MT4MTL-KD: A Multi-Teacher Knowledge Distillation Framework for Triplet Recognition

Shuangchun Gui, Zhenkun Wang, *Member, IEEE*, Jixiang Chen, Xun Zhou, Chen Zhang, and Yi Cao

***Abstract*— The recognition of surgical triplets plays a critical role in the practical application of surgical videos. It involves the sub-tasks of recognizing instruments, verbs, and targets, while establishing precise associations between them. Existing methods face two significant challenges in triplet recognition: 1) the imbalanced class distribution of surgical triplets may lead to spurious task association learning, and 2) the feature extractors cannot reconcile local and global context modeling. To overcome these challenges, this paper presents a novel multi-teacher knowledge distillation framework for multi-task triplet learning, known as MT4MTL-KD. MT4MTL-KD leverages teacher models trained on less imbalanced sub-tasks to assist multi-task student learning for triplet recognition. Moreover, we adopt different categories of backbones for the teacher and student models, facilitating the integration of local and global context modeling. To further align the semantic knowledge between the triplet task and its sub-tasks, we propose a novel feature attention module (FAM). This module utilizes attention mechanisms to assign multi-task features to specific sub-tasks. We evaluate the performance of MT4MTL-KD on both the 5-fold cross-validation and the CholecTriplet challenge splits of the CholecT45 dataset. The experimental results consistently demonstrate the superiority of our framework over state-of-the-art methods, achieving significant improvements of up to 6.4% on the cross-validation split.**

***Index Terms*— Surgical activity recognition, knowledge distillation, multi-label image classification.**

Manuscript received 30 October 2023; accepted 14 December 2023. Date of publication 21 December 2023; date of current version 3 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62106096, in part by the Shenzhen Technology Plan under Grant JCYJ20220530113013031, in part by the Characteristic Innovation Project of Colleges and Universities in Guangdong Province under Grant 2022KTSCX110, and in part by the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) under Grant pdjh2023c21602. *(Corresponding author: Zhenkun Wang.)*

Shuangchun Gui and Jixiang Chen are with the School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: 12132667@mail.sustech.edu.cn).

Zhenkun Wang is with the School of System Design and Intelligent Manufacturing and the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: wangzhenkun90@gmail.com).

Xun Zhou is with the Department of Computer Science, City University of Hong Kong, Hong Kong, SAR, China.

Chen Zhang and Yi Cao are with the Department of Radiology, the Shenzhen Third People's Hospital, the Second Affiliated Hospital of Southern University of Science and Technology, and the National Clinical Research Center for Infectious Diseases, Shenzhen 518112, China.

Digital Object Identifier 10.1109/TMI.2023.3345736

I. Introduction

VIDEO-DRIVEN fine-grained surgical action recognition aims to recognize detailed surgical activities in each video frame [1]. It can foster safety in the operating room by providing surgeons with intra-operative context-aware support [2]. As a key technology for automatically extracting information from surgical videos, it is also essential for surgical archives, postoperative recovery, and surgical education [3], [4], [5]. Among all fine-grained surgical action recognition tasks, recognizing triplets of the surgical activity is an emerging topic that delivers the finest level of granularity in surgical activity understanding. Specifically, the surgical activity is formalized as a triplet of $\langle instrument, verb, target \rangle$, which is commonly referred to as triplet recognition. Triplet recognition is a multi-label image classification problem, as multiple activities may occur in one frame. An example of triplet recognition in CholecT45 [6] is shown in Fig. 1 (a). Two triplets, $\langle hook, dissect, cystic\ plate \rangle$ and $\langle grasper, retract, gallbladder \rangle$, appear in one frame to represent the cystic plate dissection with the hook and the gallbladder retraction using the grasper, respectively.

To accurately identify triplets, it is imperative not only to recognize the involved instruments, actions, and targets effectively but also to capture the association among these three sub-tasks (*i.e.*, instrument (I), verb (V), and target (T) classifications). Existing triplet recognition methods employ the multi-task learning framework to facilitate task association learning [7], [8], [9]. The three sub-tasks are jointly optimized with the triplet (IVT) recognition task. Across these four tasks, a shared backbone is utilized for feature extraction, thereby enabling the utilization of multiple sub-task features for task association learning. For example, RDV [8] leverages the attention mechanism to recognize the sub-tasks of triplet based on ResNet-18 [10] and learn their associations. However, these methods suffer from the following two limitations:

1) *Imbalanced class distribution may lead to spurious task association learning.* For example, as in the CholecT45 dataset, the triplet $\langle grasper, retract, gallbladder \rangle$ and the triplet $\langle grasper, retract, cystic\ plate \rangle$ have similar features. Nevertheless, the former has significantly more training data than the other latter. Under the multi-task learning framework, the model may over-learn the correlation between the features concerning $\langle grasper, retract, gallbladder \rangle$. This spurious

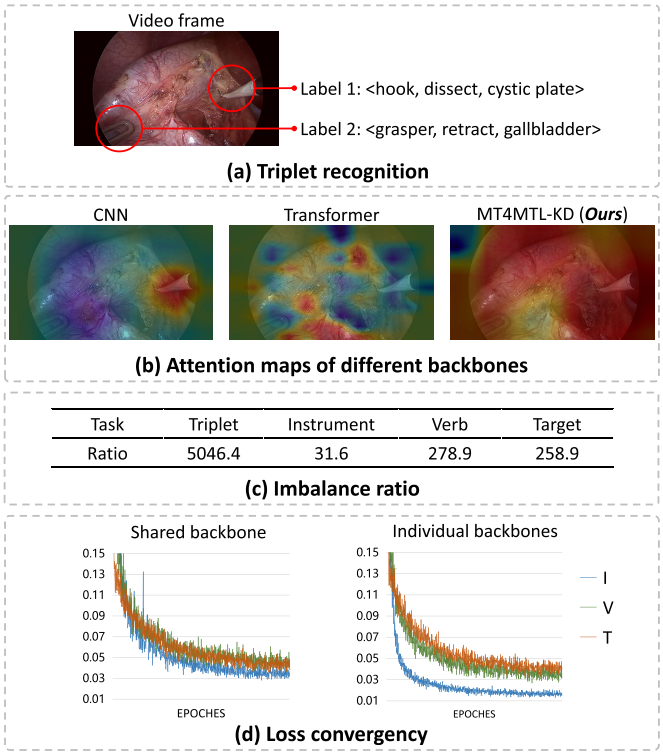

Fig. 1. (a) An introduction of triplet recognition. (b) Attention maps of different backbones. The CNN-based model possesses a limited local attention field, while the Transformer-based model presents a more extensive one. MT4MTL-KD offers a favorable attention field that facilitates both local and global context modeling. (c) The class imbalance ratios of triplet recognition and its sub-tasks. Higher values indicate a more severe class imbalance. (d) Loss convergence on a shared backbone and individual backbones. A shared backbone results in inferior performance, as it is unable to converge to an optimal point for each sub-task.

task association adversely affects the learning of target classification, making the model fail to recognize the triplet $\langle grasper, retract, cystic\ plate \rangle$.

2) *Existing backbones cannot reconcile local and global context modeling.* The backbones employed in triplet recognition can be divided into two categories: CNN-based [6], [8], [11], [12] and Transformer-based models [9], [13], [14]. CNN-based backbones excel at local context modeling and are effective at extracting features related to instruments and small targets, *e.g.*, *blood vessels*. However, CNN-based models are not adept at capturing global context information [15], which hampers the accurate recognition of verbs and large targets *e.g.*, *omentum*. In contrast, Transformer-based models are superior in global context modeling but have limitations in recognizing small structures (as demonstrated in Figure 1(b)).

To tackle the first issue, we propose to conduct knowledge distillation on the triplet's sub-tasks to achieve better multi-task learning for triplet recognition. As shown in Fig 1(c), the class distribution of sub-tasks is significantly less imbalanced than that of the triplet recognition task, which is reflected by task-imbalance ratio [16] discrepancy. According to [17], models trained on a less imbalanced data subset can perform better than those trained jointly. This motivates us to train models with respect to each sub-task as teacher models, thereafter assisting the triplet recognition learning. To address the second limitation, we choose different categories of backbones as teacher and student models to integrate both local and global context modeling [18].

In this work, we present a multi-teacher knowledge distillation framework for multi-task triplet learning (MT4MTL-KD). MT4MTL-KD first decomposes the triplet labels into three sub-task labels and performs independent teacher training with respect to each sub-task. In contrast to a shared teacher model, our experiments indicate that the teacher models independently trained are more efficient than those jointly trained (as depicted in Figure 1(d)). Once the teacher models are obtained, we conduct feature-level and prediction-level distillation to guide the training of a unified multi-task student model. Specifically, the prediction-level distillation provides additional soft supervision information to facilitate task association learning. In contrast, the feature-level distillation incorporates the Transformer's capability of grasping global context knowledge with the strength of CNN's efficient local context modeling. This is implemented by minimizing the feature discrepancy between multiple teachers and a unified student. To align the complex semantic knowledge, we introduce a novel feature attention module (FAM). For each teacher, FAM calculates its attention weights on different channels of the student feature. These weights enable this teacher to guide its highly relevant student channel, making the feature-level distillation process more accurate and effective. We conduct experiments on the official 5-fold cross-validation split and CholecTriplet challenge split of the CholecT45 dataset [6] for model evaluation. The experimental results show that our framework outperforms state-of-the-art methods. Our contributions are summarized as follows:

- We propose a novel multi-teacher knowledge distillation framework for triplet recognition (MT4MTL-KD), which utilizes knowledge distillation from sub-tasks to improve task association learning as well as transfers semantic features between Transformers and CNNs to enhance feature extraction.
- We develop a novel feature attention module (FAM) for efficient feature-level knowledge transfer between the triplet task and multiple sub-tasks.
- The experimental results on CholecT45 indicate that MT4MTL-KD consistently outperforms the state-of-the-art methods, *e.g.*, 6.4% on the cross-validation splits.

II. Related Work

A. Surgical Workflow Analysis

Developing practical deep-learning techniques for surgical procedures aims to undergo a shift from recognizing activities at a coarse level to recognizing activities with finer granularity. Phase recognition is a representative task in analyzing coarse-level surgical workflow, where prevailing methods aim to facilitate spatial-temporal modeling from continuous surgical video frames. A CNN + RNN architecture is employed by Jin et al. to acquire complementary information on visual and

temporal features in an end-to-end manner [19]. Tecno utilizes causal dilated convolutions to enable efficient long-range temporal modeling for online surgery recognition [20]. Moreover, Yi et al. point out the limitations of end-to-end training Tecno and introduce a two-stage training method [21]. To address erroneous predictions from ambiguous frames, [22] develops a segment-attentive hierarchical consistency network for surgical phase recognition. Recently, Park et al. introduce a novel loss function that enhances phase prediction accuracy. This loss function effectively addresses undesirable phase transitions and prevents over-segmentations in the predictions [13].

In [6], the Cholec45 dataset with surgical triplet annotations is presented to facilitate the fine-grained surgical activity analysis. Tripnet leverages instrument appearance cues to identify the verb and target in triplets, then projects the recognized components to learn their association [7]. Building upon this work, Nwoye et al. utilize Transformers to recognize triplets directly from surgical videos, leveraging attention mechanism at individual action triplet components and their relationships [8]. Moreover, some studies focus on learning the interaction relations among surgical triplets [14], [23], [24]. Wang et al. adopt GCN to model triplet interaction relation along the temporal dimension [23], whereas MURPHY is developed in [24] to improve feature representation by incorporating inter- and intra-relational information. However, these approaches overlook the class imbalance characteristic in surgical triplets. To address this challenge, Xi et al. integrate GCN with a classification forest to better balance triplet classes in a long-tail distribution [25]. The class imbalance characteristic is also taken into account in our method MT4MTL-KD. In contrast to [25], we propose to mitigate the negative impact of long-tail distribution in task association learning.

B. Multi-Label Image Classification

Recognizing surgical triplets is a multi-label image classification (MLIC) problem because there may be more than one surgical triplet in a single video frame. Existing MLIC methods primarily focus on addressing two key concerns: locating regions of interest and modeling label correlations. To address the first concern, early methods detect object proposals first and treat each of them as a single-label classification task [26], [27]. Wang et al. employ a spatial Transformer layer and a Long Short-Term Memory (LSTM) for locating regions with objects and subsequent score prediction, respectively [28]. Zhou et al. utilize an encoder-decoder structure to extract salient object features and align them with label concepts [29]. The other category of methods aims at modeling label correlations as prior knowledge to improve the subsequent classification. A probabilistic graphical model is employed in [30] to represent label dependencies by considering input features and labels. Hu et al. develop a generic structured CNN that leverages diverse label relations by modeling categorization at different concept layers with varied label information [31]. Recently, Chen et al. leverage GCN in learning semantic-label representations [32], [33]. They construct a graph using class-aware maps and enforce graph constraints through co-occurrence labeling. Moreover, an enhanced GCN + CNN framework is developed in [34] to capture label relationships via exploiting statistical co-occurrence information. Witnessing the success of the Transformer-based model in many computer vision tasks, Q2L [35] and TSFormer [36] employ Transformers to learn both discriminative features and their correlations, achieving comparable results. In this paper, we integrate the strengths of CNN-based and Transformer-based models via knowledge distillation and enhance task association learning to address the two concerns, respectively.

C. Knowledge Distillation

Knowledge Distillation (KD) is a technique of compressing large models into smaller ones while maintaining high performance with minimal loss, primarily used for mobile model compression [37]. KD is widely applied in the fields of computer vision and natural language processing [38]. In addition to its classical use for model compression, KD has recently been employed in MLIC to enhance performance [39], [40], [41]. Song et al. utilize distillation to mitigate the model bias towards challenging categories [39], while Xu et al. decompose the MLIC task into simpler sub-tasks and employ distillation to learn a global ensemble of full categories [40]. Liu et al. apply a distillation framework between two different tasks to facilitate cross-task knowledge transfer, *i.e.*, from weakly-supervised detection to MLIC [41]. In this paper, we leverage distillation to improve triplet recognition. Inspired from [17] and [40], we decompose triplet recognition into simpler sub-tasks and adopt multiple teachers.

III. Proposed Method

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 $\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$ is a RGB frame with height H and width W. Since each frame $\mathbf{x}_t$ may contain multiple surgical triplet classes, we use $\mathbf{y}_t = \{y_{t,k} \in \{0, 1\}\}_{k=1}^{K}$ to represent its ground-truth label, where K indicates the number of triplet classes. The objective of surgical triplet recognition is to train a multi-label classification model with minimized prediction error on the test set $\{\overline{\mathbf{X}}, \overline{\mathbf{Y}}\}$.

A. Overall Framework

Fig. 2 illustrates the workflow of our proposed MT4MTL-KD. MT4MTL-KD adopts a multi-teacher distillation framework to mitigate the adverse effects caused by spurious task association, while employing heterogeneous models as the teacher and student to integrate local and global context modeling. By feeding the video sequence and labels with respect to triplet recognition sub-tasks (*i.e.*, instrument (I), verb (V), and target (T) classifications), each teacher is trained to solve an independent sub-task. After that, task-specific knowledge from each teacher is transferred to the student model during multi-task student training. Specifically, MT4MTL-KD incorporates feature-level distillation to integrate complementary information between teacher and student models. A FAM module is developed to efficiently align the features obtained from multiple teachers with the features yielded during multi-task learning. Furthermore, we introduce

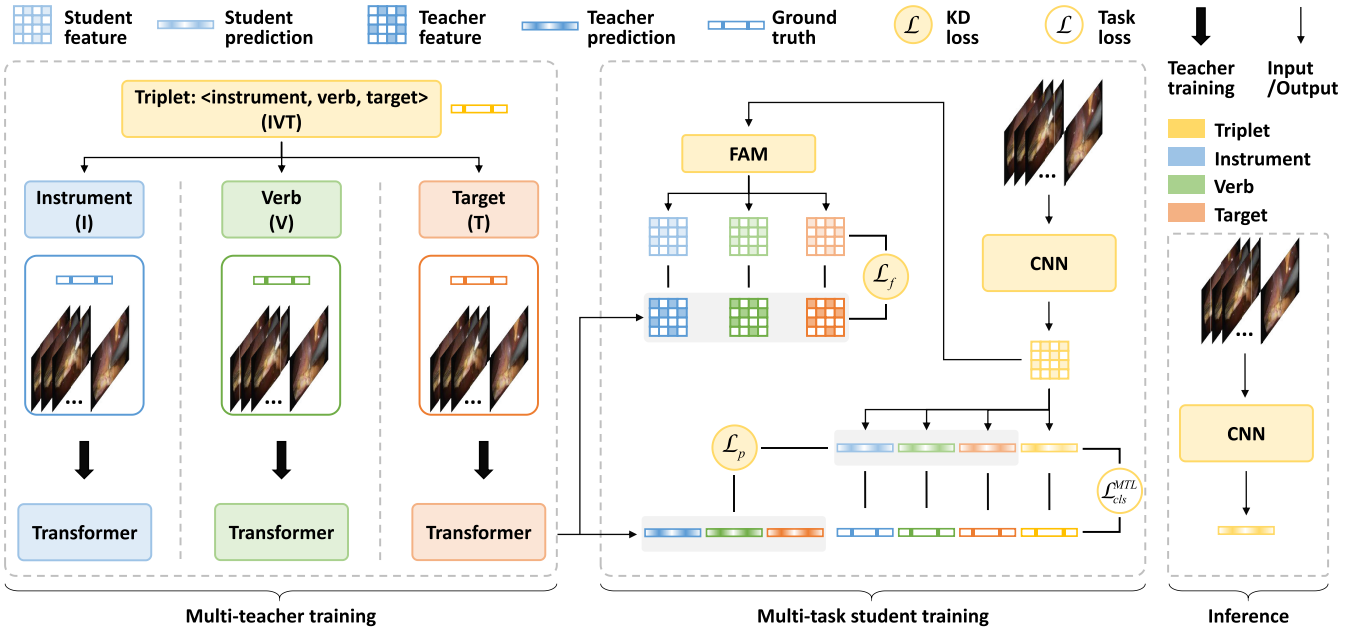

Fig. 2. The overall framework of the proposed MT4MTL-KD. From left to right: multi-teacher training, multi-task student training, and inference. Initially, MT4MTL-KD decomposes the triplet labels into three sub-task labels and performs multi-teacher training on three Transformer-based models. Then feature-level and prediction-level distillation is conducted during multi-task student training. During model inference, the student model makes triplet classification predictions without reliance on teacher models.

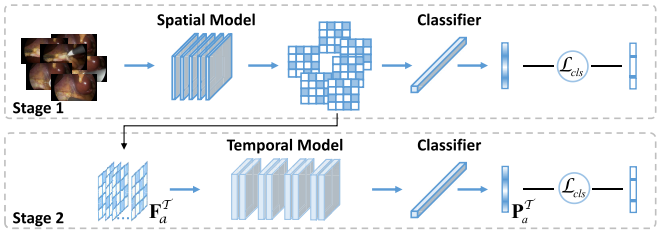

Fig. 3. Multi-teacher training for sub-task *a*. The spatial model and temporal model are trained in separate stages.

prediction-level distillation to enhance task association learning. This involves providing the soft supervision information from each teacher to its respective sub-task classifier. During inference, the student model operates independently without the reliance on teacher models, enabling a fast-response triplet recognition. Below we explain MT4MTL-KD in detail.

B. Multi-Teacher Training

As shown in Fig. 3, each teacher model is trained in a two-stage manner, where the spatial and temporal models are trained in different stages [42], [43], [44]. We denote the sub-task addressed by the teacher model as $a \in A$, where $A = \{I, V, T\}$. In the first stage, we feed input frames $\mathbf{X}$ into the spatial model $\mathcal{G}_a(\cdot)$, followed by a classifier $\mathcal{C}_{g,a}(\cdot)$. The training objective for the first stage adopts the classification loss $\mathcal{L}^a_{cls}$ and is calculated based on output predictions $\sigma(\mathcal{C}_{g,a}(\mathcal{G}_a(\mathbf{X})))$, where σ is a sigmoid function. In the second stage, the extracted spatial features are sequentially fed into two components: a temporal model and an additional classifier. We use $\mathcal{H}_a(\cdot)$ to denote the temporal model and $\mathcal{C}_{h,a}(\cdot)$ to represent the classifier. The obtained predictions can be expressed as $\sigma(\mathcal{C}_{h,a}(\mathcal{H}_a(\mathcal{G}_a(\mathbf{X}))))$. We continue to use the classification loss $\mathcal{L}^a_{cls}$ as the training objective of the second stage. In specific, we employ the weighted cross-entropy loss as $\mathcal{L}^a_{cls}$, *i.e.*,

$$\mathcal{L}^a_{cls} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K_a} W^a_k\, y^a_{t,k} \log \hat{y}^a_{t,k} + \left(1 - y^a_{t,k}\right)\log\left(1 - \hat{y}^a_{t,k}\right), \tag{1}$$

where K_a represents the number of classes in terms of sub-task a; W^a_k is a weight for class balancing; $\hat{y}^a_{t,k}$ and $y^a_{t,k}$ are the prediction and ground-truth of the k-th class of sub-task a in the video frame $\mathbf{x}_t$, respectively.

After optimizing the teacher models, we obtain the spatial feature $\mathbf{F}^{\mathcal{T}}_a$ and the prediction $\mathbf{P}^{\mathcal{T}}_a$ concerning sub-task a as

$$\begin{aligned} \mathbf{F}^{\mathcal{T}}_a &= \mathcal{G}^*_a(\mathbf{X}), \\ \mathbf{P}^{\mathcal{T}}_a &= \mathcal{C}^*_{h,a}\left(\mathcal{H}^*_a\left(\mathbf{F}^{\mathcal{T}}_a\right)\right), \end{aligned} \tag{2}$$

where $\mathbf{F}^{\mathcal{T}}_a \in \mathbb{R}^{T \times D_{\mathcal{T}}}$, $\mathbf{P}^{\mathcal{T}}_a \in \mathbb{R}^{T \times K_a}$; $D_{\mathcal{T}}$ denotes the extracted feature dimension in teacher models and the superscript * represents the model with respective optimized weights.

C. Multi-Task Student Training With Distillation

1) Student Model Structure: As shown in Fig. 4, the student training also consists of two stages. It employs the multi-task learning mechanism with four classifiers, addressing three sub-tasks and a triplet recognition task. In the first stage, the feature-level and prediction-level distillation techniques are incorporated to optimize the spatial model $\mathcal{G}(\cdot)$ and four classifiers $\mathcal{C}_g(\cdot)$. The training loss is formulated as

$$\mathcal{L} = \mathcal{L}^{MTL}_{cls} + \alpha\mathcal{L}_f + \beta\mathcal{L}_p, \tag{3}$$

Algorithm 1 Training and Inference of MT4MTL-KD

TRAINING

Input: training video squence $\mathbf{X}$ and labels $\mathbf{Y} = \{\mathbf{Y}_{IVT}, \mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T\}$.

Output: the optimized student model $\mathcal{G}^*$, $\mathcal{H}^*$ and $\mathcal{C}_h^*$.

Initialize: pre-trained models and training hyper-parameters.

Multi-Teacher Training

for a in $A = \{I, V, T\}$ **do**

$\mathcal{G}_a^*, \mathcal{C}_{g,a}^* \leftarrow \arg\min_{\mathcal{G}_a, \mathcal{C}_{g,a}} \mathcal{L}_{cls}^a(\sigma(\mathcal{C}_{g,a}(\mathcal{G}_a(\mathbf{X}))), \mathbf{Y}_a)$;

$\mathcal{H}_a^*, \mathcal{C}_{h,a}^* \leftarrow \arg\min_{\mathcal{H}_a, \mathcal{C}_{h,a}} \mathcal{L}_{cls}^a(\sigma(\mathcal{C}_{h,a}(\mathcal{H}_a(\mathcal{G}_a^*(\mathbf{X})))), \mathbf{Y}_a)$.

end for

Obtain the features $\mathbf{F}^{\mathcal{T}}$ and predictions $\mathbf{P}^{\mathcal{T}}$ by Eq. 2.

Multi-Task Student Training

Compute the classification loss $\mathcal{L}_{cls}^{MTL}(\sigma(\mathcal{C}_g(\mathcal{G}(\mathbf{X}))), \mathbf{Y})$;

Compute the distillation loss $\mathcal{L}_f(\text{FAM}(\mathcal{G}(\mathbf{X})), \mathbf{F}^{\mathcal{T}}) + \mathcal{L}_p(\mathcal{C}_g(\mathcal{G}(\mathbf{X})), \mathbf{P}^{\mathcal{T}})$;

$\mathcal{G}^*, \mathcal{C}_g^* \leftarrow \arg\min_{\mathcal{G}, \mathcal{C}_g}(\mathcal{L}_{cls}^{MTL} + \alpha\mathcal{L}_f + \beta\mathcal{L}_p)$;

$\mathcal{H}^*, \mathcal{C}_h^* \leftarrow \arg\min_{\mathcal{H}, \mathcal{C}_h} \mathcal{L}_{cls}^{MTL}(\sigma(\mathcal{C}_h(\mathcal{H}(\mathcal{G}^*(\mathbf{X})))), \mathbf{Y})$.

Return: the optimized student model $\mathcal{G}^*$, $\mathcal{H}^*$ and $\mathcal{C}_h^*$.

INFERENCE

Input: testing video sequence $\overline{\mathbf{X}}$.

Output: frame-wise predictions $\hat{\mathbf{Y}}$.

Initialize: the optimized student model.

Forward pass the student model to get $\hat{\mathbf{Y}} = \sigma\left(\mathcal{C}_h^*\left(\mathcal{H}^*\left(\mathcal{G}^*\left(\overline{\mathbf{X}}\right)\right)\right)\right)$.

Return: $\hat{\mathbf{Y}}$.

where $\mathcal{L}_{cls}^{MTL}$ denotes the multi-task learning loss; $\mathcal{L}_f$ and $\mathcal{L}_p$ denotes the feature-level and prediction-level distillation losses, respectively. α and β are coefficients used to balance the loss terms. During the second stage, the temporal model $\mathcal{H}(\cdot)$ and four classifiers $\mathcal{C}_h(\cdot)$ are trained to refine the predictions by capturing the temporal dynamics. The training process only employs $\mathcal{L}_{cls}^{MTL}$ as the loss function, which is the sum of the classification losses concerning the three sub-tasks and the triplet recognition task, represented as

$$\mathcal{L}_{cls}^{MTL} = \sum_{a \in A} \mathcal{L}_{cls}^a, \quad A = \{I, V, T, IVT\}. \tag{4}$$

During inference, we solely use the triplet recognition classifier while discarding the other three classifiers. The overall training and inference procedures of the proposed framework are summarized in Algorithm 1.

2) Feature-Level and Prediction-Level Distillation: We denote the semantic features provided by the trained teachers as $\{\mathbf{F}_I^{\mathcal{T}}, \mathbf{F}_V^{\mathcal{T}}, \mathbf{F}_T^{\mathcal{T}}\}$ and the respective features derived from FAM as $\{\mathbf{F}_I^{\mathcal{S}}, \mathbf{F}_V^{\mathcal{S}}, \mathbf{F}_T^{\mathcal{S}}\}$. The feature-level distillation process involves quantifying the discrepancies in features between the respective teachers and students using the mean squared error (MSE) as follows,

$$\mathcal{L}_f = \frac{1}{T|A|} \sum_{t=1}^{T} \sum_{a \in A} \left\| \mathbf{F}_{a,t}^{\mathcal{S}} - \mathbf{F}_{a,t}^{\mathcal{T}} \right\|_2^2, \tag{5}$$

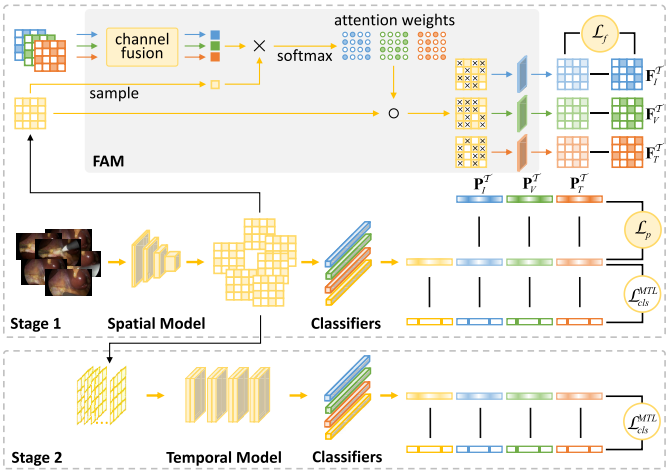

Fig. 4. The workflow of multi-task student training.

where $\mathbf{F}_{a,t}$ denotes the extracted feature for the t-th frame, $A = \{I, V, T\}$ and $|A|$ is the number of sub-tasks. In addition, we perform prediction-level distillation by measuring the differences in the distribution of soft predictions between the teacher and student, using the Kullback-Leibler (KL) divergence. To obtain the soft predictions $\widetilde{\mathbf{P}}$, we apply the softmax function with a temperature scale τ to the output predictions $\mathbf{P}$, *i.e.*,

$$\widetilde{\mathbf{P}}_t = \frac{\exp(\mathbf{P}_t/\tau)}{\sum_{k=1}^{K} \exp(\mathbf{P}_{t,k}/\tau)}, \tag{6}$$

where $\mathbf{P}_t$ represents the predictions in the t-th frame and $\mathbf{P}_{t,k}$ denotes the prediction of the k-th class in the t-th frame. The prediction-level distillation loss is formulated as

$$\mathcal{L}_p = \frac{1}{T|A|} \sum_{t=1}^{T} \sum_{a \in A} \widetilde{\mathbf{P}}_{a,t}^{\mathcal{T}} \log(\widetilde{\mathbf{P}}_{a,t}^{\mathcal{T}} / \widetilde{\mathbf{P}}_{a,t}^{\mathcal{S}}). \tag{7}$$

3) Feature Attention Module: During feature-level distillation, we propose a feature attention module (FAM) to ensure that each teacher effectively guides its highly correlated student channels. As depicted in Fig. 4, this is achieved by assigning attention weights to different channels of the student feature. To acquire these weights, FAM undertakes two steps: 1) conduct channel fusion on each teacher feature to obtain its overall semantics, and 2) calculate the correlation between each student channel and the overall semantics of different teachers. Using these weights, FAM enables each student channel to learn more relevant features from three different teacher features.

For each sub-task $a \in \{I, V, T\}$, the teacher feature is denoted as $\mathbf{F}_a^{\mathcal{T}} \in \mathbb{R}^{D_{\mathcal{T}}}$, where $D_{\mathcal{T}}$ is the teacher dimension. Firstly, we use a 1×1 convolutional layer to convert the teacher features from $D_{\mathcal{T}}$-dimensional space to $D_{\mathcal{S}}$-dimensional space, *i.e.*, $\widetilde{\mathbf{F}}_a^{\mathcal{T}} = \texttt{Conv}(\mathbf{F}_a^{\mathcal{T}}; D_{\mathcal{S}})$. Subsequently, we define $f_a \in \mathbb{R}$ as the overall semantics across all channels of $\widetilde{\mathbf{F}}_a^{\mathcal{T}}$, *i.e.*,

$$f_a = \frac{1}{\sqrt{D_{\mathcal{S}}}} \sum_{i=1}^{D_{\mathcal{S}}} \widetilde{\mathbf{F}}_a^{\mathcal{T}}(i), \tag{8}$$

where i is the channel index. For the student feature $\mathbf{F}^{\mathcal{S}} \in \mathbb{R}^{D_{\mathcal{S}}}$, the correlation between its i-th channel and the overall semantics of each teacher can be measured by $f_i \times f_a$. By applying the softmax function on different teachers, the attention weight $w_{a,i}$ can be determined as

$$w_{a,i} = \frac{\exp\left(f_i \times f_a\right)}{\sum_{j \in A} \exp\left(f_i \times f_j\right)}, \tag{9}$$

where $A = \{I, V, T\}$ denotes the set of sub-tasks. Using these attention weights, FAM adaptively scales each channel of the student feature, *i.e.*,

$$\widetilde{\mathbf{F}}_a^{\mathcal{S}} = \mathbf{w}_a \circ \mathbf{F}^{\mathcal{S}}, \tag{10}$$

where $\mathbf{w}_a = (w_{a,1}, w_{a,2}, \ldots, w_{a,D_{\mathcal{S}}})^{\intercal}$, and $\circ$ denotes the Hadamard product. Finally, the task-specific feature $\mathbf{F}_a^{\mathcal{S}}$ is obtained by a 1×1 convolutional layer, *i.e.*, $\mathbf{F}_a^{\mathcal{S}} = \texttt{Conv}(\widetilde{\mathbf{F}}_a^{\mathcal{S}}; D_{\mathcal{T}})$.

IV. Experiments

A. Implementation Details

MT4MTL-KD is developed using PyTorch [45], and all experiments are conducted on a single NVIDIA Tesla V100 32GB GPU. The source code of MT4MTL-KD[1] is available. Following [8], the input video frames for both teacher and student training undergo light data augmentations, including resizing images to 256×448, 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.

For each teacher model, we use the Q2L-SwinL [35] pre-trained on ImageNet-22k as the spatial model, with an output feature dimension of 1536. It is trained for 100 epochs, using a batch size of 16. We apply a step-wise warm-up policy, initializing the learning rate at 1e-5 and subsequently decaying it exponentially with a decay factor of $\gamma = 0.99$ after 58 epochs. MS-TCT [46] is utilized as the temporal model, with an input sequence length of 256. The training process consists of 2000 epochs with a batch size of 32, initiating with a learning rate of 1e-2, which is exponentially decayed by $\gamma = 0.999$ after 500 epochs. During multi-task student training, we leverage the ImageNet pre-trained ResNet-18 [10] model as our spatial backbone, with a 512-sized feature vector output. It is trained for 200 epochs with a batch size of 8, commencing with a learning rate of 1e-3, following the identical decay policy employed in the teacher models. 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 [22]. We train the temporal model for 1000 epochs with an initial learning rate of 1e-2, which decays exponentially after 200 epochs with the decay rate $\gamma = 0.99$.

[1]https://github.com/CIAM-Group/ComputerVision_Codes/tree/main/MT4MTLKD

B. Datasets and Evaluation Metrics

1) Datasets: This paper employs a public challenge dataset from CholecTriplet 2021 [6], 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 assess the efficacy of the proposed framework, we use the official 5-fold cross-validation split and CholecTriplet split [47] for model evaluation. In the cross-validation split, the dataset is divided into three subsets: a training set consisting of 31 videos, a validation set consisting of 5 videos, and a testing set consisting of 9 videos. Additionally, the CholecTriplet split consists of 50 videos, where 5 videos are distinct from those in CholecT45. This split follows a 40-5-5 division for training, validation, and testing, respectively. In particular, we use fold 1 of the cross-validation split to conduct ablation experiments.

2) Evaluation Metrics: In this study, we evaluate the performance of different methods using the average precision (AP) metric, which has been commonly employed in previous works [7], [8], [14]. The AP metrics employed in this study 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 in this study is AP_{IVT}, which evaluates the recognition of the complete triplets. On the other hand, AP_{IV} and AP_{IT} reflect the performance of tool-tissue interaction recognition, while AP_I, AP_V, and AP_T assess the accuracy of identifying constituent components within triplets. To obtain predictions for each component and tool-tissue interaction, we apply a filtering algorithm, as described in [7], to the triplet predictions.

C. Comparison With the State-of-the-Arts

In this section, we compare our proposed approach with state-of-the-art triplet recognition methods (SOTAs), including Tripnet [7], Attention Tripnet [8], RDV [8], RiT [11], as well as the top three performing methods from the challenge (HFUT-MedIA, SIAT CAMI, and Trequartis) as reported in [6].

1) Component and Association AP: We implement our MT4MTL-KD with different teacher-student pairs and test three configurations: 1) "SwinL→Res18": we employ Q2L-SwinL as the teacher backbone and ResNet-18 as the student backbone; 2) "Res18→SwinL": we employ ResNet-18 as the teacher backbone and Q2L-SwinL as the student backbone; 3) "Ensemble": we ensemble the trained student models in 1) and 2) to obtain final predictions. To ensure fair comparisons with experiments conducted on the CholecTriplet split, the MT4MTL-KD implementations employ an additional ensemble of six models, consisting of five models trained on cross-validation split and one model trained on the CholecTriplet split. The ensemble implementation refers to averaging the sigmoid probabilities derived from the selected models.

Table I presents the mean and standard deviation results of three types of AP on both the cross-validation split and

TABLE I
BENCHMARK TRIPLET RECOGNITION AP (%) ON CHOLECT45 DATASET. **BOLD** = BEST SCORE. UNDERLINED = BEST SCORE IN THE HE STATE-OF-ART METHODS

Method	Component detection			Triplet association		AP_{IVT}
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	
	5-fold Cross-Validation Split					
Tripnet [7]	89.9±1.0	59.9±0.9	37.4±1.5	31.8±4.1	27.1±2.8	24.4±4.7
Attention Tripnet [8]	89.1±2.1	61.2±0.6	40.3±1.2	33.0±2.9	29.4±1.2	27.2±2.7
RDV [8]	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 [11]	88.6±2.6	64.0±2.5	43.4±1.4	38.3±3.5	36.9±1.0	29.7±2.6
SwinL→Res18 (***Ours***)	91.3±2.4	69.7±1.3	46.5±4.7	43.8±5.8	43.5±1.1	36.1±1.2
Res18→SwinL (***Ours***)	93.1±2.1	71.8±3.4	48.8±3.8	44.9±2.4	43.1±2.0	37.1±0.5
Ensemble (***Ours***)	**93.9±2.0**	**73.8±2.0**	**52.1±5.2**	**46.5±3.4**	**46.2±2.3**	**38.9±1.6**
	CholecTriplet Split					
Tripnet [7]	74.6	42.9	32.2	27.0	28.0	23.4
Attention Tripnet [8]	77.1	43.4	30.0	32.3	29.7	25.5
4^{th} RDV [8]	77.5	47.5	37.7	39.4	39.6	32.7
3^{rd} HFUT-MedIA [6]	77.1	46.7	37.8	33.1	35.9	32.9
2^{nd} SIAT CAMI [6]	82.1	51.5	45.5	37.1	43.1	35.8
1^{st} Trequartista [6]	79.9	52.9	46.4	39.0	41.9	38.1
SwinL→Res18 (***Ours***)	85.3	52.4	44.4	34.2	38.9	34.2
Res18→SwinL (***Ours***)	**86.8**	**57.5**	46.3	**39.5**	41.3	37.2
Ensemble (***Ours***)	86.7	56.2	**47.0**	38.5	**43.3**	**38.5**

the CholecTriplet split. The results indicate that Tripnet, the vanilla multi-task learning method in triplet recognition, achieves inferior results in comparison to the other methods. This emphasizes the significance of task association learning for accurate triplet recognition. Although RDV and RiT incorporate task association modeling through the self-attention mechanism, their performance is limited due to the issue of spurious task association during multi-task learning. In contrast, our MT4MTL-KD approach effectively mitigates the negative impact of spurious task association by employing knowledge distillation. Additionally, it leverages the strengths of both Transformer-based models and CNN-based models to facilitate efficient feature extraction. As a consequence of these crucial aspects, MT4MTL-KD consistently outperforms previous state-of-the-art methods. By employing the same student backbone as the SOTAs (SwinL→Res18), MT4MTL-KD achieves an improvement of 6.4% AP_{IVT} on cross-validation split. Furthermore, additional improvements of 1.0% and 2.8% in AP_{IVT} are obtained by implementing the "Res18→SwinL" and "Ensemble", respectively. Apart from AP_{IVT}, MT4MTL-KD consistently showcases the advancements in component detection and triplet association. Moreover, the evaluation results on the CholecTriplet split highlight the superior performance of our method, surpassing the top submissions in the CholecTriplet 2021 challenge by a minimum margin of 0.4% in terms of AP_{IVT}.

2) Class-Wise Performances on Component APs: To explore the benefits of MT4MTL-KD in addressing class imbalance issues, we present the class-wise results obtained from 5-fold cross-validation. Regarding instrument recognition, Table II reveals that *grasper* and *hook* continue to be the most accurately recognized, primarily attributed to their high occurrence frequencies within the dataset. Notably, our approach improves the recognition accuracy on all classes of instruments except *grasper*. Specifically, our approach demonstrates a noteworthy improvement (+3.8%) in the recognition of *scissors*. For verb recognition, Table III shows that the verbs *retract*

TABLE II
PER-CLASS INSTRUMENT PRESENCE DETECTION AP (%) ON CROSS-VALIDATION SPLITS

Classes	SOTAS				***Ours***
	Tripnet	Attention Tripnet	RDV	RiT	
grasper	96.5±0.4	96.4±0.7	**96.6±0.6**	-	93.8±1.6
bipolar	88.4±4.2	86.0±4.2	87.4±4.7	-	**91.1±4.6**
hook	97.5±1.6	97.1±1.3	97.4±1.5	-	**97.7±1.4**
scissors	80.3±6.0	79.6±8.4	78.4±5.4	-	**84.1±7.6**
clipper	91.2±3.9	90.1±3.9	90.9±3.8	-	**92.7±3.5**
irrigator	86.0±4.1	85.3±2.8	84.5±6.8	-	**88.5±4.8**
Mean	89.9±1.0	89.1±2.1	89.3±2.1	88.6±2.6	**91.3±2.4**

TABLE III
PER-CLASS VERB PRESENCE DETECTION AP (%) ON CROSS-VALIDATION SPLITS

Classes	SOTAS				***Ours***
	Tripnet	Attention Tripnet	RDV	RiT	
grasp	70.5±5.8	60.5±9.9	69.8±3.7	69.6±3.9	**78.6±3.1**
retract	**90.5±5.4**	84.0±9.8	89.7±7.2	89.3±9.0	89.1±5.4
dissect	93.0±2.8	86.5±9.9	93.2±3.9	92.6±2.2	**93.4±1.9**
coagulate	67.2±6.1	56.5±9.9	68.7±5.5	68.5±6.1	**72.5±5.2**
clip	85.4±6.4	67.8±9.8	85.5±3.7	87.3±5.3	**88.4±5.0**
cut	70.5±9.1	57.7±9.9	72.0±4.8	74.9±10.9	**81.7±8.4**
aspirate	60.7±9.2	47.1±9.9	57.8±9.9	64.9±7.8	**65.3±6.1**
irrigate	29.6±8.2	17.4±9.7	25.7±5.8	22.7±9.0	**29.9±6.5**
pack	32.1±9.9	25.8±9.9	31.2±9.9	43.2±16.2	**69.1±12.3**
null-verb	23.0±2.4	21.1±5.0	24.0±4.1	26.4±4.8	**28.8±3.0**
Mean	59.9±0.9	61.2±0.6	62.0±1.3	64.0±2.5	**69.7±1.3**

TABLE IV
PER-CLASS TARGET PRESENCE DETECTION AP (%) ON CROSS-VALIDATION SPLITS

Classes	SOTAS				*Ours*
	Tripnet	Attention Tripnet	RDV	RiT	
gallbladder	93.6±1.1	91.2±6.5	**93.7±1.2**	-	93.2±02.0
cystic-plate	11.6±3.9	10.1±2.0	11.0±3.5	-	**17.2±05.3**
cystic-duct	47.2±5.8	41.9±9.9	47.1±2.8	-	**52.5±01.9**
cystic-artery	31.9±3.7	29.6±9.7	31.2±2.2	-	**39.7±10.0**
cystic-pedicle	04.0±2.4	08.7±6.0	13.4±7.8	-	**15.1±09.8**
blood-vessel	08.4±5.6	15.6±9.9	06.7±6.2	-	**28.9±23.3**
fluid	58.4±9.2	48.9±9.9	58.0±9.9	-	**65.3±06.1**
abdominal-wall/cavity	30.0±4.6	20.4±9.9	25.9±7.2	-	**32.8±11.3**
liver	71.8±5.5	65.3±9.9	72.9±2.5	-	**73.6±06.1**
adhesion	04.2±0.3	13.9±3.3	07.2±0.5	-	**31.1±12.9**
omentum	46.7±8.6	44.4±9.9	48.0±9.9	-	**53.5±14.0**
peritoneum	17.7±5.5	24.1±9.9	26.6±4.5	-	**34.8±06.4**
gut	10.7±7.7	09.6±6.9	09.5±6.9	-	**20.4±05.5**
specimen-bag	85.8±2.5	70.1±9.9	84.4±1.2	-	**91.4±01.9**
null-target	22.8±2.3	21.1±5.4	23.5±4.1	-	**28.8±03.0**
Mean	37.4±1.4	40.3±1.2	40.0±1.4	43.4±1.4	**46.5±04.7**

and *dissect* are recognized with a high AP of approximately 90%. Moreover, our proposed approach consistently surpasses the previous SOTAs in the recognition of classes with low occurrence frequency. In particular, for the recognition of *pack*, which has only 196 instances in the dataset, our MT4MTL-KD achieves an AP of 69.1%, improving over 26.9% compared to the best SOTA result. A similar effect is also observed in the results of target recognition, which are reported in Table IV. The improvement in the recognition of structures with low occurrence frequency (*e.g.*, *adhesion*, *omentum*) is significantly larger than that of predominant classes (*e.g.*, *gallbladder*, *liver*). Specifically, our method achieves a significant improvement in *adhesion* recognition, boosting the AP from 13.9% to 31.1%. Overall, MT4MTL-KD demonstrates improvements across all three sub-tasks, with more significant enhancements observed in low occurrence frequency classes compared to predominant classes. We attribute this phenomenon to the effectiveness of distillation, which facilitates enhanced sub-task learning and mitigates the spurious task association noise.

D. Ablation Study

1) Ablation on KD Components: In this subsection, we conduct ablation experiments to validate the effectiveness of different key components within our proposed knowledge distillation method: 1) Baseline-Res18: we adopt a vanilla multi-task approach without distillation. It utilizes ResNet-18 and TCN as the spatial and temporal models, respectively; 2) $+\mathcal{L}_f$: we solely use feature-level distillation to train the student model; 3) $+\mathcal{L}_p$: we exclusively employ prediction-level distillation to train the student model; 4) - 6) $+\mathcal{L}_p + \mathcal{L}_f$-AM/DM/FAM: we incorporate both prediction-level and feature-level distillation techniques, where the latter is executed through three different feature alignment approaches. Fig. 5 illustrates the framework of two alignment approaches aside from FAM. The aggregation module (AM) applies independent convolutions to each teacher feature. The obtained features are subsequently fused via an additional convolutional layer to generate the aligned student feature. The decomposition module (DM) directly feeds the student feature to three task-specific convolutions for feature alignment. According to the results in Table V, we observe that the feature-level and prediction-level distillation techniques lead to AP_{IVT} improvements of 3.69% and 2.10%, respectively. Furthermore, among the three alignment strategies for feature-level distillation, AM and DM show marginal improvements in AP_{IVT}, indicating that simplistic approaches struggle to facilitate efficient knowledge transfer. In contrast, implementing our proposed FAM module results in a substantial improvement of 2.8%, which emphasizes the significance of FAM in a multi-teacher knowledge transfer scenario.

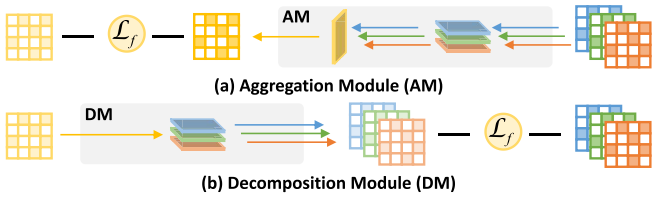

Fig. 5. Two alignment strategies for feature-level distillation.

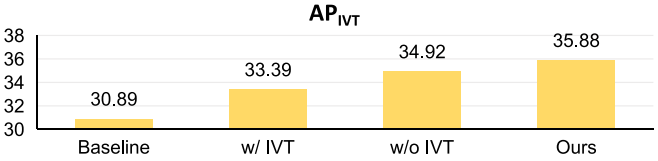

Fig. 6. Ablation on teacher training strategies.

TABLE V
ABLATION ON KD COMPONENTS

Method	Results			
	AP_I	AP_V	AP_T	$\boldsymbol{AP_{IVT}}$
Baseline-Res18	85.58	63.68	45.99	30.89
+ $\mathcal{L}_f$	89.30	68.10	50.10	34.58
+ $\mathcal{L}_p$	**90.49**	68.30	48.93	33.01
+ $\mathcal{L}_p$ + $\mathcal{L}_f$-AM	89.77	65.40	49.01	33.02
+ $\mathcal{L}_p$ + $\mathcal{L}_f$-DM	90.34	68.05	50.00	33.15
+ $\mathcal{L}_p$ + $\mathcal{L}_f$-FAM (***Ours***)	89.87	**70.60**	**50.20**	**35.88**

2) Ablation on Teacher Training Strategies: In this subsection, we conduct an ablation study on the effect of our multi-teacher models. As shown in Fig. 8, we compare the results of the baseline (Section IV-D.1) and three strategies of the teacher model training: 1) "w/ IVT": we train a shared teacher backbone that simultaneously resolves the triplet task and its sub-tasks; 2) "w/o IVT": we train a shared teacher backbone solely addressing three sub-tasks; 3) "Ours": we train three independent teacher models with respect to three sub-tasks. As shown in Fig. 8, by leveraging the teacher semantics from four tasks, the performance of the student model ("w/ IVT") is improved by 2.5% in AP_{IVT}. Moreover, removing the triplet task from the teacher model ("w/o IVT") demonstrates an additional performance gain of 1.5%. This indicates that less imbalanced sub-tasks can facilitate triplet recognition and mitigate the spurious task association. In comparison to the other two strategies, our approach utilizing multi-teacher

TABLE VI
ABLATION ON TEACHER MODELS

Teacher Models			Number	Results			
I	V	T		AP_I	AP_V	AP_T	AP_{IVT}
✗	✗	✗	0	85.58	63.68	45.99	30.89
✓	✗	✗	1	88.33	67.70	49.50	32.67
✗	✓	✗		90.15	**70.89**	51.35	33.53
✗	✗	✓		88.44	67.13	**53.53**	31.58
✗	✓	✓	2	88.69	67.98	53.03	34.85
✓	✗	✓		**90.23**	69.48	51.87	34.52
✓	✓	✗		89.63	67.06	51.86	35.01
✓	✓	✓	3	89.87	70.60	50.20	**35.88**

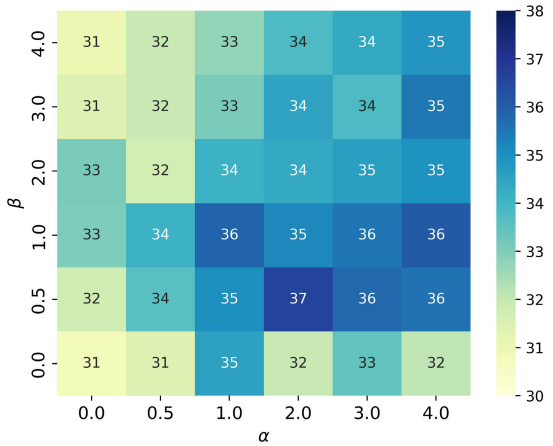

Fig. 7. Sensitivity study on distilling loss trade-off.

models achieves the highest performance with an AP_{IVT} of 35.88% for triplet recognition. This suggests that while the multi-task training regime in the student model promotes task association, it is not the optimal choice for teacher training within the distillation framework.

3) Ablation on Teacher Models: To investigate the contribution of each teacher model, we test the distillation performance over different numbers and combinations of teacher models. The results given in Table VI show that incorporating any teacher model can lead to performance improvement, and more teacher models result in more significant improvements. Specifically, as the number of teacher models increases from one to three, the average AP_{IVT} score increases from 32.59% to 35.88%. As for the impact of different teacher models on distillation performance, we can observe that the verb teacher contributes the most, while the target teacher contributes the least. In this triple recognition, the target recognition is more challenging than the other two, resulting in the target teacher's poor performance and thus leading to its less contribution in KD.

E. Sensitivity Study

1) Distillation Loss Trade-Off: In Eq. 3, α and β are used to balance the contributions of the feature-level and prediction-level distillation losses. To study the impact of different loss combinations on distillation performance, we test the performance of MT4MTL-KD using different α and β settings. All 25 combinations with $\alpha, \beta \in \{0, 0.5, 1, 2, 3, 4\}$ are examined in fold 1 of the cross-validation split. Fig. 7 shows the AP_{IVT} values achieved under these combinations. We can find that MT4MTL-KD consistently performs well with $\alpha \in [1.0, 4.0]$ and $\beta \in [0.5, 2.0]$. However, when the value of β is greater than 2, the performance of MT4MTL-KD drops significantly. It causes overemphasis on prediction-level distillation, making the student model prioritize sub-task learning over triplet recognition. Based on these experimental results, this work adopts $\alpha = \beta = 1$ for simplicity.

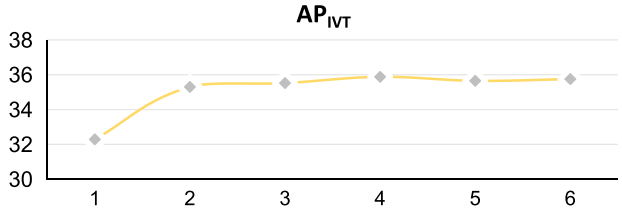

Fig. 8. Sensitivity study on distilling temperature.

TABLE VII
KD PERFORMANCE ON DIFFERENT TEACHER-STUDENT COMBINATIONS

Method	Results			
	AP_I	AP_V	AP_T	$\boldsymbol{AP_{IVT}}$ (Δ)
Baseline-Res18	85.58	63.68	45.99	30.89
Res18→Res18	90.20	65.22	50.32	33.19 (+2.30)
Res50→Res18	90.27	68.04	49.83	33.82 (+2.93)
SwinT→Res18	90.66	71.96	52.48	35.48 (+4.59)
SwinL→Res18	89.87	70.60	50.20	35.88 (+4.99)
Baseline-SwinL	87.36	71.54	51.54	33.55
SwinT→SwinL	89.95	71.83	54.57	34.45 (+0.90)
SwinL→SwinL	89.77	70.74	51.84	34.57 (+1.02)
Res18→SwinL	90.26	72.21	**54.82**	37.05 (+3.50)
Res50→SwinL	90.03	71.22	54.42	**37.25** (+3.70)
Baseline-Res50	86.19	66.66	45.10	28.01
SwinT→Res50	**92.07**	69.85	49.03	34.12 (+6.11)
SwinL→Res50	90.41	71.20	49.55	35.55 (**+7.54**)
Baseline-SwinT	88.01	69.17	53.13	32.30
Res18→SwinT	91.20	**73.44**	51.04	35.27 (+2.97)
Res50→SwinT	90.58	69.34	51.92	36.83 (+4.53)

2) Distilling Temperature: We further inspect the performance variation according to the distilling temperature $\tau \in \{1, 2, 3, 4, 5, 6\}$. This hyper-parameter controls the degree of softness across class predictions. As shown in Fig. 8, incorporating $\tau > 1$ in prediction-level distillation leads to a significant enhancement in AP_{IVT} compared to $\tau = 1$. By applying extra softening to teacher predictions, the predicted probability gap between plausible classes and the most confident class is reduced. This reinforces the influence of supplementary supervision information contained in teacher knowledge, ultimately improving triplet recognition. Furthermore, we observe the robustness over the range of $\tau \in [2, 6]$ and select the optimal setting, *i.e.*, $\tau = 4$, for our MT4MTL-KD approach.

V. DISCUSSION

A. Analysis of Different Teacher-Student Combinations

To investigate the performance of MTMTK-KD over different teacher-student combinations, we employ four different baseline models, denoted as "Baseline-Res18/50" and

TABLE VIII
MODEL COMPLEXITY OF DIFFERENT KD COMBINATIONS."TIME" IS CALCULATED FROM THE FIRST EPOCH TO THE POINT WHEN THE BEST MODEL IS SAVED."FLOPS" ARE MEASURED BY A SINGLE FRAME, WHEREAS "LATENCY" IS DETERMINED USING 200 FRAMES. †DENOTES THE RECOMMENDED MODELS, WHILE * IS OUR REPRODUCED RESULTS

Method	Time (GPU Hours)	Memory (GB)	# Params (M)	FLOPs (G)	Latency (ms)	AP_{IVT}
RiT	22.93	10.60	17.12	5.05	35.24	29.70 (28.45*)
Baseline-Res18	14.26	3.17	55.88	4.22	14.00	30.89
SwinL→Res18 (***Ours***)	19.47 × 3 + 13.73	23.74				35.88
SwinT→Res18 †	7.74 × 3 + 13.73	6.06	55.88	4.22	14.00	35.48
SwinT→Res50	7.74 × 3 + 15.02	6.06	69.77	9.50	18.68	34.12
SwinL→Res50	19.47 × 3 + 15.02	23.74				35.55
Res18→SwinT	24.36 × 3 + 15.61	6.06	117.19	7.58	22.27	34.12
Res50→SwinT †	24.71 × 3 + 15.61	6.06				36.83
Res18→SwinL	24.36 × 3 + 21.47	24.95	345.66	109.33	32.20	37.05
Res50→SwinL †	24.71 × 3 + 21.47	24.95				37.25

"Baseline-SwinT/L". "Baseline-Res18/50" utilizes ResNet-18 or ResNet-50 as the backbone, whereas "Baseline-SwinT/L" adopts Q2L-SwinT or Q2L-SwinL. The teacher-student pair is represented by "A→B", where A and B denote the teacher and student models, respectively. The experimental results are provided in Table VII. From these results, we can draw the following conclusions.

1) ***The effectiveness of MT4MTL-KD is observed across different teacher-student pairs.*** The experimental results indicate that knowledge distillation works consistently well with both balanced and unbalanced teacher-student pairs. For example, "SwinT↔Res50" and "SwinL↔Res18" can bring average performance improvements of 5.32% and 4.25%, respectively. Moreover, we find that the heterogeneous teacher-student pairs (*i.e.*, "Transformer↔CNN") always outperform the homogeneous teacher-student pairs (*i.e.*, "Transformer↔Transformer" or "CNN↔CNN").
2) ***Larger teacher models can contribute to more significant performance improvements.*** The experimental results in Table VII indicate that a more significant performance improvement can be achieved when the teacher adopts a larger model. For example, "SwinT→Res18" obtains an average performance gain of 4.59%, while "SwinL→Res18" yields 4.99%.
3) ***Transformer-based students surpass CNN-based students.*** The final performance depends heavily on the choice of the student model. "Baseline-SwinT" outperforms "Baseline-Res50" and "Baseline-Res18". Although "SwinL→Res18" can achieve a larger distillation improvement than "Res18→SwinL" (4.99 compared to 3.50), its final performance is worse than that of the latter due to the insufficient capability of the former's student model.
4) ***Deeper student models cannot achieve significant improvements.*** "Res50→SwinL" only achieves 0.42% improvement over "Res50→SwinT", while the performance of "SwinL→Res50" is even slightly worse than "SwinL→Res18". This may be due to the fact that the training data is class unbalanced, which can easily lead to overfitting of the minority classes [48]. When using a deeper model, this issue becomes more severe, and the improvement of MT4MTL-KD is limited.

B. Analysis of Model Complexity

Table VIII presents the model complexity of the SOTA method (RiT), baseline, and our KD methods. "Time" represents the cumulative training time, while "Memory" denotes the maximum value of the required GPU memory. As for the metrics related to the inference costs, *i.e.*, "# Params", "FLOPs" and "Latency", we only consider the costs of the student models for our methods. Since our student model has the same model structure as the baseline, their inference costs are the same. The results reveal that although our method requires larger memory than RiT, it is more computationally efficient than RiT in inference. For instance, "SwinL→Res18" saves 0.83 G FLOPs and has less than half the latency of RiT. In addition, our MT4MTL-KD allows for flexible teacher-student combinations that can cope with various requirements in practical deployment. As shown in Table VIII, "SwinT→Res18" has the smallest model cost, whereas "Res50→SwinL" can achieve the best AP_{IVT}. If a balance between cost and performance is required, "Res50→SwinT" is recommended.

C. Effects of Ensemble Models

To further boost the performance of MT4MTL-KD, we ensemble different types of student models. In specific, we average the video-level prediction scores obtained from different student models in Section V-A, and all the ensemble operations are implemented in the inference stage. As shown in Table IX, the results demonstrate two significant observations regarding the impact of model ensembling. First, compared with the homogeneous models (1-st and 2-nd rows), the ensemble of heterogenous models (3-rd row) exhibits a larger improvement. Second, increasing the number of ensemble models can further boost the performance. The best result is achieved by employing four distilled models, yielding an impressive AP_{IVT} of 40.11%. This demonstrates

TABLE IX
ENSEMBLE KNOWLEDGE DISTILLATION PERFORMANCE

Method	Models	Results			
		AP_I	AP_V	AP_T	$\boldsymbol{AP_{IVT}}$
Res18→Res18 + SwinL→Res18	2	91.53	69.67	51.93	37.33 (+1.5)
SwinL→SwinL + Res18→SwinL	2	90.79	73.40	55.18	37.60 (+0.6)
SwinL→Res18 + Res18→SwinL	2	91.45	72.90	53.99	39.01 (+2.0)
Res18→Res18 + SwinL→Res18 + Res18→SwinL	3	92.08	72.32	55.03	39.29 (+2.2)
Res18→Res18 + SwinL→Res18 + SwinL→SwinL + Res18→SwinL	4	**92.19**	**73.66**	**55.73**	**40.11 (+3.1)**

an improvement of 3.1% compared to the best individual constituent model. These experiments indicate that additional knowledge within the dataset might not be fully captured by our KD paradigm. This observation highlights the potential of MT4MTL-KD to achieve further improvements as more knowledge from the dataset is utilized.

VI. CONCLUSION

This paper proposes a multi-teacher knowledge distillation framework for multi-task triplet learning (MT4MTL-KD). MT4MTL-KD leverages the multi-teacher distillation framework to mitigate the adverse effects caused by spurious task association and employs heterogeneous models as the teacher and student. Specifically, our teacher models provide different context information and additional soft supervision with respect to triplet sub-tasks. In addition, the feature-level and prediction-level distillation are integrated into MT4MTL-KD to enhance feature extraction and task association learning, respectively. In feature-level distillation, a novel feature attention module (FAM) is proposed to align the semantic knowledge between the triplet task and its sub-tasks. MT4MTL-KD achieves outstanding performance on the cross-validation and CholecTriplet splits of the CholecT45 dataset. The ablation study is also conducted to validate the effectiveness of each key component of MT4MTL-KD.

REFERENCES

[1] L. Maier-Hein et al., "Surgical data science for next-generation interventions," *Nature Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.

[2] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "CAI4CAI: The rise of contextual artificial intelligence in computer-assisted interventions," *Proc. IEEE*, vol. 108, no. 1, pp. 198–214, Jan. 2020.

[3] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Med. Image Anal.*, vol. 16, no. 3, pp. 632–641, Apr. 2012.

[4] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[5] T. Blum, H. Feußner, and N. Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Beijing, China: Springer, 2010, pp. 400–407.

[6] C. I. Nwoye et al., "CholecTriplet2021: A benchmark challenge for surgical action triplet recognition," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102803.

[7] C. I. Nwoye et al., "Recognition of instrument-tissue interactions in endoscopic videos via action triplets," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 364–374.

[8] C. I. Nwoye et al., "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102433.

[9] A. Yamlahi et al., "Self-distillation for surgical action recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Vancouver, BC, Canada: Springer, 2023, pp. 637–646.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] S. Sharma, C. I. Nwoye, D. Mutter, and N. Padoy, "Rendezvous in time: An attention-based temporal fusion approach for surgical triplet recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 6, pp. 1053–1059, Apr. 2023.

[12] Y. Cheng, L. Liu, S. Wang, Y. Jin, C.-B. Schönlieb, and A. I. Aviles-Rivero, "Why deep surgical models fail: Revisiting surgical action triplet recognition through the lens of robustness," in *Proc. Int. Workshop Trustworthy Mach. Learn. Healthcare*. Virtual Event: Springer, 2023, pp. 177–189.

[13] M. Park, S. Oh, T. Jeong, and S. Yu, "Multi-stage temporal convolutional network with moment loss and positional encoding for surgical phase recognition," *Diagnostics*, vol. 13, no. 1, p. 107, Dec. 2022.

[14] L. Li et al., "SIRNet: Fine-grained surgical interaction recognition," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4212–4219, Apr. 2022.

[15] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng, and Z.-J. Zha, "A battle of network structures: An empirical study of CNN, transformer, and MLP," 2021, *arXiv:2108.13002*.

[16] E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem," *Inf. Process. Manage.*, vol. 44, no. 2, pp. 790–799, Mar. 2008.

[17] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 247–263.

[18] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[19] Y. Jin et al., "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.

[20] T. Czempiel et al., "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 343–352.

[21] F. Yi, Y. Yang, and T. Jiang, "Not end-to-end: Explore multi-stage architecture for online surgical phase recognition," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 2613–2628.

[22] X. Ding and X. Li, "Exploring segment-level semantics for online phase recognition from surgical videos," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3309–3319, Nov. 2022.

[23] S. Wang, Z. Xu, C. Yan, and J. Huang, "Graph convolutional nets for tool presence detection in surgical videos," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Hong Kong, China: Springer, 2019, pp. 467–478.

[24] S. Zhao, Y. Liu, Q. Wang, D. Sun, R. Liu, and S. Kevin Zhou, "MURPHY: Relations matter in surgical workflow analysis," 2022, *arXiv:2212.12719*.

[25] N. Xi, J. Meng, and J. Yuan, "Forest graph convolutional network for surgical action triplet recognition in endoscopic videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8550–8561, Dec. 2022.

[26] Y. Wei et al., "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[27] H. Yang, J. T. Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 280–288.

[28] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 464–472.

[29] W. Zhou, Z. Xia, P. Dou, T. Su, and H. Hu, "Aligning image semantics and label concepts for image multi-label classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2, pp. 1–23, May 2023.

[30] M. Tan et al., "Learning graph structure for multi-label image classification via clique generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4100–4109.

[31] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2960–2968.

[32] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 522–531.

[33] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5172–5181.

[34] Y. Wang et al., "Multi-label classification with label graph superimposing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12265–12272.

[35] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2Label: A simple transformer way to multi-label classification," 2021, *arXiv:2107.10834*.

[36] X. Zhu, J. Cao, J. Ge, W. Liu, and B. Liu, "Two-stream transformer for multi-label image classification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3598–3607.

[37] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," vol. 2, no. 7, 2015, *arXiv:1503.02531*.

[38] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[39] L. Song, J. Wu, M. Yang, Q. Zhang, Y. Li, and J. Yuan, "Handling difficult labels for multi-label image classification via uncertainty distillation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2410–2419.

[40] J. Xu, S. Huang, F. Zhou, L. Huangfu, D. Zeng, and B. Liu, "Boosting multi-label image classification with complementary parallel self-distillation," in *Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI)*, L. D. Raedt, Ed., Jul. 2022, pp. 1495–1501, doi: 10.24963/ijcai.2022/208.

[41] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 700–708.

[42] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1911–1923, Jul. 2021.

[43] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab, "Opera: Attention-regularized transformers for surgical phase recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Strasbourg, France: Springer, 2021, pp. 604–614.

[44] X. Gao, Y. Jin, Y. Long, Q. Dou, and P.-A. Heng, "Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Strasbourg, France: Springer, 2021, pp. 593–603.

[45] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[46] R. Dai, S. Das, K. Kahatapitiya, M. S. Ryoo, and F. Brémond, "MS-TCT: Multi-scale temporal ConvTransformer for action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20009–20019.

[47] C. I. Nwoye and N. Padoy, "Data splits and metrics for method benchmarking on surgical action triplet datasets," 2022, *arXiv:2204.05235*.

[48] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia, "ResLT: Residual learning for long-tailed recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3695–3706, Mar. 2023.