Surgical Workflow Anticipation using Instrument Interaction

Kun Yuan[1], Matthew Holden[2], Shijian Gao[3], and Won-Sook Lee[1]

[1] Faculty of Engineering, University of Ottawa
{kyuan033, wslee}@uottawa.ca
[2] School of Computer Science, Carleton University
matthew.holden@carleton.ca
[3] Department of Electrical and Computer Engineering, University of Minnesota
gao00379@umn.edu

Abstract. Surgical workflow anticipation, including surgical instrument and phase anticipation, is essential for an intra-operative decision-support system. It deciphers the surgeon's behaviors and the patient's status to forecast surgical instrument and phase occurrence before they appear, providing support for instrument preparation and computer-assisted intervention (CAI) systems. We investigate an unexplored surgical workflow anticipation problem by proposing an Instrument Interaction Aware Anticipation Network (IIA-Net). Spatially, it utilizes rich visual features about the context information around the instrument, i.e., instrument interaction with their surroundings. Temporally, it allows for a large receptive field to capture the long-term dependency in the long and untrimmed surgical videos through a causal dilated multi-stage temporal convolutional network. Our model enforces an online inference with reliable predictions even with severe noise and artifacts in the recorded videos. 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 inMAE and 2.14 v.s. 2.68 for eMAE). The code is published on https://github.com/Flaick/Surgical-Workflow-Anticipation

Keywords: Surgical Workflow Analysis · Anticipation · Temporal Convolutional Networks · Endoscopic Videos · Instrument Detection

1 Introduction

Context-aware assistance is integral for CAI systems, of which the most crucial 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 [7, 23]. The benefit of anticipation is three-fold [20]. Firstly, instrument anticipation offers a useful reference for decision making in a robotic assistance system. It helps to identify instrument usage triggering so that a robotic system can decide when to intervene. Also, for context-aware assistance, anticipating instruments such as

irrigator can help early detection and prevention of potential complications, e.g., massive haemorrhage. Thirdly, it allows real-time instruction for automated surgical coaching therefore increasing patient safety and reducing surgical errors. The anticipation of surgical phases can also provide vital input for optimizing communication in the operating room (OR).

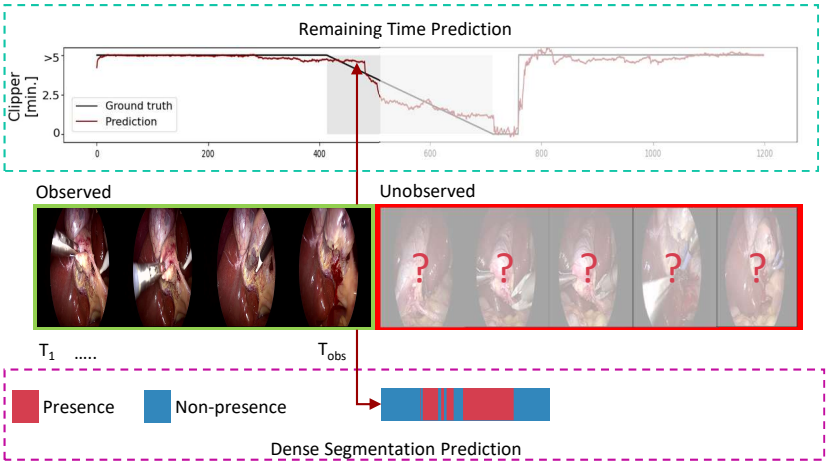

Fig. 1. Anticipation frameworks. Given an observed sequence and current time instant, T_{obs}, bottom part shows the conventional anticipation works that predicts dense segmentations. The upper part shows our strategy handling anticipation task as a real-time remaining time prediction task.

Recent models [23, 28] for surgical workflow anticipation possess spatial-temporal limitations. Spatially, they use AlexNet [17], VGG [25] and similar architectures to extract a feature vector, representing instrument/phase presence for each frame. However, they ignore the task-specific combinations present in surgical anticipation applications, i.e., instrument-instrument and instrument-surrounding interactions. This information precisely reflects the surgeon's intention and patient's anatomy status, helping models generalize to the low-quality input materials [16] and variability of patient's anatomy and surgeon style [8]. Novelly, our IIA-Net addresses instrument-instrument interaction in the form of a correlation matrix and designed geometric relations among instruments. Also, the instrument-surrounding interaction is included via the semantic segmentation map. This makes our extracted feature to be representative enough to identify the trigger event for the next instrument and phase occurrence.

Temporally, existing works have difficulty handling non-stationary 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, widely used RNNs [11] learn

a pattern from shorts segments of time series and apply it to other parts to get predictions, losing the distant observation information. Therefore, 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.

Initial works [1,4,7,9,15,19] handle anticipation as a dense segmentation prediction task, shown in the bottom of Fig. 1. They require a pre-loading process before performing anticipation, limiting their usage in online surgical applications. Specifically, [1] 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 (MSTCN) 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 Methodology

Our IIA-Net composes of two parts, a feature extractor with an Instrument Interaction Module (IIM) and a temporal model using MSTCN. Spatially, our 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 [13, 14, 21, 27], we introduce tool and phase signal to boost the feature extraction process. Temporally, we utilize causal dilated MSTCN [5] to capture long-term patterns with a large receptive field. Unlike the dense segmentation prediction, shown in Fig. 1, our IIA-Net follows [23] to handle anticipation as a real-time remaining time regression problem without any latency or pre-loading process.

2.1 Task Formulation

We process the anticipation task as a regression problem both for instrument and phase anticipation. Given a frame i from video x, we firstly extract semantic map s_i and instrument bounding boxes b_i. At the same time, we obtain the instrument presence signal t_i and phase signal p_i from the manual annotations. Given the observed sequence $\{(x_1, s_1, b_1, t_1, p_1), ...(x_{T_{obs}}, s_{T_{obs}}, b_{T_{obs}}, t_{T_{obs}}, p_{T_{obs}})\}$ from time 1 to T_{obs}, our model predicts the remaining time until the occurrence of τ/α for instrument/phase. The ground truth $r(x_{T_{obs}}, \tau/\alpha)$ for current time instant T_{obs} ranges $[0, h]$, where 0 denotes that the τ/α is currently happening and h denotes that τ/α will not happen within next h minutes.

2.2 Network Architecture

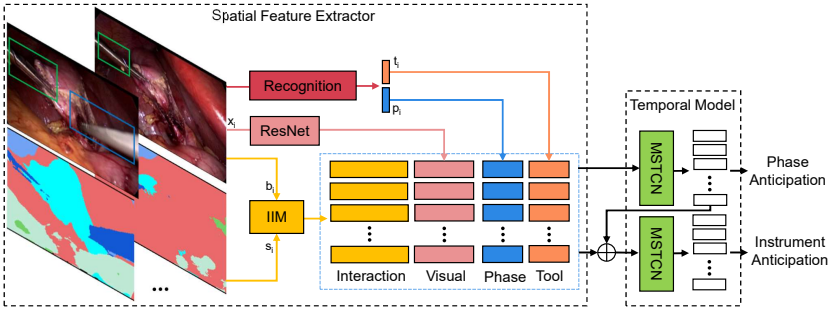

Fig. 2. Overview of the proposed model. For each frame observed, its estimated semantic map and tool detection are forwarded to instrument interaction module (IIM) to extract interaction feature. The manual annotations for phase and tool signal are fed into temporal model jointly with interaction and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

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. 2.1, of which the 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 [10] 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.

2.3 Instrument Interaction Module

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

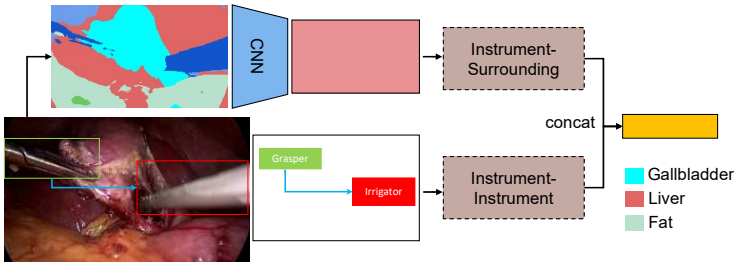

Fig. 3. Instrument interaction module. Upper: instrument-surrounding modeling uses pooled scene semantic features to encode features; Bottom: instrument-instrument modeling extracts the spatial relations between the grasper and the other instruments.

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 [18]. 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:

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

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

2.4 Multi-Stage Temporal Convolutional Network

To model temporal patterns in the anticipation task, we modify the MSTCN [5] to build a lightweight temporal network. The network 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 reduces the number of parameters.

To apply our model in an online mode, we use causal convolutions in the network. Instead of the acausal convolutions in [5] with predictions depend on both n past and n future frames, the causal convolutions ensure the prediction of current instant not relying on any n future frames but only depends on the current and previous frames.

3 Experiment Setup

3.1 Datasets and Preprocessing

Anticipation Dataset We evaluate our method on publicly available surgical workflow intraoperative video dataset, Cholec80 [26], 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 [23], separating the dataset to 60 videos for training and 20 for testing. We resize the videos spatial resolution to 224×224 to dramatically reduce the computational cost. Also, we resample the video from 25 fps to 1 fps.

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 [24] on a synthesized dataset [22], which utilizes conditional GAN [12] 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 [13] to train a YOLO [2, 6] detector. The trained model is proven to detect surgical instruments on Cholec80 dataset effectively [13].

3.2 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. Therefore, we measure the performance of 'anticipating' frames ($0 < r(x_{T_{obs}}, \tau/\alpha) < h$, using inMAE. Also, we propose to use eMAE to evaluate intervals ($0 < r(x_{T_{obs}}, \tau/\alpha) < 0.1h$) that provides the most effective support to the computer-assistance system. Also, we utilize the pMAE in [23] to measure the precision performance of our model.

Table 1. Effect of IIM and MSTCN on different feature extraction models for instrument anticipation. We report the inMAE/eMAE 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 MSTCN	1.99/4.06	1.79/3.58	1.57/2.51	1.75/2.68
1 Stage	1.62/3.74	1.57/3.29	1.42/2.22	
2 Stages	1.59/3.67	1.45/3.23	**1.40/2.14**	
3 Stages	1.53/3.64	1.60/3.31	1.48/2.15	

4 Results and Discussions

4.1 Effect of IIM and Stages in MSTCN

We conduct ablative testing to compare different feature extraction models, ResNet50 [10], 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 MSTCN stages to determine which architecture is best able to 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 inMAE and 4.06 to 2.51 in pMAE. This increase can 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 [23] even without any temporal modeling.

Tab. 1 also highlights the substantial performance improvement achieved by the MSTCN refinement stages. Those results demonstrate the ability of MSTCN to improve the performance of any feature extractor. All feature extractors achieve higher performance when only 1 stage is used. However, 2 stages model outperforms 3 stages model. This could indicate that 3 stages of refinement lead to overfitting on the training set for the limited amount of data.

4.2 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 [23] and retrain them as the baseline methods for phase anticipation.

Tab. 2 shows that IIA-Net achieves lower inMAE and pMAE error compared to the previous methods. Regarding the pMAE error, the margin increases even further. Even though [23] 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,

Table 2. inMAE/pMAE comparison. We report the mean over instrument types in minutes per metric. Ours 2D: our feature extractor without temporal training.

	Instrument			Phase		
	$h = 2$min	$h = 3$min	$h = 5$min	$h = 2$min	$h = 3$min	$h = 5$min
MeanHist	1.09/0.93	1.62/1.34	2.64/2.14	–	–	–
OracleHist (offline)	(0.92/0.83)	(1.31/1.18)	(2.01/1.73)	–	–	–
Baseline [23]	0.77/0.64	1.13/0.92	1.80/1.49	0.63/0.62	0.86/0.85	1.17/1.37
Ours 2D	0.70/0.52	1.07/0.77	1.65/**1.16**	0.70/0.53	1.04/0.76	1.40/**1.12**
Ours	**0.66**/**0.42**	**0.97**/**0.69**	**1.48**/1.28	**0.62**/**0.49**	**0.81**/**0.73**	**1.08**/1.22

Table 3. eMAE comparison. We report the mean over instrument types in minutes.

	Instrument			Phase		
	$h = 2$min	$h = 3$min	$h = 5$min	$h = 2$min	$h = 3$min	$h = 5$min
MeanHist	1.85	2.72	4.35	–	–	–
OracleHist (offline)	(1.36)	(1.93)	(2.96)	–	–	–
Baseline [23]	1.12	1.65	2.68	**1.02**	1.47	1.54
Ours 2D	1.07	1.65	2.51	1.38	1.85	2.42
Ours	**1.01**	**1.46**	**2.14**	1.18	**1.42**	**1.09**

our model trained without temporal context (Ours 2D) achieved a lower pMAE when $h = 5$. This is because the 2D model has difficulty foreseeing long-horizon occurrence and easily predicts the value that is close to $0/h$, making its prediction unsmooth. Also, our model achieves the lowest eMAE error, seen from Tab. 3. This suggests 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.

To further verify the feasibility of our model in the real-world surgical scenario, we test our model's running time performance given the online video stream. Based on the Pytorch 1.4 framework and single RTX 2080 ti, our model is able to process each frame within 0.0293s when deploying the spatial feature extractor and temporal model parallely. The running time is 10% faster than [23] taking 0.0328s, indicating our model is applicable in a real-world setup.

4.3 Limitations

The primary limitation of our current experimental setup is the incorporation of tool and phase signals. Specifically, we train the network using the signals from human annotation instead of recognition models. In real-world scenarios where this ground-truth is not available, the model's performance will likely be reduced. We conjecture, however, the degradation will be minimal using recent models which show superior performance for tool and phase recognition (95% and 88% of accuracy for tool and phase recognition). Also, [3] shows that the predicted phase signal is consistent and smooth not only within one phase, but also for the often ambiguous phase transitions. This means our IIA-Net will likely make a reliable prediction even with predicted tool and phase signals.

5 Conclusion

In this paper, we propose the IIA-Net, which incorporates existing surgical workflow analysis methods, i.e., tool detection, phase recognition, laparoscopic image segmentation, and outperforms previous works. It shows that the interaction relationship during spatial feature extraction is effective to resolve surgical workflow anticipation. Without temporal training, our model is a strong baseline for the following 2D works. Furthermore, temporal modelling using a MSTCN with causal and dilated convolution handles full temporal resolution of time series, fitting extreme long laparoscopic workflow well. Its large receptive field captures distant as well recent observations. Our multi-task learning schema provides a potential direction to jointly perform instrument and phase anticipation. Future work includes evaluation of our method on real-world phase and tool signals.

References

1. Abu Farha, Y., Richard, A., Gall, J.: 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 (2018)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 343–352. Springer (2020)
4. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: 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 (2016)
5. Farha, Y.A., Gall, J.: 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 (2019)
6. Farhadi, A., Redmon, J.: Yolov3: An incremental improvement. Computer Vision and Pattern Recognition, cite as (2018)
7. Forestier, G., Petitjean, F., Riffaud, L., Jannin, P.: Automatic matching of surgeries to predict surgeons' next actions. Artificial intelligence in medicine **81**, 3–11 (2017)
8. Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3d convolutional neural networks. International journal of computer assisted radiology and surgery **14**(7), 1217–1225 (2019)
9. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. arXiv preprint arXiv:1707.04818 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

13. Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L.: 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). pp. 691–699. IEEE (2018)
14. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. IEEE transactions on medical imaging **37**(5), 1114–1126 (2017)
15. Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9925–9934 (2019)
16. Klank, U., Padoy, N., Feussner, H., Navab, N.: Automatic feature generation in endoscopic images. International Journal of Computer Assisted Radiology and Surgery **3**(3), 331–339 (2008)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
18. Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., Fei-Fei, L.: 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 (2019)
19. Mahmud, T., Hasan, M., Roy-Chowdhury, A.K.: Joint prediction of activity labels and starting times in untrimmed videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5773–5782 (2017)
20. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)
21. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)
22. 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.: 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. pp. 119–127. Springer (2019)
23. Rivoir, D., Bodenstedt, S., Funke, I., von Bechtolsheim, F., Distler, M., Weitz, J., Speidel, S.: Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 752–762. Springer (2020)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Twinanda, A.P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Single- and multi-task architectures for surgical workflow challenge at m2cai 2016. arXiv preprint arXiv:1610.08844 (2016)
27. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)

28. Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE transactions on medical imaging **38**(4), 1069–1078 (2018)