CurConMix: A Curriculum Contrastive Learning Framework for Enhancing Surgical Action Triplet Recognition

Yongjun Jeon[1 [0009−0004−6737−9817]], Jongmin Shin[2 [0009−0009−8337−7260]], Seonmin Park[2 [0009−0003−9211−0476]], Bogeun Kim[2 [0009−0005−8693−7295]], Kanggil Park[2 [0009−0009−5525−3869]], Namkee Oh[2 [0000−0002−6594−8973]], and Kyu-Hwan Jung[1,3 [0000−0002−6626−6800]]

[1] Department of Medical Device Management and Research, Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Seoul 06355, Republic of Korea
kyuhwanjung@gmail.com
[2] Department of Surgery, Samsung Medical Center, Seoul 06351, Republic of Korea
namkee.oh@samsung.com
[3] Smart Healthcare Research Institute, Research Institute for Future Medicine, Samsung Medical Center, Seoul 06351, Republic of Korea

Abstract. Accurately recognizing surgical action triplets in surgical videos is crucial for advancing context-aware systems that deliver real-time feedback, enhancing surgical safety and efficiency. However, recognizing surgical action triplets ⟨instrument, verb, target⟩ is challenging due to subtle variations, complex interdependencies, and severe class imbalance. Most existing approaches focus on individual triplet components while overlooking their interdependencies and the inherent class imbalance in triplet distributions. To address these challenges, we propose a novel framework, **Cur**riculum **Con**trastive learning with feature **Mix**up (**CurConMix**). During pre-training, we employ curriculum contrastive learning, which progressively captures relationships among triplet components and distinguishes fine-grained variations through hard pair sampling and synthetic hard negative generation. In the fine-tuning stage, we further refine the model using self-distillation and mixup strategies to alleviate class imbalance. We evaluate our framework on the CholecT45 dataset using 5-fold cross-validation. Experimental results demonstrate that our approach surpasses existing methods across various model sizes and input resolutions. Moreover, our findings underscore the importance of capturing interdependency among triplet components, highlighting the effectiveness of our proposed framework in addressing key challenges in surgical action recognition. The official implementation is available at https://github.com/MIDAS-SurgAI/CurConMix.

Keywords: Surgical Video · Endoscopic Surgery · Surgical Action Triplet Recognition · Curriculum Learning · Contrastive Learning · Class Imbalance.

1 Introduction

Digital storage of surgical videos, made possible by endoscopic technology, has paved the way for deep learning applications with notable progress [3,4,1,11]. Context-aware deep learning systems provide real-time feedback, enhancing both surgical safety and efficiency [12]. Surgical actions are commonly represented as triplets—⟨instrument, verb, target⟩[13], but triplet recognition remains challenging due to subtle distinctions among triplets, complex interdependencies, multi-label classification, class imbalance, and limited training data. Existing approaches address each component independently [17,13,15,9,20], overlooking interdependencies and failing to adequately handle class imbalance. Although SelfD [21] mitigates class imbalance via self-distillation, it struggles to distinguish subtle differences between triplets. Similarly, TERL [5] focuses on tail classes with instance-level contrastive learning but does not fully consider triplet relationships.

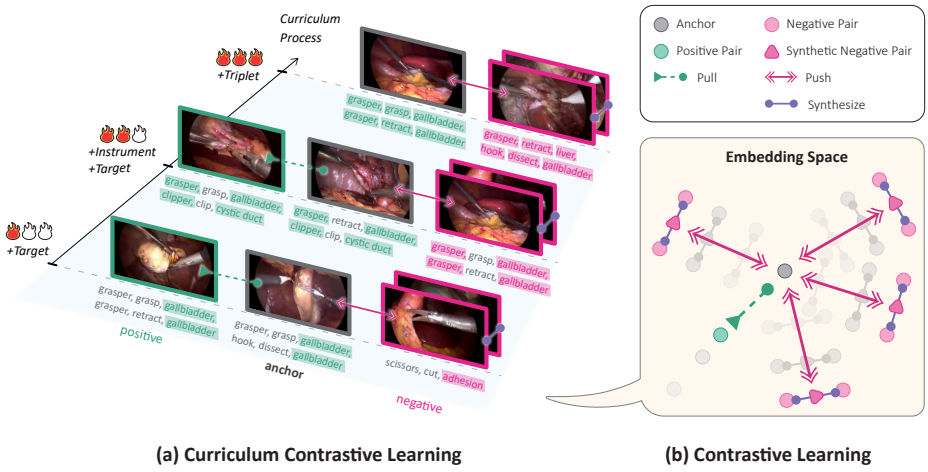

Fig. 1. Overview of CurConMix. (a) The curriculum process sequentially incorporates ⟨target⟩, ⟨instrument, target⟩, and ⟨instrument, verb, target⟩. A pair is positive only if all components match; otherwise, it is negative. (b) Hard pair sampling and feature mixup generate challenging negative examples, driving the model to pull positive pairs closer while pushing negative pairs farther apart.

To address these challenges, we propose a novel framework, **Cur**riculum **Con**trastive learning with feature **Mix**up (**CurConMix**), as illustrated in Fig. 1. Through curriculum learning [18], the model gradually trains from simple to complex tasks, enabling it to learn the interdependencies among triplet components (Fig. 1(a)). This is combined with supervised contrastive learning [8] to capture subtle differences between triplets and improves data diversity by generating various pairwise combinations within a constrained dataset. To

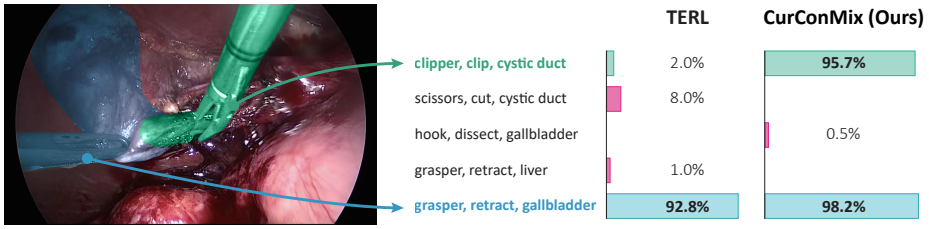

Fig. 2. Prediction comparison between the TERL method and our framework on frames containing both majority (⟨grasper, retract, gallbladder⟩) and minority (⟨clipper, clip, cystic duct⟩) classes.

facilitate effective learning, we introduce a hard pair sampling strategy based on similarity to anchor features in the embedding space and label information. Rather than using these hard negative samples directly in training, we create synthetic hard negative feature vectors through convex linear combinations in embedding space as shown in Fig. 1(b). During fine-tuning, the model incorporates self-distillation [21] and mixup [22], leveraging soft labels and further addressing class imbalance.

As a result, our framework achieves superior performance on the CholecT45 dataset [14] compared to existing methods, as validated through the official 5-fold evaluation. Notably, it consistently outperforms existing methods across various backbone sizes and resolutions. As shown in Fig. 2, severe class imbalance leads to model overconfidence in the majority class, making predictions on the minority class challenging. This challenge intensifies when the majority and minority classes appear simultaneously within the same frame. Under these conditions, considering both inter- and intra-triplet relationships makes the precise recognition of triplets increasingly difficult. Our approach effectively captures complex triplet interdependencies, mitigates class imbalance, and distinguishes subtle triplet variations, addressing critical challenges in surgical action recognition.

In summary, the contributions of our work are as follows:

1. **CurConMix** models interdependencies among triplet components via curriculum learning and identifies subtle triplet differences through contrastive learning. Additionally, hard negative sampling and synthetic hard negatives enhance representation learning by focusing on challenging instances and increasing data diversity.
2. **CurConMix** tackles the challenges of limited training samples and severe class imbalance by generating informative samples through hard pair sampling and synthetic hard negatives, also fine-tuning further refines feature representations using self-distillation and mixup with soft-label effects. This leads to more robust predictions, particularly for minority classes.
3. Our proposed framework, **CurConMix**, achieves **state-of-the-art** results on the CholecT45 dataset using official 5-fold cross-validation.

2 Methods

2.1 CurConMix

In this section, we introduce the pre-training stage of our framework. First, we provide a detailed explanation of Curriculum Contrastive Learning, which integrates curriculum learning with contrastive learning to progressively learn the relationships among triplets in a multi-label dataset. This approach enables the model to capture subtle differences between triplets more effectively. Second, we describe the proposed hard pair sampling strategy, designed to improve contrastive learning by effectively selecting informative pairs. Lastly, we leverage hard negative feature vectors in contrastive learning to enhance the model's ability to capture subtle differences between triplets while enforcing the learning of more robust feature representations.

Curriculum Contrastive Learning This curriculum learning process follows the sequence: ⟨target⟩, ⟨instrument, target⟩, and ⟨instrument, verb, target⟩, as shown in Fig. 1(a). According to [12], performance on ⟨target⟩ is the lowest among the triplet components. Furthermore, since the ⟨verb⟩ representing the action can be determined only after identifying both the ⟨instrument⟩ and ⟨target⟩, the curriculum process was designed in this order, which also achieved the best performance among various curriculum designs.

To improve the model's ability to capture subtle distinctions between triplet classes, we integrate contrastive learning with curriculum learning. Label information is used for pair generation, ensuring that a pair is positive only when all labels match exactly, as illustrated in Fig. 1(a). The following sections provide a detailed explanation of sampling hard pairs for contrastive learning.

Hard Pair Sampling In contrastive learning, the selection of training pairs plays a crucial role in determining model performance [7]. Therefore, for effective supervised contrastive learning, we propose a hard pair sampling strategy that leverages feature vectors and label information to select challenging positive and negative samples. Hard pair sampling is performed individually for each sample within a batch, meaning that each instance is assigned appropriately selected hard pair samples.

To achieve this, we first extract features from all training samples using a pre-trained model and compute the cosine similarity scores between sample pairs. Then, based on the precomputed similarity scores and label information, we identify hard negative and hard positive samples for each instance. The selection criteria for positive and negative samples vary depending on the curriculum stage. For example, as shown in Fig. 1(a), when the curriculum stage is set to <instrument, target>, the positive and negative pairs are determined based on both the instrument and target values.

For hard pair sampling, hard negative samples are selected as those most similar to a given instance while belonging to a different class. Specifically, for

each instance i in a batch (i.e., the anchor sample), we construct a set of potential negative candidates $\mathcal{H}_{\text{neg}}(i)$, which includes all samples with labels different from the anchor. From negative candidate set, we utilize the precomputed cosine similarity scores $s(i, j)$ to select the top N hardest negative samples—those with the highest similarity to the anchor, defined as:

$$\mathcal{T}_{\text{neg}}(i) = \{j \in \mathcal{H}_{\text{neg}}(i) \mid s(i, j) \text{ ranks in the top-}N\} \tag{1}$$

To ensure the model encounters diverse hard negatives across training, a randomly sampled subset of the top N hard negatives is used for training, defined as:

$$N(i) \subseteq \mathcal{T}_{\text{neg}}(i) \tag{2}$$

For hard positive sampling, for each instance i in a batch (i.e., the anchor sample), we construct a set of potential positive candidates $\mathcal{H}_{\text{pos}}(i)$, consisting of all samples that share the same label as the anchor. From positive candidate set, we select the top K hardest positive samples—those with the lowest similarity scores within the same class—defined as:

$$\mathcal{T}_{\text{pos}}(i) = \{j \in \mathcal{H}_{\text{pos}}(i) \mid s(i, j) \text{ ranks in the bottom-}K\} \tag{3}$$

A single positive pair is then randomly selected for each instance in the batch, defined as:

$$p_i \in \mathcal{T}_{\text{pos}}(i) \tag{4}$$

The selected pairs—hard positives with low similarity and hard negatives with high similarity—are used in contrastive learning to expose the model to more challenging examples. This sampling strategy ensures that the model learns from both challenging inter-class negatives and difficult intra-class positives, improving feature discriminability.

Enhancing Contrastive Learning with Hard Negatives. To enhance the effectiveness of hard pair sampling in contrastive learning, we generate synthetic hard negatives by combining selected challenging samples, as illustrated in Fig. 1(b). Each synthetic feature vector is obtained as a convex linear combination, where the combination ratio λ is drawn from a beta distribution, defined as:

$$\tilde{v}_s = \lambda v_{n_1} + (1 - \lambda) v_{n_2}, \quad n_1, n_2 \in N(i) \tag{5}$$

$$\lambda \sim \text{Beta}(\alpha, \alpha) \tag{6}$$

By integrating a novel hard pair sampling strategy and synthetic hard negative features into contrastive learning, this approach enhances the model's ability to capture subtle differences between triplets while enforcing the learning of more robust features. Moreover, by increasing the diversity of informative samples, it addresses challenges associated with limited data and mitigates overfitting caused by severe class imbalance. Consequently, the model learns more robust and discriminative feature embeddings. For supervised contrastive learning, we adopt the SupCon loss proposed in [8].

2.2 Mitigating Class Imbalance with Soft Labels

To address class imbalance and label ambiguity, we adopt a self-distillation framework, as demonstrated in [21] using the CholecT45 dataset [14]. In this framework, a teacher network $\mathcal{M}_T$ is initially trained using ground-truth multi-label annotations $\mathbf{y} \in \{0,1\}^C$ with a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{teacher}} = -\frac{1}{C}\sum_{c=1}^{C}\left[y_c \log(\hat{y}_c^{(T)}) + (1-y_c)\log(1-\hat{y}_c^{(T)})\right], \tag{7}$$

where $\hat{y}_c^{(T)} = \mathcal{M}_T(x)_c$ denotes the predicted score for class c. A student model $\mathcal{M}_S$ is then trained to mimic the teacher's soft targets using the following BCE loss:

$$\mathcal{L}_{\text{student}} = -\frac{1}{C}\sum_{c=1}^{C}\left[\hat{y}_c^{(T)} \log(\hat{y}_c^{(S)}) + (1-\hat{y}_c^{(T)})\log(1-\hat{y}_c^{(S)})\right], \tag{8}$$

where $\hat{y}_c^{(S)} = \mathcal{M}_S(x)_c$ is the student's prediction. By leveraging soft labels, this approach reduces the adverse effects of class imbalance and annotation noise, promoting smoother decision boundaries and enhanced robustness.

We further extend this framework by incorporating a simple yet effective mixup strategy [22] during fine-tuning. By applying mixup to both teacher and student models, we encourage smoother label transitions and improved generalization. In addition, mixup increases training data diversity by producing interpolated sample combinations, which is particularly advantageous in scenarios with limited data and pronounced class imbalance. This integration of mixup enhances the robustness and overall performance of the model.

3 Experiments

3.1 Dataset and Evaluation Metrics

Our study uses the CholecT45 dataset, a subset of Cholec80 [19]. CholecT45 consists of 45 videos of cholecystectomy procedures, totaling 90,489 frames. Each frame is annotated with an action triplet ⟨instrument, verb, target⟩, covering 100 distinct classes across 6 types of instruments, 10 types of verbs, and 15 target classes. As a multi-label dataset, each frame may include multiple triplets as the labels, with a severe class imbalance across labels. To evaluate the effectiveness of our framework, we use the official 5-fold validation split of CholecT45 [14]. Consistent with prior studies [13,15,21,5,16,2,6], we validate performance using triplet average precision AP_{IVT}.

3.2 Implementation Details

Model. Our experiments employ the Swin Transformer [10] in Tiny and Base configurations, with two Base model variants that differ in input size and window size. Models are implemented in the timm library and trained on an NVIDIA GeForce RTX 4090.

Table 1. Comparison of single models from different approaches on the provided 5-fold validation split of the CholecT45 dataset. **Bold** font indicates the best performance within comparable models. Results marked with † were reproduced using the official code. TERL-B(384) was reproduced with a batch size of 12 due to hardware constraints.

Method	Backbone	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RDV [15]	Res18	89.3±2.1	62.0±1.3	40.0±1.4	34.0±3.3	30.8±2.1	29.4±2.8
RiT [16]	Res18	88.6±2.6	64.0±2.5	43.4±1.4	38.3±3.5	36.9±1.0	29.7±2.6
TDN [2]	Res50	91.2±1.9	65.3±2.8	43.7±1.6	-	-	33.8±2.5
MT4MTL-KD [6]	SwinL(384)	93.1±2.1	71.8±3.4	48.8±3.8	44.9±2.4	43.1±2.0	37.1±0.5
SelfD [21]	SwinB(224)†	90.3±2.3	67.4±1.5	47.9±1.8	43.7±4.1	42.9±1.6	37.1±1.9
TERL-T [5]	SwinT(224)†	93.5±1.5	71.4±2.2	47.2±2.6	44.7±3.8	42.0±2.4	35.7±1.6
TERL-B [5]	SwinB(224)†	93.9±2.0	70.8±2.3	49.4±4.7	43.9±3.4	43.6±2.6	35.6±1.4
TERL-B [5]	SwinB(384)†	94.1±2.3	73.0±1.4	51.1±3.8	46.5±4.9	44.9±1.8	37.7±1.5
TERL-Ens [5]	Ensemble†	94.6±1.9	73.5±1.9	50.8±3.3	47.3±4.1	45.3±1.9	38.5±1.1
CurConMix-T	SwinT(224)	90.4±2.1	67.8±1.8	48.3±3.4	43.3±2.9	43.3±1.8	**37.7**±2.1
CurConMix-B	SwinB(224)	90.4±3.0	68.2±1.5	49.7±2.5	44.8±5.4	45.3±2.4	**38.8**±2.8
CurConMix-B	SwinB(384)	90.9±2.0	68.3±1.3	49.8±3.2	45.2±4.2	45.1±1.1	**39.1**±2.0
CurConMix-Ens	Ensemble	91.7±2.2	69.5±0.4	51.3±2.9	46.3±5.0	47.1±1.6	**40.7**±2.1

CurConMix. We follow [22] and set $\alpha = 0.4$ as the default setting for both feature mixup and input mixup. For hard negatives, the top $N = 1024$ are selected, following [7], with $S = 63$ negatives used during training. Positive samples are selected as $K = \min(1024, \text{samples per class})$ and the contrastive loss temperature τ is set to 0.1, as per [8]. The hyperparameters for self-distillation align with those in [21].

3.3 Comparison with Existing Methods

In this section, we evaluate the performance of our framework and existing methods, including RDV [15], RiT [16], TDN [2], SelfD [21], MT4MTL-KD [6], and TERL [5]. As shown in Table 1, our approach outperforms previous methods by a significant margin across all model backbones. For instance, compared to SelfD, our method improves AP_{IVT} metric from 37.1% to 38.8%. Notably, while MT4MTL-KD utilizes the larger Swin-Large (384) model, our method surpasses it even with the more lightweight Swin-Tiny (224) model, improving AP from 37.1% to 37.7% while requiring fewer computational resources. Additionally, our approach surpasses TERL, elevating AP from 35.7% to 37.7% on Swin-Tiny (224) model, from 35.6% to 38.8% on Swin-Base (224) model, and obtaining the highest performance of 39.1% on Swin-Base (384) model. Although TERL focuses on individual components within a triplet and thus reports strong performance when each component is considered separately, our model demonstrates superior performance in triplet prediction by effectively capturing the interdependencies among triplet components.

Furthermore, we conducted a performance comparison by ensembling Swin-Base (224) and Swin-Base (384). The ensemble method averages the class-wise

Table 2. Ablation study on the components of our framework, CurConMix, showing performance improvement as each component is added. The first row represents the baseline model, marked with an *.

Contrastive	Curriculum	Input Mixup	Feature Mixup	AP_{IVT}
				37.1*
✓				37.8
✓	✓			38.1
✓	✓	✓		38.3
✓	✓	✓	✓	**38.8**

predictions from both models to generate the final outputs. Under this ensemble setting, our method achieves 40.7%, significantly outperforming TERL's 38.5%. These results demonstrate the robustness of our framework across different model sizes and underscore the importance of capturing interdependencies among triplet components for robust performance.

3.4 Ablation Study

To validate the effectiveness of each component in CurConMix, we conducted an ablation study. The baseline for this experiment is SelfD [21], which corresponds to CurConMix without any additional components. Each component was incrementally added to assess its impact on performance. The results of the 5-fold validation are presented in Table 2. The inclusion of contrastive learning improved performance to 37.8% (+0.7%). The addition of curriculum contrastive learning further enhanced performance to 38.1% (+0.3%). Incorporating input mixup led to an increase to 38.3% (+0.2%). Finally, the integration of feature mixup provided the highest boost, achieving 38.8% (+0.5%). The progressive performance gains indicate that each component contributes to improving the model's ability to achieve precise triplet recognition. When all components are utilized, the model achieves the highest performance, highlighting its effectiveness in addressing severe class imbalance and enabling robust triplet recognition under challenging conditions.

4 Conclusions

In this paper, we address the challenges of surgical action triplet recognition, including class imbalance, subtle inter-triplet variations, and complex component interdependencies. We propose a novel framework, **CurConMix**, to effectively tackle these issues. CurConMix employs curriculum contrastive learning to progressively capture relationships between triplet components. At each step, hard pair sampling and synthetic hard negative features enhance feature robustness. During fine-tuning, self-distillation and input mixup are applied to

mitigate class imbalance and promote effective knowledge transfer. Through extensive experiments on the CholecT45 dataset, our framework consistently outperforms existing models across different backbones and resolutions. Moreover, we demonstrate that capturing the interdependencies among triplet components significantly improves recognition performance. While our framework improves triplet recognition, exploring additional techniques such as temporal modeling or unsupervised learning could further enhance performance. We believe our work offers valuable insights into surgical scene understanding and will inspire future research in complex action recognition tasks.

Acknowledgments. We thank Ms. Soyoung Lim for her professional assistance with the figure illustrations in this manuscript. This study was supported by the Samsung Medical Center Grant [SMO1250271], the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MIST) [RS-2024-00392495], and the "Future Medicine 2030 Project" of Samsung Medical Center [SMX1230771].

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

References

1. Anteby, R., Horesh, N., Soffer, S., Zager, Y., Barash, Y., Amiel, I., Rosin, D., Gutman, M., Klang, E.: Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surgical endoscopy **35**, 1521–1533 (2021)
2. Chen, Y., He, S., Jin, Y., Qin, J.: Surgical activity triplet recognition via triplet disentanglement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 451–461. Springer (2023)
3. Farhad, M., Masud, M.M., Beg, A., Ahmad, A., Ahmed, L.: A review of medical diagnostic video analysis using deep learning techniques. Applied Sciences **13**(11), 6582 (2023)
4. Guédon, A.C., Meij, S.E., Osman, K.N., Kloosterman, H.A., van Stralen, K.J., Grimbergen, M.C., Eijsbouts, Q.A., van den Dobbelsteen, J.J., Twinanda, A.P.: Deep learning for surgical phase recognition using endoscopic videos. Surgical endoscopy **35**, 6150–6157 (2021)
5. 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)
6. 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 (2023)
7. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. Advances in neural information processing systems **33**, 21798–21809 (2020)
8. 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)

9. Li, Y., Xia, T., Luo, H., He, B., Jia, F.: Mt-fist: a multi-task fine-grained spatial-temporal framework for surgical action triplet recognition. IEEE journal of biomedical and health informatics (2023)
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. Namazi, B., Sankaranarayanan, G., Devarajan, V.: A contextual detector of surgical tools in laparoscopic videos using deep learning. Surgical endoscopy pp. 1–10 (2022)
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. Medical Image Analysis **86**, 102803 (2023)
13. Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 364–374. Springer (2020)
14. Nwoye, C.I., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)
15. 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)
16. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. International Journal of Computer Assisted Radiology and Surgery **18**(6), 1053–1059 (2023)
17. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 505–514. Springer (2023)
18. Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: Curriculum learning: A survey. International Journal of Computer Vision **130**(6), 1526–1565 (2022)
19. 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)
20. Xi, N., Meng, J., Yuan, J.: Forest graph convolutional network for surgical action triplet recognition in endoscopic videos. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8550–8561 (2022)
21. 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)
22. Zhang, H.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)