Self-distillation for Surgical Action Recognition

Amine Yamlahi[1,2](✉), Thuy Nuong Tran[1,5], Patrick Godau[1,2,3,5], Melanie Schellenberg[1,2,3,5], Dominik Michael[1,2], Finn-Henri Smidt[1], Jan-Hinrich Nölke[1,5], Tim J. Adler[1,2,5], Minu Dietlinde Tizabi[1], Chinedu Innocent Nwoye[4], Nicolas Padoy[4], and Lena Maier-Hein[1,2,5,6]

[1] Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2] National Center for Tumor Diseases (NCT), NCT Heidelberg a Partnership between DKFZ and University Medical Center Heidelberg, Heidelberg, Germany
m.elyamlahi@dkfz-heidelberg.de
[3] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4] ICube Laboratory, University of Strasbourg, Strasbourg, France
[5] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg , Germany
[6] Medical Faculty, Heidelberg University, Heidelberg , Germany

Abstract. Surgical scene understanding is a key prerequisite for context-aware decision support in the operating room. While deep learning-based approaches have already reached or even surpassed human performance in various fields, the task of surgical action recognition remains a major challenge. With this contribution, we are the first to investigate the concept of self-distillation as a means of addressing class imbalance and potential label ambiguity in surgical video analysis. Our proposed method is a heterogeneous ensemble of three models that use Swin Transformers as backbone and the concepts of self-distillation and multi-task learning as core design choices. According to ablation studies performed with the CholecT45 challenge data via cross-validation, the biggest performance boost is achieved by the usage of soft labels obtained by self-distillation. External validation of our method on an independent test set was achieved by providing a Docker container of our inference model to the challenge organizers. According to their analysis, our method outperforms all other solutions submitted to the latest challenge in the field. Our approach thus shows the potential of self-distillation for becoming an important tool in medical image analysis applications. Code available at https://github.com/IMSY-DKFZ/self-distilled-swin.

Keywords: Surgical action recognition · Self-distillation · Laparoscopic surgery · Surgical workflow

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_61.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 637–646, 2023.
https://doi.org/10.1007/978-3-031-43996-4_61

1 Introduction

Surgical scene understanding is an important prerequisite for artificial intelligence (AI)-empowered surgery [12], underlying a range of application areas such as context-aware decision support, autonomous robotics, and workflow optimization. One of its key components is the fully-automatic recognition of the surgical action performed at a given point in time - a task not yet solved by state-of-the-art-methods [4,20]. To advance the field, the CholecTriplet challenge was organized in the scope of the Medical Image Computing and Computer Assisted Interventions (MICCAI) conferences 2021 and 2022. However, according to the organizers analysis [15,18,20], the task still remains unsolved. The guiding hypothesis of our work was that self-distillation could address some of the challenges in surgical action recognition, namely the high number of classes (100 in the case of CholecTriplet); high class imbalance, and label ambiguity. Self-distillation builds upon the widespread concept of knowledge distillation (KD) [6], in which the knowledge is transferred from one deep model (i.e., a teacher) to another shallow model (i.e., a student). Self-distillation diverges from traditional KD by distilling knowledge within the network itself. While KD is already used in various communities, the purpose of this work was to pioneer the concept of self-distillation in the context of surgical data science. Based on the CholecTriplet training data set, we developed a method for surgical action recognition (Fig. 2) that leverages self-distillation, Swin Transformers [11], multi-task learning, and ensembling. The following Sect. 2 presents our methodological contribution in detail. Sect. 3 presents ablation studies on the challenge data set that reveal the most important design choices as well as an external validation of our solution on an independent surgical video data set. We conclude with a brief discussion of the most relevant aspects of our work in Sect. 4.

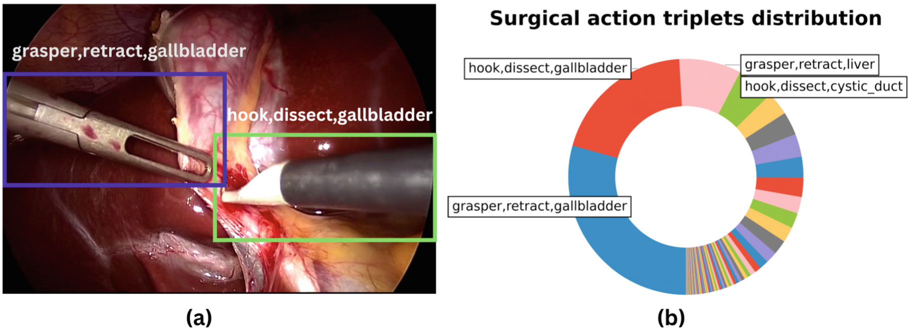

Fig. 1. Task of surgical action recognition. (a) Each action is represented by a triplet comprising instrument, verb and target. Multiple triplets can be present in one image, as shown in the example. (b) CholecTriplet training data set illustrating the heavy class imbalance. Of 100 possible triplet classes, the prevalence ranges from 0.01% to 44.6%.

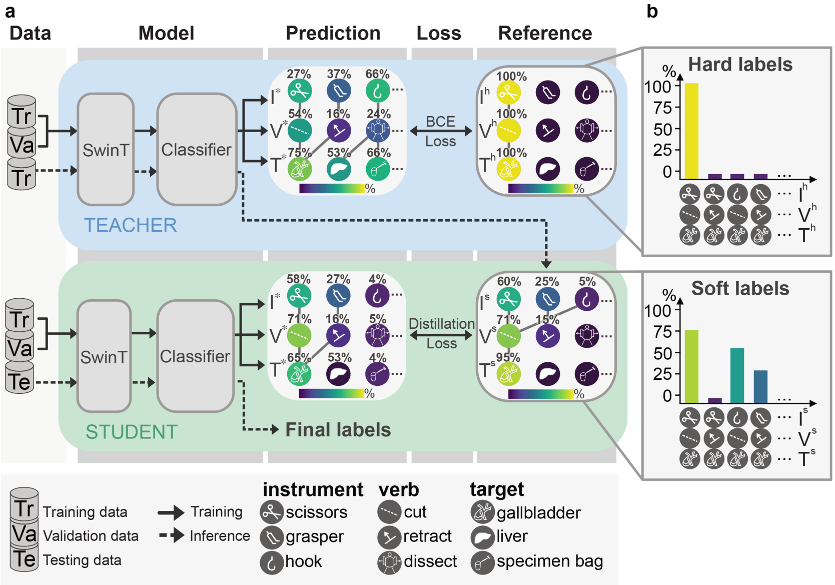

Fig. 2. Approach to surgical action triplet recognition. (a) Our architecture leverages Swin Transformer (SwinT) as a backbone and the concepts of self-distillation, multi-task learning, and ensembling as core strategies. The teacher model is trained on hard labels using binary cross-entropy (BCE) loss. Inferencing the training data, the sigmoid probabilities are used as input in the next step. The student model is trained with the noisy soft labels in a multi-task fashion to minimize the BCE loss, commonly referred to as distillation loss, between the teacher and the student's predictions. (b) Visualization of label distribution in hard and soft labels.

2 Methods

2.1 Task Description and Dataset

Our study is based on the CholecTriplet Challenge 2022 [18], which was conducted under the umbrella of the Endoscopic Vision Challenges (EndoVis) in conjunction with MICCAI. The Surgical Action Recognition task required participants to submit solutions that recognize surgical action triplets in laparoscopic videos, as illustrated in Fig. 1. The challenge granted access to the CholecT45 [19] dataset which consists of 45 video recordings of laparoscopic cholecystectomy with a total of 90,489 frames. CholecT45 is annotated with 100 action triplet classes, with one instrument, verb, and target forming a triplet. The annotations include six different instrument classes, ten verbs (denoting the action performed), and 15 targets such as organs, tissues, or foreign bodies (clip, specimen bag, etc.). The theoretical maximum of $6 \cdot 10 \cdot 15$ classes was reduced to the above-mentioned 100 based on medical relevance and prevalence. An example image from the CholecT45 dataset containing two triplet annotations can be seen in Fig. 1 (a). A chart depicting the highly imbalanced class distribution is shown in Fig. 1 (b).

2.2 Concept Overview

As illustrated in Fig. 2, our approach is based on the following key components: (1) Swin Transformer: The recently proposed Swin Transformer [11] architecture was chosen as backbone. (2) Multi-task learning: Based on the success of previous work that leveraged multi-task learning as its training paradigm, we incorporated multiple auxiliary tasks in our architecture, namely the classification of the individual components of the triplet (instrument, verb, and target) as well as the surgical phase. (3) Self-distillation: The core idea of our approach is the usage of soft labels to reduce overconfidence and address label ambiguity. (4) Ensemble: Following common successful training strategies, we implement ensembling to combine the predictions of three trained Swin Transformers of different scales.

2.3 Implementation Details

Swin Transformer. We base our method on Swin Transformer (SwinT) models of the timm [24] library and adopted the final classification layer to output the 100 triplet predictions, as well as the individual instruments, verbs, targets and surgical phase as auxiliary tasks to leverage the interconnection between them in a multi-task fashion (+Multi).

Self-distillation. The concept of self-distillation was achieved by training a teacher Swin transformer on one-hot encoded hard labels for 20 epochs, with a batch size of 64, an Adam [10] optimizer, a learning rate of 2 $\times 10^{-4}$, a cosine annealing scheduler decreasing to a minimum learning rate of 2 $\times 10^{-6}$, and a binary cross-entropy loss function. The model was trained with light augmentations that comprise resizing the images to 224×224 pixels, horizontal and vertical flips, rotation, brightness and saturation perturbations with a probability of 0.5. We trained five teacher and five student models; one for each fold of the official five-fold cross validation splitting introduced by the challenge. The teacher was trained on four of the five splits of its fold. After convergence, the soft labels (i.e., the sigmoid probabilites) for the same four splits were computed and the student was trained using these soft labels. The validation was performed on the fifth split, using hard (i.e., the original) labels for both the teacher and the student. During inference, the sigmoid probabilities of the five student models were averaged to yield the final result.

The five teacher models shared a common weight intialization seed. The five student models shared a separate weight initialization seed. The student models were trained for 40 epochs with the same augmentations as the teacher models. We saved the weights on the epoch with the best mean Average Precision (mAP) score based on the validation split for the current fold.

Ensemble. We combined three trained Swin Transformers (SwinT) of different scales (SwinT base/SwinT large) and configurations for our final ensemble (Ens) model: First, we employed a SwinT base model with multi-task learning of instrument, verb, and target and trained it using self-distillation. Second, we used a SwinT large model using the same approach, and added label smoothing to the soft labels. Third, we included phase annotations as an additional task for the multi-task training of a SwinT base model still employing self-distillation. Please note that every single model mentioned here corresponds to the five aggregated student models of the previous paragraph. All the models were trained using Nvidia GPUs Geforce RTX 3090 and Tesla V100 32GB.

3 Experiments and Results

The purpose of the experiments was to validate the performance of our method and to quantify the (potential) benefit of each individual component. To this end, we conducted (1) comprehensive ablation studies using the CholecT45 official 5-Fold cross-validation split [17], (2) an analysis of the specific benefit of soft labels, and (3) an external validation based on a Docker container submitted to the CholecTriplet 2022 challenge organizers. The Rendezvous Net [18], provided by the challenge organizers, served as the benchmark. In line with the challenge design [13], we validated the performance using mAP (following the aggregation scheme in [15]) and the top K=5 Accuracy as metrics. All scores were computed using the ivtmetrics library [17].

Ablation Studies. We designed the ablation studies as follows: We first calculated the performance of our Swin Transformer backbone as a stand-alone triplet classifier (SwinT). Next, we added multiple auxiliary targets (instruments, verbs, targets, and phases) for multi-task classification (+MultiT). As a third component, we implemented self-distillation by training a student model on soft labels, acquired by training the teacher model (+SelfD). The fourth step was the ensembling of three student model SwinT (+Ens). The results are shown in Tab. 1. A single SwinT as model backbone yields a higher mAP for triplet classification (mAP=32.3%) than the benchmark (mAP=28.8%), which corresponds to a relative improvement by 10.3%. The biggest boost was achieved by including self-distillation, which improved the Triplet mAP and top-5 accuracy by 3.8% points (pp) and 2.4pp, respectively, compared to our baseline. The final model yielded a mAP of 38.5% and a top-5 accuracy of 86.5%, which corresponds to a boost of 6.2pp in mAP and 2.7pp in top-5 accuracy compared to our own baseline, and a relative improvement by 33.7% for mAP compared to the state-of-the-art method. For transparency, we also provide per-video results, depicted in Fig. 3. With a few exceptions, our final model consistently provides the best results.

Analysis of Soft Labels. The addition of self-distillation resulted in the highest boost in performance. This holds true despite the fact that the mAP of the teacher model, trained on hard labels, was about 88% on the training set, which is sub-optimal. The question is thus why the poorer soft labels still yielded a performance improvement. While part of the answer is provided in the literature on soft/noisy labels [9,14,23,26], we also speculated that the soft labels

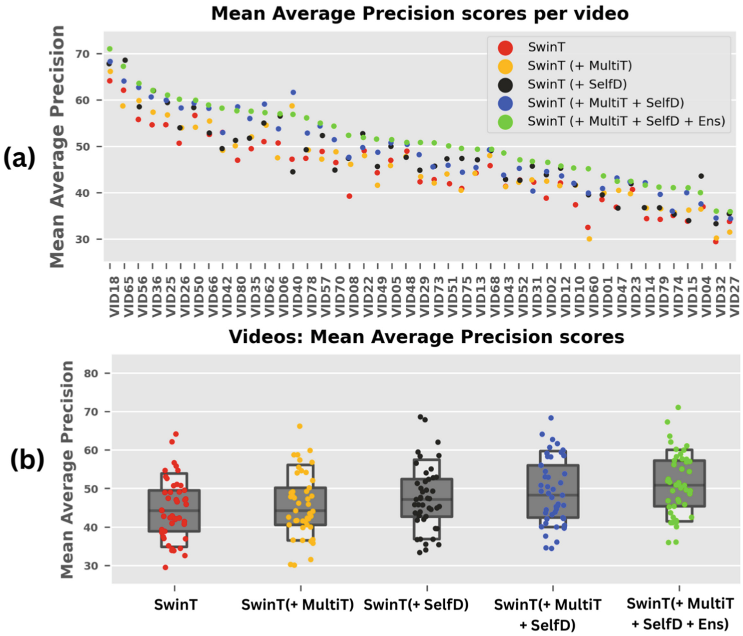

Fig. 3. Quantitative results on the validation data. (a) mean Average Precision (mAP) plotted separately for each video for five configurations of our method with increasing complexity. A Swin Transformer (SwinT) was gradually complemented by multi-task learning (MultiT), self-distillation (SelfD), and ensembling (Ens). Videos were sorted by mAP score of final ensemble performance from highest to lowest. (b) Corresponding dot- and boxplots of mAP scores, aggregated over all videos

may address the issue of ambiguous/erroneous labels in our particular use case. More specifically, we assumed that if the teacher model increases the probability of semantically close triplets in the soft labels, it could lead to an enhanced level of confidence in the student model's prediction of the ground truth, potentially accounting for the observed performance improvement. To investigate whether this is the case, we first defined a pragmatic proxy metric for semantic similarity: the number of identical triplet items (max: two for different triplets). We then selected all frames with only one unique triplet label and retrieved the top five triplets (excluding the reference) with the highest soft label score. Figure 4, depicts an example of such a comparison. The reference triplet "bipolar, dissect, cystic_plate" is shown with five soft label triplets ranked by probability. In the example, the top five triplets share an average of 1.6 components with the reference, indicating that they contain similar semantic information. We found that over all samples, the average number of component matches between reference and top five triplets is 1.0±0.002. In contrast, when comparing the reference with five triplets randomly drawn (while respecting prevalence), the agreement is 0.5±0.002.

Table 1. Main quantitative results Starting from our backbone model - a Swin Transformer (Swin T) - we gradually added individual components, namely multi-task learning (MultiT), self-distillation (SelfD), and ensembling (Ens). Each component addition leads to an increase in mAP and top-5 accuracy, in both cross-validation (left) and independent external validation (right).

	Cross Validation		External validation (CholecTriplet 2022 test set)	
	Triplet mAP [%]	Top-5 accuracy [%]	Triplet mAP [%]	Top-5 accuracy [%]
Rendezvous	29.4	79.3	32.7	69.4
Ours: SwinT	32.3	83.8	32.9	70.7
Ours: SwinT + MultiT	33.1	84.6	33.8	71.6
Ours: SwinT + SelfD	35.0	85.2	36.1	72.9
Ours: SwinT + MultiT + SelfD	36.1	86.2	37.3	73.3
Ours: SwinT + MultiT + SelfD + Ens	**38.5**	**86.5**	**37.4**	**74.0**

This shows that self-distillation specifically leads to increased scores for semantically related classes.

Independent External Validation. External validation was conducted on the CholecTriplet challenge test set. The results, shown in Table 1, confirm the results from cross-validation experiments, with the final ensemble scoring 37.4% in mAP. This equals an absolute improvement of 4.7pp and a relative improvement of 14.4% compared to the Rendezvous benchmark (mAP= 32.7%). With a previous version of the method presented in this paper (scoring slightly lower) we won the challenge in 2022 as the only team that explored the concept of self-distillation.

4 Discussion

This paper pioneers the concept of self-distillation in the medical image analysis domain. Specifically, we are the first to tackle key challenges in surgical action recognition, namely the high number of classes and class imbalance, with self-distillation. Comprehensive ablation studies combined with external validation yielded the following findings:

1. Swin Transformers, as recently introduced by the computer vision community, can serve as a strong backbone in endoscopic vision tasks. This is suggested by the fact that even our most ablated model, consisting of a single SwinT, surpasses the state-of-the-art surgical action recognition method *Rendezvouz*.
2. Multi-task learning, here using the classification of instrument, verb, and target as well as of the surgical phase as auxiliary tasks, yielded a notable increase in performance.
3. Self-distillation yielded the biggest boost in performance, suggesting that soft labels are better suited for surgical action recognition.
4. Ensembling increased performance further, as also suggested by various publications in a wide range of fields.

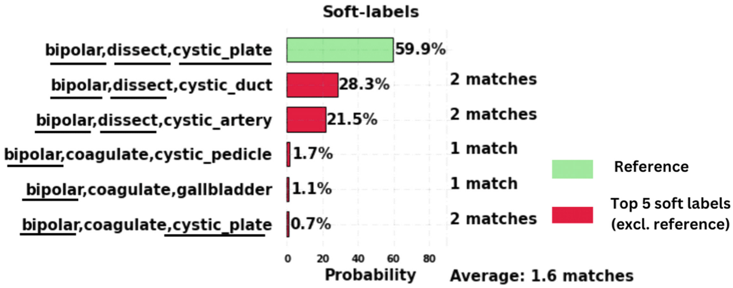

Fig. 4. Example of generated soft labels with reference (in green) and top 5 soft label triplets ranked by probability (in red). Average number of component matches between reference "bipolar, dissect, cystic_plate" and top five triplets is 1.6.

Overall, the addition of self-distillation (in combination with the SwinT as a backbone) resulted in the highest performance boost. While label noise has been shown to be beneficial in various work [9,14,23,26], the concept of self-distillation may not necessarily be intuitive; although the mAP achieved by the teacher model, trained on hard labels, is sub-optimal (32%), the teacher's noisy labels lead to an overall improvement in performance when compared to the (presumably better-quality) hard labels. In the general machine learning literature, the knowledge encoded in noisy labels is referred to as "dark knowledge" because it is not yet well-understood. Aiming to shed light on this topic, our experiments on semantic similarity suggest that soft labels may actually address the issue of ambiguous/erroneous labels. Further analyses with more sophisticated metrics for semantic similarity are, however, needed to support this finding.

Related work has so far tackled the challenge of surgical action recognition with various strategies including multi-task learning [16,21], and different attention mechanisms [3,19] incorporated into diverse architectures based on temporal convolutional networks [2,21], transformers [3,5,19], or combinations of convolutional neural networks (CNN) with recurrent neural networks (RNN) [7,8,16,25] or hidden Markov models (HMM) [22]. While our approach was particularly successful according to the challenge analysis, the overall performance is still not optimal. Advancing the methods will require more data that features a sufficient number of samples for each triplet and captures the full variability of scenes that might be encountered in practice. From a methodological perspective, future work should be directed to efficiently taking temporal context into account and addressing potential domain shifts [1].

In conclusion, our study is the first to demonstrate the benefit of self-distillation for surgical vision tasks. Based on the substantial performance boost obtained, the usage of soft labels could become a valuable tool in the endoscopic vision community.

Acknowledgements. This project was supported by a Twinning Grant of the German Cancer Research Center (DKFZ) and the Robert Bosch Center for Tumor Diseases

(RBCT). Part of this work was funded by HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator and the Helmholtz Association under the joint research school ”HIDSS4Health - Helmholtz Information and Data Science School for Health” and by French state funds managed within the Plan Investissements d’Avenir by the ANR under references: National AI Chair AI4ORSafety [ANR-20-CHIA-0029-01], Labex CAMI [ANR-11-LABX-0004], DeepSurg [ANR-16-CE33-0009], IHU Strasbourg [ANR-10-IAHU-02] and by BPI France under references: project CONDOR, project 5G-OR. Model Docker evaluation were performed with servers/HPC resources managed by CAMMA, IHU Strasbourg, Unistra Mésocentre, and GENCI-IDRIS [Grant 2021-AD011011638R1, 2021-AD011011638R2, 2021-AD011011638R3].

References

1. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nat. Commun. **11**(1), 3673 (2020)
2. Czempiel, T., et al.: TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 343–352. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_33
3. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: OperA: attention-regularized transformers for surgical phase recognition. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 604–614. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_58
4. Eisenmann, M., et al.: Biomedical image analysis competitions: The state of current participation practice. arXiv preprint arXiv:2212.08568 (2022)
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: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 593–603. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_57
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
7. Jin, Y., et al.: Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging **37**(5), 1114–1126 (2017)
8. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. IEEE Trans. Med. Imaging **40**(7), 1911–1923 (2021)
9. 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)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (October 2021)
12. Maier-Hein, L., et al.: Surgical data science-from concepts toward clinical translation. Med. Image Anal. **76**, 102306 (2022)
13. MICCAI SIG for Challenges: MICCAI registered challenges (2022). https://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/
14. Mobahi, H., Farajtabar, M., Bartlett, P.: Self-distillation amplifies regularization in hilbert space. Adv. Neural. Inf. Process. Syst. **33**, 3351–3361 (2020)

15. Nwoye, C.I., et al.: Cholectriplet 2021: a benchmark challenge for surgical action triplet recognition. arXiv preprint arXiv:2204.04746 (2022)
16. Nwoye, C.I., et al.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 364–374. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_35
17. Nwoye, C.I., Padoy, N.: Data splits and metrics for benchmarking methods on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)
18. Nwoye, C.I., Padoy, N.: Surgical action triplet detection 2022 (2022). https://cholectriplet2022.grand-challenge.org/
19. Nwoye, C.I., et al.: Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med. Image Anal. **78**, 102433 (2022)
20. Nwoye, C.I., , et al.: Cholectriplet 2022: show me a tool and tell me the triplet-an endoscopic vision challenge for surgical action triplet detection. arXiv preprint arXiv:2302.06294 (2023)
21. Ramesh, S., et al.: Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. Int. J. Comput. Assist. Radiol. Surg. **16**(7), 1111–1119 (2021). https://doi.org/10.1007/s11548-021-02388-z
22. 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 Trans. Med. Imaging **36**(1), 86–97 (2016)
23. Vu, D.Q., Le, N., Wang, J.C.: Teaching yourself: a self-knowledge distillation approach to action recognition. IEEE Access **9**, 105711–105723 (2021)
24. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
25. 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)
26. Yun, S., Park, J., Lee, K., Shin, J.: Regularizing class-wise predictions via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13876–13885 (2020)