Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# Anticipation for surgical workflow through instrument interaction and recognized signals

Kun Yuan<sup>a,\*\*</sup>, Matthew Holden<sup>a</sup>, Shijian Gao<sup>b</sup>, Wonsook Lee<sup>a</sup>

<sup>a</sup>University of Ottawa, Ottawa, K1N 6N5, Canada <sup>b</sup>University of Minnesota, Twin Cities, Minneapolis,55455, USA

# ARTICLE INFO

Article history: Received 1 May 2022 Received in final form 10 May 2022 Accepted 13 May 2022 Available online 15 May 2022

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

*Keywords:* Surgical Workflow Analysis, Anticipation, Temporal Convolutional Networks, Endoscopic Videos

# ABSTRACT

Surgical workflow anticipation is an essential task for computer-assisted intervention (CAI) systems. It aims at predicting the future surgical phase and instrument occurrence, providing support for intra-operative decision-support system. Recent studies have promoted the development of the anticipation task by transforming it into a remaining time prediction problem, but without factoring the surgeons' behaviors and the patient's status in the network design. In this paper, we propose an Instrument Interaction Aware Anticipation Network (IIA-Net) to overcome the previous deficiency while retaining the merits of two-stage models through using spatial feature extractor and temporal model. Spatially, feature extractor utilizes tooltips' movement to extracts the instrument-instrument interaction, which helps model concentrate on the surgeon's actions. On the other hand, it introduces the segmentation map to capture the rich instrument-surrounding features about the instrument surroundings. Temporally, the temporal model applies the causal dilated multi-stage temporal convolutional network to capture the long-term dependency in the long and untrimmed surgical videos with a large receptive field. Our IIA-Net enforces an online inference with reliable predictions even with severe noise and artifacts in the recorded videos and presence signals. Extensive experiments on Cholec80 dataset demonstrate the performance of our proposed method exceeds the state-of-the-art method by a large margin (1.40 v.s. 1.75 for  $MAE_{in}$ and 2.14 v.s. 2.68 for  $MAE_e$ ). For reproduction purposes, all the original codes are made public at https://github.com/Flaick/Surgical-Workflow-Anticipation.

© 2023 Elsevier B. V. All rights reserved.

# 1. Introduction

Context-aware assistance is essential to CAI systems, within which the highly relevant task is surgical workflow anticipation. It anticipates the occurrence of surgical instruments and phases before they appear, enabling the efficient instrument preparation and intelligent robot assistance system design (Rivoir et al., 2020; Forestier et al., 2017). Also, it can enhance patient safety, reduce surgical errors and facilitate communication in the operating room (OR) (Maier-Hein et al., 2017). For example, anticipating the surgical instrument's usage can provide vital input to physicians in the form of early warning in cases of deviations and anomalies. Anticipating the surgical phases can also help a robotic system identify events, such as bleeding, beforehand and decide when to intervene.

Recent models (Twinanda et al., 2018; Rivoir et al., 2020) for surgical workflow anticipation possess spatial-temporal limitations. Spatially, they use AlexNet(Krizhevsky et al., 2012), VGG(Simonyan and Zisserman, 2014) and similar architectures to extract a feature vector, representing instrument/phase





<sup>\*</sup>Part of this work has been presented in MICCAI 2021 (Yuan et al., 2021);

<sup>\*\*</sup>Corresponding author Tel.: +33-0768821690;

e-mail: kyuan033@uottawa.ca (Kun Yuan)



Fig. 1. Anticipation frameworks. Given an observed sequence and current time instant,  $T_{obs}$ , bottom part shows the conventional anticipation works that predicts dense segmentations. The upper part shows our strategy handling anticipation task as a real-time remaining time prediction task.

presence for each frame. However, they ignore the presence of task-specific combinations (i.e., instrument-instrument and instrument-surrounding interactions) in surgical anticipation applications. This information precisely reflects the surgeon's intention and patient's anatomy status, helping models generalize to the low-quality input materials (Klank et al., 2008) and variability of patient's anatomy and surgeon style (Funke et al., 2019). Novelly, our IIA-Net addresses instrument-instrument interaction in the form of a correlation matrix and designed geometric relations among instruments. Also, the instrumentsurrounding interaction is included via the semantic segmentation map. This makes our extracted feature sufficiently representative to identify the trigger event for the subsequent instrument and phase occurrence.

Temporally, existing works have difficulty handling nonstationary time series, especially for surgical workflow whose laparoscopic surgery transitions among instruments and phases are ambiguous and various. This requires the temporal modeling method to integrate recent observations with the long-range context in a computationally efficient way. However, the widely used RNNs(Hochreiter and Schmidhuber, 1997) learn a pattern from shorts segments of time series and apply it to other parts to get predictions, losing the distant observation information. In this work, we opt for dilated temporal convolutions to handle the full resolution of time series. This aids temporal pattern modeling and does not require complex computational resources.

For the anticipation formulation, initial works (Abu Farha et al., 2018; Forestier et al., 2017; Du et al., 2016; Mahmud et al., 2017) handle anticipation as a dense segmentation prediction task, shown in the bottom of Fig. 1. They require a preloading process before performing anticipation, limiting their usage in online surgical applications. Specifically, (Abu Farha et al., 2018) needs to observe at least 10%/20% of the video before it starts the prediction. Also, Fig. 1 shows an example where the predicted dense segmentation usually contains the short segments, which are ambiguous to determine the trending

of the instrument's presence.

Our contribution is four-fold: (1) Spatially, we propose a novel instrument interaction module (IIM) for the feature extraction process. (2) Temporally, we apply, for the first time, the causal dilated multi-stage temporal convolutional network (MS-TCN) structure to surgical workflow anticipation, with an accurate and fast online inference. (3) We combine spatial and temporal information to form a two-step IIA-Net for surgical workflow anticipation. (4) We propose a multi-task learning schema to jointly anticipate instrument and phase occurrence, which are important challenges in surgical workflow anticipation.

# 2. Related Work

# 2.1. Online Action Anticipation

Online action anticipation starts from the early event detection, aiming at detecting the event as soon as possible after it starts but before it ends. Also, it assumes that the target event is known and the model has already observed partially the event. (Ma et al., 2016) designed the model to jointly recognize the category and detect the start point of an action after observing only a fraction of the activity. (Sadegh Aliakbarian et al., 2017) utilizes the multi-stage LSTM to leverages context-aware and action-aware features, and introduces a novel loss function to anticipate the action as early as possible. However, these early attempts do not consider the online scenario and the formulation of online action anticipation is not fully defined. Our task is different from the early event detection as we aim to predict the future before observing any frames of the activity. Therefore, our work falls under the category of online action anticipation problem.

Recently, the rise of deep learning methods starts to tackle this problem to anticipate short-horizon of human actions. For action anticipation, early work (Lan et al., 2014) was based on traditional hand-crafted features. (Vondrick et al., 2016) predicts the feature vector of future frame and apply recognition algorithm to categorize it into actions. However, it only anticipates for single fixed time which is not desirable to represent the full future. Also, the model only take a single past frame as input instead a video clip, which is inaccurate to recognize the current status and make the prediction. Extending from the previous works, (Gao et al., 2017) apply a sliding window to anticipate a fixed sequence of future representations based on the multiple frames in a real-time manner. Also, Villegas et al. (2017) trains the network to learn a high-level pose representation and predicts the low-level pixel representation, which avoids the error propagation happening in recurrent networks. Different from predicting the sequence of visual representation, the (Abu Farha et al., 2018) directly predict the future sequential action labels either recursively with RNN or in one single step with CNN. While CNN-based method is only feasible when there is at least 10% of observed video, the RNNbased method usually suffers from the error prorogation and can not handle long-range videos. To handle the long-range video, (Sener et al., 2020) presented a temporal aggregation block to efficiently aggregate the representation from the spanning and recent observations for the long-range video action anticipation. (Ke et al., 2019) proposes to predict the long-term action anticipation in one shot by explicitly conditioning the network on time and forming a coarse-to-fine structure. Although these methods obtain good results in the natural video domain, their methods still have many limitations when processing the laparoscopic videos.

## 2.2. Surgical Workflow Anticipation

Surgical workflow anticipation, including surgical phase, instrument and action anticipation are quite important for computer-assisted intervention (CAI) applications of computer vision. While the aforementioned methods on video action anticipation have achieved state-of-the-art performance on the benchmark video datasets, such as UCF101 (Soomro et al., 2012), surgical video anticipation is rarely explored by current works due to the incomplete definition and formulation of this task. Traditional works on this problem have focused on the phase anticipation by using the Surgical process models (SPMs) to decompose the surgery into well-defined work steps and Dynamic Time Warping (DTW) for calculating the similarity between surgery procedures. For example, (Franke and Neumuth, 2015) explores an adaptive state-transition model by considering the difference between the ongoing procedure and the training set procedure instance using DTW. It jointly recognizes the current phase and predict the next phase in surgery. (Forestier et al., 2017) defined the problem as finding the optimal alignment between the partial sequence and the complete reference sequence of surgical activities, by proposing the algorithm to maximize the posteriori probability estimation. However, these are the template-based methods that are sensitive to the choice of the reference surgery procedure. Also, the need for SPMs limits the model's ability to process the frame-based laprascopic surgery. Hence, we propose to use CNN architecture jointly with recognition model to include both visual feature and activity feature in our work.

Deep learning methods have enlightened this task by the power of representational ability. For instance, (Rivoir et al.,

2020) bases on the raw surgical video and reformulates the surgical instrument anticipation task into remaining time prediction with uncertainty estimation, proving to be more applicable to real-world applications with sparse tool usage. (Ban et al., 2021) proposes and encoder-decoder predictor based on a discrete generative adversarial network to jointly predict the future surgical phase and the transitions. However, all above methods only focus either the phase or instrument anticipation task. Also, the previous works extract the feature vector for each frame solely from the raw RGB image, with little effort devoted to model surgeon's intention and patient's status for anticipation task with few exceptions, such as prediction of person behaviour under the surveillance camera (Liang et al., 2019). Therefore, inspired by these previous studies, we propose the Instrument Interaction Aware Network (IIA-Net) with multi-tasking strategy to enhance the representation ability of our feature extractor and handle phase and instrument anticipation within one model.

# 2.3. Temporal Modeling

Video can be represented by a sequence of feature vectors, requiring the temporal sequence modeling method to process the current time's prediction conditioned on the previous observations. Therefore, Hidden Markov Models (HMM) are initially applied for the online predictions of surgical phases (Padoy et al., 2012). As the rising of deep learning networks, the combination of CNNs and Recurrent neural network (RNN) has been widely used. For example, (Jin et al., 2020, 2017) propose to utilize the CNNs as feature extractors and the RNNs as the temporal modeling methods to refine the temporal output. Specifically, LSTMs are used due to its converging speed and ability to learn long-term dependency. In recent years, Lea et al. (2017) proved that CNN can be a valuable tool for sequence modeling and forecasting, give the right modification. Therefore, Temporal Convolutional Networks (TCNs) are applied and achieve an exciting result for surgical video phase segmentation (Czempiel et al., 2020). Unlike the RNNs modeling the sequence iteratively, the TCN takes the whole sequence as input at once. Also, the TCN is fully implemented by convolution layers, leading to a unified network structure when handling the video classification task. Recently, the rise of transformer networks (Vaswani et al., 2017; Gao et al., 2021; Czempiel et al., 2021) helps the deep learning methods overcome the challenges in surgical workflow analysis, thanks to their vast potential for sequential modeling in long-range sequences. Transformer can calculate the temporal relationships between current and previous frames using self-attention layers. Also, self-attention enables learning in long sequences without forgetting of previous information which often hampers LSTM-based methods. For example, (Czempiel et al., 2021) creates an attention regularization loss to encourage the model to focus on high-quality frames during training the transformer. For the time variance problem, the Dynamic Time Warping (DTW) warps the sequences into the same speed ratio and is useful for the sequence classification tasks. Recently, Lohit et al. (2019) proposes to integrate warping mechanism to shrink the intra-classes difference and enlarge the inter-classes difference. However, warping-based sequence



Fig. 2. Overview of the proposed model. For each frame observed, its estimated semantic map and tool detection are forwarded to instrument interaction module (IIM) to extract interaction feature. The recognized phase and tool signal are fed into temporal model jointly with interaction and visual features.

modeling needs the template sequence which is infeasible for most of the cases.

#### 3. Methodology

Our IIA-Net composes of two parts, a feature extractor with an Instrument Interaction Module (IIM) and a temporal model using MS-TCN. Spatially, IIA-Net models the surgeon's intention through extracting rich geometric features of the instrument-instrument interactions and semantic features of instrument-surrounding interactions. Motivated by the recognition methods (Twinanda et al., 2016; Jin et al., 2017; Padoy, 2019; Jin et al., 2018), we introduce tool and phase signal to boost the feature extraction process. Temporally, we utilize causal dilated MS-TCN (Farha and Gall, 2019) to capture longterm patterns with a large receptive field. Unlike the dense segmentation prediction, shown in Fig. 1, our IIA-Net follows (Rivoir et al., 2020) to handle anticipation as a real-time remaining time regression problem without any latency or pre-loading process.

#### 3.1. Task Formulation

Inspired by (Rivoir et al., 2020), we process the anticipation task as a regression problem both for instrument and phase anticipation, aiming to predict the remaining time until the occurrence of one of  $\tau$  surgical phases and  $\alpha$  instruments within a future horizon of *h* minutes. Given a timestamp *i* from video *x*, we firstly extract semantic map  $s_i$  and instrument bounding boxes  $b_i$  for each video frame  $x_i$ . At the same time, we obtain the instrument presence signal  $t_i$  and phase signal  $p_i$ . Given the observed sequence { $(x_1, s_1, b_1, t_1, p_1), ...(x_{T_{obs}}, s_{T_{obs}}, t_{T_{obs}}, p_{T_{obs}})$ } from time 1 to  $T_{obs}$ , IIA-Net predicts the remaining time until the phase/instrument occurrence, the ground truth is denoted as  $r_{T_{obs}}(\tau/\alpha)$ .

There is no need for extra human effort to annotate the ground truth for remaining time prediction. Specifically, the ground truth's value  $r_i(\tau/\alpha)$  for each timestamp *i* can be calculated from the existing phase/instrument presence annotations  $P^{\tau/\alpha}$ , which is a 0/1 signal from Cholec80 (Twinanda et al.,

2016) dataset. There are three categories for each frame, 'nonpresence' frames that the certain phase/instrument will not happen in next h minutes; 'anticipating' frames that the occurrence can be foreseen within next h minutes; 'presence' frames that the specific phase/instrument is current happening.

The construction method of ground truth starts from the ending point, and assigns 0 to the frame with occurrence, i.e., 'presence' frames. Then it will assign the true remaining time in minutes to the 'anticipating' frames, which are close to the occurrence. The values are truncated at *h* minutes because the the anticipation model shall not predict the future with arbitrary long interval. The ground truth ranges [0, h], where 0 denotes that the  $\tau/\alpha$  is currently happening and *h* denotes that  $\tau/\alpha$  will not happen within next *h* minutes. Also, based on the three categories, denoted as  $c_{T_{obs}}(\tau/\alpha)$ , we additionally supervise the model with a classification task to regularize the model.

#### 3.2. Network Architecture

Fig. 2 shows the overall network architecture of our IIA-Net. It is a two-step model with a feature extractor and a temporal model. The feature extractor takes five inputs  $x_i$ ,  $s_i$ ,  $b_i$ ,  $t_i$ ,  $p_i$  mentioned in Section. 3.1.  $s_i$  and  $b_i$  are used for IIM to model instrument-instrument interactions and the instrument-surrounding interactions. The frame  $x_i$  is encoded by ResNet50(He et al., 2016) into visual features, and the tool signal  $t_i$ , phase signal  $p_i$  are provided by the manual annotations from Cholec80 dataset. They are embedded into the feature space and concatenated with interaction feature and visual feature jointly for the input of the next temporal model.

For the temporal pattern modeling, we apply a multi-stage temporal convolutional network firstly for phase anticipation. Then we concatenate the above five feature vectors and the prediction of phase anticipation together as the input for the instrument anticipation. In the rest of this section, we will introduce the above modules in details.

# 3.3. Spatial Feature Extractor

#### 3.3.1. Instrument Interaction Module

In this module, we model surgeons' intention by analyzing the instrument-instrument interaction and instrumentsurrounding interaction, shown in Fig. 3. We assume each



Fig. 3. Instrument interaction module. Upper: instrument-surrounding modeling uses pooled scene semantic features to encode features; Bottom: instrument-instrument modeling extracts the spatial relations between the grasper and the other instruments.

frame is processed to obtain the spatial coordinates and bounding boxes of all instruments. Also, we extract the categorical prior for each frame, which characterizes the semantic class of a region in an image. (e.g., liver, gallbladder).

# Instrument-Instrument Encoder

This encoder explicitly models the geometric relation among instruments. Here we only consider the interaction between grasper and other instruments because the grasper is the most frequently used instrument. It provides the primary support for the other instruments during the surgery.

We encode the geometric relation  $G \in \mathbb{R}^{M \times 4}$  using Eq. 1 that is proven effective in object detection (Liang et al., 2019). Specifically, at any time instant, given the bounding box of grasper  $(x_g, y_g, w_g, h_g)$  and M other instruments in the scene  $\{(x_m, y_m, w_m, h_m) | m \in [1, M]\}$ , we encode the geometric relation into  $G \in \mathbb{R}^{M \times 4}$ , the *m*-th row of which equals to:

$$G_m = [log(\frac{|x_g - x_m|}{w_g}), log(\frac{|y_g - y_m|}{h_g}), log(\frac{w_m}{w_g}), log(\frac{h_m}{h_g})](1)$$

This encoding computes the geometric relation in terms of the geometric distance and the fraction box size. We then embed this geometric feature at each time instant into  $\mathbb{R}^{T_{obs} \times C_1}$  where  $C_1$  is the embedding size.

**Instrument-Surrounding Encoder** To encode an instrument's nearby anatomical surroundings, we first extract pixellevel scene semantic classes for each frame. Here, we use totally  $N_s = 7$  scene classes (i.e., background, liver, fat, abdominal wall, tool Shaft, tool tip, gallbladder). Then we transform the integer semantic map into  $N_s$  binary masks of the size  $T_{obs} \times h \times w$ , where h, w are spatial resolution. We apply two convolutional layers on the binary masks with a stride of 2 to get the scene CNN features. We then average the scene feature along the spatial dimensions and generate a feature vector as the encoder's output.

The generated feature vector is in  $\mathbb{R}^{T_{obs} \times C_2}$ , where  $C_2$  is the number of channels in the convolution layers. After combin-

ing the feature vectors from instrument-instrument encoder and instrument-surrounding encoder, the final feature vector outputed from IIM is in  $\mathbb{R}^{T_{obs} \times (C_1 + C_2)}$ 

# 3.3.2. Visual Features

Visual features extracted from the RGB frame  $x_i$  are the most important information to anticipate the future surgery phase and instrument usage. For instance, the usage of irrigator is often triggered by bleedings and can not be predicted solely from instrument signals. For the phase anticipation, there is mostly a significant ending activity for each phase that can be heavily recognized from visual frames. For instance, after clipping and cutting the cystic artery and duct, the hook shall be introduced again to dissect gallbladder from the liver, indicating the start of the Gallbladder-Dissection phase. Thus, IIA-Net introduces the ResNet50 to extract the visual representation. Compared to the AlexNet as the backbone, ResNet50 achieves the better result thanks to its representational ability. The design of skipconnection eases the training of ResNet50 and the pretrained version on ImageNet promotes its transfer learning ability to the other small dataset, e.g., Cholec80 in this work. Also, owing to the two-stage training, the ResNet50 does not suffer from the data leaking problem from batch-normalization layers (Ioffe and Szegedy, 2015).

# 3.3.3. Phase and Instrument Presence Signal

Surgical phase and instrument presence signals  $p_i/t_i$  have proven effective in a similar setting (Abu Farha et al., 2018) and empirically performed better than the model with only visual features. These signals could be obtained either from human annotation or from an extra recognition model, such as (Twinanda et al., 2016; Czempiel et al., 2020). In this work, we thoroughly studied these two cases and investigate the effect in the following sections. For the presence signal provided from human experts, we opt for the Cholec80 dataset's annotations. We also train the TeCNO (Czempiel et al., 2020)



Fig. 4. MS-TCN for online inference. As more frames are observed, the input and output are enlarging with the last index of output as the desired prediction.

to provide phase presence signal and YOLOv5 (Jocher, 2020) to provide instrument presence signal, that could be obtained from the real-world. Since the YOLOv5 is trained to detect the tooltip bounding boxes, the instrument presence signal can be easily transferred from its output. Noted that, YOLOv5 is not exclusively used to provide instrument presence, its output also supports the IIM.

## 3.4. Multi-Stage Temporal Convolutional Network

Surgical video is represented as a sequence of feature vector with the shape of (BS, T, D), where BS is the batch size, T is the temporal length of the video, and D is the hidden dimension. In this work, instead of processing the temporal sequence recursively like RNNs, we apply the lightweight model MS-TCN (Farha and Gall, 2019) to model the temporal pattern. The MS-TCN takes the feature vector sequence as input and apply dilated temporal convolutional layers to predicts a sequence of values with the same length as the input. As shown in the Fig. 4 in the online scenario, as the video stream is progressing, the length of the input and output will increase monotonically. However, only the last index of value from the output sequence will be considered as the true prediction for the current timestamp, framed by red in Fig. 4, while the other past predictions can be ignored for online inference.

This inference strategy is different from the RNNs and requires dynamic computational resources when the video is progressing and the input sequence is enlarging. Specifically, the cost needed for the timestamp close to the end is larger than the timestamp close to the start of the video. Therefore, we opt for the MS-TCN for its efficacy. The MS-TCN is constructed fully with dilated temporal convolutions without neither pooling layers nor fully connected layers. This design keeps the model processing the full resolution temporal sequence and reducing the number of parameters.

Also, the MS-TCN applies the multi-stage approach by adding multiple stages to the network to refine the output of the first stage multiple times. After each stage, we use a combination of smooth L1 loss and cross-entropy loss to train our model given the defined horizon h, as described in Eq. 2 and Eq. 3.

$$L^{\tau} = \sum_{\tau} \sum_{t}^{T} S \operatorname{moothL1}(f_t, r_t(\tau)) + \lambda_1 \cdot CE(p_t, c_t(\tau))) \quad (2)$$

$$L^{\alpha} = \sum_{\alpha} \sum_{t}^{I} SmoothL1(f_t, r_t(\alpha)) + \lambda_2 CE(p_t, c_t(\alpha)))$$
(3)

$$L(\tau/\alpha) = \sum_{m}^{M} (\lambda_3 \cdot L_m^{\tau} + \lambda_4 \cdot L_m^{\alpha})$$
<sup>(4)</sup>

Here,  $f_t$  is the predicted remaining time value and  $p_t$  is the classification output of each stage. The model is optimized based on the sum of these sub-losses, denoted in Eq. 4.

#### 3.4.1. Causal Modification

The original MS-TCN is proposed for the offline action segmentation, where the model is applied after the whole video is presented. Therefore, the acausal design of the temporal convolution layer makes the inference of each frame conditioned on the previous frames' and subsequent frames' information, which is not feasible for the online surgery scenario. In this work, the model shall make prediction for each frame by only observing the previous ones. Therefore, a simple and efficient method could be applied to make the network causal, through padding and shifting the output of each layer as follows:

Algorithm 1: Temporal Convolution Layer with Causal						
Modification						
<b>Data:</b> Input feature vector sequence $X \in \mathbb{R}^{(BS,T,D)}$						
<b>Result:</b> Remaining time sequence $Y \in \mathbb{R}^{(BS,T,D')}$						
kernelsize $\leftarrow 3$ ;						
layers $\leftarrow 8$ ;						
for ( $i \leftarrow 1$ to layers ) {						
dilation $\leftarrow 2^i$ ;						
if Causal is True then						
$Y \leftarrow Padding(dilation * (kernelsize - 1))(X)$						
$Y \leftarrow Conv1d(kernelsize, dilation)(Y)$						
$Y \leftarrow Y[:,:,:-(dilation * 2)]$						
end						
if Causal is False then						
$Y \leftarrow Padding(dilation)(X)$						
$Y \leftarrow Conv1d(kernelsize, dilation)(Y)$						
end						
$Y \leftarrow Y + X$						
}						

# 4. Experiment Setup

## 4.1. Datasets and Preprocessing

# 4.1.1. Anticipation Dataset

We evaluate our method on publicly available surgical workflow intraoperative video dataset, Cholec80 (Twinanda et al., 2016), which contains laparoscopic cholecystectomy procedures for the resection of the gallbladder. Cholec80 dataset consists of 80 videos ranging from 15 minutes to 90 minutes. We follow the same split as (Rivoir et al., 2020), separating the dataset to 60 videos for training and 20 for testing. We resize the videos spatial resolution to  $224 \times 224$  to dramatically reduce the computational cost. Also, we resample the video from 25 fps to 1 fps.

Table 1. Effect of IIM and MS-TCN on different feature extraction models for instrument anticipation. We report the  $MAE_{in}/MAE_e$  averaging over instrument types in minutes per metric when h = 5 min. T: tool signal feature; P: phase signal feature; IIM: interaction feature from instrument interaction module.

	ResNet50	ResNet50+T+P	ResNet50+T+P+IIM	Baseline
No MS-TCN	1.99/4.06	1.79/3.58	1.65/2.51	
1 Stage	1.62/3.74	1.57/3.29	1.42/2.22	1 75/2 69
2 Stages	1.59/3.67	1.45/3.23	1.40/2.14	1.75/2.08
3 Stages	1.53/3.64	1.60/3.31	1.48/2.15	

Table 2.  $MAE_p/MAE_p/MAE_e$  comparison. We report the mean over instrument types in minutes per metric. Ours 2D: our feature extractor without temporal training. Baseline: model from (Rivoir et al., 2020).

		Instrument		Phase			
	$h = 2\min$	$h = 3 \min$ $h = 5 \min$ $h = 2 \min$		h = 3min	$h = 5 \min$		
MeanHist	1.09/0.93/1.85	1.62/1.34/2.72	2.64/2.14/4.35	-	-	-	
OracleHist (offline)	(0.92/0.83/1.36)	(1.31/1.18/1.93)	(2.01/1.73/2.96)	-	-	-	
Baseline	0.77/0.64/1.12	1.13/0.92/1.65	1.75/1.49/2.68	0.63/0.62/1.02	0.86/0.85/1.47	1.17/1.37/1.54	
IIA-Net (2D)	0.70/0.52/1.07	1.07/0.77/1.65	1.65/ <b>1.16</b> /2.51	0.70/0.53/1.38	1.04/0.76/1.85	1.40/ <b>1.12</b> /2.42	
IIA-Net	0.66/0.42/1.01	0.97/0.69/1.46	<b>1.40</b> /1.28/ <b>2.14</b>	<b>0.62/0.49</b> /1.18	0.81/0.73/1.42	1.08/1.22/1.09	

#### 4.1.2. Detection and Segmentation Dataset

As mentioned above, we need to extract instrument bounding boxes and semantic maps for the Cholec80 dataset. However, annotating such dataset is manually unfeasible. Therefore, we opt for training the segmentation model (Ronneberger et al., 2015) on a synthesized dataset(Pfeiffer et al., 2019), which utilizes conditional GAN (Isola et al., 2017) to generate Cholec80 style laparoscopic images from simulation images. Then, we apply the trained model to infer on the Cholec80 dataset. The segmentation result can be found in the supplementary materials.

To detect the surgical instrument bounding boxes on Cholec80, we leverage the dataset from (Jin et al., 2018) to train a YOLOv5 (Jocher, 2020) detector. The trained model is proven to detect surgical instruments on Cholec80 dataset effectively (Jin et al., 2018).

#### 4.2. Implementation Detail

In this section, we provide the hyper-parameter settings and the settings to construct the IIA-Net. We train for 49 epochs for the first feature extraction process and 30 epochs for the second temporal modeling process. The batch sizes are 60 and 1 for the first and second training step, respectively. The embedding space for the  $C_1$  and  $C_2$  are both 64. For the peripheral recognition networks, we opt for the UNet for the segmentation map extraction and the YOLOv5 for the instrument bounding boxes detection. Also, the  $\lambda 1$  and  $\lambda 2$  are both equal to 0.01, while the  $\lambda 3$  and  $\lambda 4$  are 0.6 and 0.4, respectively. For the fair evaluation, we experimented three times and average them before reporting the result in the table.

## 4.3. Evaluation Metrics

Automatic instrument preparation is one of the primary tasks that benefits from surgical workflow anticipation. It does not require tools or phases to be anticipated too far in advance. Also, the preparation system should only react to the signals that indicate tool/phase is anticipating. In this work, the system outputs the specific value to indicate the remaining time of the next tool's/phase's occurrence. Therefore, we follow Rivoir et al. (2020) to opt for the frame-based evaluation metrics, mean absolute error (MAE) and its variants, i.e.,  $MAE_{in}$ ,  $MAE_p$  and  $MAE_e$ . The  $MAE_{in}$ ,  $MAE_e$ ,  $MAE_p$  can be represented in the following formulas:

$$MAE_{in} = \frac{1}{T} \sum_{i}^{T} MAE(f_i, r(\tau/\alpha)), 0 < r(\tau/\alpha) < h$$
(5)

$$MAE_e = \frac{1}{T} \sum_{i}^{T} MAE(f_i, r(\tau/\alpha)), 0 < r(\tau/\alpha) < 0.1h \quad (6)$$

$$MAE_{p} = \frac{1}{T} \sum_{i}^{T} MAE(f_{i}, r(\tau/\alpha)), 0.1h < f_{i} < 0.9h$$
(7)

. In the above,  $f_i$  is the prediction of the model and  $r(\tau/\alpha)$  represents the ground truth value for the current timestamp. In specific, we average the MAE of 'anticipating' frames using  $MAE_{in}$ , because the preparation system should only react to the signals that indicate phase/instrument is anticipating. Also, it does not require tools or phases to be anticipated too far in advance. Therefore, we propose to use  $MAE_e$  to evaluate intervals  $(0 < r_{T_{obs}}(\tau/\alpha) < 0.1h)$  that provides the most effective support to the computer-assistance system.

Also, we follow the Rivoir et al. (2020) to utilize the  $MAE_w$  and  $MAE_{out}$  to evaluate overall performance the errors outside the horizon. The  $MAE_w$  and  $MAE_{out}$  can be respectively represented as follows:

$$MAE_{out} = \frac{1}{T} \sum_{i}^{T} MAE(f_i, r(\tau/\alpha)), r(\tau/\alpha) = h$$
(8)

$$MAE_w = (MAE_{in} + MAE_{out})/2$$
<sup>(9)</sup>

Training		Noisy		Re	Real		Upper		Baseline
Testing		Noisy	Real	Noisy	Real	Noisy	Real	Upper	
Instrument Anticipation	$MAE_{in}$	1.64	1.55	1.45	1.42	1.67	1.61	1.48	1.75
	$MAE_e$	2.60	2.39	2.25	2.17	2.61	2.53	2.14	2.68
	$MAE_p$	1.32	1.34	1.32	1.32	1.39	1.39	1.28	1.49
Phase Anticipation	MAE <sub>in</sub>	1.23	1.25	1.12	1.13	1.38	1.32	1.08	1.17
	$MAE_e$	1.81	1.80	1.36	1.23	1.81	1.70	1.13	1.54
	$MAE_p$	1.29	1.27	1.25	1.26	1.42	1.38	1.22	1.37

Table 3. Model's robustness to different kinds of signals. Noisy: presence signals with artificial noises; Real: recognized signals from the real-world recognition models; Upper: presence signals from human annotations.

# 5. Results and Discussions

## 5.1. Effect of IIM and Stages in MS-TCN

We conduct ablative testing to compare different feature extraction models, ResNet50 (He et al., 2016), ResNet50 with instrument and phase features, ResNet50 with all added features, to identify a suitable feature extractor for our model. Additionally, we conduct experiments with different numbers of MS-TCN stages to determine which architecture can maximally capture temporal patterns.

As shown in Tab. 1, the ResNet50 with all features outperforms ResNet50 across the board with improvements ranging from 1.99 to 1.57 in  $MAE_{in}$  and 4.06 to 2.51 in  $MAE_p$ . The loss reduction should be attributed to the improved representation by our designed features. Among the features that we added, the IIM makes more contribution than the instrument and phase signals. This suggests that modeling the interactions signifies the surgeon's intention and the occurrence of the next situation. Interestingly, ResNet50 with all added feature achieves a comparable result with the baseline model(Rivoir et al., 2020) even without any temporal modeling.

Tab. 1 also highlights the substantial performance improvement achieved by the MS-TCN refinement stages. Those results demonstrate the ability of MS-TCN to improve the performance of any feature extractor. All feature extractors achieve higher performance when only 1 stage is used. However, the 2-stage model outperforms the 3-stage one, implying that multiple rounds of refinement may render overfitting over a limited amount of data.

# 5.2. Anticipation Results

We evaluate the model for instrument and phase anticipation on horizons of 2, 3, and 5 minutes. We remove the horizon setting of 7 minutes since anticipating the surgical workflow too early is unnecessary for instrument preparation and robot assistance. We re-implement methods from (Rivoir et al., 2020) and retrain them as the baseline for phase anticipation.

Tab. 2 shows that IIA-Net achieves lower  $MAE_{in}$  and  $MAE_p$  error compared to the previous methods. Regarding the  $MAE_p$  error, the margin increases even further. Even though (Rivoir et al., 2020) is trained in an end-to-end fashion, it is also outperformed by our IIA-Net, which is trained in a two-step process. Interestingly, our model trained without temporal context (Ours 2D) achieved a lower  $MAE_p$  when h = 5. This is because the 2D model has difficulty foreseeing the long-horizon

occurrence and is prone to outputting a near-0/h value, leading to unsmooth prediction. Also, Tab. 2 shows our model achieves the lowest  $MAE_e$  error, suggesting that our model can effectively identify instrument or phase occurrence a few seconds ahead. In real-world scenarios, this is typically the most critical time for accurate anticipation.

# 5.3. Robustness to Presence Signals

As indicated in the methodology, the model takes the peripheral presence signals and embeds them into the joint feature space for the anticipation. (Yuan et al., 2021) utilizes the Cholec80's annotation to provide presence signals for the training and testing, but how the model performs with noisy signals remains unexplored. Here, we apply the TeCNO(Czempiel et al., 2020) as the phase recognition model to generate the recognized phase presence signal and apply the YOLOv5 to generate the tool presence signal for model training. In addition, we apply random artificial noise to show the model's training and testing ability with different levels of noise. The entire experiment is conducted under the horizon as 5.

As shown in the Tab. 3, the model's performance drops when being tested with noisy signals compared to the recognized signals. Recall that the surgical phases are high-level activities and can not be easily anticipated from low-level RGB frames. Therefore, it is mostly conditioned on the phase presence signals, causing a large margin of degradation when a noisy phase signal is fed.

Also, the model trained with annotated signals is sensitive to the noise, as the  $MAE_e$  surges from 2.14 to 2.53. Tab. 3 illustrates that the model favors a cleaner signal for the testing, regardless of the signal types used for training. This indicates that the training of IIA-Net can dig the occurrence patterns even from the noisy annotations, easing the annotation burden. Also, as shown in the first two columns, the model achieves a better baseline for instrument anticipation, demonstrating IIA-Net's strong representational ability.

The model trained with the recognized signals achieves the strongest robustness against the noisy recognized signals for the instrument and phase anticipation, as shown in the second column. It has the smallest degradation when the noisy signal is fed for inference and still achieve a better result compared to the baseline. This is because the input of training has been augmented by using recognition model to generate signals. This provides the clue to perform data augmentation during the training with slightly noisy instrument and phase presence signals.

Table 4. Inference Resources. The first two colums are peripheral network to generate presence signals and tool tip detections. The spatial feature extractor contains the ResNet50 and IIM, deployed sequentially with the MS-TCN.

	YOLOv5	UNet	Feature Extractor	TeCNO	MS-TCN	Baseline(Rivoir et al., 2020)
Inference Time (s)	0.0090	0.0087	0.0142	0.193	0.0151	0.0328
Number of Parameters (M)	7.9	30	23	23.3	0.3	108.3



Fig. 5. Robustness of the model to the recognized instrument and phase signals. The model trained with human annotated presence signals is framed by blue and the model train trained with recognized signals is framed by black.

As shown in the Fig. 5, the noise from presence signals significantly influences the phase anticipation and causes a lot false positive predictions. For the instrument anticipation, the model trained with recognized presence signals can better the starting point of the specimenbag compared to that trained with annotated signals. For the other instruments, the models perform similarly, meaning the noise in instrument presence signal does not affect the anticipation too much.

# 5.4. Running Time Performance

To deploy the model into the end device, the most concerning problem is the space-time tradeoff, i.e., the balance between the GPU memory and the inference time. In this work, there are totally five networks and four of them can be deployed parallelly, namely, ResNet50(He et al., 2016), UNet(Ronneberger et al., 2015), TeCNO(Czempiel et al., 2020) and YOLOv5(Jocher, 2020). The MS-TCN will be applied for the sequence modeling after the above four networks finishing the inference. In this work, we implement the model in PyTorch and train it on a NVIDIA RTX 2080 Ti with 12GB GPU.

As shown in Tab. 4, the longest running time per frame is Max(YOLOv5, UNet, ResNet50) + MS - TCN = Max(0.0090s, 0.0087s, 0.0142s+0.193s)+0.0151s = 0.0344s, and the Max operation is introduced because multiple models can run in parallel. The speed can completely fulfill the real-time requirement in clinical tasks and the efficiency is comparable to the (Rivoir et al., 2020) taking 0.0328s via CNN+LSTM architecture.

# 5.5. Effect of Sequence Modeling

As shown in the Fig. 6, the sequence modeling of MS-TCN smooths and denoises the remaining time prediction from the



Time [sec.

Time [sec.]

Table 5. wMAE/MAE<sub>out</sub> comparison. We report the mean over instrument types in minutes per metric. Ours 2D: our feature extractor without temporal training.

Fig. 6. Effect of sequence modeling in phase anticipation. The upper part is from IIA-Net (2D), which does not contain sequence modeling method. The bottom part is the full model with MS-TCN as the sequence modeling method.

IIA-Net (2D). It significantly reduces the noise at the beginning of the surgery when the 2D model tends to output a sharp and irrational prediction. At the same time, it maintains the anticipation trending when the phase is really happening in next few minutes, as shown in the grey area. This indicates that the sequence modeling is necessary because the over-sharp outputs is not straightforward signal for the decision making and introduces ambiguity.

# 5.6. Trade-off between Inside and Outside Horizon

rrigator [min.] 2'2

SpecBag [min.] 2'2 m

IIA-Net

As the  $MAE_p$  has undesirable properties to capture the minor mistakes outside the horizon, it favors models with more erratic predictions. This makes the  $MAE_p$  metric unsuitable to measure the the overall anticipation performance. Also, the solely usage of  $MAE_{in}$  and  $MAE_e$  does not capture the 'false positive' predictions from 'non-presence' frames, leading to incomplete evaluation of the model. Therefore, we follow the Rivoir et al. (2020) and report the results of  $MAE_{out}$  and  $MAE_w$ , which are the mean square error of frames outside the horizon and the weighted average of  $MAE_{in}$  and  $MAE_{out}$ , respectively. As shown in the Tab. 5, our proposed IIA-Net achieves the superior  $MAE_w$  result compared to the previous methods for both instrument anticipation and phase anticipation. However, the MAE<sub>out</sub> of IIA-Net for instrument anticipation is comparable and sometimes worse than the baseline method Rivoir et al. (2020), indicating that our method exhibits a trade-off between the frames inside and outside the horizon. This trade-off makes IIA-Net more sensitive and generates more 'fals positive' predictions. However, given the better  $MAE_w$  and  $MAE_{in}$ , we would like to say this trade-off is acceptable for some specific surgical scenarios. Also, for the phase anticipation, the IIA-Net achieves the better result both for the  $MAE_w$  and  $MAE_out$ , confirming the better representation learning ability of this work. In the future work, the sensitivity of the model should be investigated and proposed.

# 6. Discussion

This work aims to raise the interest of surgical workflow anticipation tasks, benefiting the current computer-assisted surgery system. The anticipation model can generate a useful peripheral signal, indicating the remaining time until the next occurrence of certain instrument/phase. This signal offers fourfold prominent benefits for clinicians. Firstly, the reliable anticipation prediction provides a useful reference to robotic systems in decision making. It can support the surgeon to decide when to go into the next phase and when to introduce next instrument for the current situation. Secondly, accurately anticipating tools such as the irrigator can help the early detection and even the prevention of potential complications, for example, massive haemorrhage. Thirdly, it allows real-time instruction for automated surgical coaching therefore increasing patient safety and reducing surgical errors. Also, the anticipation of surgical phases can provide vital input to streamline communication in the operating room (OR).

While a lot of deep learning models have been proposed so far, it is still a challenging task to find an effective and robust solution to the anticipation tasks for both surgical phase and instrument. Compared with natural video action anticipation, one of the main challenges for surgical video is the task formulation design for the long-range workflow. Also, it is important to consider the different occurrence patterns between the surgical instruments and the phases to build the unified network structure for the anticipation. Specifically, the surgical phase pattern assumes that the new phase is occurring within typically one second and only its correct category is unknown. For the instrument pattern, the occurrence is sparse that certain instruments are rarely used throughout surgery. Therefore, the dense segmentation and phase trajectory prediction (Ban et al., 2021) do not fit the surgical workflow anticipation, because they can only handle the phase anticipation. As another view to handle the anticipation, remaining time prediction can process both the phase occurrence and instrument occurrence with seamless effort to modify the training objectives and schemes. Also, it could be easily deployed in a real-time manner. Therefore, we opt for this formulation and design the network around this idea.

For the remaining time prediction framework, the horizon h is used to truncate the output so that the model will neglect the frames that occur far early from the anticipating ones. However, the choice of the horizon is empirical and may significantly impact the user experience. For example, if the model is trained by setting the horizon as 5 and tested with 2, the proposed metrics will end up with different values. Currently, the horizon is set the same for both training and testing to conduct the quantitative evaluation, which can not reflect the real-life user experience.

Another improvement lies in improving the spatial-temporal representational ability of the spatial feature extraction process. Instrument interaction proves to be an effective feature to model the intention. However, the surgeon's action is modeled implicitly in this work and requires the detection model ahead. Recently, (Innocent Nwoye et al., 2021) proposes the CholecT50 dataset with human annotated action triplets for each frame, enhancing the model's representational ability explicitly. Following works can base this dataset and design the feature extraction blocks, combining more surgical expert's priors. For the sequence modeling, the emerging transformer architectures empower the long-range video processing. However, the computational cost of transformer is  $O(n^2)$  and introduces a heavier GPU burden, thus urgently calling for an more efficient transformer.

How to evaluate the model in the real-world scenario is a common issue for the clinical works. For this surgical workflow anticipation work, it is more complex because the system tells the surgeons the future in the real-time manner. Therefore, it is more challenging to evaluate this work on the real surgery. Here we provide the potential direction and hope this can stimulate more research attention to the real-world clinical evaluation.Real-time feed back is crucial. The user should not only react to the current predictions, they should also provide the post-feedback within a time limit. Then, an offline feedback based on the whole sequential prediction should be made. However, current work, including IIA-Net only consider the offline feedback, which is not feasible for the real-world evaluation.

## 7. Conclusion

In this paper, we have proposed the IIA-Net that incorporates various existing surgical workflow analysis methods (including tool detection, phase recognition and laparoscopic image segmentation) and outperforms previous works. The achieved results show that the interaction relationship during spatial feature extraction is effective to resolve surgical workflow anticipation. In the absence temporal training, the developed model can act as a decent baseline for the following 2D works. Furthermore, temporal modelling using a MS-TCN with causal and dilated convolution handles full temporal resolution of time series, fitting extreme long laparoscopic workflow well. Its large receptive field captures distant and recent observations. Our multitask learning scheme offers a promising opportunity for joint instrument performing and phase anticipation.

# References

- Abu Farha, Y., Richard, A., Gall, J., 2018. When will you do what?-anticipating temporal occurrences of activities, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5343–5352.
- Ban, Y., Rosman, G., Ward, T., Hashimoto, D., Kondo, T., Iwaki, H., Meireles, O., Rus, D., 2021. Surgical prediction gan for events anticipation. arXiv preprint arXiv:2105.04642.
- Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 343–352.
- Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N., 2021. Opera: Attention-regularized transformers for surgical phase recognition, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 604–614.
- Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L., 2016. Recurrent marked temporal point processes: Embedding event history to vector, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1555–1564.
- Farha, Y.A., Gall, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3575–3584.
- Forestier, G., Petitjean, F., Riffaud, L., Jannin, P., 2017. Automatic matching of surgeries to predict surgeons' next actions. Artificial intelligence in medicine 81, 3–11.
- Franke, S., Neumuth, T., 2015. Adaptive surgical process models for prediction of surgical work steps from surgical low-level activities, in: 6th Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI) at the 18th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), Munich, Germany.
- Funke, I., Mees, S.T., Weitz, J., Speidel, S., 2019. Video-based surgical skill assessment using 3d convolutional neural networks. International journal of computer assisted radiology and surgery 14, 1217–1225.
- Gao, J., Yang, Z., Nevatia, R., 2017. Red: Reinforced encoder-decoder networks for action anticipation. arXiv preprint arXiv:1707.04818.
- Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A., 2021. Trans-svnet: accurate phase recognition from surgical videos via hybrid embedding aggregation transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 593–603.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

- Innocent Nwoye, C., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2021. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. arXiv e-prints, arXiv=2109.
- Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448–456.
- Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
- Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L., 2018. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 691– 699.
- Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A., 2017. Svrcnet: workflow recognition from surgical videos using recurrent convolutional network. IEEE transactions on medical imaging 37, 1114–1126.
- Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.W., Heng, P.A., 2020. Multitask recurrent convolutional network with correlation loss for surgical video analysis. Medical image analysis 59, 101572.
- Jocher, G., 2020. ultralytics/yolov5: v3.1 Bug Fixes and Performance Improvements. https://github.com/ultralytics/yolov5. URL: https://doi.org/10.5281/zenodo.4154370, doi:10.5281/zenodo. 4154370.
- Ke, Q., Fritz, M., Schiele, B., 2019. Time-conditioned action anticipation in one shot, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9925–9934.
- Klank, U., Padoy, N., Feussner, H., Navab, N., 2008. Automatic feature generation in endoscopic images. International Journal of Computer Assisted Radiology and Surgery 3, 331–339.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105.
- Lan, T., Chen, T.C., Savarese, S., 2014. A hierarchical representation for future action prediction, in: European Conference on Computer Vision, Springer. pp. 689–704.
- Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165.
- Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., Fei-Fei, L., 2019. Peeking into the future: Predicting future person activities and locations in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5725–5734.
- Lohit, S., Wang, Q., Turaga, P., 2019. Temporal transformer networks: Joint learning of invariant and discriminative time warping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12426–12435.
- Ma, S., Sigal, L., Sclaroff, S., 2016. Learning activity progression in lstms for activity detection and early detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1942–1950.
- Mahmud, T., Hasan, M., Roy-Chowdhury, A.K., 2017. Joint prediction of activity labels and starting times in untrimmed videos, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5773–5782.
- Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nature Biomedical Engineering 1, 691–696.
- Padoy, N., 2019. Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies 28, 82–90.
- Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N., 2012. Statistical modeling and recognition of surgical workflow. Medical image analysis 16, 632–641.
- Pfeiffer, M., Funke, I., Robu, M.R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M.J., Gurusamy, K., Davidson, B.R., et al., 2019. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 119–127.
- Rivoir, D., Bodenstedt, S., Funke, I., von Bechtolsheim, F., Distler, M., Weitz, J., Speidel, S., 2020. Rethinking anticipation tasks: Uncertainty-aware an-

ticipation of sparse surgical instrument usage for context-aware assistance, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 752–762.

- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234– 241.
- Sadegh Aliakbarian, M., Sadat Saleh, F., Salzmann, M., Fernando, B., Petersson, L., Andersson, L., 2017. Encouraging lstms to anticipate actions very early, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 280–289.
- Sener, F., Singhania, D., Yao, A., 2020. Temporal aggregate representations for long-range video understanding, in: European Conference on Computer Vision, Springer. pp. 154–171.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Soomro, K., Zamir, A.R., Shah, M., 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
- Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36, 86–97.
- Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N., 2018. Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE transactions on medical imaging 38, 1069–1078.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.
- Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H., 2017. Learning to generate long-term future via hierarchical prediction, in: international conference on machine learning, PMLR. pp. 3560–3569.
- Vondrick, C., Pirsiavash, H., Torralba, A., 2016. Anticipating visual representations from unlabeled video, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 98–106.
- Yuan, K., Holden, M., Gao, S., Lee, W.S., 2021. Surgical workflow anticipation using instrument interaction, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 615–625.