Smarter Self-Distillation: Optimizing the Teacher for Surgical Video Applications

Amine Yamlahi[12], Piotr Kalinowski[136], Patrick Godau[1236], Rayan Younis[4], Martin Wagner[4], Beat Müller[5], and Lena Maier-Hein[123678]

[1] German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Germany m.elyamlahi@dkfz-heidelberg.de
[2] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and university medical center Heidelberg, Germany
[3] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4] Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany.
[5] University Digestive Healthcare Center Basel, Basel, Switzerland
[6] Faculty of Mathematics and Computer Science, Heidelberg University, Germany
[7] Medical Faculty, Heidelberg University, Germany
[8] Heidelberg University Hospital, Surgical Clinic, Surgical AI Research Group, Heidelberg, Germany

Abstract. Surgical workflow analysis poses significant challenges due to complex imaging conditions, annotation ambiguities, and the large number of classes in tasks such as action recognition. Self-distillation (SD) has emerged as a promising technique to address these challenges by leveraging soft labels, but little is known about how to optimize the quality of these labels for surgical scene analysis. In this work, we thoroughly investigate this issue. First, we show that the quality of soft labels is highly sensitive to several design choices and that relying on a single top-performing teacher selected based on validation performance often leads to suboptimal results. Second, as a key technical innovation, we introduce a multi-teacher distillation strategy that ensembles checkpoints across seeds and epochs within a training phase where soft labels maintain an optimal balance—neither underconfident nor overconfident. By ensembling at the teacher level rather than the student level, our approach reduces computational overhead during inference. Finally, we validate our approach on three benchmark datasets, where it demonstrates consistent improvements over existing SD methods. Notably, our method sets a new state-of-the-art (SOTA) performance on the CholecTriplet benchmark, achieving a 43.1% mean Average Precision (mAP) score and real-time inference time, thereby establishing a new standard for surgical video analysis in challenging and ambiguous environments. Code available at https://github.com/IMSY-DKFZ/self-distilled-swin.

Keywords: Self-Distillation · Surgical Action Recognition · Soft labels optimization.

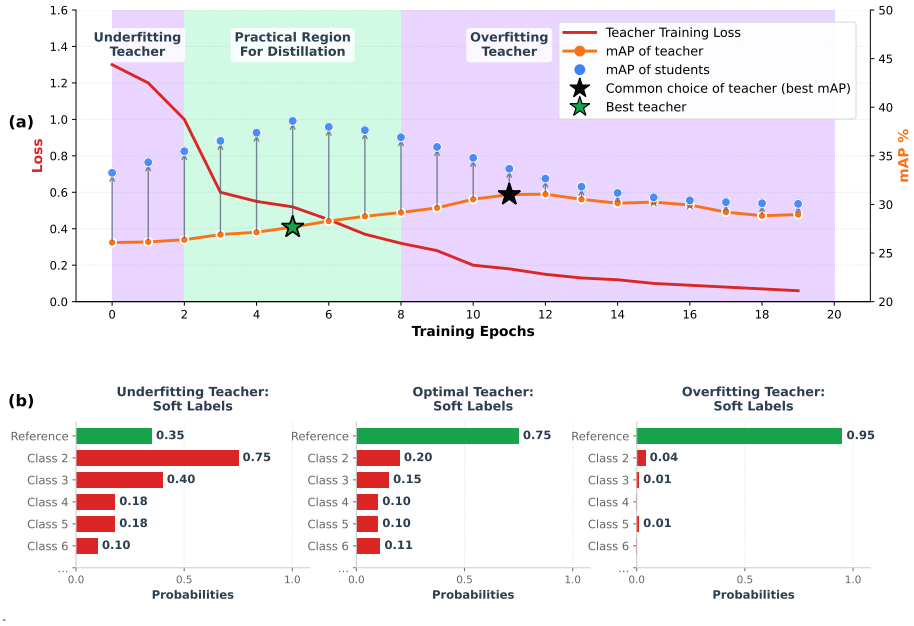

Fig. 1: **Self-distillation quality depends crucially on teacher selection.** The performance of the student models trained with the teacher's labels of a certain epoch are represented as blue circles. While the teacher for SD is commonly chosen based on the best validation performance (maximum on orange curve: black star), the actual best teacher (green star) corresponds to a different region on the loss curve. The green region represents the sweet spot for generating soft labels because the teacher is neither too uncertain (left) nor overconfident (right). (b) Representative soft labels corresponding to the three regions in (a): uncertain (left), practical (middle), and overconfident (right).

1 Introduction

Surgical workflow analysis has arisen as a fundamental prerequisite for various surgical artificial intelligence (AI) applications such as surgical quality assessment, context-aware information retrieval, and cognitive robotics. However, surgical videos come with various challenges such as poor contrast, artifacts (e.g., blur, bleeding, smoke), and limited view. Often, video analysis problems come with high ambiguity, also due to a potentially high number of similar classes [13,18]. Surgical action triplet recognition with a high number of up to 100 classes and extremely high class imbalance, exemplifies these challenges, prompting researchers to explore diverse approaches, including spatio-temporal modeling [8,4], attention mechanisms [14], tail-aware methods [3], knowledge distillation (KD) [4] and self-distillation (SD) [17]. Among these, SD has emerged as a

particularly successful technique for addressing ambiguity through its innovative use of soft labels. KD includes various model compression and knowledge transfer techniques from larger teacher to smaller student models [5]. SD, in contrast, keeps the same model architecture for teacher and student. These methods exhibit diverse transfer mechanisms, including transfers across layers, features, logits, or learned embeddings [2]. Our work focuses specifically on SD via logits-level knowledge transfer.

Recent work in surgical action triplet recognition, SD-Swin [17] has demonstrated the potential of SD to address high label ambiguity and class imbalance. However, this work left a critical gap in understanding how to optimize the soft labels used for distillation. Current approaches typically select the teacher model based on optimal validation set performance, but we argue that this strategy may not produce the most effective teacher in practice (Fig. 1). Our work addresses this fundamental knowledge gap through three core contributions:

1. Sources of variability in the quality of soft labels: We show that the quality of soft labels in SD depends crucially on the selection of the teacher. Importantly, they are highly sensitive to a number of design choices, including the epoch at which the teacher was saved and variations in random seeds or hardware.

2. Teacher optimization: We propose a new approach to compile an informative teacher in SD, based on three principles. (i) Teacher Selection: Recognizing that the effectiveness of SD varies over time, with early epochs producing noisy labels and later epochs generating overconfident, hard-label-like outputs (Fig. 1), we select teacher checkpoints based on resulting student model performance in cross-validation. (ii) Multi-Teacher Strategy: To address the high sensitivity of SD, we propose a novel multi-teacher strategy that aggregates soft labels from multiple teachers trained under varied configurations. (iii) Temporal Decoder: Only after SD we add a temporal decoder to complement our best student with temporal information.

3. New state-of-the-art (SOTA) model in surgical action triplet recognition: We validate our approach through extensive experimentation on three benchmark datasets, demonstrating consistent improvements over existing SD methods.

2 Methods

2.1 Multi-teacher self-distillation framework

Our approach builds on two key observations: Firstly, we observed that the prevailing approach of choosing the teacher with the best validation performance yields suboptimal, overconfident labels (Fig. 1). Secondly, soft labels feature limited robustness to a number of factors such as the non-determinism of neural networks (Fig. 2). We leverage these findings for a new approach to SD based on multiple teachers (Stage 1). Our framework integrates spatial learning through multi-teacher SD with temporal learning via transformer-based sequence modeling (Stage 2).

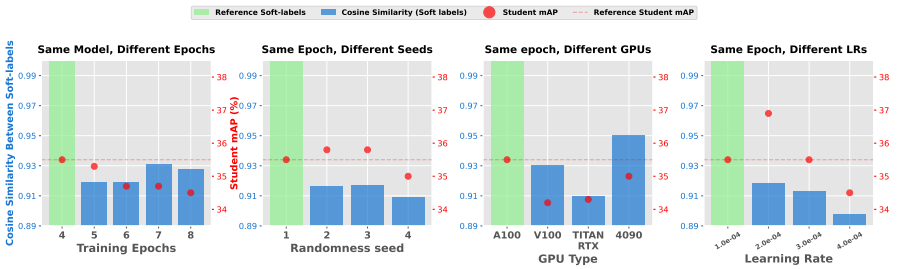

Fig. 2: **Teacher training conditions impact soft label quality.** Soft labels generated by a teacher model trained with specific parameters (epoch 4, seed 1, A100 GPU, learning rate 1e-04) serve as the reference point. The plots show cosine similarity between reference soft labels (green) and those generated under different conditions (blue), alongside corresponding student performance (red).

Stage 1: Spatial Learning employs multiple independently trained Swin transformer [10] teacher models with varying configurations selected within an optimal region of the training curve (Fig. 3). The soft labels produced by the set of teachers are then aggregated (here: across multiple epochs and random seeds). Finally, a student model is trained with the aggregated soft labels to generate frame-level embeddings and predictions.

Stage 2: To achieve temporal learning, consecutive frames are stacked to form temporal sequences (Fig. 4). A temporal decoder, comprising positional encoding, transformer encoder, and classifier layers then processes these sequences to generate final predictions that incorporate both spatial and temporal information.

Implementation details: The spatial component uses a Swin transformer backbone for both teacher and student models, configured as in SD-Swin [17]. Two variants of this backbone are used in our experiments: Swin-Base and Swin-Large. The temporal component employs a transformer encoder with 4 heads and 3 layers, with 1D embedding dimensions of 1024 (Swin-Base) and 1536 (Swin-Large). Final predictions are generated by averaging outputs from (i) the spatial encoder and (ii) the temporal decoders at multiple time resolutions of 4, 6, and 9 consecutive frames. Training duration on the CholecT45 dataset using the NVIDIA 4090 GPU was 9 hours for Swin-Base variants and 16 hours for Swin-Large variants.

2.2 Validation methodology

Datasets Our method was developed and validated on the task of surgical action triplet recognition (100 classes) using the CholecT45 dataset with the official five-fold cross-validation split [13]. Additionally, we used the independent test set from the CholecTriplet2021 [12] challenge to evaluate our approach. To assess generalizability, we further tested our framework without changing any of the hyperparameters except for the selected teacher epochs, on two independent

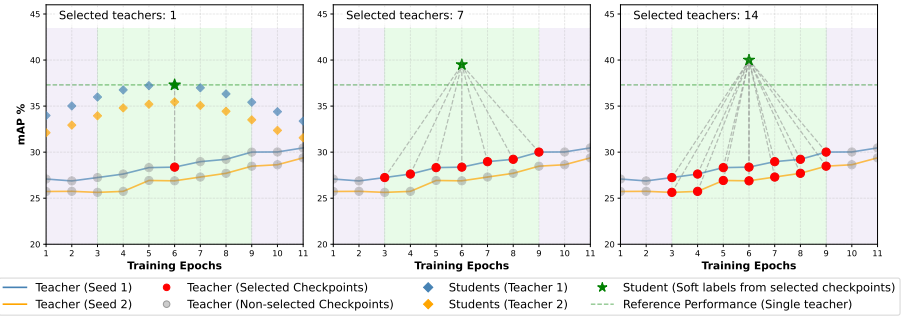

Fig. 3: **Ensembling of multiple teachers yields performance boosts (sketch).** All plots depict multiple checkpoints (red and grey circles) from two differently seeded teacher models (blue and orange curve). (Left) The corresponding mAP of the student models (blue and orange diamonds) for each teacher checkpoint and the best student mAP (green star) when using the optimal checkpoint from one teacher model for distillation; (center) Intermediate-Epochs-Ensemble (IEE): Combining soft labels from 7 different epochs of a single teacher model; (right) Intermediate-Epoch-Seed-Ensemble (IESE): Combining soft labels from 7 epochs each of teacher models trained with different random seeds.

surgical datasets: Action triplet recognition in cholecystectomy procedures (88 classes) using the HeiCholeActivity dataset [18]. We followed the recommended cross-validation split by the authors. Furthermore, we evaluated on action-target recognition in prostatectomy procedures (21 classes) using the SARAS-ESAD dataset following the training, validation and test split [1]. For this dataset, we parsed classification labels from the detection dataset. We validated the performance using the mAP metric implementation used in the CholecTriplet2021 challenge [13].

Experimental design The purpose of the experiments was to investigate the following research questions:

RQ1: What are the key factors contributing to performance variability in surgical SD? We investigated key factors in performance variability by analyzing soft label variability across multiple conditions (Fig. 2). These conditions included training duration spanning epochs 4-12, five different random initialization seeds, hardware variations, and learning rate variations at discrete values of 1e-4, 2e-4, 3e-4, and 4e-4. For our measurements, we established a reference baseline using soft labels generated by a teacher model trained with specific parameters (epoch 4 checkpoint, seed 1, A100 GPU, learning rate 1e-4). We then evaluated both the mean cosine similarity between different soft labels and the corresponding student model mAP performance.

RQ2: How can soft labels be optimized to enhance the effectiveness of SD? We explored soft label optimization through four progressive

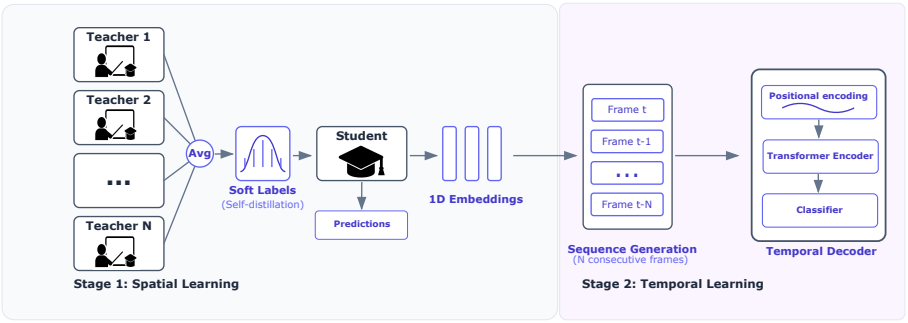

Fig. 4: **Multi-teacher self-distillation framework with two-stage spatial-temporal modeling.** The pipeline consists of SD from an ensemble of teacher models into a single student model (Stage 1: Spatial Learning), followed by temporal processing of consecutive frame embeddings using a transformer architecture (Stage 2: Temporal Learning).

teacher selection strategies. **Fully-Converged-Teacher (FCT)** selects the checkpoint with the best validation performance according to the mAP metric, while **Intermediate-Teacher (IT)** selects the single checkpoint producing best student performance – in our implementation through brute force grid search.

Building on the IT checkpoint, we define the **Practical Distillation Region (PDR)** as the range of epochs centered around the IT epoch. It's a window of epochs extending a certain number of steps both before and after the identified IT epoch. This range captures the most effective teacher checkpoints for distillation.

The radius of this window, is represented by δ. The value of δ is influenced by factors like the training configuration (such as the learning rate and total number of epochs) and the specific characteristics of the dataset being used. Following the SD-Swin [17] training setup, we empirically determined $\delta = 3$, yielding $|\text{PDR}| = 7$ teacher checkpoints.

Intermediate-Epochs-Ensemble (IEE) averages soft labels from multiple checkpoints within the PDR while Intermediate-Epoch-Seed-Ensemble (IESE) combines soft labels from multiple checkpoints and initialization seeds. We opted for averaging over 3 seeds and 7 epochs. Additionally, we enhance our IESE student model with a temporal decoder, forming our final proposed model, Temporal Optimized Distillation (TOD). The TOD-Base variant is trained with the Swin-Base backbone, while the TOD-Large variant is an ensemble of two IESE student models trained with the Swin-Base and Swin-Large backbones [10].

RQ3: Do optimized soft labels consistently improve performance across different surgical workflow analysis tasks? We developed our framework on the cross-validation dataset of CholecT45 then applied it to the corresponding test set plus two additional independent datasets, represented by the datasets HeiCholeActivity (Action Triplet) and the SARAS-ESAD (Action-Target) (see sec. 2.2).

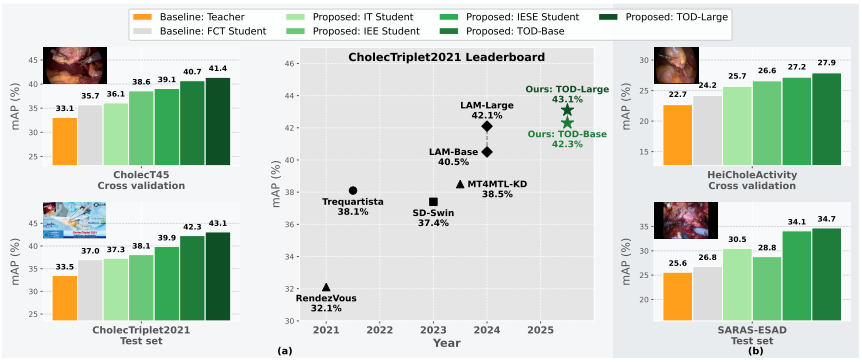

Fig. 5: **Performance gains are harmonious over three different datasets.** Intermediate-Teacher (IT) consistently outperform Fully-Converged-Teacher (FCT) across all datasets, with Temporal Optimized Distillation (TOD) approaches showing the highest gains. (a) Our TOD-Base (42.3%) and TOD-Large (43.1%) achieve new SOTA results on the CholecTriplet2021 benchmark. (b) Our conclusions hold for two independent datasets.

3 Results

RQ1: As depicted in Fig. 2, our analysis revealed significant variations in soft label distributions across all experimental conditions, with mean cosine similarities ranging from 0.89 to 0.96 relative to reference soft labels. The impact of this variability extends to student model performance, with mAP scores showing fluctuations of $\pm$1.4-1.5% across different training conditions, even when modifying only a single factor.

RQ2: According to our ablation study (Fig. 5) all of our key design choices come with a performance boost. Overall, our key observations can be summarized as follows: Firstly, there is a sweet spot occurring across specific epochs along the training of a teacher for SD, which we call the "practical region" (see Fig. 1). This region corresponds to predicted soft labels that are not overfitted, but already contain sufficient "dark knowledge" [5]. While the exact practical window is dependent on the chosen training conditions, it is not necessary to identify the single training checkpoint that produces the best soft labels (IT). By contrast, it is even beneficial to combine a variety of predictions that come from checkpoints within the practical window (IEE). This reduces the efforts that would be necessary to identify the exact optimal point on the loss curve. To prevent including checkpoints outside the practical region it is furthermore beneficial to introduce slight training variations to the teacher (see Fig. 3). Ensembling the resulting variations (IESE) can produce soft labels that further increase performance. Finally, feeding the sequences of image embeddings into a temporal decoder, provides an additional performance boost by effectively integrating both spatial and temporal information (TOD).

The impact of our approach is particularly evident in the CholecTriplet2021 leaderboard results (Fig. 5, a), where our TOD-Large model achieves SOTA performance at 43.1% mAP, with the base model variant reaching 42.3% mAP. These results surpass the previous best method (LAM-Large [8]) which achieved 42.1% mAP.

RQ3: Despite being developed initially on a single dataset, our teacher selection methods demonstrate highly consistent performance patterns across all datasets and (Fig. 5, b). The performance gain of our complete model (TOD-Base) achieves substantial improvements over the baseline (FCT): +3.7 mAP on HeiCholeActivity (24.2%, 27.9%); and +7.9 mAP on SARAS-ESAD (26.8%, 34.7%).

4 Discussion

With this paper, we challenged the prevailing assumption that choosing the teacher with the best validation performance yields optimal soft labels for SD in surgical scene analysis. We (1) uncovered the high variability in the quality of soft labels in SD and (2) proposed an advanced teacher selection strategy that aggregates soft labels from multiple complementary teachers. Strikingly, (3) our methodology leads to consistent improvement patterns across a range of datasets (Fig. 5).

Our research integrates smoothly into the state of the art outside the field of medical imaging AI. While some works suggest that soft labels act as a regularizer by providing smoothed target distributions that improve generalization by offering richer information about class similarities [19], others have shown their ability to recover useful information from corrupted labels [9] or capture the teacher's uncertainty in ambiguous cases [17]. Moreover, recent findings in KD [15] with natural images (CIFAR-100 [6] and Tiny ImageNet [7]) indicate that the best teacher is not necessarily the fully trained model; instead, intermediate teachers often lead to better student performance. We extended these findings to SD and demonstrated that intermediate teachers with inferior validation performance outperform fully trained teachers with optimal validation performance (Fig. 5). Our analysis complements Wang et al.'s work by focusing on the soft labels, showing that intermediate soft labels in surgical action triplet recognition capture the teacher's uncertainty, while those from fully trained teachers exhibit overfitting, resembling hard labels. This overfitting diminishes the effectiveness of SD, explaining the benefits of selecting intermediate teachers.

Building on this insight, we recognized that even the optimal intermediate teacher is still in a converging state, suggesting that ensembling multiple teachers from this optimal region could yield superior performance (Fig. 3). While ensemble methods have been explored in KD using complementary teachers [11], these approaches prioritize diversity without systematic teacher selection that may include suboptimal teachers. Recent work [16] introduced intermediate teacher ensembling in traditional KD by combining features through self-attention mechanisms. We extend this concept to SD with a crucial distinction: our approach

performs targeted ensembling at the soft-label level within the PDR. Unlike diversity-focused methods, our two-stage approach first identifies optimal teachers, then strategically ensembles them to capture both intra-teacher knowledge (temporal evolution within the PDR) and inter-teacher knowledge (variations across random seeds). This principled selection-then-ensemble strategy ensures that we aggregate high-quality soft labels rather than blindly combining diverse but potentially suboptimal teachers. One limitation of our approach is the need to determine the practical region which - so far - requires brute-force search. However, our method offers notable efficiency advantages at inference time because it eliminates the need for ensembling multiple student models. Compared to SD-Swin, which distills knowledge from three separately trained teachers into three corresponding students and ensembles the students at inference to achieve 37.4% accuracy at 5 fps, our IESE variant distills knowledge from 21 teachers into a single student, achieving 39.8% accuracy at 14 fps on a NVIDIA 4090 GPU.

Future work can build upon our qualitative observations on soft label distributions throughout teacher training by developing quantitative measures to identify the practical region without requiring exhaustive student training. Additionally, exploring Parameter-Efficient Fine-Tuning (PEFT) methods could be a promising direction to mitigate the extensive search for optimal teachers. This approach has already been shown to be effective in surgical action triplet recognition, as demonstrated by LAM-Large [8]. In conclusion, we have introduced a novel multi-teacher SD approach that enhances neural network performance, with potential applications extending beyond surgical workflow analysis to other complex domains.

Acknowledgments. Funding was provided by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health" and the Surgical AI Hub Germany and the National Center for Tumor Diseases (NCT), Heidelberg's Surgical Oncology Program.

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

References

1. Bawa, V.S., Singh, G., KapingA, F., Skarga-Bandurova, I., Oleari, E., Leporini, A., Landolfo, C., Zhao, P., Xiang, X., Luo, G., et al.: The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. arXiv preprint arXiv:2104.03178 (2021)
2. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
3. Gui, S., Wang, Z.: Tail-enhanced representation learning for surgical triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 689–699. Springer (2024)

4. Gui, S., Wang, Z., Chen, J., Zhou, X., Zhang, C., Cao, Y.: Mt4mtl-kd: a multi-teacher knowledge distillation framework for triplet recognition. IEEE Transactions on Medical Imaging **43**(4), 1628–1639 (2023)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
6. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images.(2009) (2009)
7. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
8. Li, Y., Bai, B., Jia, F.: Parameter-efficient framework for surgical action triplet recognition. International Journal of Computer Assisted Radiology and Surgery **19**(7), 1291–1299 (2024)
9. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: Proceedings of the IEEE international conference on computer vision. pp. 1910–1918 (2017)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
11. Morocutti, T., Schmid, F., Koutini, K., Widmer, G.: Creating a good teacher for knowledge distillation in acoustic scene classification. arXiv preprint arXiv:2503.11363 (2025)
12. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. arXiv preprint arXiv:2204.04746 (2022)
13. Nwoye, C.I., Padoy, N.: Data splits and metrics for benchmarking methods on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)
14. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
15. Wang, C., Yang, Q., Huang, R., Song, S., Huang, G.: Efficient knowledge distillation from model checkpoints. Advances in Neural Information Processing Systems **35**, 607–619 (2022)
16. Wang, C., Zhang, S., Song, S., Huang, G.: Learn from the past: Experience ensemble knowledge distillation. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 4736–4743. IEEE (2022)
17. Yamlahi, A., Tran, T.N., Godau, P., Schellenberg, M., Michael, D., Smidt, F.H., Nölke, J.H., Adler, T.J., Tizabi, M.D., Nwoye, C.I., et al.: Self-distillation for surgical action recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 637–646. Springer (2023)
18. Younis, R., Yamlahi, A., Bodenstedt, S., Scheikl, P., Kisilenko, A., Daum, M., Schulze, A., Wise, P., Nickel, F., Mathis-Ullrich, F., et al.: A surgical activity model of laparoscopic cholecystectomy for co-operation with collaborative robots. Surgical Endoscopy **38**(8), 4316–4328 (2024)
19. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3903–3911 (2020)