

Multi-Stage Dilated Convolution-Mamba Network for Surgical Action Triplet Recognition

Qingke Meng¹, Zhangyong Tang², Xuefeng Zhu¹, Xiao-Jun Wu¹, and Tao Zhou^{1,2}✉

¹ School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, Jiangsu, China.

² The Institute of Translational Medicine, Wuxi School of Medicine, Jiangnan University, Wuxi 214122, Jiangsu, China.
taozhou.ai@jiangnan.edu.cn

Abstract. Accurately recognizing surgical action triplets, comprising instrument, verb, and target, is critical for advancing intelligent health-care systems. However, it remains challenging due to the complex temporal dynamics and intricate semantic dependencies involved. Existing methods often fall short in simultaneously capturing long-range temporal dependencies and local multi-scale features, and they typically overlook explicit semantic correlations among triplet components. To address these limitations, we propose a multi-stage Dilated Convolution-Mamba Network (DCM-Net) for surgical action triplet recognition. Specifically, we propose a Dilated Convolutional Mamba (DCM) block with a parallel architecture: one branch employs a State Space Model (SSM) to capture long-range dependencies, while the other utilizes dilated convolutions with multiple dilation rates to extract multi-scale local temporal features. Moreover, to enhance semantic consistency, we present a Task-aware Graph Fusion (TGF) module, which replaces standard element-wise fusion with a graph-structured scheme. The module constructs semantic nodes from coarse- and fine-grained features, organizes them into clusters with intra-cluster full connections and inter-cluster same-semantic edges, and refines the fused representation via graph convolution to explicitly model triplet component dependencies. Extensive experiments on the CholecT45 dataset demonstrate that our method effectively captures both temporal evolution and semantic interdependencies, achieving state-of-the-art performance. The implementation code is available at <https://github.com/qikuo10/DCM-Net>.

Keywords: Surgical action triplet recognition, Dilated convolution-Mamba network, Task-aware graph fusion.

1 Introduction

Endoscopic surgical workflow analysis, particularly surgical action triplet recognition, identifying per-frame ⟨Instrument, Verb, Target⟩ combinations, is essential for computer-aided interventions, enabling intraoperative guidance, skill assessment, and automated reporting [13, 1, 3]. Unlike conventional video tasks, it

requires modeling subtle instrument–tissue interactions, complex temporal dependencies, and structured semantic relationships among components [15,10,12].

Surgical action triplet recognition poses several challenges. First, effective temporal modeling in surgical videos requires a delicate trade-off between capturing fine-grained local motions and maintaining long-range global context [13,9]. Existing methods based on convolutional neural networks or temporal attention [1] are limited to modeling only a few frames and fail to capture the full temporal dynamics of triplet evolution. Approaches like the four-stage temporal convolutional network in TERL [6] extend the receptive field to the entire video through stacked layers and convolutions with increasing dilation rates [4]. However, these methods still suffer from significant information decay over extremely long sequences, hindering comprehensive modeling of entire surgical procedures. While Transformer-based architectures offer strong global modeling capabilities [19], they incur quadratic computational complexity and often struggle to capture short-term temporal variations, such as subtle instrument vibrations or precise cutting motions. Second, the semantic interdependence among triplet components at the feature level is frequently overlooked [22,21]. Most prior work treats each component independently, neglecting the inherent correlations and distinctive characteristics that together define surgical triplets in feature space. Third, surgical datasets exhibit severe long-tailed distributions [18], where rare but safety-critical events are overshadowed by dominant classes during training, leading to poor tail-class recognition.

To this end, we propose a novel multi-stage Dilated Convolution–Mamba Network (DCM-Net) for surgical triplet recognition. Specifically, we first introduce a multi-stage temporal feature extraction module composed of stacked Dilated Convolution–Mamba (DCM) blocks. By combining the linear complexity and global receptive field of State Space Models (SSM) [5] with the local temporal modeling capabilities of dilated convolutions, the DCM blocks effectively capture both long-range procedural dependencies and short-term temporal variations. To explicitly model the semantic relationships among triplet components, we present a Task-aware Graph Fusion (TGF) module. This module takes as input the multi-stage features extracted at each level of the temporal encoder and performs top-down pairwise fusion. During this process, we construct a semantic graph comprising three node clusters (one per feature scale), each containing four nodes corresponding to Instrument, Verb, Target, and Triplet. Intra-cluster full connections enforce cross-component consistency within each scale, while inter-cluster edges connect same-semantic nodes across scales, enabling the fused representation to absorb complementary information from both coarse- and fine-grained features. **Our contributions are listed as follows: i)** We propose the DCM block, a hybrid architecture that synergizes Mamba and dilated convolutions to solve the dilemma between global context modeling and local feature retention. **ii)** We propose a Task-aware Graph Fusion (TGF) module, which replaces element-wise addition with a graph-structured fusion scheme that explicitly models semantic correlations among triplet components via node construction, intra-cluster connections, and inter-cluster same-semantic edges,

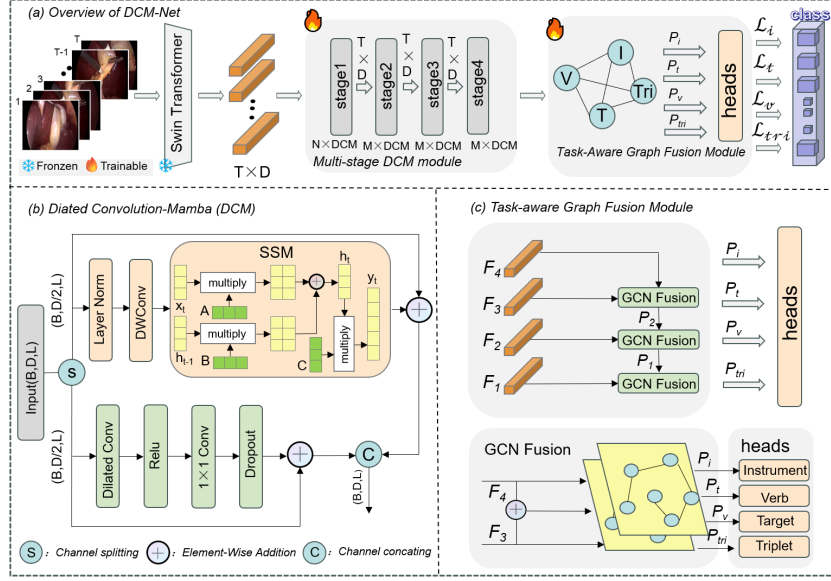


Fig. 1. The overall architecture of the proposed DCM-Net. After extracting spatial features via the backbone network, the resulting feature sequence is linearly projected into a $T \times D$ temporal representation and fed into a multi-stage DCMs for temporal feature modeling. The refined features are then enhanced by the TGF module, which produces the final recognition output.

while aligning features across different granularities. **iii)** Experimental results demonstrate our model outperforms other state-of-the-art surgical action triplet recognition approaches.

2 Proposed Method

Overview. The proposed DCM-Net captures long-range temporal dependencies and inter-task correlations for surgical triplet recognition. As illustrated in Fig. 1, the architecture comprises three main components: a backbone feature encoder, a multi-stage dilated convolution-Mamba (DCM) module, and a task-aware graph fusion (TGF) module. Given input video clips, we first extract a $T \times D$ feature sequence using a Swin Transformer backbone. This sequence is then processed by the DCM module to obtain multi-stage features, which are subsequently passed through the TGF module to predict surgical action triplets.

2.1 Multi-stage Dilated Convolution-Mamba Module

To capture temporal feature dependencies of different granularities, we adopt a multi-stage architecture. Let $F_{in} \in \mathbb{R}^{T \times D}$ represent the frame level spatial

features extracted by the backbone network. Time series extraction is performed through four consecutive stages $S = \{1, 2, 3, 4\}$. We adopted the first stage, which includes N stacked DCM blocks, for fine-grained temporal feature extraction, while the subsequent stages (stages 2 to 4) each contain M stacked DCM blocks for coarse-grained feature extraction. Let K_s represent the number of modules in the s -th stage, defined as:

$$\mathcal{K}_s = \begin{cases} N, & \text{if } s = 1 \\ M, & \text{if } s \in \{2, 3, 4\} \end{cases}. \quad (1)$$

Formally, let X_{s-1} be the input to the s -th stage (where $X_0 = F_{in}$). The output X_s is obtained by stacking K_s DCM Blocks:

$$\mathcal{X}_s = \underbrace{\Phi_{DCM}^{(K_s)}(X_{s-1})}_{K_s \text{ times}} = \text{DCM}(\dots \text{DCM}(X_{s-1}) \dots), \quad s \in \{1, \dots, 4\}. \quad (2)$$

The DCM block adopts a channel-splitting strategy for parallel processing: the SSM branch captures global contextual information, while the dilated convolution branch enhances local feature perception. The outputs are then concatenated to achieve multi-scale feature extraction. As shown in Fig. 1(b), given the input feature tensor $X \in \mathbb{R}^{B \times T \times D}$, where B is the batch size, T is the sequence length, and D is the channel dimension, we first divide it along the channel dimension into two parallel pathways. This channel-wise segmentation reduces the computational complexity of both branches and ensures that the distinct characteristics of each branch are fully utilized. Therefore, the above process can be expressed by

$$X_{glo}, X_{loc} = \text{Split}(X, \text{dim} = D), \quad (3)$$

where $X_{glo}, X_{loc} \in \mathbb{R}^{B \times T \times \frac{D}{2}}$ represent the inputs for the global (upper) branch and the local (lower) branch, respectively. The upper branch leverages the SSM to model long-range dependencies. We formulate this branch as a residual learning process as follows:

$$Y_{glo} = \mathcal{F}_{SSM}(X_{glo}) + X_{glo}, \quad (4)$$

where $\mathcal{F}_{SSM}(\cdot)$ denotes the transformation function comprising Layer Normalization, Depth-wise Convolution, and the Mamba module, with its internal computation governed by the discretized state-space recursion:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t, \quad (5)$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the discretized system matrices derived from the continuous parameters $(\mathbf{A}, \mathbf{B})$ via the zero-order hold method, allowing the model to compress historical context effectively.

The lower branch focuses on using dilated convolution to extract local temporal features. Similarly, this process can be formulated as follows:

$$Y_{loc} = \mathcal{F}_{CNN}(X_{loc}) + X_{loc}, \quad (6)$$

where $\mathcal{F}_{CNN}(\cdot)$ denotes the composite non-linear transformation. Specifically, it applies a dilated convolution with rate d followed by channel mixing:

$$\mathcal{F}_{CNN}(x) = \text{Dropout}(\text{Conv}_{1 \times 1}(\text{ReLU}(\text{D-Conv}_d(x)))). \quad (7)$$

This design effectively expands the receptive field, captures short-term and medium-term temporal features, and complements the upper branch.

Finally, the outputs from both branches are fused via channel concatenation to restore the original feature dimension, and we have

$$Y_{out} = \text{Concat}(Y_{glo}, Y_{loc}). \quad (8)$$

2.2 Task-aware Graph Fusion Module

In this module, we adopt a top-down fusion strategy to integrate coarse- and fine-grained features. The multi-stage features $\{F_i\}_{i=1}^4$ extracted from the MS-DCM are progressively fused in a top-down manner. Specifically, the high-level feature F_{i+1} is used to guide the refinement of the low-level feature F_i via a specially designed Graph Convolutional Network (GCN)-based fusion module. This mechanism ensures that features at different granularities effectively contribute complementary information to the final fused representation.

GCN Fusion. Unlike conventional fusion methods that rely on simple operations such as concatenation or element-wise addition, our GCN fusion strategy explicitly models the semantic dependencies among triplet components (Instrument, Verb, Target) during the feature integration process. Taking the fusion of F_3 and F_4 as an example, we first compute an initial fused representation via element-wise addition: $F_{fuse} = F_3 \oplus F_4$. For each of these three feature sequences, we employ four independent 1×1 convolutional projectors to decompose the representation into four semantically disentangled nodes: $\{I_1, V_1, T_1, Tri_1\}$ from F_3 , $\{I_2, V_2, T_2, Tri_2\}$ from F_4 , and $\{I_f, V_f, T_f, Tri_f\}$ from F_{fuse} , corresponding to Instrument (I), Verb (V), Target (T), and Triplet (Tri). The four projectors do not share parameters, ensuring each node learns a role-specific representation. This yields a total of 12 nodes, whose initial states $H^{(0)}$ are obtained by projecting the input features into the graph space as follows:

$$\mathcal{H}_{F_c, n}^{(0)} = \phi_n(F_c), \quad F_c \in \{F_3, F_4, F_{fuse}\}, \quad n \in \{i, v, t, tri\}, \quad (9)$$

where $\phi_n(\cdot)$ denotes a learnable linear transformation specific to semantic concept n . The 12 nodes are organized into three clusters, each grouping the four nodes derived from the same feature sequence. Within each cluster, the four nodes are fully connected (edge weight 1.0) to learn the associations and distinctions among triplet components and between components and the triplet at the feature level. Across clusters, nodes of the same semantic type are connected (edge weight 0.8), forming four vertical chains $I_1 \leftrightarrow I_2 \leftrightarrow I_f$, $V_1 \leftrightarrow V_2 \leftrightarrow V_f$, $T_1 \leftrightarrow T_2 \leftrightarrow T_f$, and $Tri_1 \leftrightarrow Tri_2 \leftrightarrow Tri_f$, which allow the fused nodes to receive precise boundary signals from shallow-scale nodes and high-level semantic context from deep-scale nodes. Information propagation is performed via a

two-layer graph convolutional network:

$$\mathcal{H}^{(l+1)} = \text{ReLU}(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}), \quad (10)$$

where $\tilde{A}$ is the adjacency matrix with self-loops and $W^{(l)}$ is the learnable weight. After two layers, the refined features of the fused nodes are extracted:

$$P_n = \mathcal{H}_{F_{use}, n}^{(2)}, \quad n \in \{i, v, t, \text{tri}\}. \quad (11)$$

Task-aware Classification Head. To promote the learning of discriminative features for each triplet component, we introduce a task-aware learning strategy. The output features $\{P_i, P_v, P_t, P_{tri}\}$ are passed to their respective classification heads for Instrument, Verb, Target, and Triplet. By jointly optimizing these tasks, the GCN module is effectively supervised to disentangle feature representations, thereby improving the accuracy of triplet recognition.

2.3 Loss Function

To optimize the recognition of triplets and their components, we employ a multi-task objective based on Binary Cross-Entropy (BCE) as follows:

$$\mathcal{L}_{BCE}(\hat{x}_i, y_i) = -[y_i \log(\sigma(\hat{x}_i)) + (1 - y_i) \log(1 - \sigma(\hat{x}_i))]. \quad (12)$$

To address class imbalance in sub-tasks (Instrument, Verb, Target), we assign positive class weights W_i based on the inverse class frequency [2] as follows:

$$W_i = \frac{N_{total} - N_{pos,i}}{N_{pos,i}}. \quad (13)$$

Consequently, the weighted loss for a specific sub-task τ is calculated by

$$\mathcal{L}_\tau = \sum_{i=1}^C W_i^T \cdot \mathcal{L}_{BCE}(\hat{x}_i, y_i). \quad (14)$$

Finally, the total objective combines the triplet loss with the weighted sub-task losses, which can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{triplet} + \lambda(\mathcal{L}_{instr} + \mathcal{L}_{verb} + \mathcal{L}_{target}). \quad (15)$$

where λ is a trade-off parameter.

3 Experiments and Results

3.1 Experimental Setup

Datasets and Evaluation Metrics. Following previous studies [6,8], we conduct comparison experiments on the CholecT45 dataset, which is a subset of Cholec80 [17]. It contains videos of 45 endoscopic cholecystectomies, recorded

Table 1. Comparison results of different methods on the CholecT45 dataset.

Method	Backbone	AP _I	AP _V	AP _T	AP _{IV}	AP _{IT}	AP _{IVT}
RDV [15]	ResNet18	89.3±2.1	62.0±1.3	40.0±1.4	34.0±3.3	30.8±2.1	29.4±2.8
RiT [16]	ResNet18	88.6±2.6	64.0±2.5	43.4±1.4	38.3±3.5	36.9±1.0	29.7±2.6
TDN [1]	ResNet50	91.2±1.9	65.3±2.8	43.7±1.6	-	-	33.8±2.5
MT4MTL-KD [7]	SwinL(384)	93.1±2.1	71.8±3.4	48.8±3.8	44.9±2.4	43.1±2.0	37.1±0.5
SelfD [20]	SwinB(224) [†]	90.3±2.3	67.4±1.5	47.9±1.8	43.7±4.1	42.9±1.6	37.1±1.9
TERL-T [6]	SwinT(224) [†]	93.5±1.5	71.4±2.2	47.2±2.6	44.7±3.8	42.0±2.4	35.7±1.6
TERL-B [6]	SwinB(224) [†]	93.9±2.0	70.8±2.3	49.4±4.7	43.9±3.4	43.6±2.6	35.6±1.4
TERL-B [6]	SwinB(384) [†]	94.1±2.3	73.0±1.4	51.1±3.8	46.5±4.9	44.9±1.8	37.7±1.5
TERL-Ens [6]	Ensemble [†]	94.6±1.9	73.5±1.9	50.8±3.3	47.3±4.1	45.3±1.9	38.5±1.1
CurConMix-T [8]	SwinT(224)	90.4±2.1	67.8±1.8	48.3±3.4	43.3±2.9	43.3±1.8	37.7±2.1
CurConMix-B [8]	SwinB(224)	90.4±3.0	68.2±1.5	49.7±2.5	44.8±5.4	45.3±2.4	38.8±2.8
CurConMix-B [8]	SwinB(384)	90.9±2.0	68.3±1.3	49.8±3.2	45.2±4.2	45.1±1.1	39.1±2.0
CurConMix-Ens [8]	Ensemble	91.7±2.2	69.5±0.4	51.3±2.9	46.3±5.0	47.1±1.6	40.7±2.1
DCM-Net-T	SwinT(224)	93.4 ± 2.3	71.7 ± 1.7	50.9 ± 2.4	45.8 ± 3.0	45.3 ± 2.9	39.6 ±3.5
DCM-Net-B	SwinB(224)	94.3±2.7	73.4±3.1	52.8±2.2	49.7±3.2	47.7±2.8	41.2 ±2.7
DCM-Net-B	SwinB(384)	95.2±2.6	74.1±2.2	52.5±2.7	50.3±3.5	48.1±3.2	42.0 ±1.6

at a frequency of 1 fps, totaling 90,489 frames. Each frame is annotated with multiple instruments, actions, targets, and triplet labels. We adopt the official 5-fold validation split of CholecT45 [14], with each fold consisting of 31 training sets, 5 validation sets, and 9 test sets, to ensure a fair comparison with other methods. Finally, we evaluate performance using the Average Precision (AP) of triples.

3.2 Implementation Details

In this study, the CurConMix[8] framework based on Swin Transformer[11] is taken as our backbone network for extracting spatial features, consisting of two configurations of network structures (“Tiny” and “Base”). The Base includes two variants with feature map sizes of 224×224 and 384×384 . We froze the backbone network and use the proposed DCM-Net for temporal feature extraction and final triplet classification. We adopt a linear transformation to extract spatial features from the backbone network, transforming the feature map size to 1×1 . Then, we input each video as a batch into the network. Our network stacks 6, 5, 5, 5 DCM Blocks in four stages, with internal dimensions configured with 768 channels. Finally, we output them to four classification heads: instruments, actions, targets, and triplets. We conduct 1000 epochs on NVIDIA GeForce RTX 3090 with an initial learning rate of $2e-2$, and exponentially decays at a rate of 0.99 after 300 epochs. The trade-off parameter λ is set to 0.1.

3.3 Comparison with State-of-the-art Methods

We provide a detailed comparison between our method and several existing approaches, including RDV [15], RiT [16], TDN [1], SelfD [20], MT4MTL-KD [7], TERL [6], and CurConMix [8]. As shown in Table 1, our method consistently

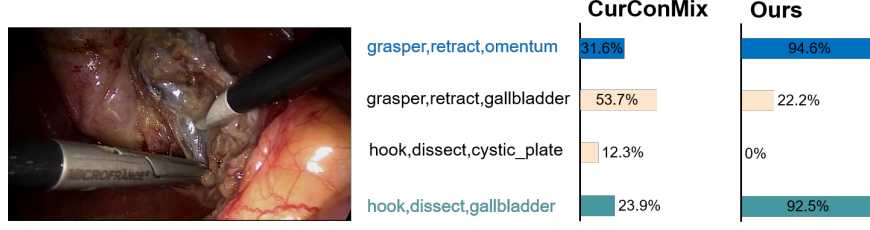


Fig. 2. Comparison results between our method and CurConMix in predicting multi-label key frames.

Table 2. Ablation studies on different components and efficiency comparison.

(a) Module Ablation							(b) DCM Depth	
Settings	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}	(N, M)	AP_{IVT}
Baseline	90.4	68.2	49.7	44.8	45.3	38.8	(3, 2)	39.7
+ MS-TCN	93.8	73.0	51.3	49.3	46.8	39.7	(6, 5)	41.2
+ MS-DCM	93.7	72.4	51.6	49.4	47.1	40.3	(9, 8)	40.5
+ MS-DCM + TGF	94.3	73.4	52.8	49.7	47.7	41.2	(11, 10)	40.2

(c) TGF Fusion Comparison			(d) Efficiency Comparison		
Method	AP_{IVT}	Params	Module	Params (M)	GFLOPs
MLP-based	40.7	5.31M	Full-attn Transformer	37.86	54.16
Transformer	40.4	11.86M	Win-attn Transformer	37.87	39.04
TGF (Ours)	41.2	5.28M	DCM (Ours)	34.21	34.04

outperforms all baselines under different settings. Notably, compared with CurConMix, our method achieves better performance with both backbone architectures. Specifically, the AP improves from 37.7 to 39.6 with Swin-Tiny, and 38.8 to 41.2 with Swin-Base. These results demonstrate the effectiveness of our temporal feature extraction and triplet-level feature fusion strategies, both of which contribute to the overall performance gains. Furthermore, as illustrated in Fig. 2, our method assigns high confidence to the correct categories, whereas CurConMix tends to produce incorrect predictions.

3.4 Ablation Study

Table 2 reports ablation results using CurConMix with Swin Base as baseline. MS-DCM improves AP_{IVT} from 38.8 to 40.3, and TGF further lifts it to 41.2, while a parameter-matched MS-TCN yields only 39.7, confirming the importance of jointly capturing long-range and local temporal features. Sub-table (b) shows the DCM depth exhibits an inverted-U pattern with (6,5) being optimal: the first stage employs six layers to capture fine-grained short-term dynamics, while the three subsequent stages each use five layers to progressively aggregate

longer-range temporal context without interfering with Mamba’s global modeling. Shallower configurations (3,2) lack temporal coverage, while deeper ones (9,8 and 11,10) introduce overfitting. Sub-table (c) shows MLP and Transformer fusion variants underperform (40.7 and 40.4) with the Transformer requiring over twice the parameters (11.86M vs. 5.28M), validating TGF’s accuracy and parameter efficiency. Sub-table (d) reports efficiency under 1000-frame input: DCM requires 34.21M parameters and 34.04 GFLOPs, substantially lower than full-attention (54.16 GFLOPs) and window-attention (39.04 GFLOPs) Transformers, validating the linear-complexity advantage of the Mamba-based design.

4 Conclusion

We propose DCM-Net for surgical triplet recognition, addressing challenges in modeling temporal dynamics and semantic dependencies. It features a DCM module that combines state space models for global context with dilated convolutions for local details, and a GCN-guided fusion module explicitly modeling instrument-verb-target relationships via top-down fusion. Experiments show our method outperforms existing approaches.

Acknowledgement . This work was in part supported by the National Natural Science Foundation of China (No. 62576153).

Disclosure of Interests . The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Y., He, S., Jin, Y., Qin, J.: Surgical activity triplet recognition via triplet disentanglement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 451–461. Springer (2023)
2. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
3. Demir, K.C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S.H.: Deep learning in surgical workflow analysis: a review of phase and step recognition. *IEEE Journal of Biomedical and Health Informatics* 27(11), 5405–5417 (2023)
4. Fang, L., Mou, L., Gu, Y., Hu, Y., Chen, B., Chen, X., Wang, Y., Liu, J., Zhao, Y.: Global–local multi-stage temporal convolutional network for cataract surgery phase recognition. *BioMedical Engineering OnLine* 21(1), 82 (2022)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
6. Gui, S., Wang, Z.: Tail-enhanced representation learning for surgical triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–699. Springer (2024)

7. Gui, S., Wang, Z., Chen, J., Zhou, X., Zhang, C., Cao, Y.: Mt4mtl-kd: A multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging* 43(4), 1628–1639 (2023)
8. Jeon, Y., Shin, J., Park, S., Kim, B., Park, K., Oh, N., Jung, K.H.: Curconmix: A curriculum contrastive learning framework for enhancing surgical action triplet recognition. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 149–158. Springer (2025)
9. Li, Y., Xia, T., Luo, H., He, B., Jia, F.: MT-FiST: a multi-task fine-grained spatial-temporal framework for surgical action triplet recognition. *IEEE Journal of Biomedical and Health Informatics* 27(10), 4983–4994 (2023)
10. Liu, D., Hu, A., Shah, M., Xu, C.: Surgical triplet recognition via diffusion model. *arXiv preprint arXiv:2406.13210* (2024)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022 (2021)
12. 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)
13. Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 364–374. Springer (2020)
14. Nwoye, C.I., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235* (2022)
15. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* 78, 102433 (2022)
16. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: 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 (2023)
17. 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)
18. Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., Santander, J., Caicedo, J., Fernández, N., Arbeláez, P.: Towards holistic surgical scene understanding. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 442–452. Springer (2022)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017)
20. Yamlahi, A., Tran, T.N., Godau, P., Schellenberg, M., Michael, D., Smidt, F.H., Nölke, J.H., Adler, T.J., Tizabi, M.D., Nwoye, C.I., et al.: Self-distillation for surgical action recognition. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 637–646. Springer (2023)
21. Zhang, H., Lu, G., Zhan, M., Zhang, B.: Semi-supervised classification of graph convolutional networks with laplacian rank constraints. *Neural Processing Letters* 54(4), 2645–2656 (2022)

22. Zou, X., Yu, D., Zheng, G.: Capturing action triplet correlations for accurate surgical activity recognition. *Computerized Medical Imaging and Graphics* p. 102604 (2025)