Label-guided Teacher for Surgical Phase Recognition via Knowledge Distillation[⋆]

Jiale Guan[⋆⋆], Xiaoyang Zou[⋆⋆], Rong Tao, and Guoyan Zheng

Institute of Medical Robotics, School of Biomedical Engineering,
Shanghai Jiao Tong University, No. 800, Dongchuan Road, Shanghai 200240, China.
`Correspondence: Guoyan Zheng (Email: guoyan.zheng@sjtu.edu.cn)`

Abstract. Automatic surgical phase recognition plays an essential role in developing advanced, context-aware, computer-assisted intervention systems. Knowledge distillation is an effective framework to transfer knowledge from a teacher network to a student network, which has been used to solve the challenging surgical phase recognition task. A key to a successful knowledge distillation is to learn a better teacher network. To this end, we propose a novel label-guided teacher network for knowledge distillation. Specifically, our teacher network takes both video frames and ground-truth labels as input. Instead of only using labels to supervise the final predictions, we additionally introduce two types of label guidance to learn a better teacher: 1) we propose label embedding-frame feature cross-attention transformer blocks for feature enhancement; and 2) we propose to use label information to sample positive (from same phase) and negative features (from different phases) in a supervised contrastive learning framework to learn better feature embeddings. Then, by minimizing feature similarity, the knowledge learnt by our teacher network is effectively distilled into a student network. At inference stage, the distilled student network can perform accurate surgical phase recognition taking only video frames as input. Comprehensive experiments are conducted on two laparoscopic cholecystectomy video datasets to validate the proposed method, offering an accuracy of 93.3±5.8% on the Cholec80 dataset and an accuracy of 91.6±9.1% on the M2cai16 dataset.

Keywords: Surgical phase recognition · Transformers · Label guidance · Knowledge distillation · Cholecystectomy.

1 Introduction

Surgical workflow analysis is a key procedure in developing computer-assisted intervention systems in modern operating room. Phase recognition, as one of the most crucial tasks, not only helps provide timely context-aware assistance and decision support during surgery [13], but also contributes to surgical video

[⋆] This study was partially supported by the National Natural Science Foundation of China via projects U20A20199.

[⋆⋆] Jiale Guan and Xiaoyang Zou contribute equally to the paper.

archiving, surgical training and surgical skill assessment [20], aiming to improve the quality and safety of modern surgery. Considering the abundant and valuable visual cues, compared to methods developed based on state signals [14], video-based surgical phase recognition has become a more prevalent solution.

The past few years have witnessed the rapid development of deep learning-based methods for surgical phase recognition, with a primary focus on temporal modeling. EndoLSTM [18] utilized long short-term memory (LSTM) network for temporal modeling. Jin et al. [8] proposed an end-to-end framework to train ResNet[6] and LSTM simultaneously. Czempiel et al. proposed TeCNO [1], which was developed based on multi-stage temporal convolutional networks. Recently, several methods used Transformer [19] for phase recognition. Czempial et al. [2] employed self-attention layers for temporal feature enhancement. Gao et al. [5] employed transformer to aggregate spatial and temporal embeddings. Zou et al. [25] proposed an auto-regressive transformer to model the inter-phase correlation implicitly by conditional probability distribution. In addition, multi-task learning methods were proposed to benefit from both phase recognition and tool detection. Jin et al. [9] proposed multi-task correlation loss to exploit the relationship between both tasks. Tao et al. [15] designed latent space-constrained transformers to learn semantic structure at video-level. Moreover, Liu et al. [11] proposed key pooling operation to record the important key information for surgical phase recognition. However, most of these methods only use ground-truth labels to supervise the final predictions, without fully leveraging valuable label information to establish feature-label correlation.

Knowledge distillation [7] has been proven to be a successful framework for transferring knowledge from a teacher network to a student network. Both knowledge distillation and teacher/student frameworks have been adopted to solve the challenging surgical phase recognition task [23,24]. Yu et al. [23] trained a teacher on a small dataset to generate synthetic annotations for a larger dataset, which were then used for student training. Zhang et al. [24] performed self-knowledge distillation by regularizing predictions of the student network to be consistent with the soft labels generated by the teacher network, which is the best model from past epochs. However, the soft labels have insufficient capacity to represent the surgical workflow compared with the ground-truth labels, leading to limited knowledge that the student network can learn from the teacher network.

In this paper, we proposed a novel label-guided teacher network for automatic surgical phase recognition via knowledge distillation. To learn a better teacher, we propose two types of label guidance in addition to supervision of final predictions, taking both video frames and ground-truth labels as input. Specifically, we introduce a label embedding-frame feature cross-attention transformer to establish correlation between ground-truth labels and feature embeddings, and a supervised contrastive learning framework to regularize the low-dimensional feature embeddings by pulling together positive feature pairs from same phase and pushing apart negative feature pairs from different phases. Benefiting from the design of the label-guided teacher network, the correlation between feature embeddings and ground-truth labels is effectively enhanced. Then, the label-guided

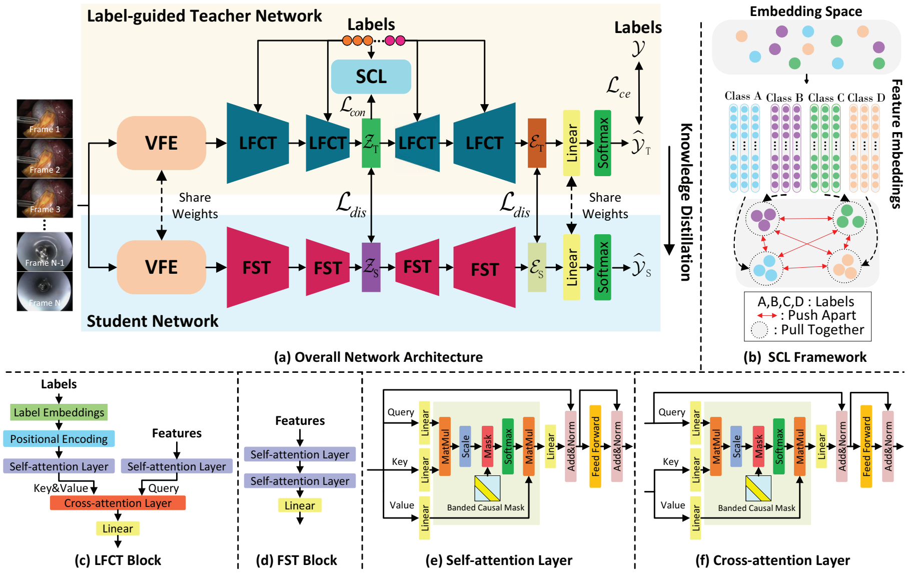


Fig. 1. A schematic overview of the proposed method, encompassing (a) the overall network architecture trained through knowledge distillation, (b) the supervised contrastive learning (SCL) framework, (c) the label embedding-frame feature cross-attention transformer (LFCT) block, (d) the frame feature self-attention transformer (FST) block, (e) the self-attention layer in detail, and (f) the cross-attention layer in detail.

knowledge learnt by the teacher network is distilled into the student network which takes only video frames to perform accurate phase recognition. In summary, the contributions of this paper are three-fold:

- We propose a knowledge distillation framework for surgical phase recognition, including a novel label-guided teacher network and a student network.
- We introduce a label embedding-frame feature cross-attention transformer and a supervised contrastive learning framework to learn a better label-guided teacher network, effectively enhancing the correlation between feature embeddings and ground-truth labels.
- Comprehensive experiments are conducted on two publicly available video datasets to validate the effectiveness of our method.

2 Methods

Fig. 1 (a) presents the overall network architecture of our method, comprising a label-guided teacher network and a student network. The label-guided teacher network is first trained to effectively exploit label information to provide valuable guidance, taking both video frames and ground-truth labels as input. Subsequently, utilizing a knowledge distillation framework, the label-guided knowledge learnt by the teacher network is transferred to a student network without label input, taking only video frames as input for accurate surgical phase recognition.

2.1 Label-guided Teacher Network

The proposed label-guided teacher network consists of a visual feature extractor (VFE), a temporal modeling module (TMM) and a linear classifier. For an input video with N frames, let C denote the number of surgical phases. The ground-truth labels can be represented as $\mathcal{Y} \in \mathbb{R}^{N\times C}$. We first trained a SwinV2-B [12] as the VFE to extract D dimensional frame-wise visual features $\mathcal{F} \in \mathbb{R}^{N\times D}$ from all video frames for further temporal modeling. To learn a better label-guided teacher network, apart from supervising the final predictions, we introduce two types of label guidance in the TMM, comprising label embedding-frame feature cross-attention transformer (LFCT) blocks for feature enhancement and a supervised contrastive learning (SCL) framework to learn better feature embeddings.

Label embedding-Frame feature Cross-attention Transformer. In the teacher network, the TMM is constructed by four LFCT blocks, which are designed to establish correlations between label embeddings and frame features by explicitly incorporating label information using cross-attention. As shown in Fig. 1 (c), each LFCT block has two input branches, a label branch and a feature branch, taking $\mathcal{Y}$ and $\mathcal{F}$ as input, respectively. Assuming the input feature dimension of the i-th LFCT block ($i \in [1, 4]$) is D_i, we follow [25] to generate D_i dimensional label embeddings with positional encoding in the label branch. Each input branch uses a self-attention layer [19] for temporal modeling, resulting in enhanced feature embeddings $\mathbf{F} \in \mathbb{R}^{N\times D_i}$ and label embeddings $\mathbf{L} \in \mathbb{R}^{N\times D_i}$. After that, to acquire label-guided knowledge, we propose using cross-attention layer [19] to facilitate interaction between enhanced feature embeddings and label embeddings, taking $\mathbf{F}$ as the query (Q) and $\mathbf{L}$ as the key (K) and value (V). Three linear projection matrices W^q, W^k, W^v are applied to Q, K, V, respectively, obtaining d dimensional vectors with 8 attention heads, where $d = D_i/8$. The label embedding-frame feature cross-attention can be calculated by:

$$Attn(Q = \mathbf{F}, K = \mathbf{L}, V = \mathbf{L}) = Softmax(Mask(\frac{\mathbf{F}W^q(\mathbf{L}W^k)^{\mathrm{T}}}{\sqrt{d}}))(\mathbf{L}W^v) \quad (1)$$

Both self-attention (Fig. 1 (e)) and cross-attention (Fig. 1 (f)) consist of a masked multi-head attention layer and a feed forward layer, with skip connection and layer normalization. Banded causal masks [15] are adopted to mask future information while providing proper temporal dependency range. Finally, a linear layer is used for dimension projection. The input feature dimension of the teacher network is D, while the output feature dimensions for the four LFCT blocks are $D/4$, $D/16$, $D/4$, D, in a down-sampling followed by up-sampling manner.

Supervised Contrastive Learning. The limited visual difference between different phases and the significant appearance variability within the same phase post great challenges for accurate surgical phase recognition [8]. To meet these challenges, inspired by [10], we introduce a SCL framework to regularize the low-dimensional feature embedding space at the bottom of the teacher network with label guidance. Different from self-supervised contrastive learning methods, we sample the positive and negative feature pairs using label information, as shown in Fig. 1 (b), aiming to bring features from the same phase closer while keep features from different phases apart, in the embedding space.

Specifically, let $\mathcal{Z}_{\mathrm{T}}$ denote the $D/16$ dimensional output features at the bottom of the teacher network. Taking an anchor feature $\mathcal{Z}_{\mathrm{T}_i}$ with phase label $\mathcal{Y}_i$ and a sample feature $\mathcal{Z}_{\mathrm{T}_j}$ with phase label $\mathcal{Y}_j$ $(i \neq j)$ for example, we regard the feature pair $\{\mathcal{Z}_{\mathrm{T}_i}, \mathcal{Z}_{\mathrm{T}_j}\}$ to be positive if $\mathcal{Y}_j = \mathcal{Y}_i$. Otherwise, the feature pair is considered to be negative if $\mathcal{Y}_j \neq \mathcal{Y}_i$. After that, we employ dot product to measure the similarity of the paired features. Therefore, the supervised contrastive loss $\mathcal{L}_{con}$ is computed by:

$$\mathcal{L}_{con} = \frac{1}{N} \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}_i|} \sum_{\mathcal{Z}_{\mathrm{T}_j} \in \mathcal{P}_i} log \frac{exp(\mathcal{Z}_{\mathrm{T}_i} \cdot \mathcal{Z}_{\mathrm{T}_j}/\tau)}{\sum\limits_{\mathcal{Z}_{\mathrm{T}_p} \in \mathcal{P}_i} exp(\mathcal{Z}_{\mathrm{T}_i} \cdot \mathcal{Z}_{\mathrm{T}_p}/\tau) + \sum\limits_{\mathcal{Z}_{\mathrm{T}_n} \in \mathcal{N}_i} exp(\mathcal{Z}_{\mathrm{T}_i} \cdot \mathcal{Z}_{\mathrm{T}_n}/\tau)} \tag{2}$$

where $\mathcal{P}_i \equiv \{\mathcal{Z}_{\mathrm{T}_p} \in \mathcal{Z}_{\mathrm{T}} \,|\mathcal{Y}_p = \mathcal{Y}_i, p \neq i\}$ and $\mathcal{N}_i \equiv \{\mathcal{Z}_{\mathrm{T}_n} \in \mathcal{Z}_{\mathrm{T}} \,|\mathcal{Y}_n \neq \mathcal{Y}_i\}$ are the collections of the positive and negative sample features for anchor feature $\mathcal{Z}_{\mathrm{T}_i}$. τ is the temperature. This design pulls together the positive pairs and pushes apart the negative pairs by maximizing the similarities of the positive feature pairs and minimizing the similarities of the negative feature pairs.

Training Objectives. For teacher network training, in addition to the supervised contrastive loss, cross-entropy loss weighted by median frequency balancing [4] is adopted to supervise the final predictions. Let λ denote the loss weight for supervised contrastive loss. The overall loss can be written as:

$$\mathcal{L}_{all} = \mathcal{L}_{ce} + \lambda \cdot \mathcal{L}_{con} \tag{3}$$

2.2 Knowledge Distillation

Considering that ground-truth labels are not available at inference stage, we propose a knowledge distillation framework to transfer the label-guided knowledge learnt by the teacher network to a student network without needing any label information. Therefore, by taking only video frames as input, the student network can perform surgical phase recognition at inference stage. To this end, the TMM in the student network is constructed by four frame feature self-attention transformer (FST) blocks. Differing from the LFCT block, each FST block comprises two self-attention layers for temporal feature enhancement, as illustrated in Fig. 1 (d), with only feature branch. Finally, the feature dimension is also transformed using linear layer. Note that the output feature dimensions for the four FST blocks are identical to those for the four LFCT blocks in the teacher network. Meanwhile, the VFE and linear classifier in the student network share the same weights as those in the teacher network.

To distill the label-guided knowledge into the student network, we extract feature embeddings at the bottom and end of the TMM and minimize feature similarity between the teacher and student networks. Specifically, let $\mathcal{Z}_{\mathrm{S}}$ and $\mathcal{Z}_{\mathrm{T}}$ represent the $D/16$ dimensional feature embeddings at the bottom of the student and teacher networks, while $\mathcal{E}_{\mathrm{S}}$ and $\mathcal{E}_{\mathrm{T}}$ represent the D dimensional feature embeddings at the end of the student and teacher networks, respectively. The knowledge distillation loss for student network training is computed as:

$$\mathcal{L}_{dis} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{Z}_{\mathrm{S}_i} - \mathcal{Z}_{\mathrm{T}_i}\|_1 + \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{E}_{\mathrm{S}_i} - \mathcal{E}_{\mathrm{T}_i}\|_1 \tag{4}$$

2.3 Implementation Details

Our VFE is initialized with weights pre-trained on ImageNet [3] and then trained for 100 epochs with 1e-4 learning rate. All video frames are resized to a resolution of 192×192 as the input of VFE. The dimension D of extracted feature embeddings is 1024. The teacher network is trained for 50 epochs with 5e-5 learning rate. For knowledge distillation, the student network is trained for 50 epochs with 1e-5 learning rate. We set λ to 0.1, τ to 0.07, and the bandwidth of banded causal mask to 500. AdamW optimizer is used for training. In each iteration, all frame features from a video are served as input. Our method is implemented in PyTorch using one NVIDIA RTX A6000 GPU.

3 Experiments

Experimental Setup. The proposed method is evaluated on two publicly available cholecystectomy video datasets recorded at 25 frames per second (fps), Cholec80 [17] and M2cai16 [16]. Cholec80 dataset contains 80 videos which are categorised into 7 surgical phases and 7 surgical tools. We use the first 40 videos for training and the rest 40 videos for testing. 8 videos in the training set are used for validation and hyper-parameters tuning. M2cai16 contains 41 videos which are categorised into 8 surgical phases. We use the first 27 videos for training and the rest 14 videos for testing. 7 videos in the training set are used for validation and hyper-parameters tuning. All videos are sampled to 1 fps for processing.

Following the previous work [5,15], to evaluate the phase recognition performance of the proposed method, we adopt four types of evaluation metrics, including the accuracy, precision, recall and Jaccard index.

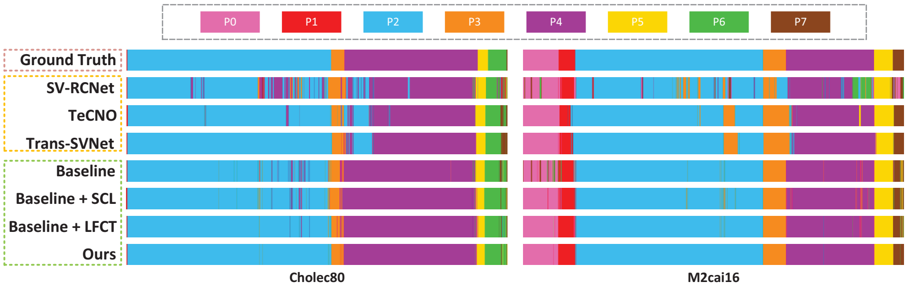


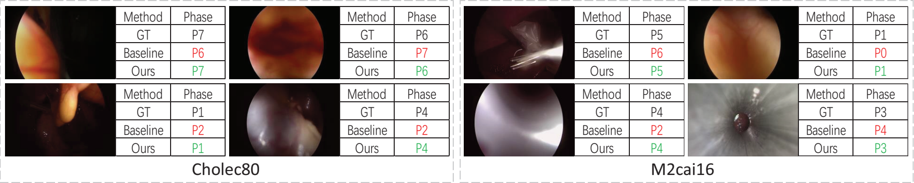
Fig. 2. Qualitative comparison results with SOTA and ablative testing results for the key components.

Table 1. Comparison with the state-of-the-arts on Cholec80 and M2cai16 datasets.

Methods	Cholec80				M2cai16			
	Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard
EndoNet* [17]	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	–	–	–	–	–
MTRCNet-CL* [9]	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9	–	–	–	–	–
LAST* [15]	93.1 ± 4.7	89.3 ± 5.5	90.1 ± 5.5	81.1 ± 7.6	91.5 ± 5.6	86.3 ± 7.5	88.7 ± 7.3	77.8 ± 10.1
PhaseNet [17,16]	78.8 ± 4.7	71.3 ± 15.6	76.6 ± 16.6	–	79.5 ± 12.1	–	–	64.1 ± 10.3
SV-RCNet [8]	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5	–	81.7 ± 8.1	81.0 ± 8.3	81.6 ± 7.2	65.4 ± 8.9
OHFM [21]	87.3 ± 5.7	–	–	67.0 ± 13.3	85.2 ± 7.5	–	–	68.8 ± 10.5
TeCNO [1]	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9	86.1 ± 10.0	85.7 ± 7.7	88.9 ± 4.5	74.4 ± 7.2
Trans-SVNet [5]	90.3 ± 7.1	90.7 ± 5.0	88.8 ± 7.4	79.3 ± 6.6	87.2 ± 9.3	88.0 ± 6.7	87.5 ± 5.5	74.7 ± 7.7
Not End-to-End [22]	92.0 ± 5.3	–	87.0 ± 7.3	77.1 ± 11.5	88.2 ± 8.5	–	$\mathbf{91.4 \pm 11.2}$	75.1 ± 10.6
Self-KD GRU [24]	93.2 ± 4.4	87.5 ± 7.4	86.9 ± 7.5	78.6 ± 9.0	–	–	–	–
Ours	$\mathbf{93.3 \pm 5.8}$	$\mathbf{91.3 \pm 6.4}$	$\mathbf{89.4 \pm 6.3}$	$\mathbf{80.6 \pm 9.7}$	$\mathbf{91.6 \pm 9.1}$	$\mathbf{90.6 \pm 6.3}$	90.6 ± 4.7	$\mathbf{80.6 \pm 9.1}$

$*$ denotes the multi-task learning methods that require both phase and tool labels.

Comparison with the state-of-the-arts. We compare the proposed method with state-of-the-art (SOTA) methods on both Cholec80 and M2cai16 datasets, including those designed for single-task and multi-task (requiring both phase and tool labels). Quantitative comparison results are presented in Table 1. When compared with the single-task SOTA methods, our method achieves the best phase recognition performance, with an average accuracy of 93.3% on Cholec80 and 91.6% on M2cai16. Although the average recall of our method on M2cai16 is slighly lower than that of [22], our method outperforms [22] by a large margin of 3.4% and 5.5% in terms of average accuracy and Jaccard index, respectively. Even when compared with the multi-task SOTA methods, our method shows comparable recognition performance to [15] on Cholec80, while outperforms [15] on M2cai16 by 4.3%, 1.9%, and 2.8% in terms of average precision, recall, and Jaccard index, respectively. We further present the qualitative comparisons, as shown in Fig. 2, which can also demonstrate that the proposed label-guided teacher network for knowledge distillation can help achieve superior surgical phase recognition performance with more reliable predictions on both datasets.



Method	Phase
GT	P7
Baseline	P6
Ours	P7

Method	Phase
GT	P6
Baseline	P7
Ours	P6

Method	Phase
GT	P1
Baseline	P2
Ours	P1

Method	Phase
GT	P4
Baseline	P2
Ours	P4

Method	Phase
GT	P5
Baseline	P6
Ours	P5

Method	Phase
GT	P1
Baseline	P0
Ours	P1

Method	Phase
GT	P4
Baseline	P2
Ours	P4

Method	Phase
GT	P3
Baseline	P4
Ours	P3

Fig. 3. Comparative analysis of predictions for challenging frames. GT: Ground Truth.

Table 2. Ablative testing results on Cholec80 and M2cai16 datasets.

Settings				Cholec80				M2cai16			
Baseline	LFCT	SCL	CA	Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard
✓	–	–	–	90.9 ± 6.4	87.1 ± 8.7	87.9 ± 7.6	75.9 ± 9.6	88.2 ± 9.4	87.7 ± 6.1	87.9 ± 5.9	75.4 ± 8.9
✓	–	✓	–	91.5 ± 6.7	88.3 ± 7.9	87.7 ± 8.5	76.6 ± 9.7	89.2 ± 10.4	88.5 ± 5.8	89.1 ± 4.8	77.4 ± 8.7
✓	✓	–	II	92.1 ± 6.7	88.4 ± 8.3	89.0 ± 5.8	78.4 ± 11.8	89.0 ± 9.5	88.0 ± 8.0	87.0 ± 6.8	75.0 ± 11.9
✓	✓	–	I	92.4 ± 6.6	90.1 ± 6.7	88.7 ± 6.7	78.9 ± 9.1	91.0 ± 9.7	90.4 ± 6.9	90.0 ± 5.4	80.1 ± 9.4
✓	✓	✓	II	92.8 ± 6.6	89.2 ± 8.0	**89.8 ± 5.6**	79.5 ± 10.8	91.1 ± 9.0	89.7 ± 7.0	88.3 ± 7.6	78.2 ± 10.4
✓	✓	✓	I	**93.3 ± 5.8**	**91.3 ± 6.4**	89.4 ± 6.3	**80.6 ± 9.7**	**91.6 ± 9.1**	**90.6 ± 6.3**	**90.6 ± 4.7**	**80.6 ± 9.1**

Ablation Study. We carry out ablation studies to investigate the effectiveness of each key component, including the baseline student network, the LFCT block and the SCL framework in label-guided teacher network. Table 2 presents the ablative testing results on both Cholec80 and M2cai16 datasets. The baseline model stands an independently trained baseline student network without knowledge distillation. Note that only when LFCT is adopted, the proposed knowledge distillation framework is used. On both datasets, one can observe that when both LFCT and SCL are adopted, the model can achieve the best phase recognition performance, which can demonstrate the effectiveness of the design of our method. Besides, when LFCT is adopted to train a label-guided teacher for knowledge distillation, an obvious performance boost can be observed compared to the baseline model without knowledge distillation. In addition, we also incorporate SCL at the bottom of the baseline model to investigate its effectiveness. The results demonstrate that SCL improves the recognition performance regardless of whether the knowledge distillation framework is used.

Furthermore, we conduct ablation studies to determine the proper setting for cross-attention (CA) in LFCT blocks. Two settings are considered: I) taking $\mathbf{F}$ as the query, and $\mathbf{L}$ as the key and value; II) taking $\mathbf{L}$ as the query, and $\mathbf{F}$ as the key and value. The results show that setting I achieves better phase recognition performance than setting II regardless of the adoption of SCL. This indicates that setting I can provide more label-guidance knowledge for feature enhancement, facilitating to learn a better teacher network.

Qualitative ablative testing results are illustrated in Fig. 2. One can see that the proposed method can achieve more accurate and smoother phase predictions compared to the ablative models. Moreover, we analyze phase predictions for challenging frames containing limited visual information. As shown in Fig. 3, it is obvious that the baseline model generates mis-recognized predictions while our method recognizes these challenging frames correctly, showing the significance of the label-guided teacher network with knowledge distillation.

4 Conclusion

In this paper, we propose a novel label-guided teacher network for surgical phase recognition based on knowledge distillation. We introduce label embedding-frame feature cross-attention transformer and supervised contrastive learning to learn a better label-guided teacher network. Experimental results on two publicly available datasets demonstrate the effectiveness of our designs, showing superior phase recognition performance compared to SOTA methods.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 343–352. Springer (2020)
2. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: Opera: Attention-regularized transformers for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 604–614. Springer (2021)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
5. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 593–603. Springer (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. IEEE transactions on medical imaging **37**(5), 1114–1126 (2017)
9. Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.W., Heng, P.A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical image analysis **59**, 101572 (2020)

10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
11. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: Skit: a fast key information video transformer for online surgical phase recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21074–21084 (2023)
12. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
13. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)
14. 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)
15. Tao, R., Zou, X., Zheng, G.: Last: Latent space-constrained transformers for automatic surgical phase recognition and tool presence detection. IEEE Transactions on Medical Imaging (2023)
16. Twinanda, A.P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Single- and multi-task architectures for surgical workflow challenge at m2cai 2016. arXiv preprint arXiv:1610.08844 (2016)
17. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
18. Twinanda, A.P.: Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. Ph.D. thesis, Strasbourg (2017)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Wu, S., Chen, Z., Liu, R., Li, A., Cao, Y., Wei, A., Liu, Q., Liu, J., Wang, Y., Jiang, J., et al.: Surgsmart: an artificial intelligent system for quality control in laparoscopic cholecystectomy: an observational study. International Journal of Surgery pp. 10–1097 (2023)
21. Yi, F., Jiang, T.: Hard frame detection and online mapping for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 449–457. Springer (2019)
22. Yi, F., Yang, Y., Jiang, T.: Not end-to-end: Explore multi-stage architecture for online surgical phase recognition. In: Proceedings of the Asian Conference on Computer Vision. pp. 2613–2628 (2022)
23. Yu, T., Mutter, D., Marescaux, J., Padoy, N.: Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. arXiv preprint arXiv:1812.00033 (2018)
24. Zhang, J., Barbarisi, S., Kadkhodamohammadi, A., Stoyanov, D., Luengo, I.: Self-knowledge distillation for surgical phase recognition. International Journal of Computer Assisted Radiology and Surgery pp. 1–8 (2023)

25. Zou, X., Liu, W., Wang, J., Tao, R., Zheng, G.: Arst: auto-regressive surgical transformer for phase recognition from laparoscopic videos. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(4), 1012–1018 (2023)