Self-Knowledge Distillation for Surgical Phase Recognition

Jinglu Zhang[1†], Santiago Barbarisi[1†], Abdolrahim Kadkhodamohammadi[1*], Danail Stoyanov[1,2] and Imanol Luengo[1]

[1]Medtronic Digital Surgery, 230 City Road, London, UK.
[2]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK.

*Corresponding author(s). E-mail(s):
rahim.mohammadi@medtronic.com;
Contributing authors: jinglu.zhang@medtronic.com;
santiago.barbarisi@medtronic.com;
danail.stoyanov@medtronic.com; imanol.luengo@medtronic.com;
[†]These authors contributed equally to this work.

Abstract

Purpose: Advances in surgical phase recognition are generally led by training deeper networks. Rather than going further with a more complex solution, we believe that current models can be exploited better. We propose a self-knowledge distillation framework that can be integrated into current state-of-the-art (SOTA) models without requiring any extra complexity to the models or annotations.

Methods: Knowledge distillation is a framework for network regularization where knowledge is distilled from a teacher network to a student network. In self-knowledge distillation, the student model becomes the teacher such that the network learns from itself. Most phase recognition models follow an encoder-decoder framework. Our framework utilizes self-knowledge distillation in both stages. The teacher model guides the training process of the student model to extract enhanced feature representations from the encoder and build a more robust temporal decoder to tackle the over-segmentation problem.

Results: We validate our proposed framework on the public dataset Cholec80. Our framework is embedded on top of four popular SOTA

approaches and consistently improves their performance. Specifically, our best GRU model boosts performance by **+3.33%** accuracy and **+3.95%** F1-score over the same baseline model.

Conclusion: We embed a self-knowledge distillation framework for the first time in the surgical phase recognition training pipeline. Experimental results demonstrate that our simple yet powerful framework can improve performance of existing phase recognition models. Moreover, our extensive experiments show that even with 75% of the training set we still achieve performance on par with the same baseline model trained on the full set.

Keywords: surgical phase recognition, knowledge distillation, self-supervised learning, surgical data science

1 Introduction

With the increasing effort in improving patient outcomes and enhancing human-computer interaction, more attention is now being paid to develop context-aware surgical (CAS) systems for the operation room [1]. Surgical phase recognition is one of the fundamental tasks in developing CAS systems, which aims at dividing a surgical procedure into segments intra-operatively with reference to its pre-defined standard steps. Applications of real time phase recognition include surgery monitoring, resource scheduling, and decision making support. However, this task is challenging due to high intra-phase and low inter-phase variance, and the long duration of surgical videos.

Some early studies used different variants of Hidden Markov Models (HMM) [2] or statistic models [3] to model phase transitions. Although these models are interpretable, for one, the handcrafted features have limited ability to represent high-dimensional surgical data. For another, these methods can only process a few local contiguous frames, thus struggling with capturing long-term temporal information. In recent decades, deep neural networks (DNNs) have become the mainstream for surgical phase recognition by showing a superior ability to learn high-level features.

EndoNet [4] was the first phase recognition model that utilized a DNN (AlexNet [5]) as its feature encoder and combined with support vector machine and HMM to detect surgical phases. With the rise in popularity of Recurrent Neural Networks (RNNs), DNNs are used both as encoder to extract latent features and temporal decoders to replace conventional graphical models. Most recent papers advanced phase recognition models by using deeper encoders, such as ResNet in TeCNO [6] and video Swin-transformer in GRU [7], and sophisticated temporal decoders, for example temporal convolutional-based TeCNO [6] and Transformer-based Trans-SVNet [8]. Designing more advanced models is important for more reliable phase recognition. However, the questions that arise here are: *Do we utilize the full capacity of these models? And should we always go for more complex solutions?*
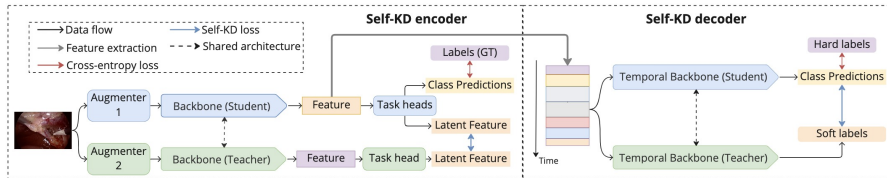
Fig. 1: An overview of our self-knowledge distillation (self-KD) framework for surgical phase recognition. Our framework is composed of a self-KD encoder and a self-KD decoder. Self-KD encoder jointly optimizes a classification loss and feature similarity loss to enhance feature representation. While the self-KD decoder use the soft labels generated by the teacher model to regularize the prediction.

We believe we can take advantage of knowledge distillation technique as a model-agnostic solution to improve the performance of surgical phase recognition. Knowledge distillation (KD) [9] is a process of transferring knowledge from a heavier and better performing teacher model to a smaller one. Recent studies [10] regularize the model by distilling the internal knowledge from the student network itself, namely self-knowledge distillation (self-KD). The self-KD framework can be combined with various backbone models without increasing network complexity.

Inspired by these observations, we propose a novel self-knowledge distillation framework built upon existing phase recognition models without adding extra trainable modules or additional annotations. Our proposed framework integrates self-KD into both encoder and decoder. At the encoder stage, we train encoders by optimizing both frame classification and feature similarity objectives. This is in contrast with recent self-supervised approaches like SimCLR [11] or BYOL [12], that add an extra pre-training step where self-supervision is the only objective. In these methods, self-supervised approaches are applied to large general domain datasets [5] where inter-phase variability is high. While in laparoscopic videos, this is not the case and inter-phase similarity is high. In our framework, frame representations are generated by the backbone of the student and forwarded into two task-specific heads for objective optimization (see self-KD encoder in Figure 1). One head optimizes the classification objective similar to Czempiel et al. and Kadkhodamoham-madi et al. [6, 13]. Concurrently, the second head optimizes a feature similarity objective. Inspired by [12], the target signal is generated by an auxiliary teacher[1] network. Moreover, student and teacher representations are obtained from differently augmented views of the same image. This enables us to obtain enhanced representations that are more robust to intra-phase variations. Figure 2 shows a low-dimension visualization of the features for different classes in a video. One can notice that self-KD helped the model to build more compact and separable representations.

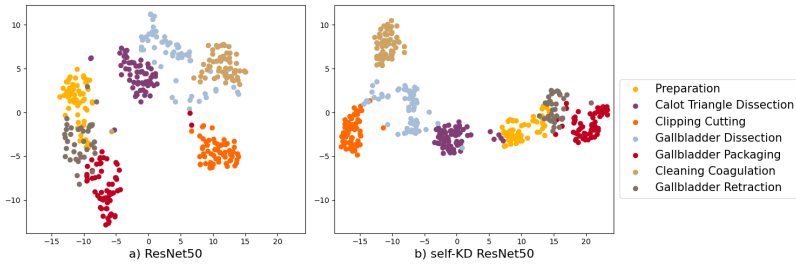[1]Student and teacher networks share the same architecture.

Fig. 2: A 2D projection of video frame features computed using t-SNE: (a) baseline ResNet50; (b) self-knowledge distillation ResNet50. One can observe that *Calot Triangle Dissection* or *Gallbladder Packaging* obtained denser and more separable representations with self-knowledge distillation, hence facilitating the detection of those phases.

At the decoder stage, we use self-KD to reduce the common over-segmentation problem [13, 14] in temporal models. For a temporal backbone model, we take its best model from preceding epochs as the teacher model to generate soft labels. The soft labels from the teacher model force the student model (current epoch) to produce more consistent predictions for the same frame with variant logits. The model is regularized by minimizing teacher and student logits using a smoothing loss.

Following SOTA approaches for both phase recognition [6, 8] and self-supervision [12], we benchmark our framework by training a ResNet50 [15] encoder. As for the backbone decoder, we select four popular and representative temporal models as our backbone, including RNN-based (GRU [7]), temporal convolutional-based (TCN [16] and TeCNO [6]), and transformer-based (Trans-SVNet [8]) models. We validate the effectiveness of our approach on Cholec80 [4] dataset. The self-KD framework shows consistent improvement over four phase recognition models. Our contribution is two-fold:

- To the best of our knowledge, we are the first to embed self-KD into surgical phase recognition architectures. It is a plug and play module. As a result, it can serve as a basic building block for any phase recognition model due to its generality and simplicity.
- With the teacher-student configuration, our method enables the encoder to extract enhanced spatial features and smooth predictions at the decoder. The extensive experiments (see subsection 4.2) demonstrate the robustness of our model with less training data.

2 Method

The architecture of our self-knowledge distillation framework (see Figure 1) for surgical phase recognition is detailed in this section, and consists of two

substructures: 1) A self-KD encoder (see subsection 2.1) to extract representative features; 2) A self-KD decoder (see subsection 2.2) integrated into existing models for reliable phase recognition. In general, a teacher-student structure is used, where a teacher network guides the student model. Both networks share the same structure, allowing the teacher to update the weights from the student, rather than optimizing them through back propagation.

2.1 Self-knowledge Distillation Encoder

Encoders in phase recognition architectures obtain frame-wise feature representations from their last average pooling layer. Features from a given video are later forwarded into a temporal decoder for final predictions. The self-KD encoder aims to generate enhanced representations. In the self-KD encoder, a student model generates two outputs from an augmented view $x_{i,1}$ of an image i: frame-based phase probabilities $\hat{y}_{i,1}$ and a latent feature representation $z_{i,1}$ (see Figure 1). An auxiliary teacher model generates a second feature representation $z_{i,2}$ from the same image i under a different augmentation $x_{i,2}$.

We optimize the student network by training two objectives simultaneously: phase recognition and feature similarity. Similar to [6, 13] we use categorical cross-entropy loss $\mathcal{L}_{CE}$ to optimize over the phase probabilities $\hat{y}_{i,1}$. The similarity objective is optimized using mean-square error loss $\mathcal{L}_{MSE}$ over the normalized[2] feature representations from the student $\bar{z}_{i,1}$ and the teacher $\bar{z}_{i,2}$. Following a batch of n images, the two losses are combined with equal weights into a single loss $\mathcal{L}_{self-KD}^{E}$ defined as:

$$\mathcal{L}_{self-KD}^{E} = \mathcal{L}_{CE} + \mathcal{L}_{MSE} = -\sum_{i=0}^{n} \log(\hat{y}_i) + \sum_{i=0}^{n} \|\bar{z}_{i,1} - \bar{z}_{i,2}\|_2^2 \qquad (1)$$

We do not back-propagate the loss through the teacher model. Rather, it is updated through Exponential Moving Average (EMA) [17], using parameters of the student model. Which means that the weights of the teacher model are a slow-moving average from the student. The teacher model is only used during training and discarded afterward.

2.2 Self-knowledge Distillation Decoder

The phase recognition decoder incorporates temporal dependencies to generate temporally smooth prediction over a video. The self-KD phase decoder aims to achieve more reliable predictions by regularizing the predictions from the backbone model. For any temporal model, we use its past predictions as an additional self-supervision signal (i.e. teacher model). The model is progressively regularized by minimizing the logits distance between student (the current predictions) and teacher model.

For an input feature sequence $x \in \mathbf{X}$ with length L: $x_{1:L} = (x_1, ..., x_L)$, we aim to assign a class label $c \in \mathbf{C}$ to each frame: $c_{1:L} = (c_1, ..., c_L)$. A temporal

---

[2]Authors in [11] show empirically that using a normalized projection improves model learning towards the downstream task.

backbone model produces the logit vectors and then a softmax function is used to calculate the predicted probabilities as $P(x) = [p_1(x), ..., p_L(x)]$. Inspired by the findings of Hinton et al. [9], both student $P^S(x)$ and teacher $P^T(x)$ probabilities are scaled by a temperature factor for better distillation.

Most KD literature [10, 18] define the Kullback-Leibler divergence as the distillation loss. We instead use the truncated mean squared error (MSE) $\mathcal{L}_{T-MSE}$ over the frame-wise student and teacher log probabilities as it has been shown to generate smoother predictions [14].

$$\mathcal{L}_{T-MSE} = \frac{1}{LC} \sum_{l,c} \tilde{\Delta}_{l,c}^2, \text{ where } \tilde{\Delta}_{l,c} = \begin{cases} \Delta_{l,c} & : \Delta_{l,c} \leq \tau \\ \tau & : otherwise \end{cases} \qquad (2)$$

where τ is the truncated threshold and we set it to 8.0. In our method, we define the teacher model as the best student model from the training history, which indicates that the teacher model dynamically evolves itself as the training proceeds. Let $P_t^S(x)$ be the predictions from the student model at t-th epoch, the MSE loss $\Delta_{l,c}$ is defined as:

$$\Delta_{l,c} = |\log p_t^S(x_{l,c}) - \log p_{i<t}^S(x_{l-1,c})| \qquad (3)$$

where i is the preceding epoch with the highest accuracy. It is worth noting that gradients are only calculated for student predictions $p^S(x_{l,c})$, while teacher predictions $p^T(x_{l-1,c})$ are not considered as model parameters. Combining with a cross-entropy loss $\mathcal{L}_{CE}$, decoder self-KD loss $\mathcal{L}_{self-KD}^D$ is defined as:

$$\mathcal{L}_{self-KD}^D = \mathcal{L}_{CE} + \lambda \mathcal{L}_{T-MSE} \qquad (4)$$

where λ is a coefficient to determine the contribution of $\mathcal{L}_{T-MSE}$ loss.

3 Experiments and Results

3.1 Experimental Settings

Dataset: We evaluated our proposed method on the public Cholec80 dataset[3]. This dataset contains 80 laparoscopic cholecystectomy procedures. Each video frame was labeled with seven phases. We used the first 40 videos as training and the rest for testing as suggested by [4]. Videos were downsampled to 1 FPS. Model performance was measured using five common metrics for surgical phase recognition, namely accuracy, precision, recall, F1-score and jaccard index. Metrics were obtained from the epoch with lowest loss at decoder training. We reported the average of metrics obtained per video with corresponding standard deviation. Moreover, all our models were trained and evaluated in online mode, which means no future information was used for current prediction.

---

[3]http://camma.u-strasbg.fr/datasets

Encoders Training details: We choose ResNet50 [15] from Pytorch repo as the backbone model for baseline and self-KD experiments, with a batch size of 64 throughout 100 epochs. We followed image augmentations operations in [12]. We used SGD to optimize the model (learning rate $1e^{-2}$ and weight decay of $1e^{-5}$). For EMA update, decay weight τ_0 started at 0.9995 and was progressively updated to a value of 1, following equation $\tau_{i+1} = 1 - \frac{1}{2}(1 - \tau_0)(\cos(\frac{\pi i}{K}) + 1)$, where i is current training step, and K is the total number of steps. The teacher model was updated after each optimization step performed over the student model.

Decoders Training details: For self-KD decoder training, we extracted the 2048-dim features using the aforementioned encoder and selected four popular state-of-the-art decoder backbones, including TCN, TeCNO, Trans-SVNet and GRU. Models were trained for 30 epochs using the Adam optimizer. The learning rate for TCN, TeCNO and GRU was set to 1e-3. Trans-SVNet is a three-stage model based on TeCNO. We trained it for 10 epochs with learning rate 1e-4. The self-KD loss weight for all models, λ, was set to 0.3.

Our framework was implemented in PyToch 1.12.1 and all models were trained on a Tesla V100 DGXS 32GB GPU.

3.2 Comparison with the SOTA

Table 1 presents the performance results of our framework compared with state-of-the-art. Our best model (self-KD GRU) achieved competitive results compared to MTRCNet-CL and TeCNO despite the fact that both of these models are multi-task and benefit from additional instrument signals, which are not always available. Both our best self-KD GRU model and self-KD TCN model surpassed [7], despite relying on a 2D encoder instead of the more complex 3D Swin-Transformer encoder. Trans-SVNet re-implemented TeCNO to only rely on phase signals. Performance metrics reported in Trans-SVNet are higher than the metrics we achieved using our implementation based on their public code. We found out that authors of Trans-SVNet considered class imbalance during metric calculation[4], which is not common in the literature. We therefore assessed both Trans-SVNet and TeCNO without weighting them based on class imbalance, which are indicated by star in Table 1. Our self-KD TeCNO and self-KD Trans-SVNet exceeded their corresponding baselines (see detailed ablative experiments in subsection 4.1). In addition, our best model self-KD GRU model reached the new SOTA results for surgical phase recognition task. It outperformed the 3D encoder based Swin+GRU by +2.35% accuracy, +2.39% precision, and +1.28% recall.

---

[4]https://github.com/YuemingJin/TMRNet

Table 1: Performance comparison against SOTA approaches, results are shown in percentage. TeCNO* and Trans-SVNet* are two models we evaluated without taking class imbalance into account.

	Acc.	Pre.	Recall	F1-score	Jaccard
MTRCNet-CL [19]	89.2	86.9	88.0	-	-
TeCNO [6]	88.56	81.64	85.24	-	-
Swin+GRU [7]	90.88	85.07	85.59	-	-
TeCNO from [8]	88.6	86.5	87.6	-	75.1
Trans-SVNet [8]	90.3	90.7	88.8	-	79.3
TeCNO*	90.20	84.12	85.34	82.65	74.07
Trans-SVNet*	89.85	84.96	84.48	82.25	74.01
Self-KD TeCNO	91.37	86.41	84.09	84.09	76.04
Self-KD Trans-SVNet	91.68	86.12	86.06	84.42	76.56
Self-KD TCN	91.78	85.88	**87.39**	85.08	77.67
Self-KD GRU	**93.24**	**87.46**	86.87	**85.61**	**78.56**

4 Discussion

4.1 Self-KD at different stages

To further investigate the efficiency of each sub-component in our framework, we conduct ablative studies with four SOTA models as follows: (1) baseline encoder + baseline decoder; (2) encoder self-KD + baseline decoder; (3) baseline encoder + decoder self-KD; (4) encoder self-KD + decoder self-KD. The results are presented in Table 2, from which we observe that:

(1) v.s. (2) and (1) v.s.(3): These experiments show the impact of applying self-KD at different stages. One can notice that applying self-KD to each stage alone improved the results. In general, we notice that adding self-KD to encoder results in higher boost in overall performance. This highlights the importance of self-KD in building more robust features.

(2) v.s. (4) and (3) v.s.(4): In general, model performance benefits from two-stage self-KD (both encoder and decoder stage), especially the accuracy. For other evaluation metrics, two-stage self-KD TCN and self-KD GRU still outperform the single-stage model. We have noticed that encoder self-KD contributes more to TeCNO and Trans-SVNet in recall, F1-score, and jaccard. Because the learning ability of self-KD relies on the solution space of the model itself. TeCNO and Trans-SVNet (TeCNO based) present a small temporal convolutional based model (117K parameters) compare with reported TCN (297K parameters parameters) or GRU (481K parameters). In this regard, the temporal model has limited capacity to distillate the knowledge.

(1) v.s. (4): Self-KD can be combined with various backbone models and it has consistently improved the results over baseline models. This highlights the effectiveness of self-KD in building more robust models. Specifically our best model, self-KD GRU, not only exceeds the performance of other three models, but boosted performance over its own baseline by +5.32% jaccard.

Moreover, we show a qualitative comparison of our self-KD GRU model in Figure 3. It can be observed that GRU baseline model mis-classified more

Table 2: Results of adding self-KD into training of architecture stages progressively, shown in percentage, for *Cholec80* test set. Metrics are obtained first per video and then average and standard deviation are reported. *TeCNO* and *Trans-SVNet* baselines results are obtained using our own implementation for fair comparison.

|	self-KD-Enc	self-KD-Dec	Acc	Pre	Rec	F1-score	JA
TCN			89.95±6.10	84.43±8.54	83.48±7.70	81.26±9.09	72.66±11.01
	✓		91.20±5.39	85.85±7.45	86.71±7.14	84.50±7.64	76.87±9.28
		✓	90.54±5.73	85.39±7.97	84.75±7.22	82.75±8.57	74.58±10.03
	✓	✓	**91.78±5.20**	**85.88±7.86**	**87.39±7.16**	**85.08±7.95**	**77.67±9.69**
TeCNO			90.20±5.76	84.12±8.89	85.34±7.18	82.65±8.76	74.07±10.85
	✓		90.78±6.32	84.85±8.34	**87.27±7.11**	**84.24±8.21**	**76.53±10.28**
		✓	90.32±5.95	84.24±8.76	85.17±6.96	82.63±8.66	74.03±10.72
	✓	✓	**91.37±5.09**	**86.41±7.67**	85.63±7.59	84.09±8.01	76.04±10.05
Trans-SVNet			89.85±6.02	84.96±8.21	84.48±7.77	82.25±9.02	74.01±10.83
	✓		91.33±5.97	85.22±7.83	**87.29±6.38**	**84.66±7.43**	**77.09±8.99**
		✓	90.81±5.59	**86.93±7.59**	85.97±6.65	84.34±8.06	76.32±10.00
	✓	✓	**91.68±4.79**	86.12±7.28	86.06±6.94	84.42±7.27	76.56 ±8.74
GRU			89.91±7.42	83.10±9.45	84.06±8.92	81.66±10.02	73.24±12.06
	✓		92.73±5.34	86.04±7.47	86.80±7.77	84.75±7.75	77.65±9.45
		✓	91.05±5.54	85.28±8.88	82.85±7.00	82.13±8.05	73.53±10.03
	✓	✓	**93.24±4.36**	**87.46±7.44**	**86.87±7.46**	**85.61±7.37**	**78.56±9.04**
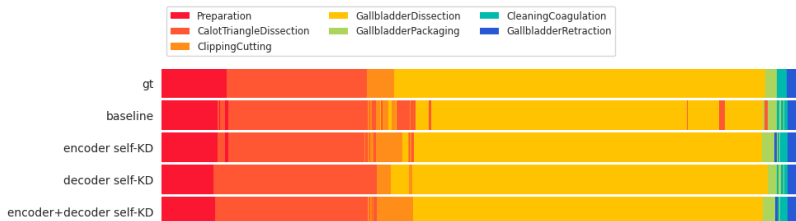
Fig. 3: Qualitative results. Predictions of a ResNet50+GRU with self-knowledge distillation (self-KD) at different encoder and decoder stages are shown against the ground truth on video 61.

frames than self-KD models and suffer from severe over-segmentation issues. Adding only the encoder self-KD generates more accurate predictions but over-segments some frames around *Clipping and Cutting*. Due to the limited feature representative ability of baseline ResNet, self-KD decoder model has smoother predictions but still some noisy prediction towards the end of the video. Combining the encoder and decoder self-KD, our method produces more reliable and smoother predictions. Both quantitative and qualitative results shows the superior performance of our proposed self-KD strategy and the indispensability of encoder and decoder self-KD in building robust models.

4.2 Training data

We conduct experiments to assess the effect of reducing training data sizes. We randomly removed videos from training set and compared with the baselines trained on full set. Videos are removed from the training set in two ways:

excluding the video from the training altogether or excluding phase ground truth labels, in other words only self-KD loss is applied over those videos. The ResNet50+GRU model is used for the experiment reported in Table 3. We refer to these experiments as: (1) models trained on full dataset; (2) baseline model trained on reduced sets of training videos; (3) self-KD GRU trained on reduced sets of training videos; (4) self-KD GRU trained on reduced sets of training videos in classification loss.

(1) v.s. (2) v.s.(3): In general reducing the size of the training set has degraded performance for all models. We however noticed the self-KD model trained on 87.5% of the data or even when it is trained on 75% of the data outperformed the baseline model trained on the full dataset. This finding is important for surgical phase recognition task, as generating annotations for surgical videos can be time consuming and more expensive than in general domain computer vision dataset.

(1) v.s. (4) and (3) v.s.(4): Less reduction in performance in (4) compared with (3), highlights the ability of the proposed framework in learning from unlabeled videos. We have observed that even reducing the size of the training set by 50%, our self-KD achieved performance on par with the same model trained on full training sets. This show the impact of using training data more effectively to generate more reliable models and hopefully paving the way towards removing the need for generating huge labeled phase datasets.

Table 3: Effect of reducing the number of training videos. Results are shown in percentage for Cholec80 test set with 40 training videos.

Encoder	# videos excluded	excluded from	Accuracy	F1-score	Jaccard
ResNet50	0		89.91±7.42	81.66±10.02	72.66±12.06
self-KD-ResNet50	0		**93.24±4.36**	85.61±7.37	**78.56±9.04**
ResNet50	5	training	87.98±9.59	78.74±1.57	69.82±12.40
	10		87.32±8.23	79.13±8.68	69.60±10.59
	20		84.31±11.35	74.39±9.42	63.86±10.79
self-KD-ResNet50	5	training	92.38±4.32	**85.82±8.47**	76.33±9.80
	10		89.45±6.78	84.42±8.48	72.78±9.95
	20		88.22±8.95	79.76±9.16	71.33 ± 10.63
self-KD-ResNet50	5	classification loss	92.44±4.82	83.39± 8.11	76.20±9.31
	10		89.99±9.10	82.54±8.48	74.26±10.67
	20		89.69±7.68	81.99±9.07	74.74±10.62

5 Conclusion

In this paper, we propose a self-knowledge distillation framework for surgical phase recognition. Rather than increasing model complexity, we embed the self-KD into existing models to utilize training data more effectively. Implicit knowledge is distilled in a single training process such that the encoder generates informative representations and the over-segmentation errors are reduced on the temporal decoder stage. We assessed our framework by applying it to four popular surgical phase recognition models. Our experiments show that our framework outperforms baselines over all evaluation metrics and our best

model self-KD GRU achieves new SOTA performance We believe the results can answer our proposed question: rather than always going to a more complex solution, training data can be utilized efficiently by our self-knowledge distillation framework in building more robust models.Due to the limitation of small dataset size, our future work will be extended to larger datasets to better study the impact of our framework.

Declarations

- Funding: This work was funded by Medtronic plc.
- Competing interests: Mr. Barbarisi, Drs. Zhang, Kadkhodamohammadi and Luengo; and Prof. Stoyanov are employees of Medtronic plc. Prof. Stoyanov is also a co-founder and shareholder in Odin Vision, Ltd.
- Ethics and consent: This article does not contain any studies with human participants or animals performed by any of the authors.
- Data, code and/or material availability: Privately held

References

[1] 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)

[2] Padoy, N., Blum, T., Feussner, H., Berger, M.-O., Navab, N.: On-line recognition of surgical activity for monitoring in the operating room. In: AAAI, pp. 1718–1724 (2008)

[3] Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B.: Real-time analysis of cataract surgery videos using statistical models. Multimedia Tools and Applications **76**(21), 22473–22491 (2017)

[4] 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)

[5] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)

[6] 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: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 343–352 (2020). Springer

[7] He, Z., Mottaghi, A., Sharghi, A., Jamal, M.A., Mohareri, O.: An empirical study on activity recognition in long surgical videos. In: Machine Learning for Health, pp. 356–372 (2022). PMLR

[8] 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, pp. 593–603. Springer, Cham (2021)

[9] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat **1050**, 9 (2015)

[10] Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation with progressive refinement of targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6567–6576 (2021)

[11] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR

[12] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

[13] Kadkhodamohammadi, A., Luengo, I., Stoyanov, D.: Patg: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos. International Journal of Computer Assisted Radiology and Surgery **17**(5), 849–856 (2022)

[14] Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3575–3584 (2019)

[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

[16] Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. Universal Language Model Fine-tuning for Text Classification (2018)

[17] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. In: Bengio, Y., LeCun, Y. (eds.) ICLR (2016)

[18] Ding, X., Liu, Z., Li, X.: Free lunch for surgical video understanding by distilling self-supervisions. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII, pp. 365–375 (2022). Springer

[19] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A.: Multitask recurrent convolutional network with correlation loss for surgical video analysis. Medical image analysis **59**, 101572 (2020)