TWiST: Temporal Weakly-Supervised Triplets Recognition in Surgical Videos (Student Abstract)

Pranshu Danani, Yash Bansal*, Parshiv Kapoor*

Indian Institute of Technology, Roorkee, Uttarakhand, India 247667
{pranshu_d, yash_b, parshiv_k}@bt.iitr.ac.in

Abstract

Deep learning is increasingly applied to intraoperative and surgical video analysis to enable real-time workflow recognition and decision support for improved surgical precision. A key direction is modeling surgical activity as triplets of instrument, action, and target, which provide a richer representation of procedures. However, existing approaches often depend on bounding-box annotations or lack temporal context. We propose TWiST (Temporal Weakly Supervised Triplet detection), a framework that combines weakly supervised instrument localization, temporal attention for triplet prediction, and grounding of triplets with detected instruments. Our experiments show that TWiST outperforms prior weakly supervised baselines.

Introduction

Surgical activity modeling focuses on the recognition of triplets, that is, recognizing interactions between instruments, actions, and anatomical structures. (Nwoye et al. 2020). Incorporating spatial localization of instruments through bounding boxes further adds spatial context to activity modeling, enabling more precise reasoning about surgical workflows. However, generating such detailed annotations requires expert knowledge, is time-consuming, and costly at scale. To reduce dependence on bounding-box annotations, weakly supervised localization methods were introduced that leverage only triplet and instrument labels for instrument localization. Approaches such as RDV-Det, IFNet, and DualMFFNet, etc. (Nwoye et al. 2023),(Nwoye et al. 2022) applied weak supervision for instrument localization, but they generally underperform because they neglect temporal dependencies critical for modeling surgical workflows. Further, (Sharma et al. 2023) applied temporal attention to improve surgical triplet prediction. However, the method lacks instrument localization and the grounding of predicted triplets with instrument detections. So, we propose TWiST(Temporal Weakly Supervised Triplet) recognition, a temporal triplet detection framework with weakly supervised instrument localization. TWiST is a compact three-stage pipeline for weakly supervised surgical triplet detection: a localization module localizes instruments through weak supervision using only instrument class label, a multi-task temporal module predicts instrument, action, target, and the final triplet. A final merging module associates detected instrument regions with the predicted triplets to output final ⟨instrument, action, target, box⟩ annotations.

*These authors contributed equally.

Methodology

Overall pipeline. The approach, as seen in figure 1, comprises a three-stage pipeline: (1) a weakly-supervised instrument counting and localization module, (2) a multi-task temporal triplet detection module predicting instrument, action, and target logits as well as final triplets, and (3) a merging module that grounds triplet predictions using instrument detections.

1. Weakly Supervised Instrument Counting and Localization (Module 1). We model instrument count prediction as a multi-output regression problem instead of a multiclass-multilabel classification, where each of the six surgical instrument classes may appear with multiple instances per frame. Let $\mathbf{x}$ denote an input frame and $\mathbf{y} = [y_1, y_2, \ldots, y_6]^\top$ represent the corresponding ground-truth counts for each instrument class. A ResNet-50 backbone with a final fully connected regression head, $f_\theta(\mathbf{x})$, predicts $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$. The model is trained using mean squared error (MSE). This constitutes a weakly supervised instrument counting and localization head, requiring only class-level instrument labels. To obtain spatial cues, we apply Gradient-weighted Class Activation Maps (Grad-CAM) to the feature maps of the last bottleneck block of ResNet-50 to produce bounding boxes for each instrument instance.

2. Multi-task Temporal Triplet Detector (Module 2). We employ a multi-task temporal model for surgical triplet detection by finetuning a pretrained ResNet-50 backbone to extract per-frame features. Features are processed through two multi-head attention layers to capture short-range temporal dependencies across five consecutive frames. Thus, by leveraging multi-head attention across consecutive frames, we incorporate temporal attention modeling to capture motion-related cues. Three parallel branches predict the logits for instrument (6 classes), action (10 classes), and target (15 classes), which are then concatenated and fed to a triplet branch (100 classes) with the initial extracted image fea-

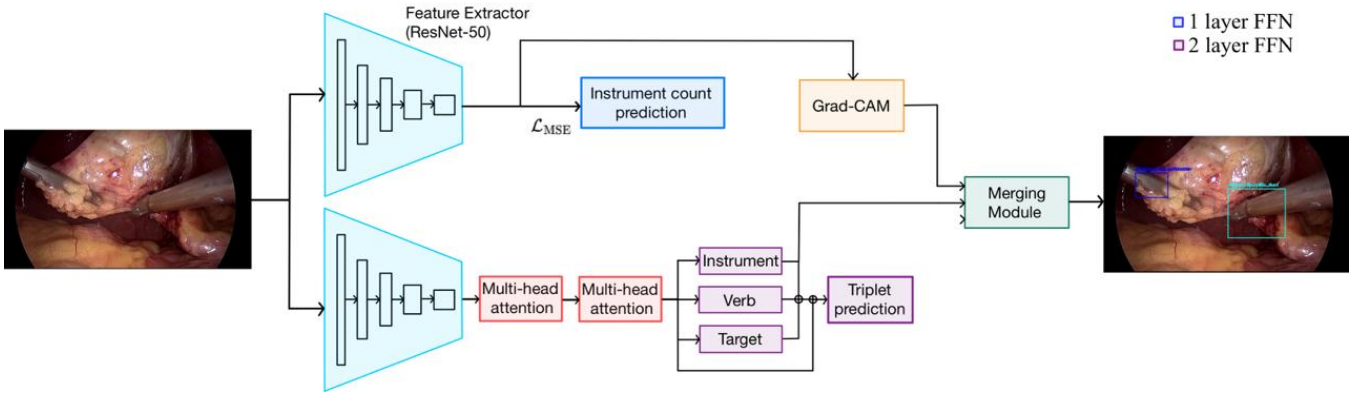

Figure 1: Overview of the proposed TWiST architecture

Model	mAP_i	mRec_i	mAP_{ivt}	mRec_{ivt}	mPre_{ivt}
MTTT*	10.98	**21.15**	1.47	**3.65**	7.18
DualMFFNet	4.57	6.66	0.36	0.73	2.29
RDV-Det	3.00	8.20	0.24	0.86	1.59
DATUM	0.23	2.70	0.07	0.39	0.65
IF-Net	0.70	3.60	0.22	0.92	–
Atom-TKD	0.90	2.40	0.15	0.32	–
SurgeNet*	10.80	19.60	1.30	3.90	–
TWiST (Ours)	**12.07**	19.15	**1.74**	3.14	**7.92**
TWiST (1 attention layer)	5.52	9.50	0.59	1.49	4.31
TWiST (No attention)	4.91	9.15	0.54	1.35	3.02

Table 1. Comparison with existing methods and ablations.

*Methods trained on external CholecT datasets

Bold: Best score, Underline: Best score w/o using other CholecT datasets.

tures. The model is optimized with a weighted binary cross-entropy loss.

3. Merging and Grounding (Module 3). Module 3 receives the triplet predictions with their probabilities from Module 2 and the instrument detections from Module 1. If no instruments are detected, triplet outputs are set to absent (zero probability, coordinates -1). Otherwise, each detected bounding box of the instrument is linked to the top-ranked triplet(s), and the triplet with highest probability is chosen for each instrument. This enforces consistency and grounds triplet predictions. The module outputs triplets with their probabilities, instrument IDs, and bounding box coordinates.

Experiments and Results

Dataset and Experimental Setup : We used a subset of the CholecT45 dataset, specifically 10 videos (out of 45) for training and 5 for testing due to computational limits. For training, there were a total of 17,651 frames sampled at 1 frame per second, while the testing set contained 10,375 frames. We used 2 T-4 GPUs to train our model and conduct ablations.

Experimental Results We evaluated the proposed method, TWiST, against existing surgical action triplet recognition and weakly supervised instrument localization approaches. For in-depth implementation details of these methods, refer to -(Nwoye et al. 2023). (Methods utilizing fully supervised-pretraining on external datasets have not been included in Table1). As seen in Table1, through extensive ablation studies, we observed that TWiST consistently outperformed all methods using no external datasets and achieved competitive results closely trailing methods pretrained on external CholecT datasets. Furthermore, ablation studies on the attention layers indicate that using temporal attention enhances performance by effectively capturing temporal dependencies.

Future Work

Future research can explore advanced temporal attention architectures specifically designed for video-based learning, such as TimeSformer or Video Swin Transformer, which leverage spatiotemporal self-attention for long-range dependency modeling. Additionally, extending weakly supervised learning frameworks to incorporate 3D volumetric feature representations and spatiotemporal localization, as seen in works like WS3D-Net, can further improve joint learning of action and anatomical context in surgical video analysis.

References

Nwoye, C. I.; Gonzalez, C.; Yu, T.; Mascagni, P.; Mutter, D.; Marescaux, J.; and Padoy, N. 2020. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *International conference on medical image computing and computer-assisted intervention*, 364–374. Springer.

Nwoye, C. I.; Yu, T.; Gonzalez, C.; Seeliger, B.; Mascagni, P.; Mutter, D.; Marescaux, J.; and Padoy, N. 2022. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78: 102433.

Nwoye, C. I.; Yu, T.; Sharma, S.; Murali, A.; Alapatt, D.; Vardazaryan, A.; Yuan, K.; Hajek, J.; Reiter, W.; Yamlahi, A.; et al. 2023. Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for sur-

gical action triplet detection. *Medical Image Analysis*, 89: 102888.

Sharma, S.; Nwoye, C. I.; Mutter, D.; and Padoy, N. 2023. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. *International Journal of Computer Assisted Radiology and Surgery*, 18(6): 1053–1059.