from the video with Gaussian pre-processing; A Gaussian sphere with a sigma of 150 pixels around the midpoint of the bounding box around the Enseal tool is added to the frame, which results in a spotlight effect. (c) shows a frame of the video without Gaussian pre-processing. (d) shows a frame from the video with Gaussian pre-processing; No Enseal is detected, so a Gaussian sphere with a sigma of 200 pixels is added around the midpoint of the frame.

3.3.7. Evaluation metrics

To evaluate the performance of the network, the metrics described by the authors of the MSTCN++ article[73] will be utilized. This includes frame-wise accuracy (Acc), segmental edit distance(edit), and segmental F1 score, which are calculated at overlapping thresholds of 10%, 25%, and 50%, detonated by F1@10,25,50. Although frame-wise accuracy is commonly used as a metric for action segmentation, this measure is not sensitive to over-segmentation errors.[77] Therefore, segmental edit distance and segmental F1 score will also be used as a metric. The segmental edit distance between two segmentations is the minimum number of insertions, deletions, and substitutions required to transform one segmentation into the other.[78] The segmental F1 score will be calculated by determining whether each predicted action segment is a true or false positive, comparing its IoU with the corresponding ground truth using threshold t (10, 25, or 50%), followed by the computation of the F1 score.[79]

Also, the number of activations predicted, number of correct predicted activations, number of incorrect predicted activations, and number of missed activations are calculated. The number of activations predicted is calculated by finding the peaks, and thus activation, in the output. The number of correct predicted activations is calculated by finding if a frame was activated in the ground truth during an activation in the output. An activation was considered correct if the sum of Ground truth[start_activation_output:end_activation_output] > 0. The number of incorrectly predicted activations was calculated as the difference between the total number of predicted activations and the number of correctly predicted activations. Lastly, the number of missed activations was computed by subtracting the number of correctly predicted activations from the total number of activations present in the ground truth.

A 5 k-fold cross-validation will be performed to ensure robustness and assess the generalization capa-

bility of the network.[80]

3.4. Experiments and Results

3.4.1. Proposed method

This section describes the network hyperparameters and the network's results. The I3D network available on Github[71] is employed as the feature extractor. The step size has been set to 1 and the stack size has been set to 50. As for the network itself, the implementation of the MSTCN++ paper[73] available on GitHub is utilized and modified accordingly.

The hyperparameters used for training the baseline network were determined through a Bayesian hyperparameter optimization, optimizing the following parameters: number of epochs (75), number of layers in the prediction generation stage (8), number of layers in the refinement stages (20), number of refinement stages (5), weight of classification loss (1/number of frames of class c), and learning rate(0.0005).

A 5-fold cross-validation technique was used to evaluate the baseline network's performance. The dataset was divided into five subsets of two videos, with each subset serving as a test set once and the remaining four subsets serving as training sets. Training of the network was done five times, and the mean and standard deviation of the results were calculated.

The evaluation metrics used to quantify the performance of the action detection system include accuracy (Acc), edit distance (Edit), and F1 score at various IoU thresholds. The evaluation results for the 5-fold cross-validation are summarized in the Table below:

Metric	Mean (%)	std(%)
Acc	90.74	1.53
Edit	86.86	4.93
F1@0.10	73.99	8.10
F1@0.25	70.17	8.69
F1@0.50	54.90	10.25

Table 6: Evaluation Results for 5-fold Cross-Validation. Acc= accuracy, Edit= Segmental edit distance, F1@0.10= segmental F1 score with IoU threshold of 10%, F1@0.25= segmental F1 score with IoU threshold of 25%, F1@0.50= segmental F1 score with IoU threshold of 50%.

Figure 25 depicts the output of the network for a part of a video (15% of the whole video), with the network's output represented in orange and the ground truth in light blue. The dark blue areas highlight the regions where the network's predictions overlap with the ground truth, indicating accurate action detection.
Also, the number of activations predicted, number of correct predicted activations, number of incorrect predicted activations, and number of missed activations were calculated and the results can be seen in Table 7.

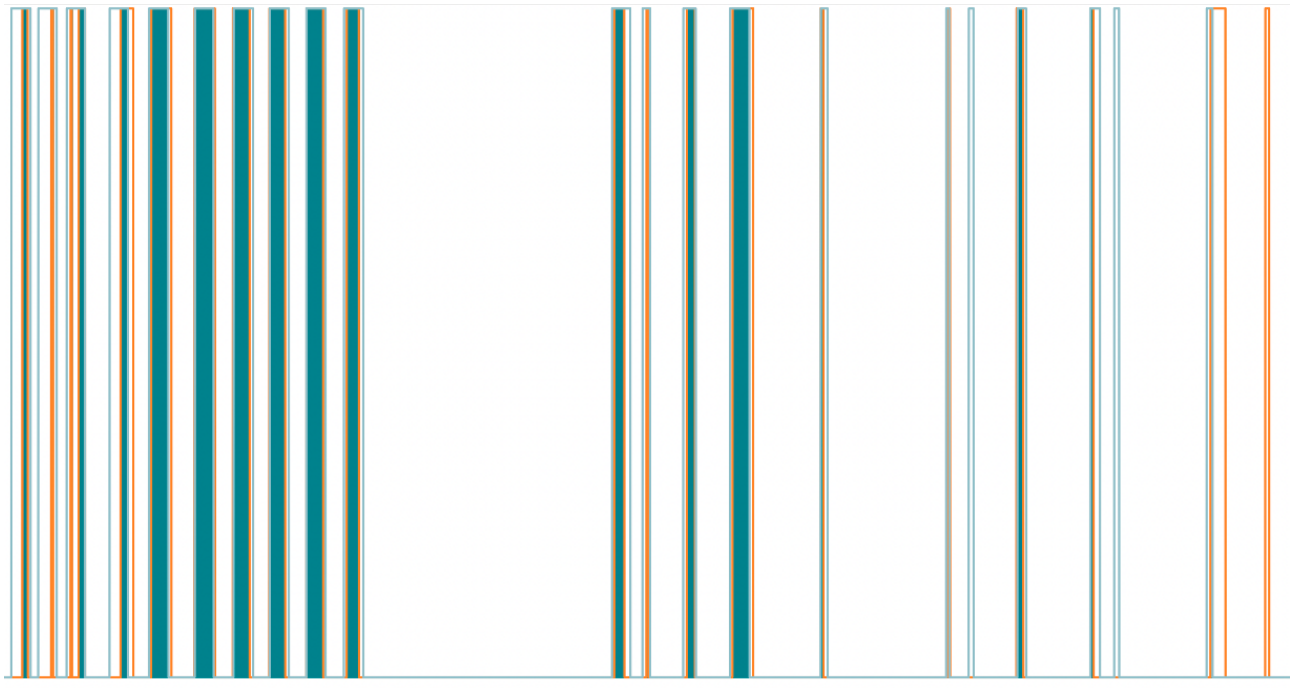


Figure 25: Visualization of action detection results: Network output (orange), ground truth (light blue), and overlapping regions (dark blue) where the network's predictions align with the ground truth.

Category	Mean	Std
Number of activations predicted	67.5	37.3
Number of correct predicted activations	43.3	14.9
Number of incorrect predicted activations	24.2	24.0
Number of missed activations	13.7	12.3

Table 7: A Table of the mean and standard deviation of the number of total activations, correct, incorrect, and missed activations in the prediction. The mean and standard deviation were taken of all patients during 5 k-fold validation.

3.4.2. Experiments

Effectiveness of stack size To show the effectiveness of using a large number for stack size, the network is trained with a stack size of 10, 25, and 50. The training set and the rest of the remaining parameters were the same during both trainings. The outcomes are presented in Table 8. This experiment demonstrates that the network trained with a larger stack size outperforms the networks with a smaller stack size. As expected, the network's performance improves with a higher stack size, as it can gather information from more video frames within a single feature map.

Stack size	Acc	Edit	F1@0.10	F1@0.25	F1@0.50
10	90,54	86,34	73,85	69,64	54,74
25	90,24	82,44	73,80	69,43	53,79
50	**90,74**	**86,86**	**73,99**	**70,17**	**54,90**

Table 8: Results of the experiments for different stack sizes, presented as mean percentages from a 5-fold cross-validation. Acc= accuracy, Edit= Segmental edit distance, F1@0.10= segmental F1 score with IoU threshold of 10%, F1@0.25= segmental F1 score with IoU threshold of 25%, F1@0.50= segmental F1 score with IoU threshold of 50%.

Effectiveness of position of the ground truth frame To demonstrate the significance of selecting the optimal position of the ground truth frame, the network is trained with a dataset with the first, middle, and last frame as the position of the ground truth frame (as described in Section 3.3.2). The training set and the rest of the remaining parameters were the same during both trainings. The results of the experiments are presented in Table 9. Interestingly, the findings indicate that the choice of position does not significantly impact the overall performance of the network. Also, manual inspection of the activation patterns shows similar results, indicating that the performance is comparable. Therefore, regardless of the position of the ground truth frame, the networks produce nearly identical results.

position	Acc	Edit	F1@0.10	F1@0.25	F1@0.50
first	**90,78**	84,47	73,23	69,97	**56,13**
middle	90,74	**86,86**	73,99	**70,17**	54,90
last	90,28	86,58	**74,11**	69,91	55,25

Table 9: Results of the experiments for different positions of the ground truth frame, presented as mean percentages from a 5-fold cross-validation. Acc= accuracy, Edit= Segmental edit distance, F1@0.10= segmental F1 score with IoU threshold of 10%, F1@0.25= segmental F1 score with IoU threshold of 25%, F1@0.50= segmental F1 score with IoU threshold of 50%.

Effectiveness of pre-processing the data
The performed experiment aimed to investigate the impact of adding a Gaussian sphere to the frame on the network's performance. The study compared a network trained with a pre-processed dataset, where a Gaussian sphere was added around the middle point of the Enseal in the video frames, against a network trained without this pre-processing (as described in Section 3.3.6). The training set and the rest of the remaining parameters were the same during both trainings. The results of this experiment are shown in Table 10.

pre-processing	Acc	Edit	F1@0.10	F1@0.25	F1@0.50
Gaussian pre-processing	90,09	**88,35**	**74,25**	70,08	53,13
no pre-processing	**90,74**	86,86	73,99	**70,17**	**54,90**

Table 10: Results of the experiment for different positions of the ground truth frame, presented as mean percentages from a 5-fold cross-validation. Acc= accuracy, Edit= Segmental edit distance, F1@0.10= segmental F1 score with IoU threshold of 10%, F1@0.25= segmental F1 score with IoU threshold of 25%, F1@0.50= segmental F1 score with IoU threshold of 50%.

Surprisingly, the results of the study indicated that pre-processing did not lead to improved performance. The network trained without pre-processing achieved comparable results. These findings suggest that the addition of the Gaussian sphere may not be beneficial for this particular action detection task, and alternative pre-processing techniques or approaches may need to be explored.

3.5. Discussion and conclusion

During the course of this research, the following research question is addressed: To what extent can the activation of Enseal (an electrosurgical tool) be recognized in laparoscopic fundoplication videos with the use of artificial intelligence? To accomplish this, I developed an efficient and reliable method for video labeling, leveraging audio processing from the Gen11 during surgical procedures. These labels

and the feature maps from the I3D feature extractor are used to train a MSTCN++. The neural network results show promising performance based on several evaluation metrics. With a frame-wise accuracy of 90.74, segmental edit distance of 86.86, and segmental F1@0.1 of 73.99, F1@0.25 of 70.17, and F1@0.5 of 54.90.

When comparing the results of this research to those of similar studies mentioned in the related work section, it becomes evident that the outcomes of this study are comparable or better. For example, Czempiel et al.[70] showed promising results on the Cholec80 dataset, with an accuracy of 88.56, and I achieved an accuracy of 90.74. However, you can not really compare these two studies, because the dataset utilized in this research is quite different. In this study, the dataset consists of short actions with only one type of activation. In contrast, the datasets used in the other studies focus on detecting different phases in laparoscopic videos, which involves identifying long actions in the videos and classifying them into multiple classes.
In terms of practical applications in the future, comparing the number of activations and full cycle activations among surgeons provides valuable insights. This analysis can contribute to evaluating surgeons' performance, identifying areas for improvement, and assessing the difficulty level of different patients. Such insights have the potential to significantly enhance surgical training, assessment, and ultimately patient safety. The results, as shown in Table 7 do not meet the required standards for implementation in a hospital setting yet, but there is potential for improvement. For example, increasing the dataset, improving the feature extractor, adding data augmentation, and adding a fourth layer to the video for Enseal detection. These improvements will now be elaborated on.

One limitation of the study is the relatively small dataset consisting of only 10 patients. This sample size may not provide enough variation and diversity to train a neural network effectively. It is important to acknowledge that a larger dataset could improve the performance and generalizability of the network. For example, in the following case: An interesting example was discovered during the analysis of the results. In one video, the audio data indicated that the Enseal device was activated. However, this activation did not occur when tissue sealing was taking place, making it invisible both to humans and possibly neural networks. This error could potentially affect the performance of the network. However, as more data is collected, the impact of this data on the network's performance is expected to decrease. Additionally, while the current research focused on a specific surgery, involved three surgeons in one hospital, and targeted the detection of actions related to the Enseal tool, the generalizability of the findings is limited. Expanding the dataset to include data from multiple surgeries, surgeons, hospitals, and different sealing tools is crucial. By doing so, the network's performance can be validated in diverse surgical contexts, enhancing its generalizability and practical utility.

The following limitation is using a feature extractor trained on the kinetics 400 dataset, which presents some difficulties for this particular dataset. The feature extractor is primarily designed to identify labels such as "salsa dancing" or "grinding meat," which are not directly applicable to the surgical context. Consequently, the extracted features may not accurately represent the actions in the laparoscopic surgery videos. To overcome this limitation, an I3D network can be trained with a dataset of laparoscopic videos. This would enable the feature extractor to learn more relevant and discriminative features specifically tailored to the actions in the surgical context. However, there are also advantages to using a pre-trained feature extractor. For example, pre-trained feature extractors trained on large and diverse datasets (such as kinetics400), have learned general features that are likely to be useful across various domains. By leveraging such pre-trained models, you can benefit from their learned representations and potentially save training time and computational power.[81] Also, it reduces overfitting issues on the specific dataset, which is particularly beneficial when working with smaller datasets, as demonstrated in this research study.[81] Future work should consider training the feature extractor with the available laparoscopic surgery video data to address the advantages and disadvantages of using a feature extractor trained on different types of data.

Additionally, it is worth noting that data augmentation, although not incorporated in this study, can

play a significant role in improving the performance of the network. The network can learn from a more diverse range of examples by artificially expanding the dataset through techniques such as image rotation, flipping, and scaling.[82] Data augmentation was not incorporated in this research because this requires a lot of time. You have to do feature extraction for each data augmentation that you add, and feature extraction can take up more than a couple of days for all videos. Nevertheless, future studies should consider integrating data augmentation techniques to further enhance the network's performance.

Based on the findings of the experiment, it is evident that incorporating a Gaussian spotlight around the Enseal tool does not improve the network's performance. This outcome was a surprise because, upon analyzing the results of the network trained without the Gaussian pre-processing step, it was observed that the network detected actions in a section of the video where only a grasper was present, exhibiting movements similar to the Enseal activation. Excepted is that this action will not be detected when the network only focuses on the Enseal. However, this outcome may be explained by the Enseal detection network's inability to identify the Enseal in all video frames correctly. To address this issue, a strategy was implemented where, in cases where no Enseal was detected, the midpoint of the Gaussian spotlight was positioned at the center of the video frame. However, it is possible that this approach may not have been the most efficient or ideal solution. An alternative method to still recognize the Enseal and have the benefits of this pre-processing step is to add a fourth layer to the input video. This fourth layer consists of a mask of the bounding box around the sealing tool. This could provide additional spatial information and improve the accuracy of action detection. However, the current feature extraction process does not support the extraction of features from 4D data. When the feature extractor is trained with custom data in the future, the inclusion of a fourth layer could be explored to capture more detailed information about the sealing tool's actions.

This research also raises ethical concerns. During the study, audio recordings were made, which had the potential to capture voices if individuals spoke loudly in the operating room. Although these voices and sounds are filtered from the audio in subsequent steps, the complete audio recordings still exist, posing a potential invasion of privacy. This ethical dilemma highlights the significance of the research. Because the network developed during this research eliminates the need for audio recordings in the future. One possible solution to mitigate privacy concerns is to record and directly process audio using a separate device, following the principle of privacy by design.

In conclusion, the results of the neural network for action detection in laparoscopic surgery videos demonstrate promising performance. However, to further enhance the network's accuracy and generalizability, it is crucial to gather more data. Also, training the feature extractor with laparoscopic surgery data, incorporating data augmentation techniques, and exploring the inclusion of additional spatial information are essential steps for future improvements in the field of action detection in laparoscopic surgery.

4. Final conclusion and future improvements

In the previous chapters, I addressed the research question: "To what extent can surgical tool location and activation be recognized in laparoscopic fundoplication videos with the use of artificial intelligence?". From Chapters 2 and 3, we can conclude that tool detection is possible with high precision of 0.901, recall of 0.808, F1 of 0.852, mAP@0.5 of 0.810, and mAP@0.5:0.95 of 0.577. The method for action detection also shows promise, with a frame-wise accuracy of 90.74, segmental edit distance of 86.86, and segmental F1@0.1 of 73.99, F1@0.25 of 70.17, and F1@0.5 of 54.90. But more data is needed to achieve higher accuracy.

As discussed in previous chapters, there are also some improvements that can be made to the previous methods. For tool detection, adding a temporal aspect to the network would enhance its performance. In the case of action detection, multiple improvements are recommended. Primarily, it is essential to acquire more data including data from diverse hospitals, multiple types of surgeries, and a wider range of surgeons. Additionally, further investigation is necessary to optimize the utilization of the tool detection network within this method. For instance, improving the method for creating a Gaussian sphere around the tool in the frames. Finally, training the network to differentiate between complete and incomplete cycles could give more insight into energy device usage.

When the tool and action detection networks are improved, the following step in the research into energy devices can be taken. Referring back to Figure 1 in the introduction, the following step could be to combine the tool and action detection methods with active bleeding detection, which could evaluate the effectiveness of the activation. There has already been some research into bleeding detection, for example, Hua et al.[83], use a faster RCNN with optical flow input and had a precision rate of 0.8373, recall rate of 0.8034, and average precision of 0.6818. Additionally, Hong et al.[84] conducted research into image segmentation-guided intraoperative active bleeding detection. When exploring active bleeding detection, these prior studies could be valuable, but it's also worth considering the use of MSTCN, similar to its application in this master's thesis. The main challenge will be that there are very few instances of active bleeding during fundoplication surgery. As a result, annotating the dataset becomes difficult and the dataset will be unbalanced.

Additionally, exploring alternative data acquisition methods such as energy monitoring devices (shown in light blue in Figure 1) could also provide information about energy usage. Meeuwsen et al.[6] have already studied the differences in the usage of electrosurgical devices with a sensor. They employed a sensor to record the magnitude of electric current delivered to an electrosurgical device at a frequency of 10 Hz. The study revealed that experienced surgeons have a longer activation time than residents (3.01 vs 1.41 seconds, $P < .001$) and a lower number of activations (102 vs 123). These findings underscore the importance of such research. An existing challenge is that this research still relies on the use of a sensor. However, this master thesis suggests that a deep learning network could extract this information solely from surgical videos. Utilizing this technique would make it easier to implement and enhance scalability for widespread adoption. However, an advantage of using an energy monitoring device is that the amount of energy delivered to the patient can be acquired. This data can be correlated with tool activation and effectiveness, ultimately contributing to surgeons' self-improvement and benchmarking.

This research demonstrates the potential of utilizing deep learning networks for the detection of energy devices and their corresponding actions. In the future, this can be used for automatic objective assessment of energy device usage.

5. Appendix

5.1. Appendix A: YOLOv7 improvements

The YOLOv7 network is based on previous YOLO models, but it includes several modifications. These modifications will be explained in the following sections. Firstly, the two architectural changes will be described, followed by an elaboration on the two trainable 'Bag of Freebies' improvements. The term 'Bag of Freebies' refers to enhancements made to the model with the aim of increasing accuracy without incurring additional training costs.[43]

Extended Efficient Layer Aggregation Networks (E-ELAN)

The computational block in the YOLOv7 backbone is named E-ELAN, standing for Extended Efficient Layer Aggregation Networks. The E-ELAN architecture of YOLOv7 enables the model to learn better by using "expand, shuffle, merge cardinality" to achieve the ability to continuously improve the learning ability of the network without destroying the original gradient path.[12]

Model scaling for concatenation-based models

Model scaling is a technique to adapt the size of an existing model to fit different computing devices. This is done by changing key factors such as the resolution of the input image, the number of layers (depth), the number of channels (width), and the number of feature pyramids (stages). The goal is to strike a balance between network parameters, computation, inference speed, and accuracy. Traditionally, model scaling is done by independently adjusting each factor. However, in models based on concatenation, such as DenseNet[85] and VoVNet[86], changing the depth of the model also affects the width of some layers. Therefore, YOLOv7 introduces compound model scaling for a concatenation-based model. When scaling the depth factor of a computational block, the change of the output channel of that block is calculated as well. Then, width factor scaling with the same amount of change on the transition layers is performed, and the result is shown in Figure 26. YOLOv7's compound scaling method can maintain the properties that the model had at the initial design and maintains the optimal structure.[12]

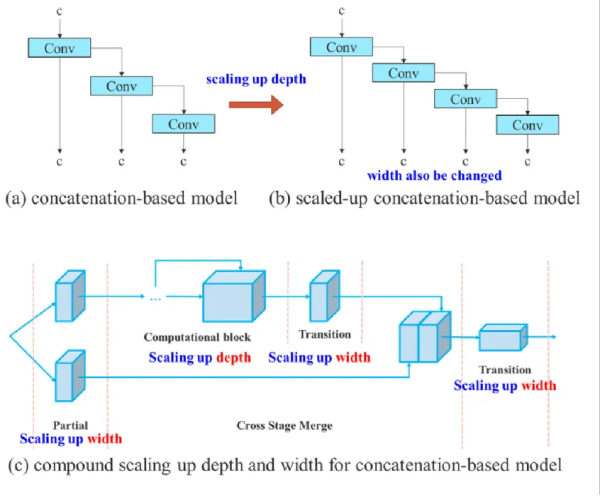


Figure 26: This figure shows model scaling for concatenation-based models. (a) and (b) shows the result of depth scaling, which will cause the input width of the subsequent transmission layer to increase. The YOLOv7 method (c) shows a method in which only the depth needs to be scaled. The remaining of transmission layer is performed with corresponding width scaling. This figure shows model scaling for concatenation-based models. As seen in (a) and (b), increasing the depth of a computational block leads to an increase in the input width of the subsequent transmission layer. The YOLOv7 approach in (c) solves this problem. Only the depth in a computational block needs to be scaled, and the remaining transmission layer is performed with corresponding width scaling. [12]

Planned re-parameterized convolution

Model re-parameterization is a technique used to merge multiple computational modules into one at the time of making predictions. This technique can be seen as an ensemble method and can be divided into two categories: module-level ensemble and model-level ensemble.
In the ***model-level re-parameterization***, there are two common ways to create the final prediction model. One way is to train multiple identical models with different training data, and then average their weights. Another way is to do a weighted average of the weights of models at different iteration numbers.
Module-level re-parameterization is a more recent area of research. This method splits a module into multiple identical or different branches during training and then combines these branches into a single equivalent module during inference. However, not all proposed re-parameterized modules can be applied to different architectures. The YOLOv7 authors used gradient flow propagation paths to analyze the best way to re-parameterize the network. RepConv [87] is a type of re-parameterization that combines 3x3 convolution, 1x1 convolution, and identity connection in one layer. For YOLOv7, RepConv without identity connection is used to design the architecture of planned re-parameterized convolution. [12]

Coarse for auxiliary and fine for lead loss

Deep supervision is a technique used to train deep networks. [88] It adds extra auxiliary heads to the middle layers of the network, and the shallow network uses these auxiliary losses as guidance. The final output is produced by the lead head, and the auxiliary head helps in training. Lead head predictions are used to create coarse-to-fine hierarchical labels because it has a strong learning ability and can provide a better representation of the relationship between the input and target data. This approach can be seen as a form of generalized residual learning, as the auxiliary head learns information that the lead head has already learned, freeing up the lead head to focus on learning residual information. During the process, two different sets of soft labels are generated: a coarse label and a fine label. The fine label is the same as the soft label generated by the lead head's guided label assigner. The coarse label is created by loosening the constraints of the positive sample assignment process, allowing more grids to be treated as positive targets. This is because the learning ability of the auxiliary head is not as strong as the lead head, so to prevent losing important information, the focus is on optimizing the recall of the auxiliary head in object detection tasks. A representation of this method can be found in

Figure 27.[12]

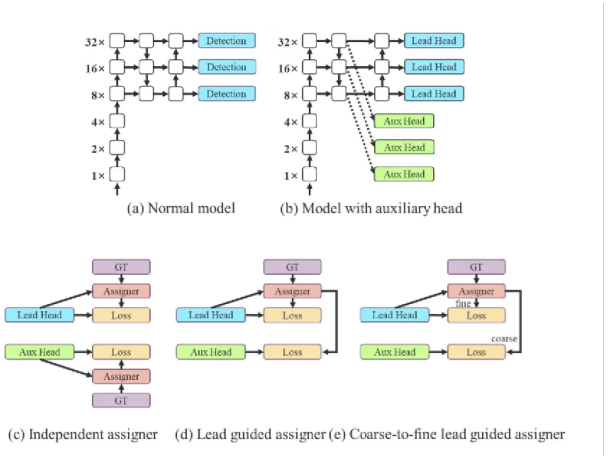


Figure 27: Coarse for auxiliary and fine for lead head label assigner. Compare with normal model (a), the schema in (b) has auxiliary head. Different from the usual independent label assigner (c), we propose (d) lead head guided label assigner and (e) coarse-to-fine lead head guided label assigner. The proposed label assigner is optimized by lead head prediction and the ground truth to get the labels of the training lead head and auxiliary head at the same time.[12]

5.2. Appendix B: hyperparameters during training

lr0: 0.01 # initial learning rate (SGD=1E-2, Adam =1E-3)
lrf: 0.1 # final OneCycleLR learning rate (lr0 * lrf)
momentum: 0.937 # SGD momentum/Adam beta1
weight decay: 0.0005 # optimizer weight decay 5e-4
warmup epochs: 3.0 # warmup epochs (fractions ok)
warmup momentum: 0.8 # warmup initial momentum
warmup biaslr: 0.1 # warmup initial bias lr
box: 0.05 # box loss gain
cls: 0.5 # cls loss gain
cls pw: 1.0 # cls BCELoss positive weight
obj: 0.7 # obj loss gain (scale with pixels)
obj pw: 1.0 # obj BCELoss positive weight
iou t: 0.20 # IoU training threshold
anchor t: 4.0 # anchor-multiple threshold
fl gamma: 0.0 # focal loss gamma (efficientDet default gamma=1.5)
hsv h: 0.015 # image HSV-Hue augmentation (fraction)
hsv s: 0.7 # image HSV-Saturation augmentation (fraction)
hsv v: 0.4 # image HSV-Value augmentation (fraction)
degrees: 20.0 # image rotation (+/- deg)
translate: 0.2 # image translation (+/- fraction)
scale: 0.5 # image scale (+/- gain)
shear: 0.0 # image shear (+/- deg)
perspective: 0.0 # image perspective (+/- fraction), range 0-0.001
flipud: 0.0 # image flip up-down (probability)
fliplr: 0.5 # image flip left-right (probability)
mosaic: 1.0 # image mosaic (probability)
mixup: 0.5 # image mixup (probability)
copy paste: 0.0 # image copy paste (probability)
paste in: 0.0 # image copy paste (probability), use 0 for faster training
loss ota: 1 # use ComputeLossOTA, use 0 for faster training

5.3. Appendix C: Results YOLOv7 for tool detection

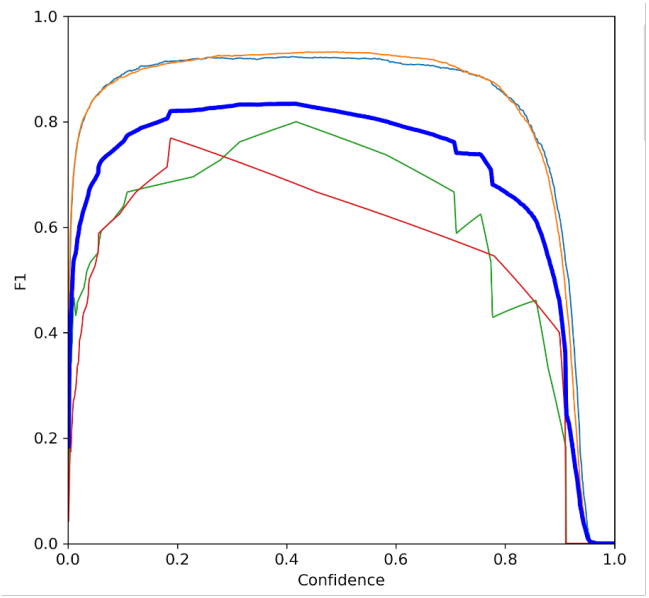


Figure 28: Comparison of F1 Scores for tool detection: Grasper (Orange), Enseal (Light Blue), Irrigator (Green), Scissors (Red), Overall (Dark Blue)

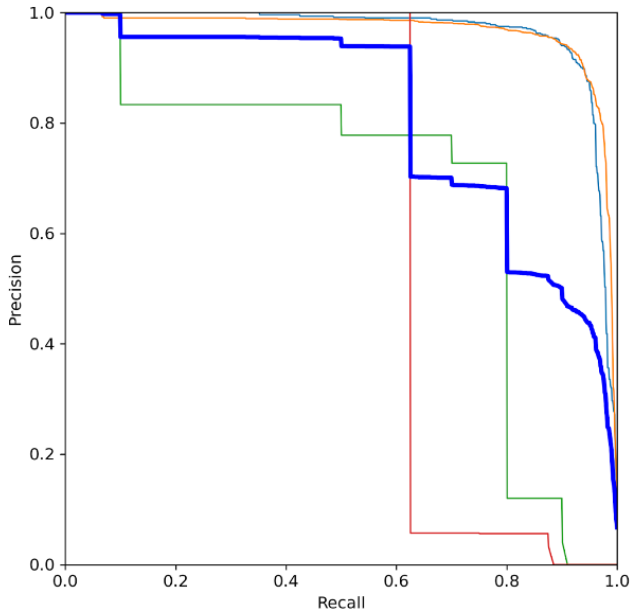


Figure 29: Precision-Recall graph: Grasper (Orange), Enseal (Light Blue), Irrigator (Green), Scissors (Red), Overall (Dark Blue)

References

[1] N. J. Van De Berg, J. J. Van Den Dobbelsteen, F. W. Jansen, C. A. Grimbergen, and J. Dankelman, "Energetic soft-tissue treatment technologies: An overview of procedural fundamentals and safety factors," *Surgical Endoscopy*, vol. 27, no. 9, pp. 3085–3099, 4 2013. [Online]. Available: https://link-springer-com.ezproxy2.utwente.nl/article/10.1007/s00464-013-2923-6

[2] A. I. Brill, "Bipolar electrosurgery: Convention and innovation," *Clinical Obstetrics and Gynecology*, vol. 51, no. 1, pp. 153–158, 3 2008. [Online]. Available: https://journals-lww-com.ezproxy2.utwente.nl/clinicalobgyn/Fulltext/2008/03000/Bipolar_Electrosurgery__Convention_and_Innovation.17.aspx

[3] M. P. Wu, C. S. Ou, S. L. Chen, E. Y. Yen, and R. Rowbotham, "Complications and recommended practices for electrosurgery in laparoscopy," *The American Journal of Surgery*, vol. 179, no. 1, pp. 67–73, 1 2000.

[4] A. Karayiğit, B. Karakaya, D. B. Özdemir, H. Dizen, Özer, B. Ünal, and Özdemir, "Is electrosurgery a revolution? Mechanism, benefits, complications and precautions," *Journal of Pharmaceutical Technology*, vol. 1, no. 3, pp. 60–64, 12 2020. [Online]. Available: https://dergipark.org.tr/en/pub/jpharmtech/issue/56675/830166

[5] "ETHICON GEN11 Generator | Ethicon | J&J MedTech EMEA." [Online]. Available: https://www.jnjmedtech.com/en-EMEA/product/ethicon-gen11-generator

[6] F. C. Meeuwsen, A. C. Guédon, E. A. Arkenbout, M. Van Der Elst, J. Dankelman, and J. J. Van Den Dobbelsteen, "The Art of Electrosurgery: Trainees and Experts," *Surgical Innovation*, vol. 24, no. 4, pp. 373–378, 8 2017. [Online]. Available: https://journals-sagepub-com.ezproxy2.utwente.nl/doi/10.1177/1553350617705207

[7] I. Alkatout, T. Schollmeyer, N. A. Hawaldar, N. Sharma, and L. Mettler, "Principles and safety measures of electrosurgery in laparoscopy," *JSLS : Journal of the Society of Laparoendoscopic Surgeons*, vol. 16, no. 1, pp. 130–139, 1 2012. [Online]. Available: https://pubmed-ncbi-nlm-nih-gov.ezproxy2.utwente.nl/22906341/

[8] "ENSEAL® X1 | Bipolar Device | Jaw Tissue Sealer | Ethicon." [Online]. Available: https://www.jnjmedtech.com/en-US/product/enseal-x1-large-jaw-tissue-sealer

[9] J. E. Richter, "Gastroesophageal Reflux Disease Treatment: Side Effects and Complications of Fundoplication," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 5, pp. 465–471, 5 2013.

[10] C. L. McKnight and B. Burns, "Pneumothorax," *StatPearls*, 2 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK441885/http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1745823

[11] M. L. Wani, A. G. Ahangar, F. A. Ganie, S. N. Wani, and N. D. Wani, "Vascular Injuries: Trends in Management," *Trauma Monthly*, vol. 17, no. 2, p. 266, 2012. [Online]. Available: /pmc/articles/PMC3860641//pmc/articles/PMC3860641/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3860641/

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors."

[13] N. N. Massarweh, N. Cosgriff, and D. P. Slakey, "Electrosurgery: history, principles, and current and future uses," *Journal of the American College of Surgeons*, vol. 202, no. 3, pp. 520–530, 3 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16500257/

[14] J. M. Davison and N. M. Zamah, "Electrosurgery: Principles, Biologic Effects and Results in Female Reproductive Surgery," *The Global Library of Women's Medicine*, 2009. [Online]. Available: http://www.glowm.com/section-view/heading/Electrosurgery:Principles, BiologicEffectsandResultsinFemaleReproductiveSurgery/item/21

[15] A. I. Brill, "Electrosurgery: principles and practice to reduce risk and maximize efficacy," *Obstetrics and gynecology clinics of North America*, vol. 38, no. 4, pp. 687–702, 12 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22134017/

[16] G. A. Vilos and C. Rajakumar, "Electrosurgical Generators and Monopolar and Bipolar Electrosurgery," *Journal of Minimally Invasive Gynecology*, vol. 20, no. 3, pp. 279–287, 5 2013.

[17] D. J. Hay, "Electrosurgery," *Surgery (Oxford)*, vol. 26, no. 2, pp. 66–69, 2 2008.

[18] P. A. Sutton, S. Awad, A. C. Perkins, and D. N. Lobo, "Comparison of lateral thermal spread using monopolar and bipolar diathermy, the Harmonic Scalpel and the Ligasure," *The British journal of surgery*, vol. 97, no. 3, pp. 428–433, 3 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20101674/

[19] E. G. Chekan, M. A. Davison, D. W. Singleton, J. Z. Mennone, and P. Hinoul, "Consistency and sealing of advanced bipolar tissue sealers," *Medical Devices (Auckland, N.Z.)*, vol. 8, p. 193, 4 2015. [Online]. Available: /pmc/articles/PMC4408972//pmc/articles/PMC4408972/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4408972/

[20] "Coagulation electrosurgical unit - GEN11 - Ethicon - thermofusion / gynecological surgery / ultrasonic." [Online]. Available: https://www.medicalexpo.com/prod/ethicon/product-74984-531999.html

[21] "Acid Reflux (GER & GERD) in Adults," *National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)*, 11 2015. [Online]. Available: https://www.niddk.nih.gov/health-information/digestive-diseases/acid-reflux-ger-gerd-adults/all-content

[22] P. J. Kahrilas, N. J. Shaheen, and M. F. Vaezi, "American Gastroenterological Association Institute technical review on the management of gastroesophageal reflux disease," *Gastroenterology*, vol. 135, no. 4, pp. 1392–1413, 2008.

[23] M. Parker, "Book Review: Krause's Food and Nutrition TherapyMahanLKEscott-StumpS. Krause's Food and Nutrition Therapy. 12th ed. Philadelphia: Saunders; (2007). 1376 pp, $$149.95. ISBN: 978-1-4160-3401-8." *Nutrition in Clinical Practice*, vol. 25, no. 3, pp. 314–314, 6 2010. [Online]. Available: http://dx.doi.org/10.1177/0884533610362901

[24] Mayo Clinic, "Gastroesophageal reflux disease (GERD)," 2022.

[25] P. O. Katz, L. B. Gerson, and M. F. Vela, "Guidelines for the diagnosis and management of gastroesophageal reflux disease," *American Journal of Gastroenterology*, vol. 108, no. 3, pp. 308–28, 2013.

[26] "Los Angeles Classification of Reflux Esophagitis Accurate ... | GrepMed." [Online]. Available: https://www.grepmed.com/images/13202/egd-losangeles-esophagitis-diagnosis-grading

[27] W. A. Draaisma, J. P. Ruurda, R. C. Scheffer, R. K. Simmermacher, H. G. Gooszen, H. G. Rijnhart-De Jong, E. Buskens, and I. A. Broeders, "Randomized clinical trial of standard laparoscopic versus robot-assisted laparoscopic Nissen fundoplication for gastro-oesophageal reflux disease," *British Journal of Surgery*, vol. 93, no. 11, pp. 1351–1359, 10 2006. [Online]. Available: https://dx.doi.org/10.1002/bjs.5535

[28] G. W. Holcomb and S. D. St. Peter, "Error traps and safety steps when performing a laparoscopic Nissen fundoplication." *Semin Pediatr Surg.*, vol. 28, no. 3, pp. 160–163, 6 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:150096013#id-name=S2CID

[29] R. C. Minjarez and B. A. Jobe, "Surgical therapy for gastroesophageal reflux disease," *GI Motility Online*, 5 2006. [Online]. Available: https://www.nature.com/gimo/contents/pt1/full/gimo56.html

[30] J. A. Broeders, F. A. Mauritz, U. A. Ali, W. A. Draaisma, J. P. Ruurda, H. G. Gooszen, A. J. Smout, I. A. Broeders, and E. J. Hazebroek, "Systematic review and meta-analysis of laparoscopic Nissen (posterior total) versus Toupet (posterior partial) fundoplication for gastro-oesophageal reflux disease," *The British journal of surgery*, vol. 97, no. 9, pp. 1318–1330, 9 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20641062/

[31] N. Linzberger, S. V. Berdah, P. Orsoni, D. Faucher, J. C. Grimaud, and R. Picaud, "Fundoplicature postérieure cœlioscopique pour reflux gastro-œsophagien : Résultats à moyen terme," *Annales de Chirurgie*, vol. 126, no. 2, pp. 143–147, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11291677/

[32] J. Mardani, L. Lundell, and C. Engström, "Total or posterior partial fundoplication in the treatment of GERD: results of a randomized trial after 2 decades of follow-up," *Annals of surgery*, vol. 253, no. 5, pp. 875–878, 5 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21451393/

[33] X. Du, J. M. Wu, Z. W. Hu, F. Wang, Z. G. Wang, C. Zhang, C. Yan, and M. P. Chen, "Laparoscopic Nissen (total) versus anterior 180° fundoplication for gastro-esophageal reflux disease: A meta-analysis and systematic review," *Medicine*, vol. 96, no. 37, 9 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28906412/

[34] "Nissen fundoplication - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Nissen_fundoplication

[35] K. Seeras, K. Bittar, and M. A. Siccardi, "Nissen Fundoplication," *StatPearls*, 7 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK519521/

[36] R. Garry, "Laparoscopic surgery," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 20, no. 1, pp. 89–104, 2 2006.

[37] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 275–279, 12 2019. [Online]. Available: https://onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/full/10.1049/htl.2019.0064https://onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/abs/10.1049/htl.2019.0064https://ietresearch-onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/10.1049/htl.2019.0064

[38] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, vol. 2018-January, pp. 691–699, 5 2018.

[39] G. Humm, R. L. Harries, D. Stoyanov, and L. B. Lovat, "Supporting laparoscopic general surgery training with digital technology: The United Kingdom and Ireland paradigm," *BMC surgery*, vol. 21, no. 1, p. 123, 3 2021. [Online]. Available: https://bmcsurg.biomedcentral.com/articles/10.1186/s12893-021-01123-4

[40] A. Chan-Hon-Tong, "Deep learning pipeline for grid classification, segmentation, detection and related problems," 2017. [Online]. Available: https://hal.science/hal-01412086v4

[41] M. Wu, H. Yue, J. Wang, Y. Huang, M. Liu, Y. Jiang, C. Ke, and C. Zeng, "Object detection based on RGC mask R-CNN," *IET Image Processing*, vol. 14, no. 8, pp. 1502–1508, 6 2020. [Online]. Available: https://onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/full/10.1049/iet-ipr.2019.0057https:

//onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/abs/10.1049/iet-ipr.2019.0057https://ietresearch-onlinelibrary-wiley-com.ezproxy2.utwente.nl/doi/10.1049/iet-ipr.2019.0057

[42] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10 778–10 787, 11 2019. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1911.09070v7

[43] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 4 2020. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/2004.10934v1

[44] B. Choi, K. Jo, S. Choi, and J. Choi, "Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 1756–1759, 9 2017.

[45] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2 2016. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1602.03012v2

[46] Y. Yang, Z. Zhao, P. Shi, and S. Hu, "An Efficient One-Stage Detector for Real-Time Surgical Tools Detection in Robot-Assisted Surgery," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12722 LNCS, pp. 18–29, 2021. [Online]. Available: https://link-springer-com.ezproxy2.utwente.nl/chapter/10.1007/978-3-030-80432-9_2

[47] "Cholec80 Dataset | Papers With Code." [Online]. Available: https://paperswithcode.com/dataset/cholec80

[48] A. Vardazaryan, J. Marescaux, and N. Padoy, "Weakly-Supervised Learning for Tool Localization in Laparoscopic Videos."

[49] S. Targ, D. Almeida, and K. L. Enlitic, "Resnet in Resnet: Generalizing Residual Architectures," 3 2016. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1603.08029v1

[50] M. Ali, G. Ochoa-Ruiz, and S. Ali, "A semi-supervised Teacher-Student framework for surgical tool detection and localization." [Online]. Available: https://github.com/Mansoor-at/Semi-supervised-surgical-tool-detection.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 779–788, 6 2015. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1506.02640v5

[52] ——, "You Only Look Once: Unified, Real-Time Object Detection." [Online]. Available: http://pjreddie.com/yolo/

[53] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 4 2018. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1804.02767v1

[54] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A New Backbone That Can Enhance Learning Capability of CNN," pp. 390–391, 2020.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 9 2015.

[56] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path Aggregation Network for Instance Segmentation," pp. 8759–8768, 2018.

[57] "A Comprehensive Guide To Object Detection Using YOLO Framework — Part I | by Pratheesh Shivaprasad | Towards Data Science." [Online]. Available: https://towardsdatascience.com/object-detection-part1-4dbe5147ad0a

[58] "Real-time Object Detection with YOLO, YOLOv2 and now YOLOv3 | by Jonathan Hui | Medium." [Online]. Available: https://jonathan-hui.medium.com/real-time-object-detection-with-yolo-yolov2-28b1b93e2088

[59] "Open Source Data Labeling | Label Studio." [Online]. Available: https://labelstud.io/

[60] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. J. Cardoso, V. Cheplygina, B. Cimini, G. S. Collins, K. Farahani, B. Glocker, P. Godau, F. Hamprecht, D. A. Hashimoto, D. Heckmann-Nötzel, M. M. Hoffmann, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, A. Karargyris, A. Karthikesalingam, B. Kainz, E. Kavur, H. Kenngott, J. Kleesiek, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K. G. M. Moons, H. Müller, F. Nickel, J. Petersen, G. Polat, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C. S. Gutierrez, J. Schroeter, A. Saha, S. Shetty, B. Stieltjes, R. M. Summers, A. A. Taha, S. A. Tsaftaris, B. van Ginneken, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, A. Kopp-Schneider, P. Jäger, and L. Maier-Hein, "Common Limitations of Image Processing Metrics: A Picture Story," 4 2021. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/2104.05642v4

[61] "Transfer learning with frozen layers - Ultralytics YOLOv8 Docs." [Online]. Available: https://docs.ultralytics.com/yolov5/tutorials/transfer_learning_with_frozen_layers/

[62] "COCO - Common Objects in Context." [Online]. Available: https://cocodataset.org/#home

[63] H. Alqaysi, I. Fedorov, F. Z. Qureshi, and M. O'nils, "A Temporal Boosted YOLO-Based Model for Birds Detection around Wind Farms," *Journal of Imaging 2021, Vol. 7, Page 227*, vol. 7, no. 11, p. 227, 10 2021. [Online]. Available: https://www.mdpi.com/2313-433X/7/11/227/htmhttps://www.mdpi.com/2313-433X/7/11/227

[64] C. A. Ronao and S. B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 10 2016.

[65] M. Ullah, H. Ullah, S. D. Khan, and F. A. Cheikh, "Stacked Lstm Network for Human Activity Recognition Using Smartphone Data," *Proceedings - European Workshop on Visual Information Processing, EUVIP*, vol. 2019-October, pp. 175–180, 10 2019.

[66] S. R. Sekaran, Y. H. Pang, G. F. Ling, and O. S. Yin, "MSTCN: A multiscale temporal convolutional network for user independent human activity recognition," *F1000Research*, vol. 10, p. 1261, 5 2021. [Online]. Available: /pmc/articles/PMC9989544//pmc/articles/PMC9989544/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9989544/

[67] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3281–3288, 2011. [Online]. Available: https://paperswithcode.com/dataset/gtea

[68] "50 Salads | CVIP." [Online]. Available: https://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/

[69] "Serre Lab » The Breakfast Actions Dataset." [Online]. Available: https://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/

[70] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12263 LNCS, pp. 343–352, 2020. [Online]. Available: https://link-springer-com.ezproxy2.utwente.nl/chapter/10.1007/978-3-030-59716-0_33

[71] "I3D (RGB + Flow) - Video Features Documentation." [Online]. Available: https://iashin.ai/video_features/models/i3d/

[72] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," 5 2017. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1705.06950v1

[73] S. Li, Y. A. Farha, Y. Liu, M.-M. Cheng, and J. Gall, "MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 3570–3579, 6 2020. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/2006.09220v2

[74] H. Phan, K. Yamamoto, T. H. Phan, and K. Yamamoto, "Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses," 6 2020. [Online]. Available: https://arxiv.org/abs/2006.01413v1

[75] N. Petra, D. De Caro, V. Garofalo, E. Napoli, and A. G. Strollo, "Truncated binary multipliers with variable correction and minimum mean square error," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 6, pp. 1312–1325, 2010.

[76] "Voice Memos on the App Store." [Online]. Available: https://apps.apple.com/us/app/voice-memos/id1069512134

[77] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka, "Alleviating Over-segmentation Errors by Detecting Action Boundaries." [Online]. Available: https://github.com/yiskw713/asrf

[78] D. Pucher and W. G. Kropatsch, "Segmentation Edit Distance," *Proceedings - International Conference on Pattern Recognition*, vol. 2018-August, pp. 1175–1180, 11 2018.

[79] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection."

[80] Y. Jung and J. Hu, "A K-fold Averaging Cross-validation Procedure," *Journal of nonparametric statistics*, vol. 27, no. 2, p. 167, 4 2015. [Online]. Available: /pmc/articles/PMC5019184//pmc/articles/PMC5019184/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5019184/

[81] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 1 2021.

[82] R. Aluvalu, K. Kotecha, N. Cauli, and D. Reforgiato Recupero, "Survey on Videos Data Augmentation for Deep Learning Models," *Future Internet 2022, Vol. 14, Page 93*, vol. 14, no. 3, p. 93, 3 2022. [Online]. Available: https://www.mdpi.com/1999-5903/14/3/93/htmhttps://www.mdpi.com/1999-5903/14/3/93

[83] S. Hua, J. Gao, Z. Wang, P. Yeerkenbieke, J. Li, J. Wang, G. He, J. Jiang, Y. Lu, Q. Yu, X. Han, Q. Liao, and W. Wu, "Automatic bleeding detection in laparoscopic surgery based on a faster region-based convolutional neural network," *Annals of Translational Medicine*, vol. 10, no. 10, pp. 546–546, 5 2022. [Online]. Available: /pmc/articles/PMC9201197//pmc/articles/PMC9201197/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9201197/

[84] S. Hong, S. Hong, J. Jang, K. Kim, J. Hyung, M.-K. C. Visionai, and S. Korea, "Amplifying action-context greater: Image segmentation-guided intraoperative active bleeding detection," 2022. [Online]. Available: https://sghong977.github.io/bleeding/

[85] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, 8 2016. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1608.06993v5

[86] Y. Lee, J. W. Hwang, S. Lee, Y. Bae, and J. Park, "An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2019-June, pp. 752–760, 4 2019. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1904.09730v1

[87] M. Soudy, Y. Afify, and N. Badr, "RepConv: A novel architecture for image scene classification on Intel scenes dataset," *International Journal of Intelligent Computing and Information Sciences*, vol. 0, no. 0, pp. 1–11, 4 2022.

[88] C. Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," *Journal of Machine Learning Research*, vol. 38, pp. 562–570, 9 2014. [Online]. Available: https://arxiv-org.ezproxy2.utwente.nl/abs/1409.5185v2