Tail-Enhanced Representation Learning for Surgical Triplet Recognition

Shuangchun Gui and Zhenkun Wang *

School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, China

Abstract. Surgical triplets recognition aims to identify instruments, verbs, and targets in a single video frame, while establishing associations among these components. Since this task has severe imbalanced class distribution, precisely identifying tail classes becomes a critical challenge. To cope with this issue, existing methods leverage knowledge distillation to facilitate tail triplet recognition. However, these methods overlook the low inter-triplet feature variance, diminishing the model's confidence in identifying classes. As a technique for learning discriminative features across instances, contrastive learning (CL) shows great potential in identifying triplets. Under this imbalanced class distribution, directly applying CL presents two problems: 1) multiple activities in one image make instance feature learning to interference from other classes, and 2) limited training samples of tail classes may lead to inadequate semantic capturing. In this paper, we propose a tail-enhanced representation learning (TERL) method to address these problems. TERL employs a disentangle module to acquire instance-level features in a single image. Obtaining these disentangled instances, those from tail classes are selected to conduct CL, which captures discriminative features by enabling a global memory bank. During CL, we further conduct semantic enhancement to each tail class. This generates component class prototypes based on the global bank, thus providing additional component information to tail classes. We evaluate the performance of TERL on the 5-fold cross-validation split of the CholecT45 dataset. The experimental results consistently demonstrate the superiority of TERL over state-of-the-art methods. Our code is available at https://github.com/CIAM-Group/ComputerVision_Codes/tree/main/TERL.

Keywords: Surgical Videos · Triplet Recognition · Multi-label classification · Imbalanced Class Distribution · Prototype Learning.

1 Introduction

Triplet recognition aims to identify fine-grained surgical activities in a video frame [14]. It can foster safety in the operating room by providing surgeons with intra-operative context-aware support [22]. As a key technology for automatically extracting information from surgical videos, it is also essential for surgical archives, postoperative recovery, and surgical education [18,21,1]. In this task, each surgical activity is represented as a triplet of $\langle instrument, verb, target \rangle$. To accurately identify the triplet, it is crucial

* Corresponding Authors: wangzhenkun90@gmail.com

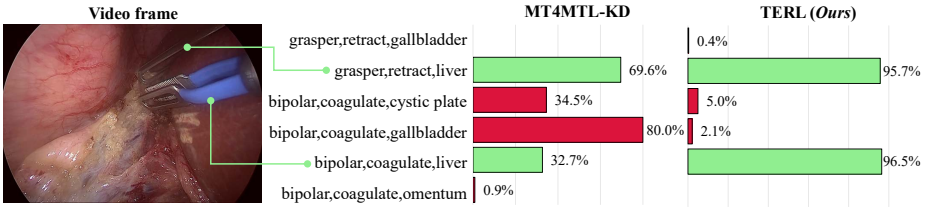

Fig. 1: Top 5 predictions from the SOTA method (*i.e.*, MT4MTL-KD [7]) and our method. Ground-truth labels are highlighted in green.

not only to recognize the involved instrument, verb, and target but also to capture the associations among these components.

Existing methods employ the multi-task learning (MTL) framework to jointly optimize three component tasks and a triplet recognition task [15,16,4,24,13,5]. Across these four tasks, a shared backbone is utilized for feature extraction, thereby enabling the utilization of multiple component features for component association learning. For example, RDV [16] leverages the attention mechanism to learn associations based on features yielded from ResNet-18 [9]. Since the triplet task has an imbalanced class distribution, these MTL models may be over-confident in head classes, resulting in insufficient tail class learning. To solve this issue, many works leverage knowledge distillation to facilitate tail triplet recognition [25,7]. For instance, Yamlahi *et al*. [25] employ soft supervision to mitigate over-confidence, while Gui *et al*. [7] train models with respect to each component task, thereafter assisting the multi-task student training. However, these works overlook inter-class feature variance modeling and are under-confident in recognizing triplets. As shown in Fig. 1, tail triplets of $< bipolar, coagulate, cysticplate >$, $< bipolar, coagulate, gallbladder >$, and $< bipolar, coagulate, liver >$ have limited inter-class feature variance. This overlook may lead to a low confidence score in recognizing $< bipolar, coagulate, liver >$. Under imbalanced classes, this issue is more severe, as capturing tail semantics is more difficult than that of balanced classes.

As a technique for learning instance discriminative features, CL shows great potential in medical image analysis [2,26,23,3]. This motivates us to conduct CL between tail triplet classes, thereby enhancing their importance in model training. However, directly applying CL to an imbalanced triplet task has two problems: 1) multiple activities in one image make instance feature learning to interference from other classes, and 2) limited training samples of tail classes may lead to inadequate semantic capturing.

In this paper, we propose a tail-enhanced representation learning (TERL) method to address these problems. TERL first incorporates a disentangle module into MTL triplet recognition, thus acquiring instance-level features in a single image. Obtaining these features, those from tail classes are selected to conduct CL. It enables a memory bank to store tail instances, thereafter learning discriminative features across classes. During CL, we further conduct semantic enhancement for each tail class. This generates component class prototypes based on the bank, thus providing additional component information to facilitate tail triplet semantic capturing. We conduct experiments on an

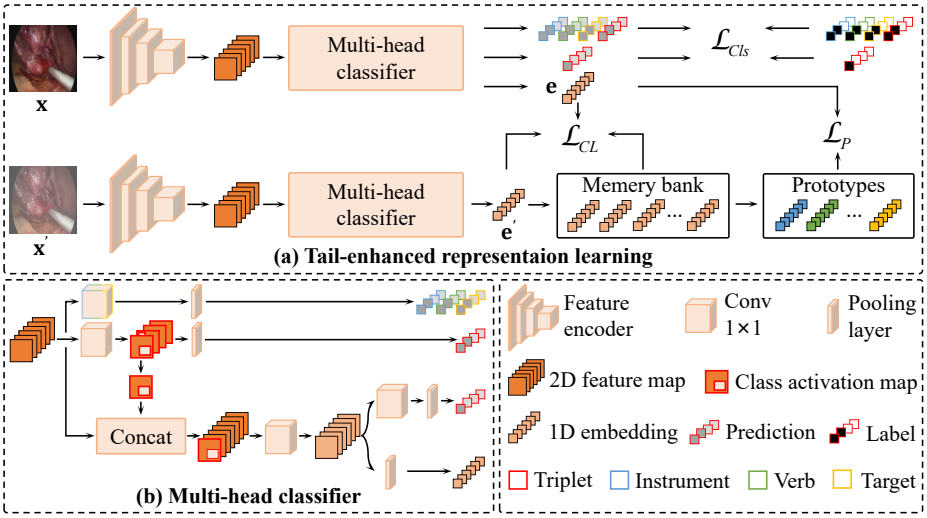

Fig. 2: Overall workflow of TERL.

official 5-fold cross-validation split of the CholecT45 dataset [14]. The experimental results show that TERL outperforms state-of-the-art methods.

2 Methodology

In this paper, we denote each training sample as $\{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{x}_t, \mathbf{y}_t\}_{t=1}^{T}$, where $\mathbf{X}$ is a video sequence comprised of T frames and $\mathbf{x}_t$ is a RGB frame with height H and width W. We use a one-hot vector $\mathbf{y}_t = \{y_{t,k} \in \{0,1\}\}_{k=1}^{K}$ to represent its ground-truth label, where K indicates the number of classes. The objective of surgical triplet recognition is to train a multi-label classification model with minimized prediction error on the test set $\{\overline{\mathbf{X}}, \overline{\mathbf{Y}}\}$. Following [7], the classification model is trained in a two-stage manner, where the spatial and temporal models are trained in different stages. We first employ TERL to train a spatial model. Its output features are sequentially stacked and fed to the temporal model. This model is trained in a multi-task manner, and we employ the weighted cross-entropy loss to optimize the parameters. During the inference stage, the optimized spatial and temporal models are fixed for video recognition.

2.1 Overall Workflow of TERL

As shown in Fig. 2 (a), TERL adopts the MTL framework for tail-enhanced triplet recognition (TETR), while employing instance-level CL (ILCL) to model inter-triplet feature variance. During CL, we further develop prototype-based semantic enhancement (PBSE) to facilitate tail semantic capturing. By feeding the video frames and labels with respect to component (*i.e.*, instrument (I), verb (V), and target (T)), triplet

(IVT), and tail triplet (Ta) classification tasks, a feature encoder followed by a multi-head classifier (MHC) are trained to conduct TETR. After that, the tail embeddings derived from MHC are stored in a global memory bank for ILCL. Obtaining the memory bank, component class prototypes are generated to enhance the component semantics within each tail embedding. The overall training objective is defined as

$$\mathcal{L} = \mathcal{L}_{Cls} + \alpha\mathcal{L}_{CL} + \beta\mathcal{L}_{P}, \tag{1}$$

where $\mathcal{L}_{Cls}$, $\mathcal{L}_{CL}$ and $\mathcal{L}_{P}$ denote the training objectives of TETR, ILCL, and PBSE, respectively. $\alpha = \beta = 1$ are coefficients used to balance the loss terms.

2.2 Tail-Enhanced Triplet Recognition

Model Structure. As shown in Fig. 2 (a), TETR feeds the input frame $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ into a feature encoder, followed by MHC. The training objective for TETR adopts a multi-task classification loss $\mathcal{L}_{Cls} = \sum_{a\in A}\mathcal{L}^{a}_{Cls}$, where $A = \{I, V, T, IVT, Ta\}$. In specific, we employ a simplified asymmetric loss [19] as $\mathcal{L}^{a}_{Cls}$, *i.e.*,

$$\mathcal{L}^{a}_{Cls} = \frac{1}{K_a}\sum_{k=1}^{K_a}\begin{cases}(1-\hat{y}^{a}_{k})^{\gamma+}\log(\hat{y}^{a}_{k}), & y^{a}_{k}=1,\\ (\hat{y}^{a}_{k})^{\gamma-}\log(1-\hat{y}^{a}_{k}), & y^{a}_{k}=0,\end{cases} \tag{2}$$

where $\hat{y}^{a}_{k}$ and y^{a}_{k} are the prediction and ground-truth of the k-th class of task a in the video frame $\mathbf{x}$, respectively. The total loss is computed by averaging this loss over all samples in the current training batch. γ^{+} and γ^{-} are focusing parameters of positive and negative values, respectively. We empirically set $\gamma^{+} = 0$ and $\gamma^{-} = 2$.

Multi-Head Classifier. As shown in Fig. 2 (b), MHC consists of two branches, where a multi-task branch solves the triplet and its three component tasks, and a tail branch disentangles the tail embedding and solves the tail triplet task. We denote the feature maps output from the feature encoder as $\mathbf{F} \in \mathbb{R}^{D\times\frac{H}{32}\times\frac{W}{32}}$. In the multi-task branch, we feed the feature maps $\mathbf{F}$ into four task heads, each of which comprises a 1×1 convolution layer and an average pooling layer. The feature maps from the convolution layer of the triplet head are referred to as class activation maps (CAMs), with task predictions being output from the pooling layer. To disentangle the instance-level feature for tail class k, we derive the k-th channel of CAMs, which is denoted $\mathbf{C}_k \in \mathbb{R}^{1\times\frac{H}{32}\times\frac{W}{32}}$. In the tail branch, we concatenate $\mathbf{C}_k$ with the image-level feature $\mathbf{F}$ and utilize a 1×1 convolutional layer to convert the concatenated features from $(D+1)$-dimensional space to D-dimensional space, obtaining the disentangled feature

$$\widetilde{\mathbf{F}}_k = \texttt{Conv}(\texttt{Concat}(\mathbf{C}_k, \mathbf{F}); D). \tag{3}$$

Finally, the tail triplet prediction and embedding are obtained by a tail head and an average pooling layer, respectively.

2.3 Instance-Level Contrastive Learning

We employ Moco [8] to implement ILCL. ILCL captures semantics by maximizing the similarity between positive pairs, *i.e.*, embeddings from the same tail triplet class, and minimizing it between negative ones, *i.e.*, embeddings from different tail triplet classes. As shown in Fig. 2, $\mathbf{x}$ and $\mathbf{x}^{'}$ are the input frame and its augmented version, respectively. In ILCL, they are passed through two networks, of which each consisting of a feature encoder and the proposed MHC. The two disentangled tail embeddings are denoted as query embedding $\mathbf{e}$ and key embedding $\mathbf{e}^{'}$, respectively. Following [8], the key embeddings produced by previous batches are stored in a memory bank $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_L\}$, and this branch employs momentum-updating strategy. The positive key embeddings is denoted as $\mathbf{E}_p = \{e_i \in \mathcal{B} : y_i = y_q\}$, where y_q is the triplet label of $\mathbf{e}$. Under the imbalanced class distribution, we adopt the loss function of k-positive CL [11] as the training objective following [10]. This function employs the InfoNCE [17] loss, and keeps the number of positive embeddings equal for each class:

$$\mathcal{L}_{CL} = \frac{1}{k+1} \sum_{\mathbf{e}_p \in \mathbf{E}_p^k \cup \mathbf{e}'} \log \frac{\exp\left(\mathbf{e}^\top \mathbf{e}_p / \tau\right)}{\exp\left(\mathbf{e}^\top \mathbf{e}' / \tau\right) + \sum_{j=1}^{L} \exp\left(\mathbf{e}^\top \mathbf{e}_j / \tau\right)}, \tag{4}$$

where $\mathbf{E}_p^k$ is a subset of $\mathbf{E}_p$, comprising k randomly drawn embeddings. The temperature hyperparameter τ and the number of positive embeddings k are set to default values of 0.07 and 7, respectively.

2.4 Prototype-Based Semantic Enhancement

PBSE guides the semantic learning by maximizing the similarity between the tail embedding and corresponding component class prototypes. For query embedding $\mathbf{e}$, we denote the tail triplet label and the corresponding component task labels as $\mathbf{y}_q^{Ta}$ and $\{\widetilde{\mathbf{y}}_q^a \in \mathbb{R}^{K_a} : a \in \{I, V, T\}\}$, respectively. For component task a, we find the set of embeddings that has the same component labels with $\mathbf{y}_q^{Ta}$, denoted as $\mathcal{B}^a = \{\mathbf{e}_i \in \mathcal{B} : \widetilde{\mathbf{y}}_i^a = \widetilde{\mathbf{y}}_q^a\}$, where $\widetilde{\mathbf{y}}_i^a$ is the component label of $\mathbf{e}_i$. By averaging each embedding channel within $\mathcal{B}^a$, the prototype channel is obtained as

$$p_{q,d}^a = \frac{1}{|\mathcal{B}^a|} \sum_{i=1}^{|\mathcal{B}^a|} e_{i,d}, \tag{5}$$

where $e_{i,d}$ denotes the d-th channel of $\mathbf{e}_i$, and $|\mathcal{B}^a|$ is the set length of $\mathcal{B}^a$. The corresponding class prototype of component a is acquired by concatenating all channels, denoted as $\mathbf{p}_q^a = (p_{q,1}^a, p_{q,2}^a, ..., p_{q,D}^a)^\top$. The training objective is formulated as

$$\mathcal{L}_P = \sum_{a \in \{I,V,T\}} \log \frac{\exp\left(\mathbf{e}^\top \mathbf{p}_q^a\right)}{\sum_{j=1}^{K_a} \exp\left(\mathbf{e}^\top \mathbf{p}_j^a\right)}. \tag{6}$$

3 Experiments

3.1 Datasets and Evaluation Metrics

This paper employs a public challenge dataset from CholecTriplet 2021 [14], which is referred to as CholecT45. This dataset comprises 45 laparoscopic cholecystectomy video sequences recorded at a frequency of 1 fps, resulting in a total of 100.9K frames and 161K triplet instance labels. Each frame is annotated with 100 binary action triplets, consisting of 6 instruments, 10 verbs, and 15 targets. To evaluate the effectiveness of TERL, we adopt the official 5-fold cross-validation strategy for model evaluation. This strategy involves a 31-5-9 split for training, validation, and testing, respectively. Following [7], we utilize Fold 1 to perform ablation and sensitivity studies. The performances are evaluated by using the average precision (AP) metrics, which are commonly employed in previous works [16,12,15]. AP metrics comprise three key aspects: triplet average precision (AP_{IVT}), association average precision (AP_{IV} and AP_{IT}), and component average precision (AP_I, AP_V, and AP_T). The main metric is AP_{IVT}, which evaluates the recognition of complete triplets.

3.2 Implementation Details

Following [7], the input video frames for spatial modeling undergo light data augmentations, including resizing images to 224×224, flips, rotations, brightness, and saturation perturbations with the probability of 0.5. Throughout all training stages, models are optimized using stochastic gradient descent (SGD) with a momentum of 0.95. In TERL, triplet class indexes 17, 19, and 60 are head classes, whereas the remaining classes are considered as the tail. The head class is selected if the category has more than 10,000 samples. We employ the Swin Transformer pre-trained on ImageNet-22k as the spatial model, with the output feature dimension $D = 768$. It is trained for 20 epochs with a learning rate of 1e-5 and a batch size of 16. All hyperparameters in ILCL are consistent with those in [10]. For the temporal model, we utilize a four-stage TCN, with each stage consisting of 12 dilated convolution layers, and the hidden layer dimension is set to 512. Our temporal model takes the entire video sequence as input following the specification in [6]. We train the temporal model for 1000 epochs with an initial learning rate of 1e-2, which decays exponentially after 200 epochs with the rate of 0.99.

3.3 Comparison with the State-of-the-Arts

We compare our TERL with state-of-the-art triplet recognition methods (SOTAs), including RDV [16], RiT [20], Chen *et al*. [4], Yamlahi *et al*. [25], and MT4MTL-KD [7]. We implement TERL with different backbones, where TERL-T and TERL-B represent Swin Transformer Tiny and Base, respectively. Moreover, we ensemble the trained models in TERL-T and TERL-B to obtain final predictions of TERL-Ens. This is implemented by averaging sigmoid probabilities derived from the selected models. Table 1 presents the mean and standard deviation results of AP on the cross-validation split. The results indicate that TERL achieves comparable performance to MT4MTL-KD using a single model. Moreover, an additional improvement of 1.5% AP_{IVT} is observed by implementing TERL-Ens.

Table 1: Benchmark triplet recognition AP (%) on CholecT45 dataset. **bold** = best score. underlined = best score in the state-of-art methods.

Method	Backbone	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	$\boldsymbol{AP_{IVT}}$
RDV [16]	Res18	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 [20]	Res18	88.6±2.6	64.0±2.5	43.4±1.4	38.3±3.5	36.9±1.0	29.7±2.6
Chen *et al*. [4]	Res50	91.2±1.9	65.3±2.8	43.7±1.6	-	-	33.8±2.5
Yamlahi *et al*. [25]	SwinB×2+SwinL	-	-	-	-	-	38.5±0.0
MT4MTL-KD [7]	Res18+SwinL	<u>93.9±2.0</u>	<u>73.8±2.0</u>	<u>52.1±5.2</u>	<u>46.5±3.4</u>	<u>46.2±2.3</u>	<u>38.9±1.6</u>
TERL-T	SwinT	93.1±2.4	71.1±1.7	48.9±3.9	44.9±4.4	41.9±3.1	35.7±2.3
TERL-B	SwinB	93.5±2.4	72.8±2.8	51.3±3.8	47.0±5.6	45.7±2.8	38.9±2.5
TERL-Ens	SwinT+SwinB	**94.5±2.2**	**74.0±1.6**	**52.9±4.9**	**47.6±5.2**	**46.3±2.1**	**40.4±2.4**

Table 2: Ablation on key modules within TERL.

Method	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	$\boldsymbol{AP_{IVT}}$
Baseline-SwinT	90.0	65.7	47.9	37.7	41.4	32.6
+ TETR	89.6	67.5	47.3	38.4	42.2	33.6
+ TETR + PBSE	90.6	69.0	49.5	40.3	43.7	34.9
+ TETR + ILCL	90.6	70.7	49.1	40.9	42.9	36.1
+ TETR + ILCL + PBSE	**91.5**	**71.3**	**54.0**	**43.2**	**46.1**	**39.0**

3.4 Ablation Study

Ablation on key modules within TERL. We conduct ablation experiments to validate the effectiveness of TETR, ILCL, and PBSE. Initially, we employ a vanilla multi-task approach (Baseline-SwinT) based on SwinT. This configuration excludes the tail branch defined in MHC and includes the triplet and its three component tasks. Subsequently, we integrate key modules by adding the corresponding loss terms. Table 2 presents the results obtained in Fold 1. According to these results, we observe that adding a tail branch results in only a marginal improvement of 1.0% AP_{IVT}. Moreover, PBSE and ILCL yield additional improvements of 2.3% and 3.5% AP_{IVT}, respectively. Our TERL, which combines both ILCL and PBSE, achieves the highest AP_{IVT} of 39.0%.

Ablation on key modules within TETR and PBSE. In TETR, the tail branch concatenates CAM with the image-level feature, thereby disentangling tail instances. To investigate the contribution of CAM, we test the TERL performance with or without this map. As shown in the left of Fig. 3, by leveraging the CAM, AP_{IVT} achieved by TERL is improved from 37.5% to 39.0%. We further conduct ablation experiments to investigate the contribution of each component enhancement within PBSE. These TERL performances are tested over different numbers and combinations of component tasks. As shown in the right of Fig 3, enhancing any component task can lead to performance improvement, and more component refinements result in more significant improvements. As for the impact of enhanced components on TERL performance, the instrument task contributes the most, while the target contributes the least. In this triple

Methods	w/ CAM	w/o CAM
AP_I	91.4	90.6
AP_V	71.2	70.7
AP_T	53.9	50.8
AP_{IVT}	39.0	37.4

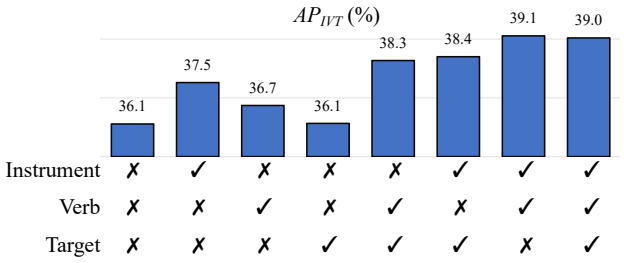

Fig. 3: **Left:** Ablation on CAM in TETR. **Right:** Ablation on components in PBSE.

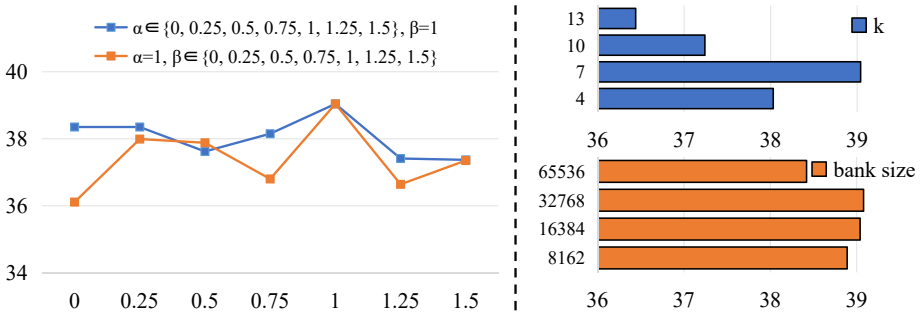

Fig. 4: **Left:** AP_{IVT} scores under different balance weights of ILCL (α) and PBSE (β). **Right:** AP_{IVT} scores under different hyper-parameters used in ILCL.

recognition, the target recognition is more challenging than the other two, resulting in the poor performance of target prototypes and thus leading to less contribution in TERL.

3.5 Sensitivity Studies

We conduct sensitivity studies on four hyper-parameters, including two hyper-parameters used in ILCL and two balance weights of ILCL and PBSE. In Eq. 1, α and β are used to balance the contributions of ILCL and PBSE losses. To study the loss impact, we test the performance of TERL using $\alpha = 1, \beta \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$, and $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}, \beta = 1$. The left of Fig. 4 shows the AP_{IVT} values achieved on Fold 1. We can find that TERL is sensitive to these hyper-parameters, and achieves the best in $\alpha = \beta = 1$. As for hyper-parameters within ILCL, we test the performance using different numbers of positive embeddings k and bank sizes L in Eq. 4. Results in the right of Fig. 4 demonstrate TERL prefers low k values and is robust in $L \in \{8162, 16384, 32768\}$.

4 Conclusion

This paper proposes a tail-enhanced representation learning method for multi-task triplet recognition (TERL). TERL incorporates a disentangle module into MTL triplet recogni-

tion, thus acquiring instance-level features in a single image. Obtaining these features, those from tail classes are selected to conduct CL. This is implemented by capturing discriminative features across classes, which are stored in a global memory bank. During CL, we further conduct semantic enhancement for each tail class. This generates component class prototypes based on the bank, thus providing additional component information to facilitate tail triplet semantic capturing. TERL achieves outstanding performance on the cross-validation split of the CholecT45 dataset. The ablation studies are also conducted to validate the effectiveness of each key module.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 62106096), the Natural Science Foundation of Guangdong Province (Grant No. 2024A1515011759), the National Natural Science Foundation of Shenzhen (Grant No. JCYJ20220530113013031).

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

References

1. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–407. Springer (2010)
2. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in neural information processing systems **33**, 12546–12558 (2020)
3. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis **87**, 102792 (2023)
4. 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)
5. Cheng, Y., Liu, L., Wang, S., Jin, Y., Schönlieb, C.B., Aviles-Rivero, A.I.: Why deep surgical models fail?: Revisiting surgical action triplet recognition through the lens of robustness. In: International Workshop on Trustworthy Machine Learning for Healthcare. pp. 177–189. Springer (2023)
6. Ding, X., Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. IEEE Transactions on Medical Imaging **41**(11), 3309–3319 (2022)
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 (2023)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

10. Hou, C., Zhang, J., Wang, H., Zhou, T.: Subclass-balancing contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5395–5407 (2023)
11. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2020)
12. Li, L., Li, X., Ding, S., Fang, Z., Xu, M., Ren, H., Yang, S.: Sirnet: fine-grained surgical interaction recognition. IEEE Robotics and Automation Letters **7**(2), 4212–4219 (2022)
13. 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 (2023)
14. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Cholectriplet2021: a benchmark challenge for surgical action triplet recognition. Medical Image Analysis **86**, 102803 (2023)
15. 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)
16. 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)
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis **16**(3), 632–641 (2012)
19. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
20. 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 pp. 1–7 (2023)
21. 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)
22. Vercauteren, T., Unberath, M., Padoy, N., Navab, N.: Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. Proceedings of the IEEE **108**(1), 198–214 (2019)
23. Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J.: Federated contrastive learning for volumetric medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 367–377. Springer (2021)
24. Xi, N., Meng, J., Yuan, J.: Chain-of-look prompting for verb-centric surgical triplet recognition in endoscopic videos. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5007–5016 (2023)
25. 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)
26. Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y.: Positional contrastive learning for volumetric medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,

Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 221–230. Springer (2021)