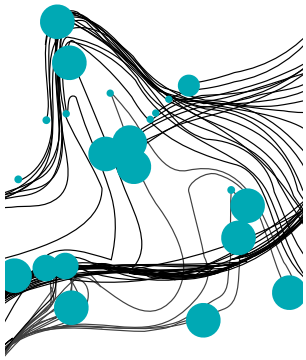
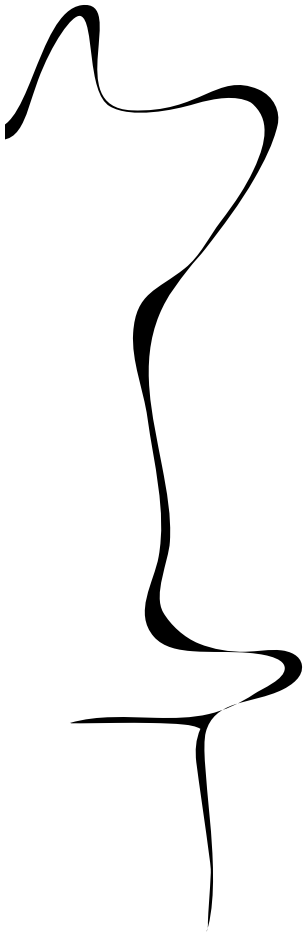


.72498

DMB

DATA MANAGEMENT AND BIOMETRICS



AUTOMATIC PHASE RECOGNITION FOR SURGICAL VIDEO ANALYSIS: A CROSS-MODAL MULTI-VISUAL CUE APPROACH

Jelise Schokker

MASTER'S ASSIGNMENT

Committee:
dr. E. Talavera Martínez
dr.ing. G. Englebienne

January, 2025

2025DMB0002
Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Automatic Phase Recognition for Surgical Video Analysis: A Cross-Modal Multi-Visual Cue Approach

Jelise Schokker
n.j.schokker@student.utwente.nl
University of Twente
The Netherlands

ABSTRACT

Surgical phase recognition is an important field in medical image analysis for improving surgical safety, efficiency, and training. Phase recognition involves predicting the different phases of a surgery using machine learning methods. This research proposes a framework for improving phase recognition using a cross-modal multi-visual cue approach. The proposed model leverages the video frames in combination with the extracted descriptors of tool presence, segmentation masks, and action labels. The ablation study conducted in this study shows the best configuration of visual cues is image data and action triplets, achieving an accuracy of 0.826 and F1 score of 0.871 on the Cholec80 dataset, compared to 0.792 and 0.844 for the baseline model. The model also outperforms state-of-the-art models on the HeiChole benchmark dataset with an F1 score of 0.796 and an accuracy of 0.732. These results demonstrate the effectiveness of integrating multi-visual cues for phase recognition, offering a promising direction for improving surgical workflow analysis.

1 INTRODUCTION

Surgical phase recognition is an important area of research in the field of medical image analysis and computer-assisted intervention (CAI). It has several benefits that include making surgeries safer, more efficient, and more effective [56]. Phase recognition involves using machine learning methods to predict the phases and steps in surgical videos and can offer real-time support and feedback for laparoscopic surgeries. Around 15 million laparoscopic surgeries are performed each year, so improving this procedure could have a great potential benefit [2].

There are various applications where the online recognition of the phases in surgical videos offers important benefits. By recognizing and tracking the progression of surgical steps, surgical phase recognition systems can alert the surgical team about deviations or anomalies from the standard workflow [21, 56, 69]. This can potentially reduce errors and increase the patient's safety. Recognizing deviations from the standard workflow additionally allows standardizing surgical procedures across different surgeons and hospitals [69]. Surgical phase recognition also offers a structured way to evaluate and quantify surgical performance, which is very useful for training new surgeons [16]. It can give them a clear and comprehensive view of the surgical process and provide instant feedback on how they are doing [32, 33, 66]. Postoperative analysis and feedback are two other benefits of surgical phase recognition. The data gathered by phase recognition models during surgeries can be reviewed afterwards in order to see what went well and what could be improved, helping surgical teams to get better over time [69]. By managing and predicting the workflow efficiently, phase recognition can help shorten surgeries and better organize OR scheduling. This can lead to more efficient usage of operating rooms decreasing overall healthcare costs [56].

There are several features that can be used for the representation of surgical videos. These features can be used as the input for a phase recognition network to improve the performance compared to generic feature extraction methods. Tool presence is one of these features. It involves recognizing and identifying which tools are present in a frame and sometimes also includes segmentation of the location of different tools. Tool presence can provide valuable information on the current phase in videos, since the different phases often use distinct tools. Semantic segmentation in surgical videos is another descriptor that can be used as input to a recognition model. This focuses on segmenting the entire surgical field instead of only the tools present in the frame. Although the surgical field does not change significantly during laparoscopic surgeries, the position of the organs can be useful. Especially in surgeries such as cholecystectomies or hysterectomies, in which a specific organ is removed, segmentation could help the phase prediction [50]. Information on the current actions of the surgeons can also be a valuable input feature for phase recognition. Each surgical phase usually has certain actions that are performed within it, so recognizing these actions can help the final prediction network. Action recognition is usually performed by creating an action triplet: three words describing the subject, verb, and object of the action.

Attention mechanisms are a popular tool in machine learning. Attention is designed to help models focus on the most relevant parts of an input [58]. It was originally developed for natural language processing tasks but has proved very effective in fields like image and video analysis as well [27]. By focusing on key features while minimizing distractions, attention allows models to handle complex and variable data more effectively. In surgical phase recognition, attention can be especially useful. Surgical videos often include challenges like motion blur, changes in lighting, and overlapping actions. Attention mechanisms can help focus on the important details while ignoring irrelevant or noisy information. This focus can make predictions more reliable for surgical video analysis tasks. Cross-modal attention takes this a step further by bringing together information from different sources, like tool usage or segmentation maps.

1.1 Problem Statement

Current approaches to surgical phase recognition often use a single modality as input. However, many potential input modalities exist that can possibly improve the performance of phase recognition networks. While single-modality approaches have shown

good performance, they each have their limitations. Surgical videos are complex, with challenges such as changes in lighting, camera movements, and overlapping actions. Single-modality methods sometimes fail to capture contextual information that can be useful to accurately recognize surgical phases [41]. A method that combines these different types of input modalities in a structured way could provide models with a more complete understanding of the surgical progress. This research explores an approach that uses multiple inputs together, aiming to improve the accuracy and reliability of surgical phase recognition.

1.2 Research Questions

The goal of this research is to address the following research questions:

(1) For the task of automatic phase recognition in surgical videos, what computer vision-based methods exist and which input features do these methods use?

This question investigates existing methodologies for phase recognition in surgical videos. The first step involves exploring the prediction networks and input features used for this task.

(2) How can we leverage different visual descriptors extracted from surgical videos in a multi-visual cue framework for surgical phase recognition?

To address this question, this research presents the pipeline of a multi-visual cue model. The model integrates visual descriptors like tool presence detection, segmentation maps, and action recognition. The process involves creating multiple model configurations to determine the optimal design, input features, and architectures.

(3) What effect does the inclusion of different descriptors in a multi-visual cue model have on the performance of automatic phase recognition in surgical videos compared to single visual cue models and the state of the art?

Finally, to evaluate the performance of the multi-visual cue model, the performance for each descriptor will be compared to a baseline single visual modality model in an ablation study. By comparing it with the baseline model, this analysis can show whether incorporating multiple input features is indeed a valuable addition to surgical phase prediction models. The best configuration of the multi-visual cue model will be compared to the performance of state-of-the-art models to show how it compares in the field and whether it offers significant improvements over existing approaches for the task of surgical phase prediction. This model configuration will also be evaluated on an additional dataset to show the robustness of the model.

1.3 Proposed Approach

To answer the research questions, this study proposes a method for surgical phase recognition that combines multiple visual modalities, in order to try to address the limitations of single-modality models. The approach involves building upon an existing phase recognition network, MTRCNet [24], and integrating tool presence, segmentation maps and action labels. The different modalities are combined using a cross-modal attention mechanism. This mechanism allows the model to connect relevant information from different sources, focusing on key details while ignoring irrelevant parts of the input. For example, segmentation maps can provide information about organ positions, while tool presence and action labels add context about the current surgical tasks. Together, these modalities form a comprehensive representation of the surgical workflow. The proposed model is evaluated on the Cholec80 dataset, comparing its performance against existing single-modality methods. This comparison aims to show whether combining multiple visual cues improves phase recognition accuracy and robustness. The ablation study is conducted to analyze the contributions of each modality to the overall performance of the model.

The main contributions of this research include:

(1) Design of a cross-modal attention-based framework for surgical phase recognition that integrates multiple visual cues, including tool presence, segmentation maps, and action triplets.
(2) Ablation study showing the impact of combining different input modalities on phase recognition performance, including an ablation study to analyze the contribution of each individual modality.

1.4 Thesis Outline

This report starts by exploring the scientific background in Section 2. This includes the medical background, the relevant terminologies and an overview of available datasets. Section 2.4 discusses the existing methods for surgical phase recognition and a comparison of their results. The methodology for the proposed approach and the extraction of each input modality is outlined in Section 3. Section 4 discusses the implementation details, training procedure, baseline models and relevant evaluation metrics. The results of the ablation studies as well as the comparisons with state-of-the-art models are presented in Section 5. Finally, Section 6 contains a comprehensive discussion of all results and recommendations for future work, followed by a conclusion of this research in Section 7.

2 SCIENTIFIC BACKGROUND

2.1 Medical Background

Laparoscopic surgeries are minimally invasive surgeries where surgical instruments are inserted into the stomach or pelvic area through several small incisions [26]. A laparoscopic camera is also inserted and gives the surgeons real-time footage of the surgical field without making large incisions. Because laparoscopic surgeries are done using a camera, it makes them very accessible for surgical phase recognition, as the camera is already present and a great deal of footage exists.

There are several challenges when it comes to using laparoscopic surgery videos for automatic phase recognition. The laparoscopic camera is not static, so there is a great deal of fluctuation in the scenes that can be seen during the surgery, including motion blur [56]. Videos can also contain fast camera movements, smoke blocking the view, and an array of different tools. Additionally, blood stains on the lens can obscure or distort the image that the laparoscopic camera captures. Different types of surgery can be performed laparoscopically, common procedures include cholecystectomy (removal of the gallbladder) and hysterectomy (removal of the uterus and cervix) [10, 38].

2.1.1 Terminology. When it comes to the task of recognizing the different phases in a surgery, several terms are used across different papers. Some papers refer to it simply as *phase recognition* [11, 53, 56, 63, 65]. Other authors use the term *workflow recognition* to describe the recognition of the phases in surgery [23, 25, 56, 69]. A couple of papers also refer to the process as *step recognition* [18, 52]. However, this term can be ambiguous because it is also used to describe the recognition of actions, or a shorter period within a larger phase.

In this report, the term *workflow recognition* will be used to refer to the overlapping research field which focuses on gaining insights into the processes that happen during surgeries. The actual task of predicting the phases in a surgical video will be referred to as *surgical phase recognition*, in line with the terminology used by Twinanda et. al [56].

2.2 Multi-Visual Cue Models

Multi-visual cue models for phase recognition are models that use various types of visual information or cues to predict the current phase of a video. The usage of multiple visual inputs can potentially increase the performance of prediction models, especially in surgical videos which sometimes include complex or blurry frames. The features discussed in Section 2.4.6 can be used as input to a multi-visual cue model. There are several choices that are important for the performance of such a model.

Firstly, it is important to decide when the different features will be combined for the final prediction. In multi-modal models, such as a multi-visual cue model, there are two main categories for data fusion: early fusion and late fusion [43]. Late fusion involves first training a model for each individual feature, and then combining the outputs at the end, either by concatenating the outputs or averaging them in classification tasks. In early fusion, the features are combined before being fed into the prediction model, and this way only a single model is trained.

2.3 Datasets

Various benchmark datasets that are commonly used for automated phase recognition in surgical videos include Cholec80 [56], HeiChole [59], M2CAI16 [53, 56] and AutoLaparo [60]. These datasets are discussed in the following sections. Figure 1 shows several example frames from the videos in each dataset. Table 1 shows a brief overview of the datasets and their technical details together with the class labels for each phase in the datasets.

2.3.1 Cholec80. The Cholec80 dataset [56] is a publicly available dataset which contains 80 videos of laparoscopic cholecystectomy surgeries. Laparoscopic cholecystectomy is a surgery often used for surgical phase recognition because it is relatively common with a highly standardized process that is not entirely linear [53]. This means that some of the phases of the surgery can be performed in a different order depending on the surgeon and the specific situation. The videos in the Cholec80 dataset are labeled with phase annotations for seven distinct phases. The presence of 7 different tools is also annotated in the videos, provided by a senior surgeon at Strasbourg Hospital. These tools are *Bipolar*, *Clipper*, *Grasper*, *Hook*, *Irrigator*, *Scissors*, and *Specimen bag*. An example of each tool is shown in Figure 2a. The videos are captured at 25 frames per second (FPS) and have been downsampled to 1 FPS. They have a frame resolution of 1920x1080 or 854x480 pixels, and an average duration of 39 minutes [33].

Figure 3 shows the distribution of frames per video in the cholec80 dataset. It can be seen that there is a large variety in the number of frames in each video. Some videos are shorter with around 1000 frames, while longer videos have up to 6000 frames. Figure 4 shows the number of frames per phase in the surgery. This graph shows that *Calot triangle dissection* (P2) and *Gallbladder dissection* (P4) are the phases with the most frames in the dataset, and the remaining phases have a more similar number of total frames.

2.3.2 HeiChole. The HeiChole benchmark dataset [59] was created for the 2019 Endoscopic Vision sub-challenge for surgical workflow and skill analysis. It consists of 33 laparoscopic cholecystectomy videos, captured at three different surgical centers. 24 of these videos are part of the publicly available training set, and the remaining 9 videos belong to a private test set. Each video has been annotated with a phase, tool, and action label, and parts of the video also include annotations for skill classification. The phase labels are the same as the phases used in the Cholec80 dataset.

2.3.3 M2CAI16. The M2CAI16 workflow dataset [53] was created for the M2CAI challenges in 2016 and consists of 41 cholecystectomy surgeries. Some of the videos are taken from the Cholec80 dataset. Similar to the Cholec80 dataset, these videos are captured at 25 FPS with a resolution of 1920x1080. All videos in the dataset have been segmented into eight phases by experienced surgeons at the Hospital of Strasbourg and the Hospital Klinikum Rechts der Isar in Munich. The M2CAI16 tool dataset was created for the same challenge and consists of 15 laparoscopic cholecystectomy videos. The tools annotated in this dataset are the same as those in the Cholec80 dataset.

2.3.4 AutoLaparo. The AutoLaparo dataset [60] consists of 21 videos of laparoscopic hysterectomy surgeries. Each video is labeled with phase annotations for 7 phases. A senior gynaecologist and a specialist with experience in hysterectomies performed these annotations. Tools and key anatomy are also annotated in this dataset. The key anatomy is the uterus, and the 4 annotated tools are *Grasping forceps*, *LigaSure*, *Dissecting and grasping forceps*, and *Electric hook*. An example of each tool can be found in Figure 2b. The videos have a standard resolution of 1920x1080 pixels and are captured at 25 FPS. Because of the different levels of difficulty involved in the surgeries, the videos range in length from 27 to 112 minutes, with an average duration of 66 minutes. Similar to cholecystectomy surgeries, hysterectomies also have a relatively long duration which can make the task of phase recognition difficult.

2.4 Related Works

The task of surgical phase recognition has already been discussed in many papers. The different types of models that were previously used for the task are discussed, and several different architectures for each of the model types are presented.

2.4.1 CNN-based Models for Surgical Phase Recognition. Convolution is a linear operation used for feature extraction, where a matrix of numbers called a kernel is applied across the input [62].

Dataset	Cholec80 [56]	HeiChole [59]	M2CAI16 [53]	AutoLaparo [60]
Nr. of videos	80	33	41	21
Annotations	Phase, tool	Phase, action, tool, skill	Phase	Phase, motion, tool
FPS	25	25	25	25
Resolution	1920x1080, 854x480	960x540, 1920x1080, 720,576	1920x1080	1920x1080
Phases	1. Preparation	1. Preparation	1. Trocar placement	1. Preparation
	2. Calot triangle dissection	2. Calot triangle dissection	2. Preparation	2. Dividing ligament and peritoneum
	3. Clipping and cutting	3. Clipping and cutting	3. Calot's triangle dissection	3. Dividing uterine vessels and ligament
	4. Gallbladder dissection	4. Gallbladder dissection	4. Clipping and cutting of cystic duct and artery	4. Transecting the vagina
	5. Gallbladder packaging	5. Gallbladder packaging	5. Gallbladder dissection	5. Specimen removal
	6. Cleaning and coagulation	6. Cleaning and coagulation	6. Gallbladder packaging	6. Suturing
	7. Gallbladder retraction	7. Gallbladder retraction	7. Cleaning and coagulation of liver bed (haemostasis)	7. Washing
			8. Gallbladder retraction	

Table 1: An overview of the technical details and class labels of the benchmark datasets for surgical phase recognition. For each dataset, the general statistics are summarized and the phases are explicitly stated.

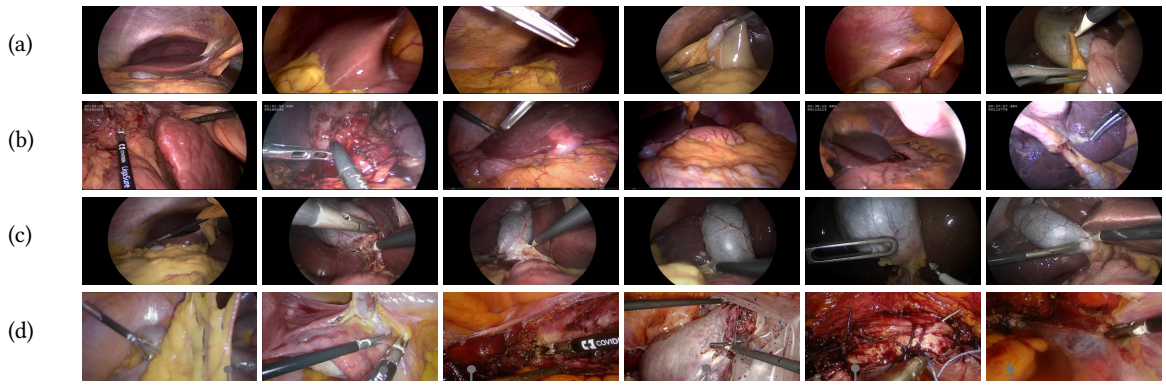


Figure 1: Example frames of videos in the dataset in (a) Cholec80 [56], (b) M2CAI16-tool [53], and (c) AutoLaparo [60].

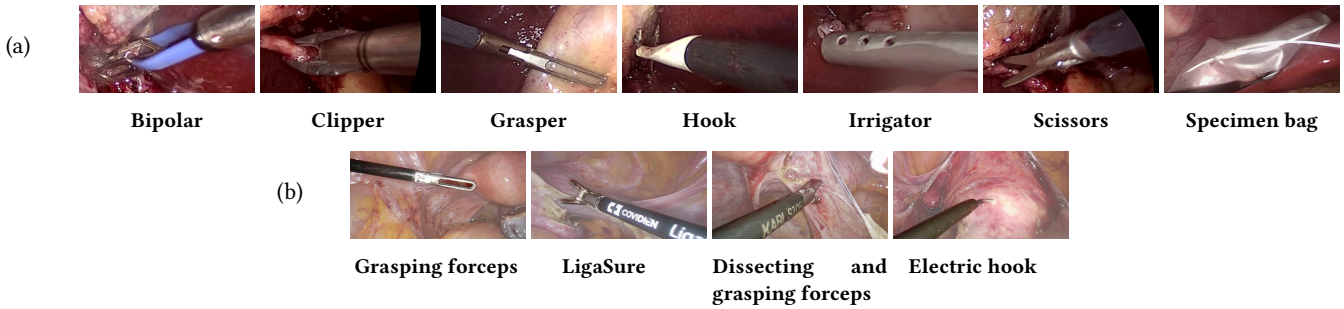


Figure 2: Example of each of the tools annotated in (a) Cholec80 [56] and M2CAI16-tool [53] and (b) AutoLaparo [60].

This operation is used in a Convolutional Neural Network (CNN), a network designed to recognize patterns using convolutional layers, pooling layers [39]. CNNs perform better than other types of networks for many image-related tasks. The main advantages of CNNs over fully connected networks are the use of shared weights in the form of kernels, and the translation invariance. The pooling layers reduce the dimensionality of these maps. This makes the network more computationally efficient and improves the feature extraction. Finally, the fully connected layers calculate the output classifications.

Twinanda et. al [56] used a CNN to perform the task of workflow recognition in surgical videos, a novelty at the time. Their proposed *EndoNet* architecture is an extension of the AlexNet architecture [28]. The approach fine-tunes the recognition network in a multi-task manner, simultaneously carrying out the task of phase recognition and tool presence. After the main set of convolutional layers, the network is temporarily split into two simultaneous branches. The tool detection branch of the architecture identifies the presence of specific surgical tools and outputs a binary vector for each tool. The second branch performs phase recognition by integrating features extracted by the CNN with the tool detection output to classify the surgical phase in each frame. It uses a Hierarchical HMM to learn the temporal information from the videos. The branches are trained together in an end-to-end manner.

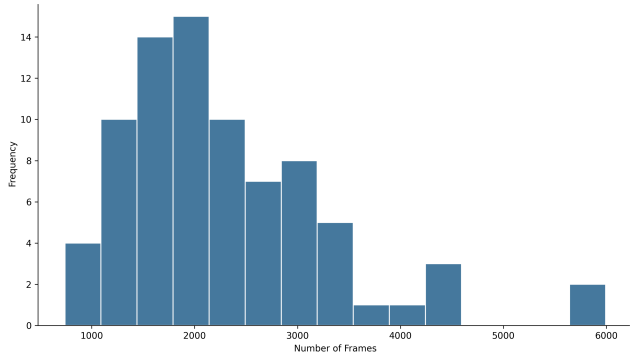


Figure 3: Histogram showing the distribution of video frames per video in the Cholec80 dataset [56]. (Bin size is 15)

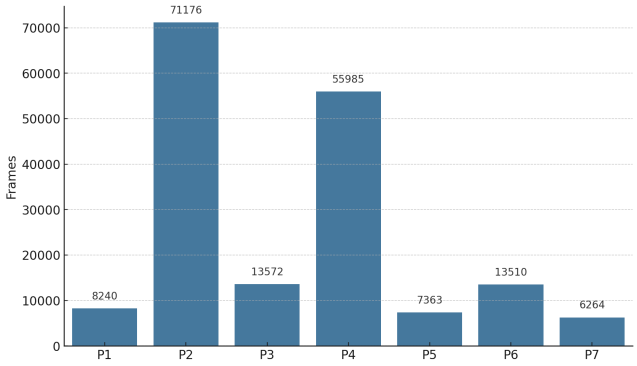


Figure 4: Distribution of video frames per phase in the Cholec80 dataset [56], for the phases *Preparation* (P1), *Calot triangle dissection* (P2), *Clipping and cutting* (P3), *Gallbladder dissection* (P4), *Gallbladder packaging* (P5), *Cleaning and coagulation* (P6), and *Gallbladder retraction* (P7).

2.4.2 RNN-, LSTM-, and GRU-based Models for Surgical Phase Recognition. Recurrent Neural Networks (RNNs) are a type of neural network that is very useful for sequential data processing like text and speech [51]. The RNN architecture uses cycles that transmit information back into itself. This allows them to keep information from previous inputs, which makes RNNs very useful for tasks where context and structure are important. The looped network structure of an RNN processes every element of the input sequence while maintaining a memory of previous inputs in its internal state. This memory influences the current output and the output of future states. RNNs are often used in natural language processing, speech recognition and time series analysis, because the data in these tasks is usually sequential. However, RNNs have several weaknesses, including their susceptibility to vanishing gradients, which makes it hard for them to learn long-term dependencies in sequences. Also, RNNs are inherently sequential in nature, limiting their ability to be parallelized during training, which can lead to longer training times compared to most other neural network architectures.

Long Short-Term Memory (LSTM) networks [19] were introduced to solve several of the problems encountered with standard RNNs. LSTM networks consist of memory cells that can store information over longer sequences. Each memory cell has three components. The input gate specifies how much of the new input will be stored in the memory cell. The forget gate determines which information should be removed from the cell. Finally, the output gate chooses how much data the memory cell should use to compute the hidden state. Together, these components make sure only the important information is kept and everything else is discarded. Because of the special memory cells, they are better at keeping track of contextual information spread out over longer sequences than RNNs are.

Gated Recurrent Units (GRUs) [6] are a model similar to LSTMs but with a simpler architecture [51]. They only consist of two gates: the update gate and the reset gate [51]. The update gate determines which information to keep and the reset gate controls what information is discarded. Although the architecture is simpler than an LSTM, GRUs are still good at capturing long-term dependencies.

Zisimopoulos et al. [69] propose the *DeepPhase* architecture for tool and phase recognition. The model first extracts tool presence information from the surgical videos using a CNN-based architecture with residual connections, specifically the ResNet-152. Then, two types of data are used in the phase recognition step: the binary tool classification and tool features gathered from the last pooling layer of the ResNet-152. The purpose of training on the tool features as well was to capture tool motion and orientation data and visual cues such as colour and lighting that could improve phase identification. Two RNN-based networks are trained for phase recognition, an LSTM and a GRU network, both using cross-entropy loss.

The *SV-RCNet* (Surgical Video Recurrent Convolutional Network) [23] also uses LSTMs for surgical phase recognition. It starts by extracting discriminative visual descriptors using ResNet-50. These descriptors are then fed into the LSTM network to capture the temporal information. The SV-RCNet is trained in an end-to-end manner and the ResNet and LSTM network are optimized jointly. The authors also propose the Prior Knowledge Inference (PKI) system, which can improve the accuracy and consistency of the phase predictions by considering the structured and predictable nature of surgical workflows.

Jin et al. present the *MTRCNet-CL* (Multi-Task Recurrent Convolutional Network with Correlation Loss) [24]. The architecture is similar to other methods using LSTMs for phase recognition. It starts with a 50-layer residual convolutional network to extract visual features. Because the network uses a multi-task approach, it is split into two branches. The first branch performs tool presence recognition using a single fully connected layer and sigmoid activation. The phase recognition branch consists of an LSTM network. The entire model is trained end-to-end, jointly optimizing the tool and phase predictions.

Another method that uses an LSTM for surgical phase recognition is the *State-Preserving LSTM* [48]. In contrast with many other prediction models, this approach focuses mainly on learning phase transitions. This is done by first extracting tool presence information using ZIBNet [47]. The state-preserving LSTM is then trained on the tool predictions to learn the evolution of tool transitions between surgical phases. This approach focuses only on the presence of tools as the primary data source, but suggests tool localisation could be a valuable addition.

Jalal et al. [22] proposed a *CNN and LSTM model*, a spatio-temporal deep learning approach for surgical phase recognition. The network architecture which consists of a CNN model and three LSTM models. The CNN part uses a ResNet-50 to extract visual features from the input frames. These features are fed into the first LSTM model, LSTM-clip, which extracts temporal information within short clips. LSTM-video and LSTM-phase then capture the temporal dependencies across the entire video, and use fully-convolutional layers to predict the phases.

2.4.3 Temporal Convolutional Network-based Models for Surgical Phase Recognition. Temporal Convolutional Networks (TCNs) were introduced by Lea et al. [30]. As opposed to previous methods that often used CNNs for spatial information and RNN-like architectures for temporal information, the TCN architecture captures both levels hierarchically. TCNs use an encoder-decoder framework. The encoder uses one-dimensional convolutional layers followed by activation and max-pooling layers. After pooling, channel-wise normalisation is applied. The decoder is similar to the encoder structure but uses upsampling instead of pooling. The convolutions in a TCN are all causal, which means they do not use information from future time-steps. The input to a TCN can be a sequence of variable length.

The *TeCNO* architecture [7] used a dilated, causal Multi-Stage Temporal Convolutional Network (MS-TCN) for the first time in surgical workflow analysis. The network starts by using a ResNet-50 for feature extraction, followed by stacked predictor stages. These predictor stages use dilated, causal convolutions for a large receptive field and online phase prediction respectively. The model is trained using cross-entropy loss after each prediction stage. The use of multiple stages allows the model to refine the prediction of the early stages.

2.4.4 Transformer-based Models for Surgical Phase Recognition. The paper *Attention is All You Need* [58] originally introduced the Transformer architecture. Their architecture uses self-attention to map queries, keys, and values to outputs across long input sequences. This solves the problem encountered in earlier models like RNNs and LSTMs, which struggled with capturing relations over long sequences. Because the order of data points in a sequence is very important when dealing with sequential data, Transformers use positional encoding. Positional encoding adds relative or absolute positional information to each token in the sequence.

Gao et al. [12] present *Trans-SVNet*, a transformer-based model that uses both spatial and temporal information to predict the surgical phase. The network first uses ResNet-50 [17] to extract the spatial embeddings. These spatial embeddings are then used to extract temporal embeddings using a Temporal Convolutional Network (TCN) called TeCNO [7]. In the next step, the network outputs a prediction based on the temporal and spatial embeddings. This is done using an aggregation model consisting of two Transformer layers. The aggregation model is trained using cross-entropy loss.

Zhang et al. [65] propose a Transformer-based architecture for phase recognition in surgical videos. Their model is called C-ECT (Cross-Enhancement Causal Transformer) and is created by modifying the ASFormer [64]. The network first performs feature extraction using EfficientNetV2 [55]. The features at each timestep are saved and used as the input to the C-ECT model. This model is similar to the ASFormer with a few modifications. The values V in the cross-attention layer in the decoder come from the self-attention of the corresponding encoder layer. This aligns the decoder with the encoder's self-attention layer so the network can learn both global and local information continuously. The model is trained using cross-entropy loss and smooth loss. C-ECT is a causal transformer model because it uses a method that makes sure that the attention operation is only performed on past information, and not on future information.

Another method that uses Transformers for phase recognition is *LoViT* (Long Video Transformer) [32]. The approach starts by extracting spatial embeddings from the input frames. These spatial embeddings are used as the input to the first Transformer layer. This layer captures small-local features. The second Transformer layer uses these small-local features to obtain the large-local features. Then, a global temporal feature aggregator captures the long-range dependencies more efficiently using an Informer module [68] Finally, all features are combined in a fusion head, which outputs a phase transition map and a phase label for each frame.

The *SKiT* architecture [33] introduces the novel *Key-recorder* which can efficiently capture global information. The Key-recorder provides a way to record key events in a video sequence. First, the spatial feature extractor extracts spatial features from the input frames. These features are put into a Transformer-based local temporal feature aggregator that captures local temporal features, and the Global Key-recorder which records key events by pooling information across frames using a max operation. Finally, a fusion head integrates the local and global features to perform the final phase prediction.

In *Friends Across Time: Multi-Scale Action Segmentation Transformer for Surgical Phase Recognition* [66], the authors introduce the MS-ASCT (MultiScale Action Segmentation Causal TransformerMS-ASCT) for online surgical phase recognition. The model builds upon the ASFormer [64], a Transformer for action segmentation. The self-attention layer in the encoder and the cross-attention layer in the decoder of the ASFormer are modified into multi-scale temporal self-attention and multi-scale temporal cross-attention. Because the MS-ASCT model has to perform online surgical phase recognition, causal sliding window attention is used. Similar to the C-ECT model, MS-ASCT is also trained using cross-entropy and smooth loss.

2.4.5 Vision Transformer-based Models for Surgical Phase Recognition. Vision Transformers (ViT) are an adapted version of the Transformer model in section 2.4.4. The architecture proposed in the paper *An image is worth 16x16 words: transformers for image recognition at scale* [9] is mainly aimed at image-related tasks, whereas the original Transformers were often used for tasks such as natural language processing. The model architecture of a Vision Transformer is very similar to that of the original Transformer, but the main difference is the input. In a Transformer the input is usually a piece of text converted into tokens. Because images cannot be directly converted into tokens, they are split into fixed-sized patches. These patches are flattened and serve as the network's input tokens. After this, positional embedding is added to the tokens. The rest of ViT architecture uses the original transformer blocks and stacks them.

It is important to note that ViT is an encoder only model, whereas the original Transformer architecture is an encoder-decoder model.

Liu et al. [34] introduced the Swin (Shifted window) Transformer as a more efficient and powerful adaptation of the vision transformer. Swin Transformers are more efficient than regular transformers because of the shifted window-based approach. The model architecture consists of Swin Transformer blocks, which are similar to the regular ViT blocks but they use shifted local self-attention over iterations. In each iteration the local attention is shifted. The model is divided into stages, and each stage consists of two successive Swin Transformer blocks. Between stages they apply a technique called patch merging, which reduces the number of tokens by a factor of four while doubling the embedding size, effectively reducing the input size by a factor of two.

Pan et al. [40] propose a novel method for surgical phase prediction that combines a Swin Transformer with an LSTM network. Using the remote dependencies captured by the Swin Transformer and the temporal information from the LSTM, the *TSTNet* can extract spatiotemporal features with more contextual information. First, the Swin Transformer is trained based on the Imagenet-22k dataset. The Swin Transformer then extracts multi-scale visual features from the input frames. The LSTM network models the temporal information from the input sequences. The Swin Transformer and LSTM are trained in an end-to-end matter.

2.4.6 Surgical Video Representation. There are several features that can be used for the representation of surgical videos. These features can be used as the input for a phase recognition network to improve the performance compared to generic feature extraction methods. In this section, five of these representations will be discussed: tool presence recognition, action recognition, semantic segmentation, video captioning, and optical flow.

Tool presence is one of the features that can be used as input for phase prediction networks. It involves recognizing and identifying which tools are present in a frame and sometimes also includes segmentation of the location of different tools. Tool presence can provide valuable information on the current phase in videos, since the different phases often use distinct tools. In several papers, automatic tool recognition has already been used to successfully predict phases [29, 56]. These works show that including these signals as an input feature can improve the network performance.

Information on the current *actions* of the surgeons can also be a valuable input feature for phase recognition. Each surgical phase usually has certain actions that are performed within it, so recognising these actions can help the final prediction network. Action recognition is usually performed by creating an action triplet: three words describing the subject, verb, and object of the action. Tripnet [36], Attention Tripnet [37], Rendezvous [37], and MT-FiST [31] are the state of the art models for action triplet recognition.

Semantic segmentation in surgical videos is another descriptor that can be used as input to a recognition model. This focuses on segmenting the entire surgical field instead of only the tools present in the frame. Although the surgical field does not change significantly during laparoscopic surgeries, the position of the organs can be useful. Especially in surgeries such as cholecystectomies or hysterectomies, in which a specific organ is removed, segmentation could help the phase prediction.

Video captioning in the context of surgical videos refers to the automatic generation of descriptive text for the actions and events within surgical videos [3]. This task not only recognises the actions or tools in a video but also aims to understand the events that are happening. This can provide a comprehensive description of the surgical procedure in text format. Video captioning usually involves a combination of computer vision and natural language processing (NLP) to translate visual information into suitable sentences. Sequence-to-sequence models, which generally use a combination of CNNs for feature extraction and LSTMs for sequence generation, are often used for this task because they can generate accurate and contextually relevant captions for surgical scenes [44].

The computation of the motion of objects in a video is known as *optical flow estimation*. The optical flow is estimated based on the movement of pixels or features in an image between frames. The surgeon's motions and the transitions between these motions are often generic for various surgical tasks [49]. This makes optical flow recognition a valuable input feature for phase recognition, as each phase often has distinct tasks and motions.

Five possible input features have been discussed in this section. Three of the most promising features are: tool presence detection, semantic segmentation, and action triplets. Together, these descriptors can give valuable information on the actions performed by the surgeon, the state of the surgical scene, and the instruments being used. All of this information can be useful for the prediction of the current surgical phase, and are therefore valuable visual cues to implement in a multi-visual cue model.

2.5 Datasets for Surgical Video Representation

For training and evaluating the different descriptors for surgical video representation, three datasets are used. CholecSeg8K [20] is used for segmentation tasks and contains pixel-wise annotations of surgical scenes. Cholec80 [56], a widely used dataset for tool recognition, consists of 80 videos of laparoscopic cholecystectomy surgeries annotated with surgical tool labels. CholecT50 [37] is used for action recognition and contains triplet annotations consisting of an instrument, verb, and target label.

3 METHODOLOGY

For the task of automatic phase recognition, a cross-modal approach is proposed, presented in Figure 5. The model architecture consists of a main phase recognition network and three additional branches to extract descriptors from the input frames.

3.1 Video Characterization

The first step in the pipeline is video characterization (**I**). This is done using three models, one to extract segmentation maps (**A**), one for action labels (**B**) and one for tool labels (**C**). The models are trained separately, and the trained models are used to generate the input data for the phase recognition network.

A) The segmentation model is based on the MaskFormer Swin Base ADE model [5], which combines the MaskFormer network with a Swin Transformer [34]. The model is pre-trained on the ADE20K dataset [67]. For this project, the model was fine-tuned on the CholecSeg8k [20] dataset to make it more effective for surgical scene segmentation. The trained and fine-tuned model is used to

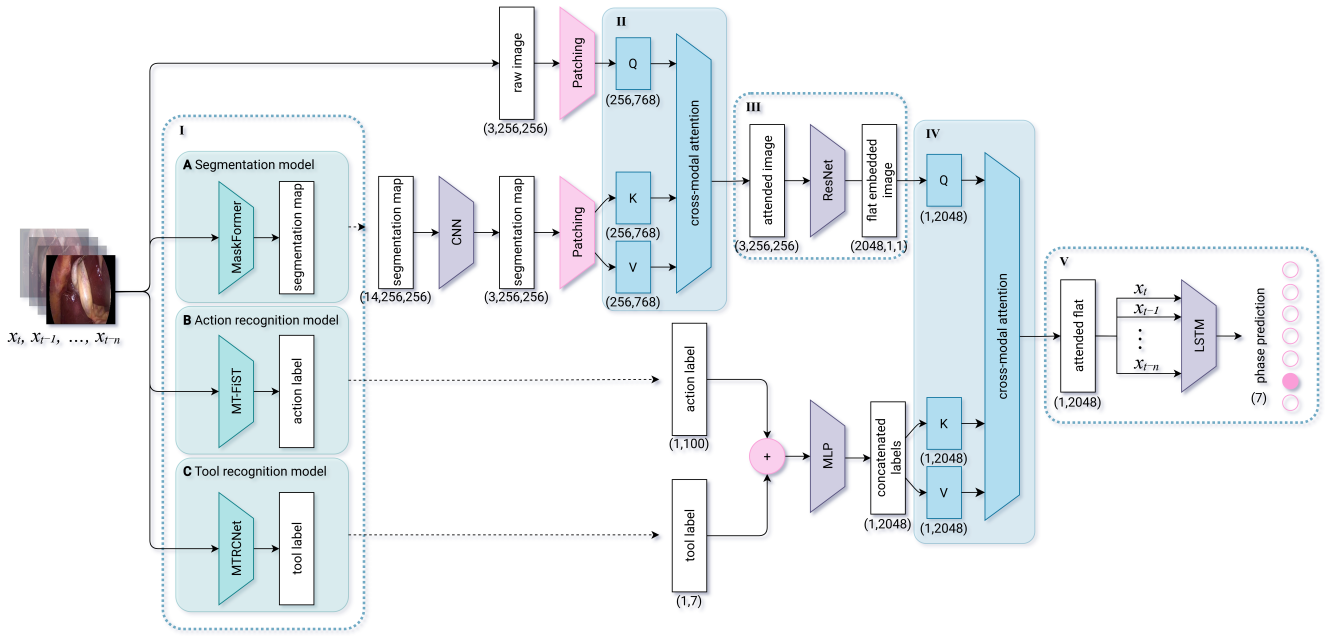


Figure 5: Diagram showing the full phase prediction pipeline including each model for extracting the descriptors (I). Method (A) uses a fine-tuned version of the MaskFormer model presented in [5]. Method (B) employs the pre-trained action recognition model MT-FiST [31]. Method (C) is a tool recognition model based on the tool branch of MTRCNet [24]. The phase prediction model uses two separate cross-attention blocks, one to combine global descriptors with the segmentation maps (II). The output is fed into the ResNet (III) and a second cross-attention block further combines the tool and action labels (IV). This is followed by an LSTM (V) and finally the model outputs the predicted phase. The possible phases are *Preparation*, *Calot Triangle Dissection*, *Clipping Cutting*, *Gallbladder Dissection*, *Gallbladder Packaging*, *Cleaning Coagulation*, and *Gallbladder Retraction*. The labels underneath each block represent the shape of the data.

predict segmentation maps for the frames of the Cholec80 dataset [56]. The MaskFormer model uses the per-pixel classification loss and mask classification loss as defined in [5]. The architecture of the MaskFormer is presented in Appendix A.3.

B) For the action descriptors, the Multi-Task Fine-grained Spatial-Temporal framework (MT-FiST) [31] is used. The network uses a ResNet-50 and an LSTM module, and four branches for recognizing instruments, actions (verbs), targets, and action triplets, which are a combination of the instrument, verb, and target label. The MT-FiST model is trained on the CholecT50 dataset [37] and used to generate the action labels for the Cholec80 dataset. The architecture of the MT-FiST model is presented in Appendix A.2. For the action labels, only the predicted triplets are used as a descriptor in the proposed approach. The MT-FiST model uses the binary cross entropy loss which is defined as:

$$L_{\text{BCE}} = -\sum_{2}^{c=1} y_c \log(p_c), \tag{1}$$

where y_c is the ground truth label for class c and p_c is the predicted probability for that class.

C) The Multi-Task Recurrent Convolutional Network(MTRCNet) [24] is used to extract the tool descriptors from the input frames. The tool branch of MTRCNet uses a ResNet to extract features, followed by a fully connected layer to predict which tools are present in the current frame. The tool label predictions for the Cholec80 dataset are used as the fourth descriptor for the phase recognition network. The tool recognition model also uses the binary cross entropy loss as defined in Equation 1. Appendix A.1 shows the architecture of the tool branch of the MTRCNet.

3.2 Multi-modal Phase Prediction with Cross-Attention

The next step in the pipeline consists of the phase recognition network based on the MTRCNet [24]. The MTRCNet consists of both a tool recognition branch and a phase recognition branch. The phase recognition branch, consisting of a ResNet and an LSTM module, is used as the backbone of the main pipeline of the proposed phase prediction network. The four different descriptors from each extraction method are combined using two cross-attention blocks. In the first cross-attention block (**II**), the model aligns the global image descriptors with the generated segmentation maps. The global features act as the query (Q), and the segmentation map features act as the key (K) and value (V).

The attention block uses multi-head attention [58], which is calculated using:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

where head_i is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

and:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

The block's output is an attended representation of the image that shows the important regions based on the segmentation input. This attended image forms the input to the ResNet module (**III**) from the MTRCNet model.

The second cross-attention block (**IV**) further combines the output of the ResNet with the tool and action descriptors. Here, the output of the ResNet serves as the query. A concatenated label consisting of the action and tool descriptors is used as the key and value. The same formulas shown in Equation 2 and 4 are used to calculate the attention. The output of this block is an attended representation that combines the relevant aspects of the tool and action features with the original attended image.

After the second cross-attention block, an LSTM (**V**) performs the final phase prediction based on the combined inputs. The model uses *cross-entropy loss* as the final loss function. Cross entropy is a commonly used loss in multi-class classification tasks. Using the predicted phase labels and the ground truth labels, the cross-entropy loss L_{CE} is computed as follows:

$$L_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{5}$$

where y_c is the ground truth label for class c and p_c is the predicted probability for that class.

4 EXPERIMENTAL SETUP

This section discusses the data preprocessing steps and data augmentation used for the training of each network. It discusses the implementation details and training procedure of all experiments in this study. The baseline models used for comparison with the proposed model is also described in this section. Finally, it presents the evaluation metrics used for measuring the performance.

4.1 Data Preprocessing and Augmentation

To fine-tune the segmentation model, the CholecSeg8K dataset [20] was used. Random horizontal flipping is used to add variation to the data. For the action recognition model, the dataset CholecT50 [37] was used. Frames in this dataset are resized to 250 x 250. Random horizontal flipping and random rotation are applied, as described in the MT-FiST paper [31].

For the training of both the phase recognition network and the tool prediction network, the Cholec80 [56] was used. This dataset consists of 80 videos, as described in Section 2.3. First, the video in the Cholec80 dataset were converted to frames using the same method as was used for the MTRCNet [24]. Then, the videos are downsampled from 25 fps to 1fps. Finally, the orignal frames are resized to a resolution of 250 x 250. For both networks, data augmentation was implemented by including random horizontal flipping, and for the phase prediction model random cropping of 224 × 224 was used for additional augmentation. The same methods for data augmentation are used for the baseline model MTRCNet.

4.2 Implementation Details

All experiments presented in this study are carried out using PyTorch [42]. The models are trained on a single GPU using the EEMCS HPC Cluster of the University of Twente. Weights and Biases [1] was used to keep track of the progress and results of all experiments during this research.

For the segmentation model, the `HuggingFace` library is used. The `facebook/maskformer-swin-base-ade` model is loaded and fine-tuned using the CholecSeg8K dataset. The loss function `BCEWithLogitsLoss` with reduction `sum` was used for the training of the action recognition model. The tool recognition model uses the same loss function with reduction `mean`.

For applying the cross-attention between images and segmentation and between images and action and tool labels in the phase recognition model, PyTorch's `MultiheadAttention` is used. A pre-trained `ResNet50` from the `torchvision` library extracts the features from the image data in both the phase and tool recognition models. A custom `ResNetBlock` layer preprocesses segmentation maps to match the input format and channel dimensions before applying cross-attention. Dropout of 0.1 is used in the Resnet blocks. The LSTM and the fully connected layers use Xavier weight initialization [13], which can improve stability and make convergence faster during training. Segmentation maps are one-hot encoded.

4.3 Training Procedure

4.3.1 Action Recognition Model. For the action recognition, the pre-trained MT-FiST model as described in [31] is used. For the training of this model, the first 36 videos of the CholecT50 dataset were used. The parameters were chosen based on validation of the last 9 videos. The weights of the pre-trained model are loaded and used for inference on the Cholec80 dataset, and the action triplets predicted by the model are saved. Because the CholecT50 dataset consists of 45 videos that are also in the Cholec80 dataset, the inference is only done on the remaining 35 videos that were not used for training or validation of the MT-FiST model.

4.3.2 Segmentation Model. The test set consists of the same 35 videos as the test set for the pre-trained action recognition model described in the previous section. The training set includes all videos from the CholecSeg8K dataset that do not appear in the chosen test set, which is 10 videos in total. Of these videos, 85 percent of the frames are used for training and the remaining 15 percent are used for validation.

The segmentation model uses the weights of the pre-trained MaskFormer Swin Base ADE [5]. The fine-tuning of the model is done using a batch size of 32 and the Adam optimizer with a learning rate of 5e-5. The model is trained for 100 epochs, and hyperparameters were chosen based on the results of the validation set. The parameters of the experiment with the highest IoU scores are chosen for the final training. The mean IoU of the predicted segmentation maps is saved at every epoch. The fine-tuned model is saved at every epoch, and after training the model with the highest IoU score is used for inference on the test set. The predicted segmentation maps of the test set are saved as a NumPy array, to be used as input to the phase prediction model.

4.3.3 Tool Recognition Model. The Cholec80 dataset is used to train the tool recognition model. The data is split into a train and validation set the same way as was done for the action recognition model. The same 36 videos are used for training and 9 videos for validation. This ensures that the remaining videos have not been used in the training of any of the separate input models for the phase recognition network, and can therefore be used for training, validating and testing that network.

The training is done using a batch size of 100, and Adam optimizer with a learning rate of 1e-3 is used. The model was trained for 50 epochs. The average accuracy and the individual accuracy for each tool are saved every epoch. After training, the saved model is used for inference on the remaining videos of the Cholec80 dataset, and the predicted tool labels are saved.

4.3.4 Phase Recognition Model. As explained in Section 4.3.1, the CholecT50 dataset consists of 45 videos from the Cholec80 dataset. Those 45 videos have been used to train and test the pre-trained model used for action triplet recognition, which serves as one of the inputs to the phase recognition model. Because the phase recognition network is trained using the Cholec80 dataset, this could potentially cause data leakage, if the same videos are used again. In order to prevent this data leakage, only the remaining 35 videos are used for the training, validation and testing of the phase recognition network.

These available 35 videos are split into a train, validation, and test set, where 20 percent of the data is used to create a test set. The phase model is trained using 5-fold cross-validation, training on 80 percent of the remaining data and validating on the 20 percent for each fold. The data split for the train and test sets is done at a video level, to ensure that subsequent frames from a single video do not appear in the different subsets. This approach guarantees that the test dataset consists entirely of actual unseen data. However, due to the fact that data is randomly split into a train and validation set in each fold, this split was performed at a frame level. Figure 3 in Section 2.3 showed that the number of frames per video varies greatly. A random split at the video level would therefore result in large variations of training and validation set, which is why the frame-wise split was implemented instead.

Many hyperparameters were tested through experimentation, and the best subset of hyperparameters was used for all the final trainings. During training, a batch size of 32 is used for both training and validation. The model uses the Adam optimizer with a learning rate of 3e-5. The running loss, accuracy, and F1 score are calculated and saved every epoch for both training and validation. A sequence length of 4 is used as the input into the model. Each fold is trained for 50 epochs, and the model weights are saved for the epochs with the highest validation accuracy. After the cross-validation, the model with the highest validation accuracy is evaluated on a separate test set. This evaluation phase gives a final test accuracy and F1 score that show the model's performance across both validation and unseen test data.

4.4 Baseline Model

The pipeline shown in Figure 6 shows the baseline model that is used to compare the results of the cross-modal approach. This pipeline consists of a similar structure as the proposed architecture but does not include the cross-attention blocks. It is a straightforward network that starts with a ResNet. The output of the ResNet is fed into an LSTM layer followed by a fully connected layer to predict the surgical phases. Similar to the proposed model, the baseline model is also trained using *cross-entropy loss*.

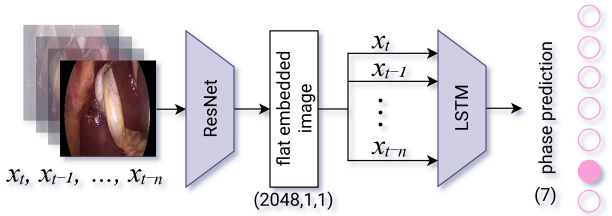


Figure 6: Pipeline of the baseline model without cross-attention for phase prediction, based on the phase branch of MTRCNet [24].

4.5 Ablation Study

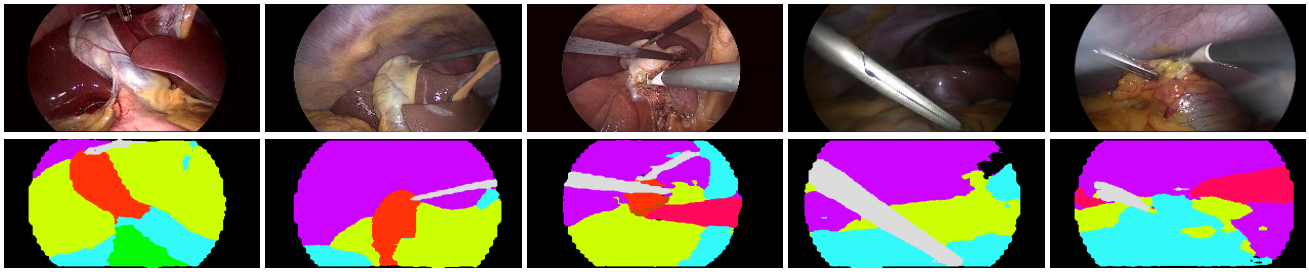
An ablation study is conducted to show the impact of the different descriptors. In each experiment, a different combination of descriptors was used for the cross-modal attention, in order to see how each of them contributes to the final phase prediction results. The first configuration uses all descriptors, so segmentation, action, and tool, along with the global image features. In the next experiment, cross-attention is limited to only segmentation features and global features. The third configuration uses only action descriptors for the cross-attention, and in the final experiment, only tools are combined with the global descriptors in the cross-attention block. Each of these ablation tests shows different insights on how the individual descriptors contribute to the predictions of the model. A full ablation table is presented in Table 5 in Section 5.

4.6 Complete Dataset

The initial experiments were conducted using a subset of Cholec80 to avoid data leakage from the pre-trained models for action triplet recognition and the segmentation model, as they were partially trained on Cholec80 videos. This ensured that there was no overlap between the data used for pre-training the descriptor models and the data used for evaluating the proposed model. To show the performance of the proposed model without any data constraints, an additional ablation study was conducted using the complete Cholec80 dataset. The training procedure followed the same steps as the original ablation as described in Section 4.5. However, a different data-splitting method is used, as there are more videos in the training data for this experiment. The data split used in the MTRCNet paper [24] was used, with 40 videos for training and 40 for testing. The results of these experiments are presented in Table 7 in Section 5.

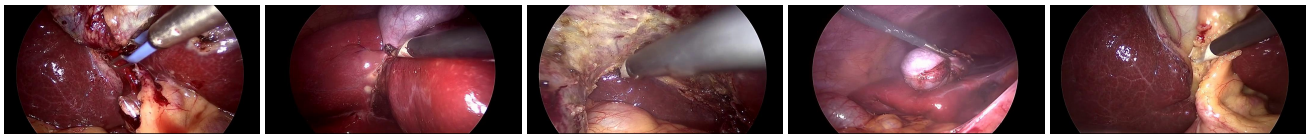
4.7 State of the Art Comparison

The baseline model and the best-performing model found in the ablation studies are then compared to state-of-the-art on two different datasets, Cholec80 [56] and HeiChole [59]. For the Cholec80 results,



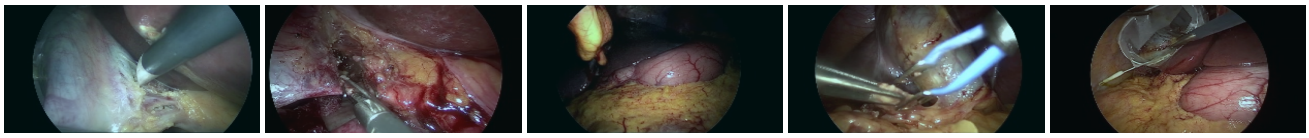
(a) (b) (c) (d) (e)

Figure 7: Examples of the predicted segmentation masks along with the original images.



<bipolar, liver> | coagulation, <hook, gallbladder> | cutting, <grasper, abdominal wall> | coagulation, <grasper, gallbladder> | retraction, <hook, large intestine> | coagulation,

Figure 8: Examples of the action triplets predicted by the model.



(a) hook (b) scissors (c) no_tool (d) bipolar, grasper (e) specimen_bag

Figure 9: Examples of the tools predicted by the model, along with the corresponding input frame.

the same training procedure and implementation were used for the ablation study with the full Cholec80 dataset. For the HeiChole results, the training procedure is also the same as the ablation study, but the data splitting method is different. Since the official test set for the HeiChole dataset is not publicly available, a custom data-split is used to create a test set from the training data, as detailed in [45].

4.8 Evaluation Metrics

The different evaluation metrics that measure the performance of each individual model are described in the following paragraphs.

4.8.1 Segmentation. For the segmentation model, the mean Intersection over Union (mIoU) is used, which measures the overlap between the predicted segmentation map and ground truth masks. The mean IoU is calculated as follows:

$$\text{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{6}$$

$$\text{Mean IoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c \tag{7}$$

where C is the total number of classes, TP_c (True Positives) is the count of pixels correctly classified as class c, FP_c (False Positives) is the count of pixels incorrectly classified as class c (they belong to another class but are predicted as c), FN_c (False Negatives) is the count of pixels belonging to class c but incorrectly classified as a different class. Mean IoU is often used in semantic segmentation tasks and indicates how well the model predicts the different surgical regions in the segmentation map.

The Dice similarity coefficient (DSC) [8], or Dice score, is also used to evaluate the performance of the segmentation model. It measures the overlap between the predicted and ground truth masks. It is calculated as:

$$\text{Dice}_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \tag{8}$$

$$\text{Mean Dice} = \frac{1}{C}\sum_{c=1}^{C} \text{Dice}_c \tag{9}$$

The Dice score ranges from 0 to 1, where a higher score shows better overlap between the predicted and actual segmentation. The main difference with the mean IoU is that Dice score weighs correct predictions higher.

4.8.2 Action Recognition. For action recognition, the same evaluation method as described in the MT-FiST framework for surgical action triplet recognition [31] is used. The mean Average Precision (mAP) is calculated on the instrument-verb-target triplets as detailed in the MT-FiST paper.

4.8.3 Tool Recognition. The performance of the tool recognition network is evaluated using the average accuracy and the individual

Method	Mean IoU
Mask2Former [4]	0.691
HRNet + SP-TCN [14, 54]	0.6537
Swin base + SP-TCN [14, 34]	0.6938
MaskFormer [5] (fine-tuned on CholecSeg8k [20])	**0.863**

Table 2: Surgical scene segmentation result comparison for Mask2Former, HRNet + SP-TCN and Swin base + SP-TCN. The models are evaluated on the CholecSeg8K dataset [20].

Method	mAP_I	mAP_V	mAP_T	mAP_{IVT}
Tripnet [36]	74.6	42.9	32.2	23.4
Attention Tripnet [36]	77.1	43.4	30.0	25.5
Rendezvous [37]	32.2	47.5	37.7	32.7
MT-FiST [31]	**82.1**	**51.5**	**45.5**	**35.8**

Table 3: Action triplet recognition result comparison for Tripnet, Attention Tripnet, Rendevous and MT-FiST. The models are evaluated on the CholecT50 dataset [35].

accuracy per tool class. The average accuracy is expressed as:

$$\text{Average Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{10}$$

For individual tool accuracy, each tool t has its accuracy calculated as:

$$\text{Accuracy for Tool} = \frac{\text{Correct Predictions for Tool}}{\text{Total Instances of Tool}} \tag{11}$$

This metric is calculated for every tool in the set `['Grasper', 'Bipolar', 'Hook', 'Scissors', 'Clipper', 'Irrigator', 'SpecimenBag']` to calculate the model's performance for each tool individually.

4.8.4 Phase recognition. To evaluate the performance of the phase recognition network, the accuracy and the F1 score are used. Accuracy measures the ratio of correct predictions out of the total number of predictions, and gives an indication of the overall model performance. It is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{12}$$

The F1 score is the harmonic mean of precision and recall and shows the model's accuracy on the positive class. It is calculated using the following formula:

$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}, \tag{13}$$

where TP, FP, and FN are the true positives, false positives, and false negatives, respectively.

5 RESULTS

This section showcases the results of the individual models for descriptor extraction, the ablation studies of the phase recognition models, and the comparisons with state-of-the-art models.

5.1 Video Characterization

The segmentation model MaskFormer [5], after being fine-tuned using the CholecSeg8k dataset [20], achieved a mean IoU of 0.863. This score shows that the model can successfully generate segmentation masks that closely match the ground truth masks. Table 2 shows the mean IoU of the fine-tuned MaskFormer compared to the results of three other models: Mask2Former [4], HRNet + SP-TCN [14, 54], and Swin base + SP-TCN [14, 34]. The fine-tuned MaskFormer model achieves the highest mean IoU score out of all of these models. This result seems to indicate that using an already well-performing pre-trained model as the base, and fine-tuning this on surgical video data, has improved the result of the segmentation model. However, certain inputs such as very dark frames or frames with motion blur in it still present some issues for the model. Figure 7 shows some examples of the predicted segmentation masks to illustrate the model performance qualitatively. Figure 7a, 7b and 7c show some of the better outputs of the segmentation model. The predicted masks are clear, most of the borders are accurate and the masks seem to match the input images well. Figure 7d shows an example of a dark frame. The tool in the bright foreground seems to be segmented well, while the segmentation of the dark background matches the input image less well. Finally, in Figure 7e, a frame with motion blur is shown. The predicted segmentation for this frame also shows less clear boundaries and incorrectly segmented areas on the left and right where the scene is blurred.

Tool	EndoNet [56]	FCN_ESP_Msk [57]	MTRCNet [24]
Grasper	0.844	**0.967**	0.820
Bipolar	0.869	0.955	**0.990**
Hook	0.956	**0.996**	0.932
Scissors	0.586	0.500	**0.993**
Clipper	0.801	0.823	**0.991**
Irrigator	0.744	0.943	**0.969**
Specimen bag	0.868	0.935	**0.976**
Mean	0.810	0.874	**0.953**

Table 4: Tool presence detection result comparison for EndoNet, MTRCNet-CL and FCN_ESP_Msk. The models are evaluated on the Cholec80 dataset [56].

After re-running the experiments using the pre-trained model, the results for the action recognition model are the same as presented in the MT-FiST paper [31]. The individual *mAP* for the instrument, verb and target labels are 82.1, 51.5, and 45.5 respectively. The *mAP* for the full triplets is 35.8. Table 3 shows these results compared to three state of the art models: Tripnet [36], Attention Tripnet [37], and Rendezvous [37]. The results of these models are presented in Table 3. mAP_I, mAP_V, mAP_T and mAP_{IVT} denote the mean average precision of the instrument, verb, target, and triplet recognition tasks. While the scores for the full action triplets are not incredibly high, Table 3 shows that the MT-FiST model obtained the highest scores on action triplet recognition out of these three state of the art models. Figure 8 presents several examples of images and their corresponding predicted action triplet label to show the results of the model on the Cholec80 dataset.

For the tool recognition branch, the MTRCNet model was trained. Table 4 shows the accuracy of this model on the individual tool classes, along with two other state of the are models for comparison. The MTRCNet achieved the highest accuracy for almost all seven classes, except for the classes *Grasper* and *Hook*. For these classes,

the model FCN_ESP_Ms showed the best results. The class *Scissors* shows a large difference between the score of the MTRCNet and the other two models, with accuracies of 0.993, 0.586 and 0.500 respectively. This gap gives a mean accuracy for the MTRCNet of 0.953, which is significantly higher than the other models (0.810 and 0.874). The class *Grasper* was the hardest for the model to predict correctly with an accuracy of 0.8195. Figure 9 shows some examples of images in the Cholec80 dataset along with their predicted tool label. Even in the case where there is some motion blur present in the frame, such as in Figure 9d, the model can accurately predict the tools present in the image.

5.2 Ablation Study

Table 5 shows the results of the ablation study for the phase recognition model described in Section 4.5 after training on the subset of Cholec80 [56]. On the validation set, Model 4 has both the lowest accuracy and F1-score, with scores of 0.966 and 0.976 respectively. Model 5, the model that incorporates cross-attention with segmentation, action triplets and tool labels, has the best performance on the validation set with an accuracy of 0.972 and an F1-score of 0.981. All models have higher accuracies and F1-scores on the validation set than on the test set. This will be discussed in detail in Section 6.

The results on the test set are distributed differently than on the validation set. Model 3 outperformed the other ablation models with an accuracy of 0.823 and an F1-score of 0.739 on the test set. The model using all three visual cues, Model 5, consistently performs worse than most of the ablation models. Models incorporating cross-attention between images and only one other visual cue seem to achieve higher accuracies and F1 scores in this ablation study. It is worth mentioning that both Model 2 and 3 outperform the baseline Model 1 when it comes to the F1 score, with scores of 0.732, 0.739 respectively compared to 0.726 for Model 1. Only Model 3 outperforms the baseline model in accuracy. In general, Model 3 shows the most promising results out of the 5 ablation models presented in Table 5.

Table 6 shows the individual accuracies of these ablation models for each phase. For all models, the highest accuracy is achieved on the phase *Gallbladder Dissection*. The phase *Preparation* shows the lowest scores across almost all models. An interesting observation is that Model 2, which incorporates segmentation maps, is the only model to achieve an accuracy above 0.5 for this phase (0.627). There is no model that consistently outperforms the others in every phase. Model 4 outperforms the other models in most phases, achieving the highest accuracies in *Calot Triangle Dissection* (0.839), *Cleaning Coagulation* (0.800), and *Gallbladder Retraction* (0.733). Model 3 achieves the best performance in *Clipping Cutting* with an accuracy of 0.730, while Model 5 performs best in *Gallbladder Packaging* at 0.805. Although the baseline model (Model 1) achieves the highest accuracy for Gallbladder Dissection (0.909), it generally underperforms compared to models incorporating multiple visual cues.

Figure 10 shows confusion matrices for the baseline Model 1, and the best-performing model in the ablation study, Model 3. For both models the diagonal values, representing correctly classified phases, are higher compared to the values outside of the diagonal for each phase. Figure 10b shows that Model 3 misclassified 6 percent of *Gallbladder Dissection* as *Calot Triangle Dissection*, compared to only 4 percent by Model 1. Similarly, both Model 1 and 3 incorrectly predict the label *Gallbladder Dissection* for 15 and 14 percentof ground truth label *Calot Triangle Dissection* respectively, and 26 and 18 percent of *Clipping Cutting* respectively. Both models also shows a high number of misclassifications for the *Preparation* phase. Model 1 incorrectly predicted 46 percent as *Calot Triangle Dissection* compared to 44 and correctly predicted labels Model 3 has a slightly lower number of misclassification for this class, with 42 percent incorrect predictions compared to 48 correct ones. Overall, the confusion matrix of Model 3 shows less significant misclassification than the percentages in Model 1.

Table 7 presents the results of the ablation models trained and evaluated on the complete Cholec80 dataset, with no data constraints. The baseline Model 1, which uses only image data, achieved an accuracy of 0.792 and an F1 score of 0.844. Of the models incorporating cross-attention with multiple visual cues, Model 3, which integrates action triplets with image features, achieved the highest accuracy and F1 score of 0.826 and 0.871, respectively. Model 5, which combines all three visual cues (segmentation maps, action triplets, and tool labels), achieved a slightly lower performance, with an accuracy of 0.801 and an F1 score of 0.852. Only models 3 and 5 outperform the baseline Model 1 on both the accuracy and F1 score.

Figure 11 shows several example frames from the Cholec80 dataset along with the predicted label and the ground truth label.

5.3 State of the Art Comparison

Table 8 shows the accuracy and F1 score on the Cholec80 dataset for the models presented in Section 2.4 and for the baseline model and the best ablation model proposed in this research. The two best-performing models on the Cholec80 dataset are EffNetV2 C-ECT+CAFF and EffNetV2 MS-ASCT. Both of these models are Transformer-based methods as discussed in Section 2.4.4. When it comes to both the accuracy and F1 score, the proposed model does not outperform the state-of-the-art models. The best ablation model proposed in this research achieved an accuracy of 0.826 and an F1 score of 0.871, which, while not outperforming the state-of-the-art models, falls within the error margins of the highest reported F1 score, 0.928 ± 0.061.

Table 9 shows an overview of the accuracy and F1 score of the model on the HeiChole benchmark dataset [59]. The best-performing existing model, MuST, achieved an F1 score of 0.773. The best ablation model proposed in this research achieved an F1 score of 0.796, outperforming the performance of state-of-the-art models CUHK, HIKVision, and CAMMA 1, which had F1 scores of 0.650, 0.654, and 0.688, respectively. It is important to note that the models marked with an asterisk (*) in the table were evaluated on a separate test set, which is not publicly available.

6 DISCUSSION

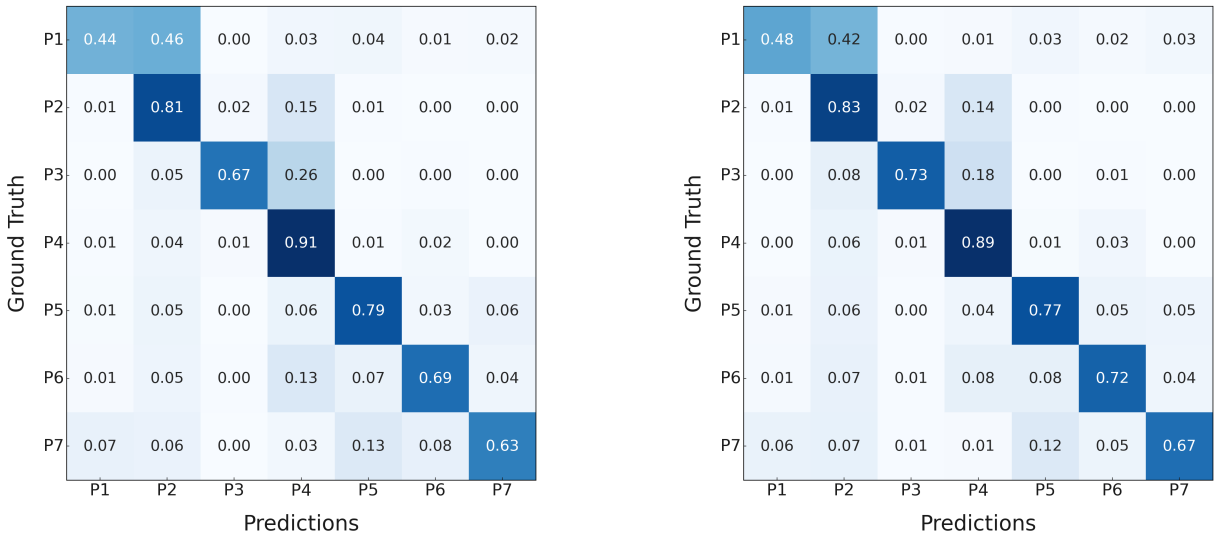
This section discusses the obtained results of the ablation study and the state of the art comparisons in detail. It also answers the research questions for this study, and gives recommendations for future work.

Model	Visual cues				Validation Set		Test Set	
	Image	Segment-ation	Action triplets	Tools	Accuracy	F1 Score	Accuracy	F1 Score
1	✓				0.970	0.979	0.816	0.726
2	✓	✓			0.970	0.979	0.786	0.732
3	✓		✓		0.971	**0.981**	**0.823**	**0.739**
4	✓			✓	0.966	0.976	0.804	0.719
5	✓	✓	✓	✓	**0.972**	**0.981**	0.796	0.720

Table 5: Accuracy and F1 score for the ablation models trained on a subset of Cholec80 [56]. Model 1 is the baseline model using only image data as input. Models 2, 3, and 4 use cross attention between images and segmentation masks, action triplets and tools respectively. Model 5 incorporates cross attention between images and the three visual cues.

Phase	P1	P2	P3	P4	P5	P6	P7
Model 1	0.443	0.815	0.674	**0.909**	0.788	0.694	0.634
Model 2	**0.627**	0.728	0.623	0.906	0.788	0.672	0.630
Model 3	0.480	0.826	**0.730**	0.888	0.775	0.717	0.675
Model 4	0.348	**0.839**	0.602	0.864	0.725	**0.800**	**0.733**
Model 5	0.456	0.790	0.627	0.881	**0.805**	0.746	0.583

Table 6: Individual accuracies for the ablation models on the test set. The phases are *Preparation* (P1), *Calot triangle dissection* (P2), *Clipping and cutting* (P3), *Gallbladder dissection* (P4), *Gallbladder packaging* (P5), *Cleaning and coagulation* (P6), and *Gallbladder retraction* (P7).



(a) Model 1 **(b) Model 3**

Figure 10: Confusion matrices showing the predictions of the baseline Model 1 and Model 3 (images and action triplets). The phases are *Preparation* (P1), *Calot triangle dissection* (P2), *Clipping and cutting* (P3), *Gallbladder dissection* (P4), *Gallbladder packaging* (P5), *Cleaning and coagulation* (P6), and *Gallbladder retraction* (P7). Colours and values in the matrix are normalized based on the number of class instances.

6.1 Ablation Study

The results on the validation set in Table 5 show no clear immediate difference in performance between the different descriptors. The results do show quite a large difference between the performance on the test set and the validation set. The most probable reason for this contrast is the way the data is split for both sets. As explained in Section 4.3, the train and validation set are split in a frame-wise matter, and the test set is split at a video level. This means that the test set results show the performance on entirely unseen data, while the validation results might show the performance on frames

	Visual cues					
Model	Image	Segment-ation	Action triplets	Tools	Accuracy	F1 Score
1 (baseline)	✓				0.792	0.844
2	✓	✓			0.771	0.827
3	✓		✓		**0.826**	**0.871**
4	✓			✓	0.777	0.831
5	✓	✓	✓	✓	0.801	0.852

Table 7: Accuracy and F1 score for the ablation models trained on the full Cholec80 dataset [56]. Model 1 is the baseline model using only image data as input. Models 2, 3, and 4 use cross-attention between images and segmentation masks, action triplets and tools respectively. Model 5 incorporates cross-attention between images and the three visual cues.

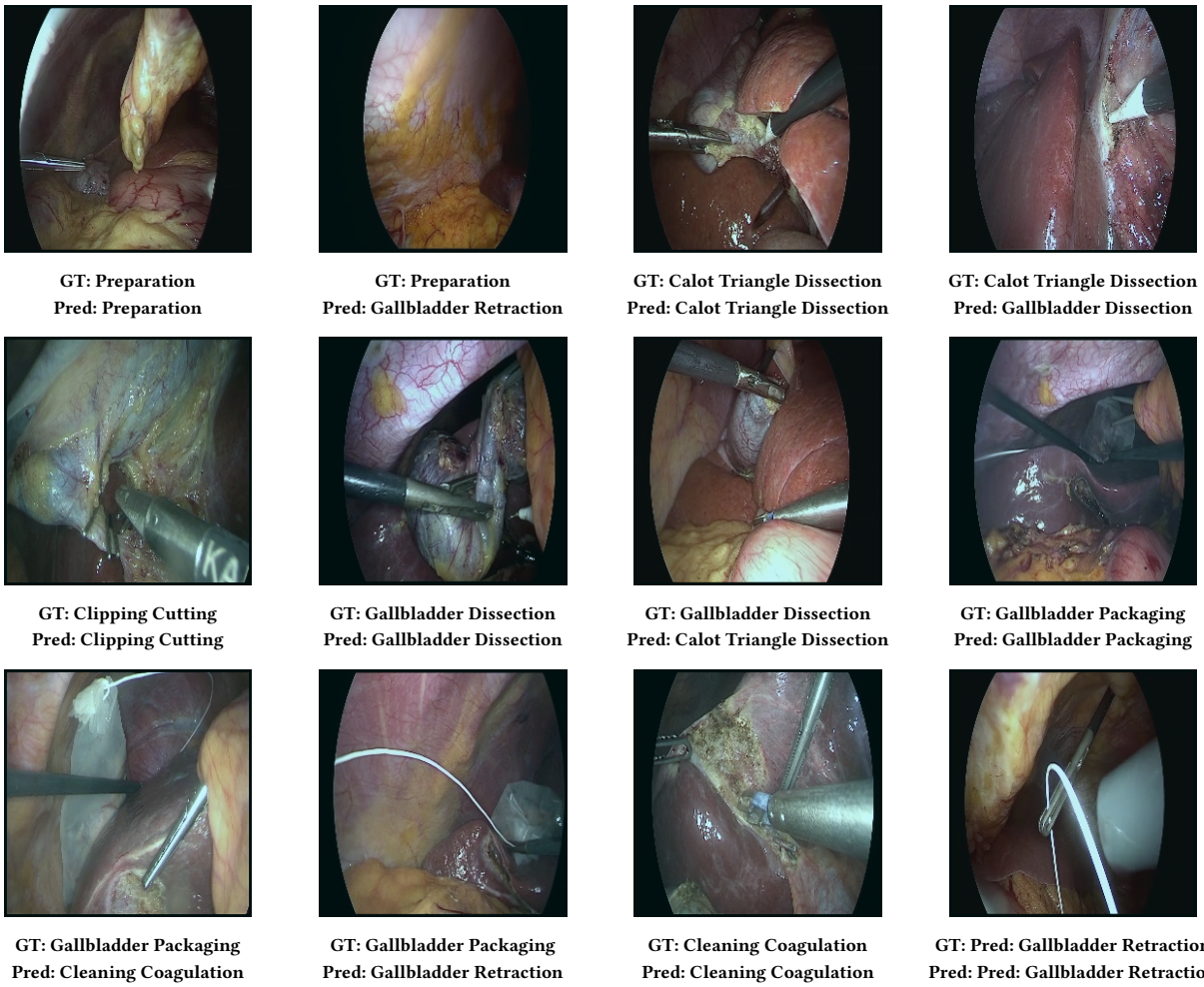


Figure 11: Examples of the predicted phases along with the original image and the ground truth phase label. All predictions are made by ablation Model 3 on the Cholec80 dataset [56]. All images are normalized and resized to 256x256, as per the training and testing transforms.

adjacent to those seen in the training set. In this study, this was mainly done because k-fold cross-validation was used for training, and randomly splitting at a video level would result in large varieties between dataset sizes. However, splitting all datasets at this level could give more realistic and robust results of the performance of the models.

The results on the test set in Table 5 show that while the performance is similar, cross-attention does not always give higher

Method	Year	Accuracy	F1 Score
EndoNet [56]	2016	0.817 ± 0.042	0.765 ± 0.055
SV-RCNet [23]	2018	0.853 ± 0.073	0.821 ± 0.072
MTRCNet-CL [24]	2019	0.892 ± 0.076	0.874 ± 0.056
MTRCNet [24]	2019	0.859 ± 0.076	0.823 ± 0.108
MTRCNet SingleNet [24]	2019	0.853 ± 0.069	0.816 ± 0.110
State-Preserving LSTM [48]	2020	-	0.814
TeCNO [7]	2020	0.886 ± 0.078	0.871 ± 0.069
Trans-SVNet [12]	2021	0.903 ± 0.071	0.897 ± 0.062
CNN+LSTM [22]	2021	0.929	0.875
LoVit [32]	2023	0.924 ± 0.063	0.903 ± 0.052
TSTNet-PRA [40]	2023	0.928 ± 0.086	0.906 ± 0.086
SKiT [33]	2023	0.934 ± 0.052	0.913
EffNetV2 C-ECT+CAFF [65]	2023	0.949 ± 0.040	**0.928 ± 0.061**
EffNetV2 MS-ASCT [66]	2024	**0.953 ± 0.04**	0.925 ± 0.071
MTRCNet SingleNet (own training)	2024	0.792	0.844
Proposed model	2024	0.826	0.871

Table 8: Accuracy and F1 scores of the models discussed in this research on the Cholec80 dataset [56], the baseline model MTRCNet SingleNet (own training) and the best-performing ablation model, Model 3.

Method	Year	F1 Score
CUHK* [59]	2021	0.650
HIKVision* [59]	2021	0.654
CAMMA 1* [46]	2021	0.688
MuST [45]	2024	0.773
MTRCNet SingleNet (own training)	2024	0.761
Model 3	2024	**0.796**

Table 9: F1 score of Model 3 and the state-of-the-art models on the HeiChole benchmark dataset [59]. Models marked with * were evaluated on a separate test set that is not publicly available.

accuracies or F1 scores when compared to the baseline Model 1. This can partially be explained by the use of trained models for descriptor extraction, as mentioned earlier in this section. As shown by the results of the three separate branches for feature extraction in Section 5, the models do not perform perfectly. The extracted tool and action labels and segmentation maps are a prediction, and this prediction can be wrong. This means sometimes the descriptors do not match the input frames and can lead to the model receiving conflicting information. For example, if the tool recognition model incorrectly predicts the tool label, or the action model misclassifies an action, the cross-attention mechanism might focus on wrong or irrelevant features. This could decrease the effectiveness of the cross-attention block. Possible improvements for this will be discussed in future work in Section 6.4.

An interesting observation from the ablation study is the higher F1 scores achieved on the test set by Model 2 and Model 3 (0.732 and 0.739 respectively) compared to both Model 1 and Model 5 (0.726 and 0.720 respectively). This indicates that segmentation maps and action triplets seem to provide more valuable information for phase recognition compared to tool features. The relatively lower F1 score of Model 5 suggests that adding tool features may introduce noise or confusion to the model that negatively affects the performance when combined with the other descriptors. Overall, these findings show that segmentation and action triplets are the most critical visual cues for understanding surgical workflows, as they outperform the single visual cue model as well as the model with all visual cues combined. Improving the segmentation and action triplet extraction methods could therefore likely have the most significant effect on the overall model performance.

The accuracies of the individual phases in Table 6 gave a more in-depth look into the performance of the ablation models. Even though the accuracy and F1 score for Model 3 were the highest, the individual accuracies do not outperform the other ablation models for each class. Table 6 shows that only for the phase *Clipping Cutting* (P3) this model achieves the highest individual class accuracy. This discrepancy between the best total accuracy and best individual accuracies can be explained by the class imbalance. Phases with more frames, such as *Calot Triangle Dissectio*n (P2) and *Gallbladder Dissection* (P4), contribute more to the overall accuracy and F1 score Models can perform well in these phases despite not consistently achieving the highest individual accuracy for all phases. However, the fact that Model 3 also achieved the highest F1 score indicates that it balances precision and recall across all classes.

The pie charts in Figure 16 in Appendix C provide further insights into the challenges the model faces, especially for Model 4 and 5. These models incorporate tool data, but Figure 16 shows that there is a significant overlap in the tool usage across phases. The phases *Calot Triangle Dissection* and *Gallbladder Dissection* are often confused by the model, and the charts show that both of these phases mainly use the tools *Grasper* and *Hook*. Other phases that show very similar tool usage are *Gallbladder Packaging* and *Gallbladder Retraction*. As shown in the confusion matrices of Model 4 and 5 presented in Figure **??** in Appendix D, these phases also show some misclassification by the models, but not as much as the phases *Calot Triangle Dissection* and *Gallbladder Dissection*. The reason for

this could be that visually, *Gallbladder Packaging* and *Gallbladder Retraction* look more different because a completely different action is performed, and the other two phases perform the same action (dissection), just on a different target.

The confusion matrices in Figure 10 presented which phases the baseline Model 1 and best-performing Model 3 performed well on and which phases they struggled with the most. It shows that, for Model 3 mainly in *Preparation* (P1), *Calot Triangle Dissection* (P2) and *Clipping Cutting* (P3) there are a lot of misclassifications. For the *Preparation* phase, this is because many frames are classified as *Calot Triangle Dissection* frames. These are two subsequent phases and share several similarities. *Calot Triangle Dissection* is also one of the longer phases in the surgery, while *Preparation* only has very few frames in the Cholec80 dataset, as shown in Figure 4.

The two other phases that show a decent amount of misclassifications are *Calot Triangle Dissection* and *Clipping Cutting* (P3). Model 3 predicted *Calot Triangle Dissection* (P2) as *Gallbladder Dissection* (P4) in 14 percent of the cases. *Clipping Cutting* is predicted as *Gallbladder Dissection* 18 percent of the time. The confusion matrix for the baseline Model 1 shows very similar results. The most often misclassified phases are also *Preparation*, *Calot Triangle Dissection* and *Clipping Cutting*, as seen in Figure 10b. This confusion of the models is likely due to the similarities between the phases *Calot Triangle Dissection*, *Clipping Cutting* and *Gallbladder Dissection*. *Calot Triangle Dissection* and *Gallbladder Dissection* are both dissection phases, and actions in phase *Clipping Cutting* are very similar to dissecting. This could explain the fact that the model sometimes struggles to correctly predict these specific phases.

The results for the ablation study using the full Cholec80 dataset, presented in Table 7, shows the potential of using multiple visual cues in an attention mechanism framework, especially when enough data is available. The fact that Model 5, which combines all visual cues, outperforms the baseline model indicates that using additional input modalities improves the model's performance. However, similar to the ablation study on the subset of the data, the lower performance of Model 5 compared to Model 3 suggests that it is important to be selective when it comes to choosing the visual cues, otherwise the input can be confusing to the model and decrease the effectiveness of the cross-modal attention.

The higher scores for Model 3 compared to the other ablation models also reinforce the claim that action triplets are the most valuable modality in the multi-visual cue model. This can be explained by the amount of information present in the action labels. Including the full triplet (instrument, verb and target) in the label gives the full context of the current surgical activity taking place, compared to, for example, only using a tool label. This allows the model to better differentiate between phases and achieve a higher performance on the recognition task.

6.2 State of the Art Comparison

The best performing model in the ablation studies, Model 3, was compared to the state-of-the-art models discussed in Section 2.4 as well as the baseline Model 1. Table 8 shows that Model 3 achieves an accuracy of 0.826 and an F1 score of 0.871, outperforming earlier methods such as EndoNet (accuracy: 0.817, F1: 0.765) and MTRCNet SingleNet (accuracy: 0.792, F1: 0.844). These results further indicate that incorporating action triplets in a multi-visual cue framework has positive effects on the performance on the phase recognition task. However, the model's performance does not reach the scores of transformer-based architectures like EffNetV2 MS-ASCT (accuracy: 0.953, F1: 0.925) and SKiT (accuracy: 0.934, F1: 0.913). These transformer models can understand long-term patterns and spatial features, which explains their higher performance [27].

The improvement of Model 3 over the baseline does demonstrate the potential of integrating multiple visual cues for surgical phase recognition, even if it does not outperform the best current models. The baseline model itself was already one of the older, lower-performing models. Applying the cross-modal attention mechanisms proposed in this research to the best-performing state-of-the-art models could result in an even higher performance of the models.

Table 9 presents the comparison between the state-of-the-art models trained on the HeiChole dataset [59]. Model 3 achieves an F1 score of 0.796, outperforming MuST (0.773) and the MTRCNet SingleNet trained in this research (0.761). This result shows the robustness of the proposed approach. Compared to earlier methods like CUHK (0.650) and HIKVision (0.654), the best-performing ablation model, Model 3, shows significant improvement. Although these models were evaluated on a private test set, the difference in performance clearly shows the advantage of the proposed framework for phase recognition.

6.3 Research Questions

Based on the obtained results, the answers to the research questions in Section 1.2 are described below.

(1) For the task of automatic phase recognition in surgical videos, what computer vision-based methods exist and which input features do these methods use?

This research described several different computer vision methods for surgical phase recognition in Section 2.4. These methods including CNN-based, RNN-, LSTM-, and GRU-based models, as well as Transformer-based architectures and Vision Transformer-based networks. Each of these methods has its own strengths in processing the spatial or temporal data of surgical videos. Transformer-based models showed the highest performances, likely because of their ability to recognize patterns over longer sequences.

Input features commonly used include tool presence, semantic segmentation maps, and action triplets. Tool labels and action triplets can give contextual information about the current surgical steps, while segmentation maps can add important spatial information. These features can give a good representation of the surgical scene in a video, but have not been studied together. This study contributes by investigating the combination of these features.

(2) How can we leverage different visual descriptors extracted from surgical videos in a multi-visual cue framework for surgical phase recognition?

The proposed multi-visual cue model for surgical phase recognition incorporates three additional descriptors, along with the input frame. These descriptors are tool presence, segmentation maps, and action triplets. They are combined using cross-modal attention mechanisms. The descriptors are extracted from input frames

using trained models. The model pipeline uses two separate cross-attention blocks, one to align the raw image with the segmentation maps, and one to further integrate the action and tool features.

Cross-modal attention has become increasingly important in research because of its ability to capture relationships between different types of data, leading to improved performance in complex tasks. By using cross-modal attention, the proposed framework aligns and combines information from multiple visual cues, allowing a better representation of surgical procedures. This work aims to further explore the use of cross-modal attention within for task of surgical phase recognition, contributing to the ongoing research that uses this technique to enhance multi-modal learning. By focusing on the combination of multiple visual cues through cross-modal attention, the proposed approach addresses the limitations of single-cue models and builds upon existing methods for multi-modal learning.

(3) What effect does the inclusion of different descriptors in a multi-visual cue model have on the performance of automatic phase recognition in surgical videos compared to single visual cue models and the state of the art?

The proposed multi-visual cue framework using cross-attention achieved mixed results. The results presented in Tables 5 and 7 showed that mainly Model 3, which combines image data and action triplet labels through cross-attention, achieves competitive scores on the task of automatic phase recognition. This model outperforms the baseline, single visual cue Model 1 when trained and evaluated on the Cholec80 dataset. It does not, however, outperform the best state-of-the-art models on this dataset. As discussed in Section 6.2, the chosen baseline model could be one of the reasons for this, as that model already performs lower than the newer state-of-the-art models presented in this research. This model was chosen because it is the best-performing open-source model. When comparing Model 3 to the state-of-the-art models on the Heihole benchmark dataset, the results are even more promising. Model 3, the best configuration of the cross-modal attention framework as shown in the ablation study, improves upon existing methods and achieves a higher F1 score than the best performing state-of-the-art model MuST [45].

Another important finding in this research is the fact that when using cross-modal attention, it is important to be selective with the choice of input visual cues. The ablation studies showed that while the combination of multiple visual cues could improve the performance of the model, using the wrong descriptors or using too many descriptors could result in lower results than the single-visual cue baseline model. Especially the use of action triplets as additional visual cue improved the model performance according to the experiments conducted in this study.

6.4 Future Work

This research presented several limitations that can be addressed in future work. First of all, the same training setup and hyperparameters of the baseline model were used for all experiments of the phase prediction network, and also for the proposed model which uses attention mechanisms. This includes the number of epochs the models were trained for, as well as the batch size, sequence length and the learning rate. This decision was made to keep the comparison between the baseline model and the proposed model fair. Future work could investigate whether a new training setup and additional hyperparameter tuning would benefit the cross-attention in the proposed model. This could focus in particular on the training duration, as the model was only trained for 25 epochs due to hardware and time constraints. This is the same amount of epochs as the baseline model uses in the MTRCNet paper [24]. However, as attention mechanisms can often benefit from longer training, future work could include training the proposed model for more epochs to see if this positively affects the performance.

The models used for the extraction of the descriptors also presented limitations. It could be useful to further improve these input models, to make the descriptors more accurate. This way, the model can train with high-quality data, and future work could investigate if this enhances model performance. Additionally, an end-to-end framework could also be implemented instead of using separate models for descriptor extraction.

Another possibility for future work lies in applying the cross-modal attention mechanism to newer, better-performing state-of-the-art models. The proposed framework was based on an older baseline model and did not outperform the newest models on this task. The best-performing models discussed in this research are transformer-based methods, and cross-attention aligns better with these architectures since they already have attention mechanisms [58]. Combining the use of multiple visual cues with such a model as the baseline could therefore result in an even higher-performing model.

Another relevant area of future work is further investigating which of the modalities is the most relevant for the model. The ablation study presented in Section 5.2 is a start, but other methods can be used to better explain exactly which modality provides the most valuable information for the model. Works on the importance of specific modalities within multi-modal models already exist, as well as methods for determining the effect of each modality [15, 61]. Implementing such approaches can give a more comprehensive evaluation of different modalities and help in the development of more efficient and interpretable multi-modal models.

7 CONCLUSION

This study introduced a multi-visual cue method framework using cross-modal attention for automatic phase recognition in surgical videos. By using descriptors such as tool presence, segmentation maps, and action triplets, the proposed pipeline addresses certain limitations of single-modality models and provides a robust network for phase prediction. The results demonstrate that while the multi-visual cue model can represent the surgical scene in more detail, its performance is very dependent on the choice and combination of descriptors.

The best-performing configuration in the ablation study conducted in this research was the model using image data and action triplets, which outperformed the baseline model in accuracy and F1 score on the Cholec80 dataset. However, the model did not outperform other state-of-the-art, transformer-based models on this dataset. The results on the HeiChole dataset do show an improvement compared to the existing state-of-the-art methods. The model using action triplets achieved a higher F1 score than MuST, CUHK,

HIKVision, and CAMMA 1, indicating that the proposed approach performs well across different datasets.

Despite these results, the approach still has several limitations. The training duration was constrained by computational resources, and hyperparameters were kept constant across the different models in order to make a fair comparison. Future work should explore longer training durations, more specific hyperparameter tuning, and the use of better-performing descriptor extraction methods. Additionally, integrating the proposed approach with state-of-the-art transformer models instead of the current baseline model could further improve the model performance, since transformers naturally align with cross-attention mechanisms.

Overall, this study presents a comprehensive analysis of surgical phase prediction models, and contributes to this field by introducing a cross-modal attention-based framework that integrates tool presence, action triplets, and segmentation maps. The field shows promising results that, after some refinements, can improve surgical workflow management, error reduction, and feedback during surgeries in real-life applications.

REFERENCES

[1] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/

[2] Linn Boberg, Jagdeep Singh, Agneta Montgomery, and Peter Bentzer. 2022. Environmental impact of single-use, reusable, and mixed trocar systems used for laparoscopic cholecystectomies. *PloS One* 17, 7 (2022), e0271601. https://doi.org/10.1371/journal.pone.0271601

[3] Zhen Chen, Qingyu Guo, Leo K. T. Yeung, Danny T. M. Chan, Zhen Lei, Hongbin Liu, and Jinqiao Wang. 2023. Surgical Video Captioning with Mutual-Modal Concept Alignment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (Lecture Notes in Computer Science)*, Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (Eds.). Springer Nature Switzerland, Cham, 24–34. https://doi.org/10.1007/978-3-031-43996-4_3

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. http://arxiv.org/abs/2112.01527 arXiv:2112.01527 [cs].

[5] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. https://doi.org/10.48550/arXiv.2107.06278 arXiv:2107.06278.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. http://arxiv.org/abs/1406.1078 arXiv:1406.1078 [cs, stat].

[7] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. 2020. TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks. Vol. 12263. 343–352. https://doi.org/10.1007/978-3-030-59716-0_33 arXiv:2003.10751 [cs, eess].

[8] Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26, 3 (July 1945), 297–302. https://doi.org/10.2307/1932409 Publisher: John Wiley & Sons, Ltd.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. http://arxiv.org/abs/2010.11929 arXiv:2010.11929 [cs].

[10] Neda Farhangmehr and Donald Menzies. 2021. Laparoscopic cholecystectomy: from elective to urgent surgery. *Laparoscopic Surgery* 5, 0 (Jan. 2021). https://doi.org/10.21037/ls-20-46 Number: 0 Publisher: AME Publishing Company.

[11] Germain Forestier, Laurent Riffaud, and Pierre Jannin. 2015. Automatic phase prediction from low-level surgical activities. *International Journal of Computer Assisted Radiology and Surgery* 10, 6 (June 2015), 833–841. https://doi.org/10.1007/s11548-015-1195-0

[12] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. 2021. Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. https://doi.org/10.48550/arXiv.2103.09712 arXiv:2103.09712 [cs].

[13] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 249–256. https://proceedings.mlr.press/v9/glorot10a.html ISSN: 1938-7228.

[14] Maria Grammatikopoulou, Ricardo Sanchez-Matilla, Felix Bragman, David Owen, Lucy Culshaw, Karen Kerr, Danail Stoyanov, and Imanol Luengo. 2023. A spatio-temporal network for video semantic segmentation in surgical videos. http://arxiv.org/abs/2306.11052 arXiv:2306.11052 [cs].

[15] Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. 2023. Modality Influence in Multimodal Machine Learning. https://doi.org/10.48550/arXiv.2306.06476 arXiv:2306.06476 [cs].

[16] Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. 2018. Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery* 268, 1 (July 2018), 70. https://doi.org/10.1097/SLA.0000000000002693

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385 arXiv:1512.03385 [cs].

[18] Shruti R. Hegde, Babak Namazi, Niyenth Iyengar, Sarah Cao, Alexis Desir, Carolina Marques, Heidi Mahnken, Ryan P. Dumas, and Ganesh Sankaranarayanan. 2024. Automated segmentation of phases, steps, and tasks in laparoscopic cholecystectomy using deep learning. *Surgical Endoscopy* 38, 1 (Jan. 2024), 158–170. https://doi.org/10.1007/s00464-023-10482-3

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[20] W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih. 2020. CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80. http://arxiv.org/abs/2012.12453 arXiv:2012.12453 [cs].

[21] Arnaud Huaulmé, Pierre Jannin, Fabian Reche, Jean-Luc Faucheron, Alexandre Moreau-Gaudry, and Sandrine Voros. 2020. Offline identification of surgical deviations in laparoscopic rectopexy. *Artificial Intelligence in Medicine* 104 (April 2020), 101837. https://doi.org/10.1016/j.artmed.2020.101837

[22] Nour Aldeen Jalal, Tamer Abdulbaki Alshirbaji, Paul D. Docherty, Thomas Neumuth, and Knut Moeller. 2021. A Deep Learning Framework for Recognising Surgical Phases in Laparoscopic Videos. *IFAC-PapersOnLine* 54, 15 (Jan. 2021), 334–339. https://doi.org/10.1016/j.ifacol.2021.10.278

[23] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. 2018. SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. *IEEE Transactions on Medical Imaging* 37, 5 (May 2018), 1114–1126. https://doi.org/10.1109/TMI.2017.2787657

[24] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. 2019. Multi-Task Recurrent Convolutional Network with Correlation Loss for Surgical Video Analysis. http://arxiv.org/abs/1907.06099 arXiv:1907.06099 [cs, eess].

[25] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. 2021. Temporal Memory Relation Network for Workflow Recognition from Surgical Video. https://doi.org/10.48550/arXiv.2103.16327 arXiv:2103.16327 [cs].

[26] William E. Kelley. 2008. The Evolution of Laparoscopy and the Revolution in Surgery in the Decade of the 1990s. *JSLS : Journal of the Society of Laparoendoscopic Surgeons* 12, 4 (2008), 351–357. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3016007/

[27] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. *Comput. Surveys* 54, 10s (Jan. 2022), 1–41. https://doi.org/10.1145/3505244 arXiv:2101.01169 [cs].

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[29] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin. 2012. A Framework for the Recognition of High-Level Surgical Tasks From Video Images for Cataract Surgeries. *IEEE Transactions on Biomedical Engineering* 59, 4 (April 2012), 966–976. https://doi.org/10.1109/TBME.2011.2181168

[30] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. http://arxiv.org/abs/1608.08242 arXiv:1608.08242 [cs].

[31] Yuchong Li, Tong Xia, Huoling Luo, Baochun He, and Fucang Jia. 2023. MT-FiST: A Multi-Task Fine-Grained Spatial-Temporal Framework for Surgical Action Triplet Recognition. *IEEE Journal of Biomedical and Health Informatics* 27, 10 (Oct. 2023), 4983–4994. https://doi.org/10.1109/JBHI.2023.3299321

[32] Yang Liu, Maxence Boels, Luis C. Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2023. LoViT: Long Video Transformer for Surgical Phase Recognition. http://arxiv.org/abs/2305.08989 arXiv:2305.08989 [cs].

[33] Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2023. SKiT: a Fast Key Information Video

Transformer for Online Surgical Phase Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 21017–21027. https://doi.org/10.1109/ICCV51070.2023.01927

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. http://arxiv.org/abs/2103.14030 arXiv:2103.14030 [cs].

[35] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, Derong Yu, Guoyan Zheng, Xiaotian Duan, Neil Getty, Ricardo Sanchez-Matilla, Maria Robu, Li Zhang, Huabin Chen, Jiacheng Wang, Liansheng Wang, Bokai Zhang, Beerend Gerats, Sista Raviteja, Rachana Sathish, Rong Tao, Satoshi Kondo, Winnie Pang, Hongliang Ren, Julian Ronald Abbing, Mohammad Hasan Sarhan, Sebastian Bodenstedt, Nithya Bhasker, Bruno Oliveira, Helena R. Torres, Li Ling, Finn Gaida, Tobias Czempiel, João L. Vilaça, Pedro Morais, Jaime Fonseca, Ruby Mae Egging, Inge Nicole Wijma, Chen Qian, Guibin Bian, Zhen Li, Velmurugan Balasubramanian, Debdoot Sheet, Imanol Luengo, Yuanbo Zhu, Shuai Ding, Jakob-Anton Aschenbrenner, Nicolas Elini van der Kar, Mengya Xu, Mobarakol Islam, Lalithkumar Seenivasan, Alexander Jenke, Danail Stoyanov, Didier Mutter, Pietro Mascagni, Barbara Seeliger, Cristians Gonzalez, and Nicolas Padoy. 2023. CholecTriplet2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis* 86 (May 2023), 102803. https://doi.org/10.1016/j.media.2023.102803 arXiv:2204.04746 [cs].

[36] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2020. Recognition of Instrument-Tissue Interactions in Endoscopic Videos via Action Triplets. Vol. 12263. 364–374. https://doi.org/10.1007/978-3-030-59716-0_35 arXiv:2007.05405 [cs, eess].

[37] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2022. Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos. *Medical Image Analysis* 78 (May 2022), 102433. https://doi.org/10.1016/j.media.2022.102433 arXiv:2109.03223 [cs].

[38] Katherine A. O'Hanlan, Suzanne L. Dibble, Anne-Caroline Garnier, and Mirjam Leuchtenberger Reuland. 2007. Total Laparoscopic Hysterectomy: Technique and Complications of 830 Cases. *JSLS : Journal of the Society of Laparoendoscopic Surgeons* 11, 1 (2007), 45–53. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3015816/

[39] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. http://arxiv.org/abs/1511.08458 arXiv:1511.08458 [cs].

[40] Xiaoying Pan, Xuanrong Gao, Hongyu Wang, Wuxia Zhang, Yuanzhen Mu, and Xianli He. 2023. Temporal-based Swin Transformer network for workflow recognition of surgical video. *International Journal of Computer Assisted Radiology and Surgery* 18, 1 (Jan. 2023), 139–147. https://doi.org/10.1007/s11548-022-02785-y

[41] Bogyu Park, Hyeongyu Chi, Bokyung Park, Jiwon Lee, Hye Su Jin, Sunghyun Park, Woo Jin Hyung, and Min-Kook Choi. 2023. Visual modalities-based multimodal fusion for surgical phase recognition. *Computers in Biology and Medicine* 166 (Nov. 2023), 107453. https://doi.org/10.1016/j.compbiomed.2023.107453

[42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (Oct. 2017). https://openreview.net/forum?id=BJJsrmfCZ

[43] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. 2023. Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors* 23, 5 (Jan. 2023), 2381. https://doi.org/10.3390/s23052381 Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[44] A. Preethi and P. Dhanalakshmi. 2023. Video Captioning using Pre-Trained CNN and LSTM. In *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*. 1–7. https://doi.org/10.1109/IConSCEPT57958.2023.10170131

[45] Alejandra Pérez, Santiago Rodríguez, Nicolás Ayobi, Nicolás Aparicio, Eugénie Dessevres, and Pablo Arbeláez. 2024. MuST: Multi-scale Transformers for Surgical Phase Recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (Eds.). Springer Nature Switzerland, Cham, 422–432. https://doi.org/10.1007/978-3-031-72089-5_40

[46] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, Georgios Exarchakis, Alexandros Karargyris, and Nicolas Padoy. 2023. Dissecting Self-Supervised Learning Methods for Surgical Computer Vision. https://doi.org/10.48550/arXiv.2207.00449 arXiv:2207.00449 [cs].

[47] Manish Sahu, Anirban Mukhopadhyay, Angelika Szengel, and Stefan Zachow. 2017. Addressing multi-label imbalance problem of surgical tool detection using CNN. *International Journal of Computer Assisted Radiology and Surgery* 12, 6 (June 2017), 1013–1020. https://doi.org/10.1007/s11548-017-1565-x

[48] Manish Sahu, Angelika Szengel, Anirban Mukhopadhyay, and Stefan Zachow. 2020. Surgical phase recognition by learning phase transitions. *Current Directions in Biomedical Engineering* 6 (Sept. 2020), 20200037. https://doi.org/10.1515/cdbme-2020-0037

[49] Duygu Sarikaya and Pierre Jannin. 2020. Surgical Gesture Recognition with Optical Flow only. http://arxiv.org/abs/1904.01143 arXiv:1904.01143 [cs].

[50] Paul Maria Scheikl, Stefan Laschewski, Anna Kisilenko, Tornike Davitashvili, Benjamin Müller, Manuela Capek, Beat P. Müller-Stich, Martin Wagner, and Franziska Mathis-Ullrich. 2020. Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. *Current Directions in Biomedical Engineering* 6, 1 (May 2020). https://doi.org/10.1515/cdbme-2020-0016 Publisher: De Gruyter.

[51] Robin M. Schmidt. 2019. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. http://arxiv.org/abs/1912.05911 arXiv:1912.05911 [cs, stat].

[52] Nisarg A. Shah, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. 2023. $$\textsf{GLSFormer}$$: Gated - Long, Short Sequence Transformer for Step Recognition in Surgical Videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (Lecture Notes in Computer Science)*, Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (Eds.). Springer Nature Switzerland, Cham, 386–396. https://doi.org/10.1007/978-3-031-43996-4_37

[53] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab. 2017. The TUM LapChole dataset for the M2CAI 2016 workflow challenge. https://doi.org/10.48550/arXiv.1610.09278 arXiv:1610.09278 [cs].

[54] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. 2019. High-Resolution Representations for Labeling Pixels and Regions. http://arxiv.org/abs/1904.04514 arXiv:1904.04514 [cs].

[55] Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. http://arxiv.org/abs/2104.00298 arXiv:2104.00298 [cs].

[56] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2016. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. http://arxiv.org/abs/1602.03012 arXiv:1602.03012 [cs] version: 2.

[57] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2018. Weakly-Supervised Learning for Tool Localization in Laparoscopic Videos. http://arxiv.org/abs/1806.05573 arXiv:1806.05573 [cs].

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs].

[59] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Hannes G. Kenngott, Felix Nickel, Moritz von Frankenberg, Franziska Mathis-Ullrich, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. 2021. Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark. https://doi.org/10.48550/arXiv.2109.14956 arXiv:2109.14956 [eess].

[60] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. 2022. AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparoscopic Hysterectomy. http://arxiv.org/abs/2208.02049 arXiv:2208.02049 [cs].

[61] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing multimodal cooperation via sample-level modality valuation. https://doi.org/10.48550/arXiv.2309.06255 arXiv:2309.06255 [cs].

[62] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging* 9, 4 (Aug. 2018), 611–629. https://doi.org/10.1007/s13244-018-0639-9 Number: 4 Publisher: SpringerOpen.

[63] Fangqiu Yi and Tingting Jiang. 2019. Hard Frame Detection and Online Mapping for Surgical Phase Recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Lecture Notes in Computer Science)*, Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (Eds.). Springer International Publishing, Cham, 449–457. https://doi.org/10.1007/978-3-030-32254-0_50

[64] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. 2021. ASFormer: Transformer for Action Segmentation. https://doi.org/10.48550/arXiv.2110.08568 arXiv:2110.08568 [cs].

[65] Bokai Zhang, Alberto Fung, Meysam Torabi, Jocelyn Barker, Genevieve Foley, Rami Abukhalil, Mary Lynn Gaddis, and Svetlana Petculescu. 2023. C-ECT: Online Surgical Phase Recognition with Cross-Enhancement Causal Transformer. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, Cartagena, Colombia, 1–5. https://doi.org/10.1109/ISBI53787.2023.10230841

[66] Bokai Zhang, Jiayuan Meng, Bin Cheng, Dean Biskup, Svetlana Petculescu, and Angela Chapman. 2024. Friends Across Time: Multi-Scale Action Segmentation Transformer for Surgical Phase Recognition. https://doi.org/10.48550/arXiv.

2401.11644 arXiv:2401.11644 [cs].

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, 5122–5130. https://doi.org/10.1109/CVPR.2017.544

[68] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. http://arxiv.org/abs/2012.07436 arXiv:2012.07436 [cs].

[69] Odysseas Zisimopoulos, Evangello Flouty, Imanol Luengo, Petros Giataganas, Jean Nehme, Andre Chow, and Danail Stoyanov. 2018. DeepPhase: Surgical Phase Recognition in CATARACTS Videos. https://doi.org/10.48550/arXiv.1807.10565 arXiv:1807.10565 [cs, stat].

A MODEL DIAGRAMS

This section presents the model diagrams of the three models used for feature extraction: MTRCNet [24], MT-FiST [31], and MaskFormer [5].

A.1 Tool Recognition Model

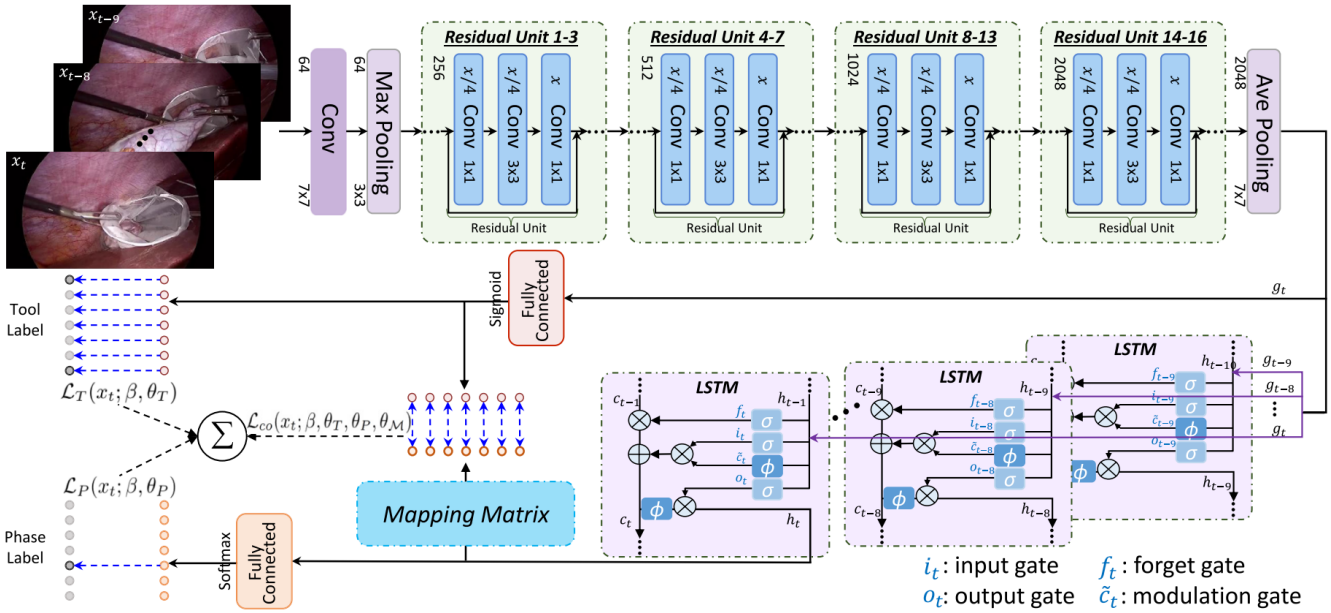


Figure 12: Architecture of the MTRCNet for tool presence detection [24]. Only the tool branch is used for tool label extraction.

A.2 Action Recognition Model

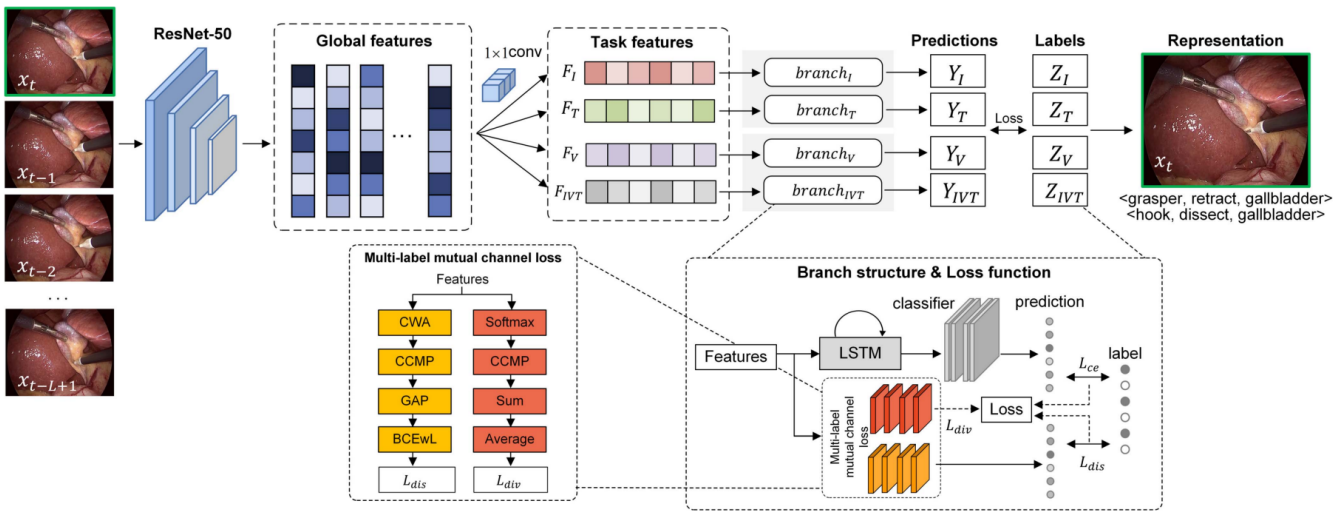


Figure 13: Architecture of the MT-FiST model used for action triplet detection [31].

A.3 Segmentation Model

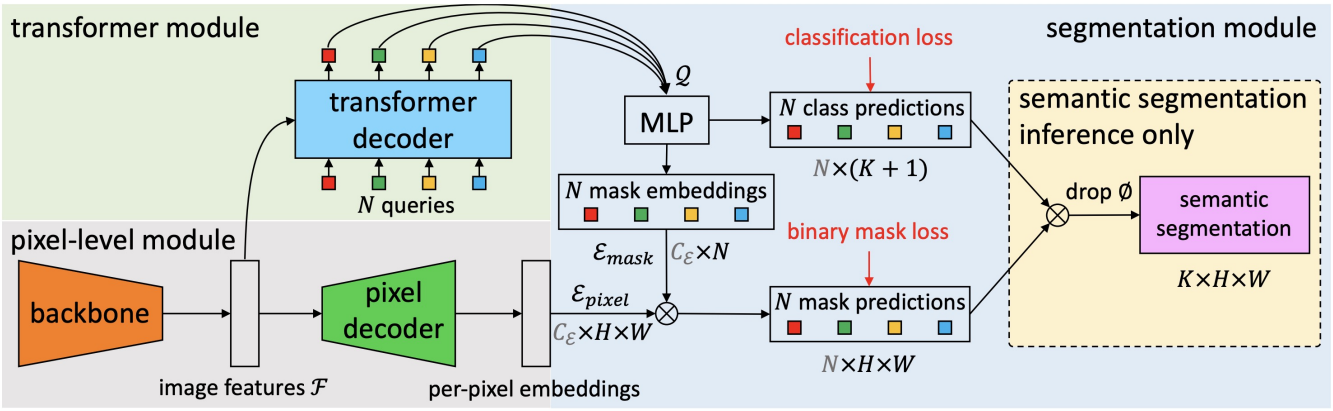


Figure 14: Architecture of the MaskFormer used for segmentation [5].

B VALIDATION PLOTS

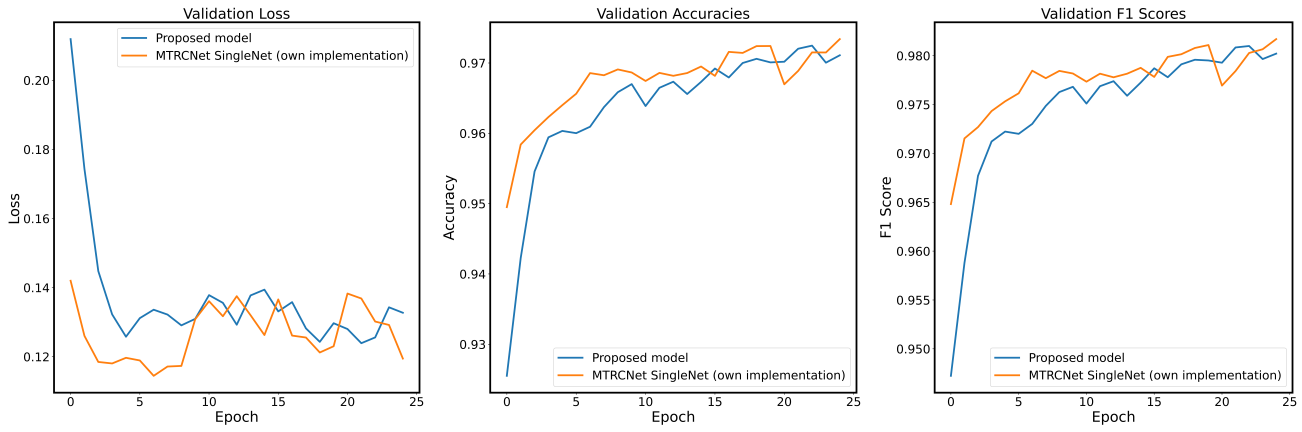


Figure 15: Graphs showing the validation loss, accuracy and F1 scores of the proposed model and the MTRCNet (own implementation) for a single fold of the training.

C CHOLEC80 TOOL DISTRIBUTION

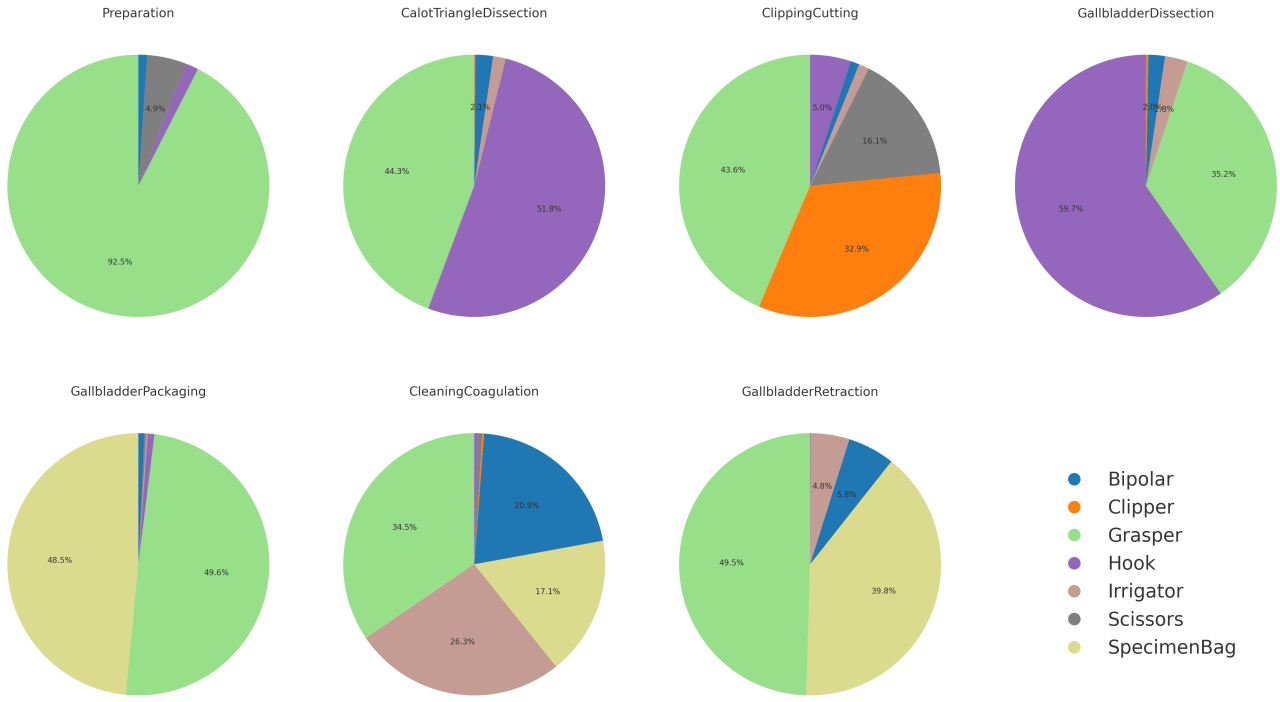
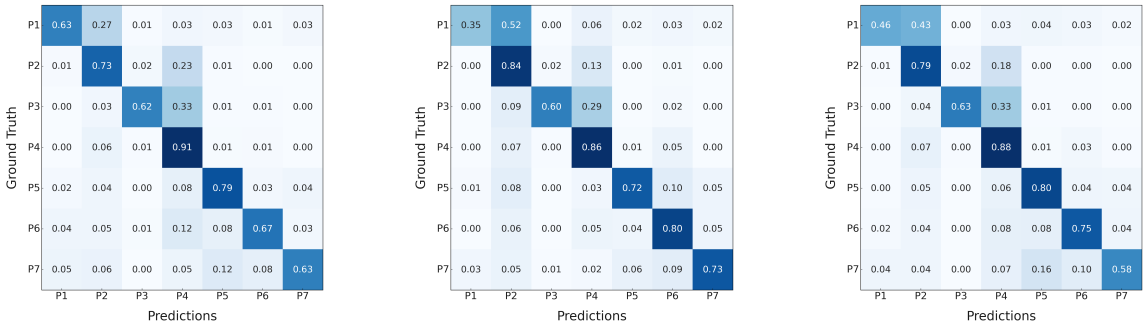


Figure 16: Pie charts showing the appearance of each tool per phase in the Cholec80 dataset [56].

D CORRELATION MATRICES



(a) Model 2 **(b) Model 4** **(c) Model 5**

Figure 17: Confusion matrices showing the predictions of the remaining models in the ablation study presented in Table 5. The phases are *Preparation* (P1), *Calot triangle dissection* (P2), *Clipping and cutting* (P3), *Gallbladder dissection* (P4), *Gallbladder packaging* (P5), *Cleaning and coagulation* (P6), and *Gallbladder retraction* (P7). Colours and values in the matrix are normalized based on the number of class instances.