Surgical Activity Triplet Recognition via Triplet Disentanglement

Yiliang Chen[1], Shengfeng He[2], Yueming Jin[3], and Jing Qin[1(✉)]

[1] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
yiliang.chen@connect.polyu.hk, harry.qin@polyu.edu.hk
[2] Singapore Management University, Singapore, Singapore
[3] National University of Singapore, Singapore, Singapore

Abstract. Including context-aware decision support in the operating room has the potential to improve surgical safety and efficiency by utilizing real-time feedback obtained from surgical workflow analysis. In this task, recognizing each surgical activity in the endoscopic video as a triplet <*instrument, verb, target*> is crucial, as it helps to ensure actions occur only after an instrument is present. However, recognizing the states of these three components in one shot poses extra learning ambiguities, as the triplet supervision is highly imbalanced (positive when all components are correct). To remedy this issue, we introduce a triplet disentanglement framework for surgical action triplet recognition, which decomposes the learning objectives to reduce learning difficulties. Particularly, our network decomposes the recognition of triplet into five complementary and simplified sub-networks. While the first sub-network converts the detection into a numerical supplementary task predicting the existence/number of three components only, the second focuses on the association between them, and the other three predict the components individually. In this way, triplet recognition is decoupled in a progressive, easy-to-difficult manner. In addition, we propose a hierarchical training schedule as a way to decompose the difficulty of the task further. Our model first creates several bridges and then progressively identifies the final key task step by step, rather than explicitly identifying surgical activity. Our proposed method has been demonstrated to surpass current state-of-the-art approaches on the CholecT45 endoscopic video dataset.

Keywords: triplet disentanglement · surgical activity recognition · endoscopic videos

1 Introduction

Surgical video activity recognition has become increasingly crucial in surgical data science with the rapid advancement of technology [1–3]. This important task provides comprehensive information for surgical workflow analysis [4–7] and surgical scene understanding [8–10], which supports the implementation of safety

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 451–461, 2023.
https://doi.org/10.1007/978-3-031-43996-4_43

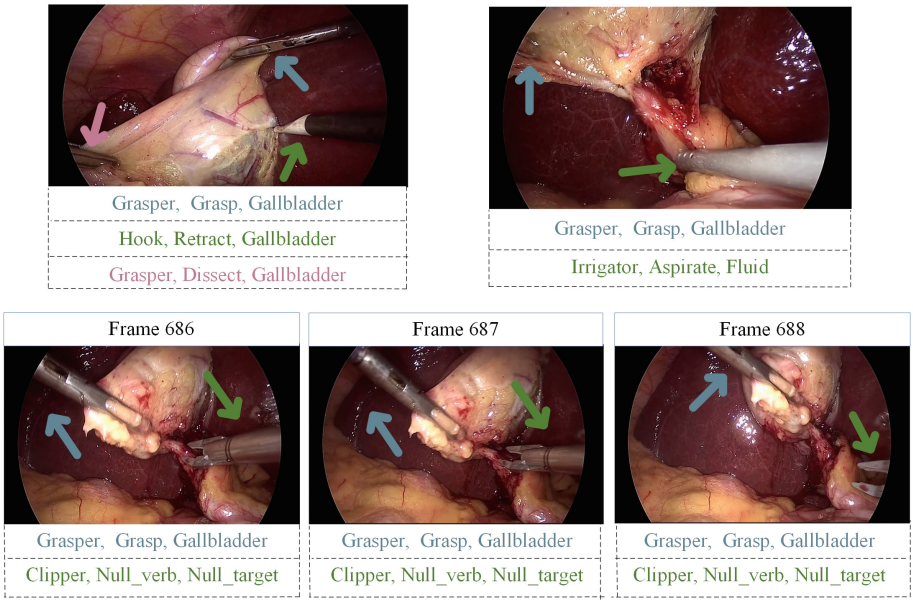

Fig. 1. First Line: Left: Multiple surgical triplets appearing at the same time. **Right:** Instruments located near the boundary. **Second Line:** In consecutive frames of an endoscopic surgery video, an unrelated action (green arrow) is labeled as null. (Color figure online)

warning and computer-assisted systems in the operating room [11]. One of the most popular surgical procedures worldwide is laparoscopic cholecystectomy [12, 13], which is in high demand for the creation of an effective computer-assisted system. Therefore, automated surgical activity triplet recognition is increasingly essential, and learning-based methods are promising solutions to address this need.

Most current works in the field of surgical video analysis primarily focus on surgical phase recognition [6,14–18]. However, only a small portion of the literature is dedicated to surgical activity recognition. The first relevant study [19] dates back to 2020, in which the authors built a relevant dataset and proposed a weakly supervised detection method that uses a 3D interacting space to identify surgical triplets in an end-to-end manner. An updated version Rendezvous (RDV) [20] employs a Transformer [21] inspired semantic attention module in their end-to-end network. Later this method is extended to include the temporal domain, named Rendezvous in Time (RiT) [22] with a Temporal Attention Module (TAM) to better integrates the current and past features of the verb at the frame level. Significantly, benchmark competitions such as the Cholectriplet2021 Challenge [28] have garnered interest in surgical action triplet recognition, with its evaluation centering on 94 valid triplet classes while excluding 6 null classes.

Although significant progress has been made in surgical activity triplet recognition, they suffer from the same limitation of ambiguous supervision from the triplet components. Learning to predict the triplet in one shot is a highly imbalanced process, as only samples with correct predictions across all components are considered positive. This objective is particularly challenging in the following scenarios. Firstly, multiple surgical activities may occur in a single frame (see Fig. 1), with instruments appearing at the edge of the video or being obscured or overlapped, making it difficult to focus on them. Secondly, similar instruments, verbs, and targets that do not need to be recognized can have a detrimental impact on the task. As illustrated in the second row of Fig. 1, an obvious movement of the clipper may be labeled as null in the dataset because it is irrelevant to the recognition task. However, this occurrence is frequent in real situations, and labeling all surgical activities is time-consuming. Furthermore, not all surgical activities are indispensable, and some only appear in rare cases or among surgeons' habits. Moreover, the labels of instruments and triplets for this task are binary, and when multiple duplicate instruments or triplets appear, recognizing them directly only determines whether their categories have appeared or not. To improve recognition results, these factors should also be considered. Lastly, most of the previous methods are end-to-end multi-task learning methods, which means that training may be distracted by other auxiliary tasks and not solely focused on the key task.

To solve the above problems, we propose a triplet disentanglement framework for surgical activity triplet recognition. This approach decomposes the learning objectives, thereby reducing the complexity of the learning process. As stated earlier, surgical activity relies on the presence of tools, making them crucial to our mission. Therefore, our approach concentrates on a simplified numerical representation as a means of mitigating these challenges. Initially, we face challenges such as multiple tools appearing at the same time or irrelevant surgical activities. Therefore, we adopt an intuitive approach to first identify the number/category of tools and whether those activities occur or not, instead of directly recognizing the tool itself. This numerical recognition task helps our network roughly localize the tool's location and differentiate irrelevant surgical activities. Subsequently, we employ a weakly supervised method to detect the tools' locations. However, unlike [20], we extend our architecture to 3D networks to better capture temporal information. Our approach separates different types of instruments using the class activation map (CAM) [23] based on the maximum number of identified instrument categories, allowing our model to slightly minimize the probability of mismatching and reduce the learning difficulty after separation when multiple surgical activities occur simultaneously. Additionally, we propose a hierarchical training schedule that decomposes our tasks into several sub-tasks, starting from easy to hard. This approach improves the efficiency of each individual task and makes training easier.

In summary, this work makes the following contributions: 1) We propose a triplet disentanglement framework for surgical action triplet recognition, which decomposes the learning objectives in endoscopic videos. 2) By further exploit-

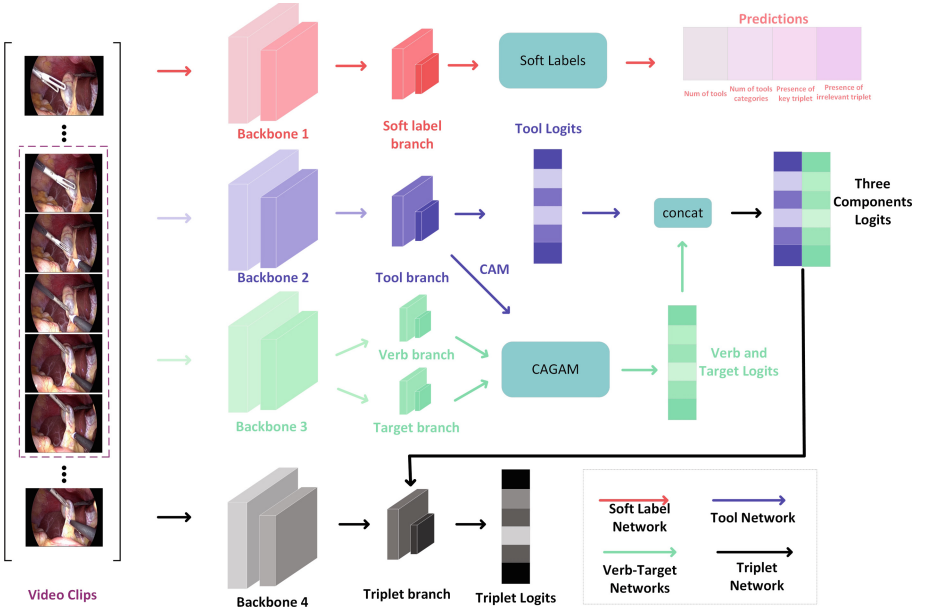

Fig. 2. Overview of Our triplet Disentanglement Network. The parameters of the soft label network are used for the initialization of the other three networks.

ing the knowledge of decomposition, our network is extended to a 3D network to better capture temporal information and make use of temporal class activation maps to alleviate the challenge of multiple surgical activities occurring simultaneously. 3) Our experimental results on the endoscopic video dataset demonstrate that our approaches surpass current state-of-the-art methods.

2 Methods

Our objective is to recognize every triplet at each frame in each video. Let $X = \{X_1, ..., X_n\}$ be the frames of endoscopic videos and $Y = \{Y_1, ..., Y_n\}$ be the set of labels of triplet classes where n is the number of frames and the sets of triplet classes. Moreover, each set of labels of triplet classes can be denoted as $Y = \{Y^I, Y^V, Y^T\}$ where I, V, and T are indicated as Instrument, Verb, and Target. Figure 2 shows the overview of our method.

2.1 Sampling Video Clips

Unlike RDV [20] method which identifies surgical activities in individual frames, the video is cut into different segments to identify them. At the beginning of our process, short video clips are obtained from lengthy and differently-sized videos. Each sample clip drawn from the source videos is represented as $X_{\mathrm{s}} \in {}^{H \times W \times 3 \times M}$, while X_{t} represents a sample clip from the target. It should be noted that all sample clips have identical specifications, namely a fixed height H, a fixed width W, and a fixed number of frames M.

2.2 Generating Soft Labels

We generate four different soft labels for the number of instruments, the number of instrument categories, the presence of an unrelated surgical activity, and the presence of a critical surgical activity, which are labeled as $<N_I, N_C, U_A, S_A>$ respectively. As for N_I, our goal is to know the number of instruments occurrences because the appearance of an instrument also means the appearance of a surgical activity. For N_C, it is employed to differentiate situations where multiple instruments of the same type are present. In terms of the other two soft labels, they are binary labels, which refer to presence or absence, respectively. In addition, labels with the format *<instrument, null, null>* are marked as irrelevant surgical activity. In contrast, those with the format *<instrument, verb, target>* are marked as critical surgical activity. For example, in the case shown in the second line of Fig. 1, our soft label will be marked as *<2, 2, presence, presence>*.

2.3 Disentanglement Framework

Our encoder backbone is built on the RGB-I3D [26] network (preserving the temporal dimension), which is pre-trained on both ImageNet [25] and Kinetics dataset [26]. Prior to training, each video is segmented into T non-overlapping segments consisting of precisely 5 frames, which are then fed into the soft label network to recognize their corresponding soft labels. This allows our network to more easily identify the location of the tool and to differentiate between crucial and irrelevant actions, which ultimately aids the network's comprehension of triplet associations. Once the network is trained, its parameters are stored and transferred for initialization to the backbones of other networks. In the second stage, we divide our triplet recognition task into three sub-networks: the tool network, verb network, and target branch network. Notably, our backbone operates at the clip level, whereas RDV is frame-based. The features extracted by the backbone from video clips are fed into the three sub-networks. The tool classifier recognizes the tool class and generates its corresponding Class Activation Map (CAM). The features of the verb and target networks, along with the CAMs of the corresponding tool network, are then passed into the Class Activation Guided Attention Mechanism (CAGAM) module [20]. Our CAGAM, a dual-path position attention mechanism, leverages the tool's saliency map to guide the location of the corresponding verb and target. Subsequently, the individual predictions of the three components are generated, and the last objective is to learn their association. Therefore, the CAMs and logits of the three components are aligned into the triplet network to learn their association and generate the final triplet prediction. It is important to note that the tool network, verb-target networks, and triplet network are all initialized by the soft label network and contain branches to predict soft labels.

2.4 Hierarchical Training Schedule

Our framework is a multi-task recognition approach, but training so many tasks simultaneously poses a huge challenge to balancing hyper-parameters. To address

this, we propose a hierarchical training schedule method that divides training into different stages. Initially, we train only our soft label network to recognize soft labels, storing its parameters once training is complete. In the next stage, video clips are fed into the tool network to recognize tool categories while simultaneously identifying soft labels. After successful training, the parameters of the tool network are frozen. In the subsequent stage, the verb and target networks identify their respective components and soft labels. At this point, the tool network passes its class activation map to the verb-target networks without updating its parameters. Besides, following previous Tripnet [20], which masks out impossible results, we also mask out improbable outcomes for different components. For instance, predefined masks for the tool are based on possible combinations, while the verb and target masks follow the tool's predictions, excluding scissors and clippers; subsequently, the masking results undergo further refinement for the instrument. Finally, we train the final triplet network using the CAMs and output logits of the three components. Similarly, the parameters of the three components' networks are not updated at this stage. This approach allows us to break down the complexity of the task and improve the accuracy of each individual component at each stage, ultimately leading to higher overall accuracy.

2.5 Separation Processing

Our framework provides a unique approach to address the impact of multiple tool categories that may be present simultaneously. It enables the handling of different surgical instrument categories individually, which reduces the level of complexity for learning. In contrast to RDV [20] and RiT [22], we extend the Grad-CAM [27] approach to 3D-CNN by utilizing the input of 3D tensor data instead of a 2D matrix, resulting in a more precise class activation map. Based on the maximum number of instrument categories (K) predicted by the soft label and the scores obtained by summing each channel along the CAM, we isolate the top-K CAMs. Each isolated CAM is then used to guide the corresponding features of verbs and targets in our CAGAM module to generate individual predictions of the verb and target. Finally, the different predictions of the verb and target are combined. However, differentiating between various instruments, as we do, can help match them with the correct verbs and targets, especially when multiple tool categories appear at the same time, which may be challenging to do without this approach, such as in the RDV method [20].

2.6 Loss Function

For the number of tools and the categories of tools, softmax cross-entropy loss is adopted, while for other soft labels, three components, and triplet, we employ sigmoid cross-entropy losses. Taking sigmoid cross-entropy loss as an example:

$$L = \sum_{c=1}^{C} \frac{-1}{N} (y_c \log_c(\sigma(\hat{y_c})) + (1 - y_c) \log(1 - \sigma(\hat{y_c})),$$

where σ is the sigmoid function, while y_c and $\hat{y_c}$ are the ground truth label and the prediction for specific class c. Besides, the balanced weights are also adopted based on previous works [20].

3 Experiments

Dataset. Following previous works [20], an endoscopic video dataset of the laparoscopic cholecystectomy, called CholecT45 [20], is used for all the experiments. The database consists of a total of 45 videos and is stripped of triplet classes of little clinical relevance. There are 6 instruments, 10 verbs, 15 targets, and 100 triplet classes in the CholecT45 dataset. For all the videos, each triplet class is always presented in the format of <*instrument, verb, target*>. In addition, following the data splits in [24], the K-fold cross-validation method was used to divide the 45 videos in the dataset into 5 folds, each containing 9 videos.

Metric. The performance of the method is evaluated based on the mean average precision (mAP) metric to predict the triplet classes. In testing, each triplet class is computed its own average precision (AP) score, and then the AP score of a video will be calculated by averaging the AP scores of all the triplets in this video. Finally, the mean average precision (mAP) of the dataset is measured by averaging the AP scores of all tested videos.

Besides, Top-N recognition performance is also adopted in our evaluation, which means that given a test sample X_i, a model made a correctness if the correct label y_i appears in its top N confident predictions $\hat{Y_i}$. We follow previous works [20] and measure top-5, top-10, and top-20 accuracy in our experiment.

Implementation Details. I3D (Resnet50) [26] is adopted as a backbone network in our framework, which is pre-trained on the ImageNet [25] and the Kinetics dataset [26]. As for the branches of the different networks, they will be slightly modified to fit their corresponding subtasks. We use SGD as an optimizer and apply a step-wise learning rate of $1e{-}3$ for all sub-tasks, but for the soft-labeled task branches that need to be finetuned, their learning rates start from $1e{-}6$. Our batch size is 32, and there is no additional database for surgical detection.

3.1 Results and Discussion

Quantitative Evaluation. We demonstrate our experimental results with other state-of-the-art approaches on the CholecT45 [20] dataset.

Table 1 presents the benchmark results on the CholecT45 cross-validation split, and compares our model with current state-of-the-art methods [19,20,22] using the mean Average Precision (mAP) metric. The key mission of surgical activity recognition is to calculate the average precision (AP) score of the triplet, denoted as AP_{IVT}. We present tool recognition, verb recognition, and target recognition as AP_I, AP_V, and AP_T, respectively. Compared to previous methods [19,20], our hierarchical training schedule method enables us to achieve

Table 1. Quantitative Results of the Proposed Method Compared to State-of-the-art.

Method	AP_I	AP_V	AP_T	AP_{IVT}
Tripnet [19]	89.9 ± 1.0	59.9 ± 0.9	37.4 ± 1.5	24.4 ± 4.7
Attention Tripnet [20]	89.1 ± 2.1	61.2 ± 0.6	40.3 ± 1.2	27.2 ± 2.7
RDV [20]	89.3 ± 2.1	62.0 ± 1.3	40.0 ± 1.4	29.4 ± 2.8
RiT [22]	88.6 ± 2.6	64.0 ± 2.5	43.4 ± 1.4	29.7 ± 2.6
Our method	**91.2 ± 1.9**	**65.3 ± 2.8**	**43.7 ± 1.6**	**33.8 ± 2.5**

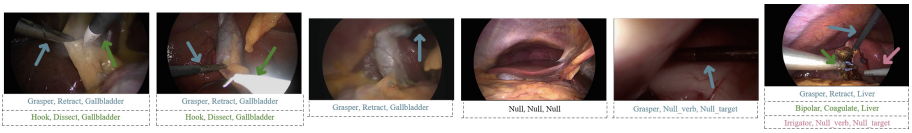

Fig. 3. Qualitative and representative results obtained from our triplet activity recognition model. Color-coded arrows are used to highlight multiple instruments in examples.

better results on individual sub-tasks because we can tune each task individually to achieve the best results. For AP_{IVT}, our framework improves the current SOTA method RiT [22] by 4.1%, demonstrating that decomposing our tasks helps to improve the final result. In addition, we include Fig. 3 in our paper to illustrate the qualitative and representative results obtained from our triplet activity recognition model. Although we are unable to compare our results with other methods as their trained models have not been released, our method can successfully predict many difficult situations, such as the simultaneous appearance of multiple surgical activities, the influence of brightness, and the tool located at the edge.

Following the previous experiments [20], we compare the Top N accuracy of the triplet predictions with different methods. However, the previous method [20] performed this metric only on the CholecT50 [20] dataset, and this dataset has not been published yet. Besides, the source codes of RiT [22] have not been released yet. However, the performance between RDV and RiT is very close according to Table 1. Hence, we try to reproduce the RDV method on this metric. As shown in Table 2, our framework outperforms the RDV method [20] by 8.1%, 5.9% and 2.0% in top-5, top-10, and top-20 respectively.

Ablation Study. In this section, we conduct ablation studies to showcase the effectiveness of each module in our model. As shown in Table 3, the final triplet prediction outcomes experienced a slight decrease of 0.4% and 1.1% in the absence of the separation process or the soft label module, respectively. Additionally, a marginal decrease was observed in the individual sub-tasks. In conclusion, these results demonstrate that the inclusion of these modules can contribute to the overall performance of our model.

Table 2. Top N Accuracy of the Triplet Predictions among Different Methods.

Method	Top-5	Top-10	Top-20
RDV [20]	73.6	84.7	93.2
Our method	**81.7**	**90.6**	**95.2**

Table 3. Comparison of the Different Modules in Our Framework. SC: Separating the category of the instrument in Sect. 2.6; SL: Predicting soft labels in Sect. 2.3.

Method	AP_I	AP_V	AP_T	AP_{IVT}
our method	91.2	65.3	43.7	33.8
our method w/o SC	91.2	64.7	43.1	33.4
our method w/o SL	90.6	63.6	42.8	32.7

4 Conclusion

In this paper, we introduce a novel triplet disentanglement framework for surgical activity recognition. By decomposing the task into smaller steps, our method demonstrates improved accuracy compared to existing approaches. We anticipate that our work will inspire further research in this area and promote the development of more efficient and accurate techniques.

Acknowledgments. The work described in this paper is partly supported by a grant of Hong Kong RGC Theme-based Research Scheme (project no. T45-401/22-N) and a grant of Hong Kong RGC General Research Fund (project no. 15218521).

References

1. Maier-Hein, L., et al.: Surgical data science: enabling next-generation surgery. ArXiv Preprint ArXiv:1701.06482 (2017)
2. Nowitzke, A., Wood, M., Cooney, K.: Improving accuracy and reducing errors in spinal surgery-a new technique for thoracolumbar-level localization using computer-assisted image guidance. Spine J. **8**, 597–604 (2008)
3. Yang, C., Zhao, Z., Hu, S.: Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature. Comput. Assist. Surg. **25**, 15–28 (2020)
4. Zhang, Y., Bano, S., Page, A., Deprest, J., Stoyanov, D., Vasconcelos, F.: Retrieval of surgical phase transitions using reinforcement learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 497–506. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_47
5. Twinanda, A., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**, 86–97 (2016)

6. Zisimopoulos, O., et al.: DeepPhase: surgical phase recognition in CATARACTS videos. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018, Part IV. LNCS, vol. 11073, pp. 265–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_31
7. Nakawala, H., Bianchi, R., Pescatori, L., De Cobelli, O., Ferrigno, G., De Momi, E.: "Deep-Onto" network for surgical workflow and context recognition. Int. J. Comput. Assist. Radiol. Surg. **14**, 685–696 (2019)
8. Valderrama, N., et al.: Towards holistic surgical scene understanding. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 442–452. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_42
9. Lin, W., et al.: Instrument-tissue interaction quintuple detection in surgery videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 399–409. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_38
10. Seidlitz, S., et al.: Robust deep learning-based semantic organ segmentation in hyperspectral images. Med. Image Anal. **80**, 102488 (2022)
11. Franke, S., Meixensberger, J., Neumuth, T.: Intervention time prediction from surgical low-level tasks. J. Biomed. Inform. **46**, 152–159 (2013)
12. Pucher, P., et al.: Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis. Surg. Endosc. **32**, 2175–2183 (2018)
13. Alli, V., et al.: Nineteen-year trends in incidence and indications for laparoscopic cholecystectomy: the NY State experience. Surg. Endosc. **31**, 1651–1658 (2017)
14. Kassem, H., Alapatt, D., Mascagni, P., AI4SafeChole, C., Karargyris, A., Padoy, N.: Federated cycling (FedCy): semi-supervised federated learning of surgical phases. IEEE Trans. Med. Imaging (2022)
15. Ding, X., Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. IEEE Trans. Med. Imaging **41**, 3309–3319 (2022)
16. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: OperA: attention-regularized transformers for surgical phase recognition. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part IV. LNCS, vol. 12904, pp. 604–614. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_58
17. Jin, Y., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging **37**, 1114–1126 (2017)
18. Sahu, M., Szengel, A., Mukhopadhyay, A., Zachow, S.: Surgical phase recognition by learning phase transitions. Curr. Direct. Biomed. Eng. **6** (2020)
19. Nwoye, C.I., et al.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part III. LNCS, vol. 12263, pp. 364–374. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_35
20. Nwoye, C., et al.: Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med. Image Anal. **78**, 102433 (2022)
21. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. Adv. Neural. Inf. Process. Syst. **34**, 15908–15919 (2021)
22. Sharma, S., Nwoye, C., Mutter, D., Padoy, N.: Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. Int. J. Comput. Assist. Radiol. Surg. **18**, 1053–1059 (2023)
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)

24. Nwoye, C., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. ArXiv Preprint ArXiv:2204.05235 (2022)
25. 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)
26. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
27. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
28. Nwoye, C., et al.: CholecTriplet 2021: a benchmark challenge for surgical action triplet recognition. Med. Image Anal. **86**, 102803 (2023)