VideoCutMix: Temporal Segmentation of Surgical Videos in Scarce Data Scenarios

Rohan Raju Dhanakshirur[1], Mrinal Tyagi[1], Britty Baby[2], Ashish Suri[2], Prem Kalra[1], and Chetan Arora[1] ⋆

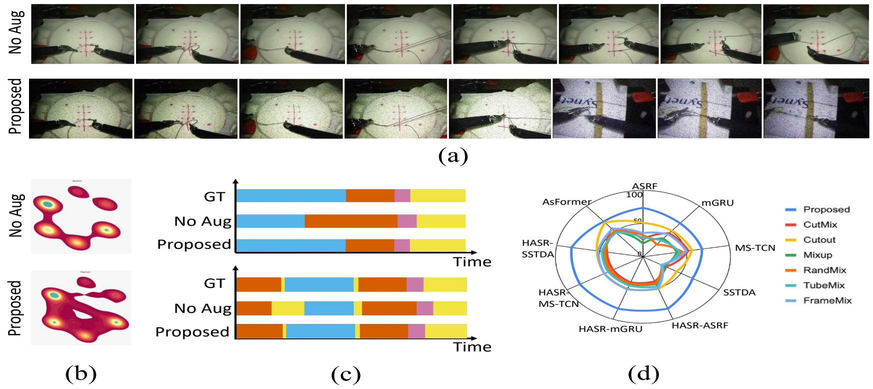
[1]Indian Institute of Technology Delhi [2]AIIMS Delhi

Abstract. Temporal Action Segmentation (TAS) of a surgical video is an important first step for a variety of video analysis tasks such as skills assessment, surgical assistance and robotic surgeries. Limited data availability due to costly acquisition and annotation makes data augmentation imperative in such a scenario. However, extending directly from an image-augmentation strategy, most video augmentation techniques disturb the optical flow information in the process of generating an augmented sample. This creates difficulty in training. In this paper, we propose a simple-yet-efficient, flow-consistent, video-specific data augmentation technique suitable for TAS in scarce data conditions. This is the first augmentation for data-scarce TAS in surgical scenarios. We observe that TAS errors commonly occur at the action boundaries due to their scarcity in the datasets. Hence, we propose a novel strategy that generates pseudo-action boundaries without affecting optical flow elsewhere. Further, we also propose a sample-hardness-inspired curriculum where we train the model on easy samples first with only a single label observed in the temporal window. Additionally, we contribute the first-ever non-robotic Neuro-endoscopic Trainee Simulator (NETS) dataset for the task of TAS. We validate our approach on the proposed NETS, along with publicly available JIGSAWS and Cholec T-50 datasets. Compared to without the use of any data augmentation, we report an average improvement of 7.89%, 5.53%, 2.80%, respectively, on the 3 datasets in terms of edit score using our technique. The reported numbers are improvements averaged over 9 state-of-the-art (SOTA) action segmentation models using two different temporal feature extractors (I3D and VideoMAE). On average, the proposed technique outperforms the best-performing SOTA data augmentation technique by 3.94%, thus enabling us to setup a new SOTA for action segmentation in each of these datasets. The dataset and the complete source-code is available at: https://aineurosurgery.github.io/VideoCutMix.

1 Introduction

Temporal action segmentation. Given a video V with n frames $(f_1, \ldots, f_n)$, temporal action segmentation (henceforth referred to as TAS) can be defined as

⋆ corresponding author



(a)

(b) (c) (d)

Fig. 1. The proposed video-specific augmentation technique is illustrated in (a), generating pseudo-action boundaries while preserving flow consistency. The strategy significantly improves the performance of all existing temporal segmentation techniques. The density plots for each class before and after augmentation are shown in (b). Effectiveness is demonstrated in (c), and (d) showcases the performance against various state-of-the-art (SOTA) augmentation techniques with different action segmentation models for the JIGSAWS [7] dataset.

the task of labelling every frame of a video with the corresponding action label chosen from a fixed set. TAS is inherently different from action recognition, which is a video classification problem where one classifies a given video clip based on the action being performed in the video (single label for the whole clip).

TAS for neuro-endoscopic skills training. Neuro-endoscopic skills training is generally achieved using a box-based trainer, in which a trainee performs pick and place tasks under the guidance of an endoscope [18]. However, the lack of a well-curated dataset does not allow TAS of these videos and thereby limits the performance of automated evaluation, surgical training, etc. Hence, there is a pressing need for a well-curated dataset for the problem.

TAS in a data-constrained setting. In a data-constrained scenario like surgical skills evaluation [13], the typical size of the video samples available is merely 176 (approx. 58 min dataset) for the JIGSAWS dataset [7]! In contrast, a natural video dataset like Breakfast [11] has 1712 videos (approx. 77 hours of data). The situation becomes even more critical with increasing specialization of the task [17], or safety and privacy considerations [15]. Hence, the use of effective data augmentation becomes imperative.

Data augmentation for video tasks. Data augmentation enhances deep neural network (DNN) training, starting from initial models like AlexNet [10]. Early methods introduced image corruptions (e.g., blur, colour jitter), and recent techniques involved mixing samples through masking [23] or generic convex combination [25]. Generally, video datasets are much smaller than their image counterparts. Despite that, surprisingly, there are very few techniques available for video augmentation. Researchers have explored temporal (alter the speed of the video, playback sound, etc.) [16], spatial (alter the pixels of the frames in a video) [23], and appearance augmentation (frame blurring, frame rotation, etc.) [4]. However, most of these techniques were predominantly developed for image classification tasks and have limited utility for video analysis [24]. Recently, researchers have introduced video action-recognition-specific data augmentation techniques, such as RandMix, FrameMix and TubeMix [21]. These techniques work well on video recognition and large-scale TAS. However, little to no effort has taken place in the data augmentation techniques for small dataset scenarios, as is the case in most surgical video analysis problems.

Key insights. We observe that current video augmentation techniques are highly motivated by their image-based counterparts. For example, inspired from CutMix [23], a recent technique suggests mixing of pixel-tubes from two video samples [21]. On the other hand, optical flow is a critical and often most differentiating cue for a video sample. However, extending directly from an image-based augmentation strategy, almost all video augmentation techniques fail to maintain flow consistency. For instance, on average, augmented videos of RandMix [21] have an 80.9% deviation in optical flow from the base videos of the JIGSAWS dataset [7], TubeMix [21] with 67.54%, and FrameMix [21] with 48%. Therefore, the generated augmented video often does not belong to the original training distribution. Training a model with these samples, especially in scarce data scenarios, can lead to a model memorizing the samples rather than learning the intended distribution.

Contributions. (1) We contribute the first-ever non-robotic neurosurgery-specific Neuro-endoscopic Trainee Simulator (NETS) dataset, which can be used for the TAS and related problems. The proposed dataset is annotated for temporal segmentation by a consensus of a group of 3 expert neurosurgeons (30+ years of experience) from different medical schools. The dataset opens up the doors for automated skills evaluation for neurosurgery and many other applications. **(2)** This is the first augmentation for data-scarce TAS in surgical scenarios. We propose a video-specific data augmentation technique that does not disturb the critical optical flow information in a video. Hence, instead of mixing a frame spatially, we propose to augment a video segment temporally only. This helps generate new pseudo-action boundaries, which are usually scarce in a video. **(3)** Most TAS techniques process a temporal window at a time. Since, action/event boundaries are scarce in a video, this makes the windows containing two actions (at the boundary) also scarce, and thus difficult to learn for a deep neural network (DNN). Recognizing this problem, we present a hardness-specific curriculum specifically for the video analysis tasks, which only trains on single action-label samples initially and moves to multi-label and augmented samples in the later epochs. (Fig. 1) **(4)** Compared to without use of any data augmentation, we report an average improvement of 7.89%, 5.53%, 2.80% on our NETS dataset, JIGSAWS [7] and Cholec T-50 [14] respectively, in terms of edit score using our technique. The reported numbers are improvements averaged over 9 SOTA action segmentation models using two different temporal feature representations (I3D and VideoMAE). The proposed technique outperforms the best-performing SOTA

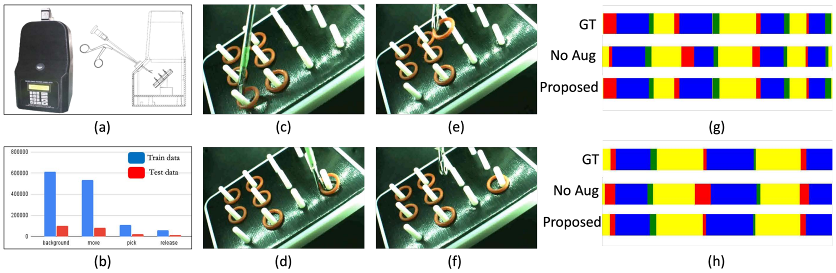


Fig. 2. Proposed NETS dataset. (a) shows the box-based trainer, (b) shows the class distribution, (c)-(f) shows a sample frame for each of the classes, pick, move, release, and background (g),(h) shows the performance of the proposed action segmentation algorithm on two of the samples

data augmentation technique by 3.94%, thus enabling us to setup a new `SOTA` for action segmentation in each of these datasets. The dataset and the complete source-code will be publicly released post-publication.

2 Proposed methodology

Proposed NETS dataset. We contribute the first ever and the largest surgeon-generated neurosurgical endoscopic `TAS` dataset, namely the Neuro-endoscopic Trainee Simulator (`NETS`) Dataset for `TAS` problem. 70 neurosurgery trainees from 14 hospitals across 3 countries performed the task of "pick and place" the rings from one peg to another in 6 box-based trainers [18] in a period of 5 years as shown in Fig. 2. An auxiliary camera was placed to capture the endoscope and the tool movements of the trainee neurosurgeon. The videos captured from the auxiliary camera constitute the proposed dataset. The proposed dataset consists of 174 videos, with each video spanning 60 secs on average. We identify four activities in the dataset, viz pick, move, release and background. Each of the activities is defined as follows: **(1) Pick:** The set of frames encompassing the period from when the forceps made contact with the ring until the moment when the ring separated from the base. **(2) Move:** The period from when the ring and forceps are tightly in contact with each other and the ring is not present on the base. **(3) Release:** The set of frames covering the period from where the angle between the two teeth of the forceps starts increasing to the time when the ring is no longer in contact with the forceps. **(4) Background:** Any other frame that does not fit into the definition of Pick, Move or Release. Each frame was annotated independently by three annotators, and the discrepancies were resolved through consensus. The annotation was then verified by three expert neurosurgeons. Class label distribution is shown in Fig. 2

Problem formulation. Let a video sample V contains n frames, $\{f_1, \dots, f_n\}$. In a non augmentation scenario, for any given frame $f_i \in V$, we consider δ neighboring frames from each side, $\{f_{i-\delta}, \dots, f_{i+\delta}\}$ to create a temporal window and use

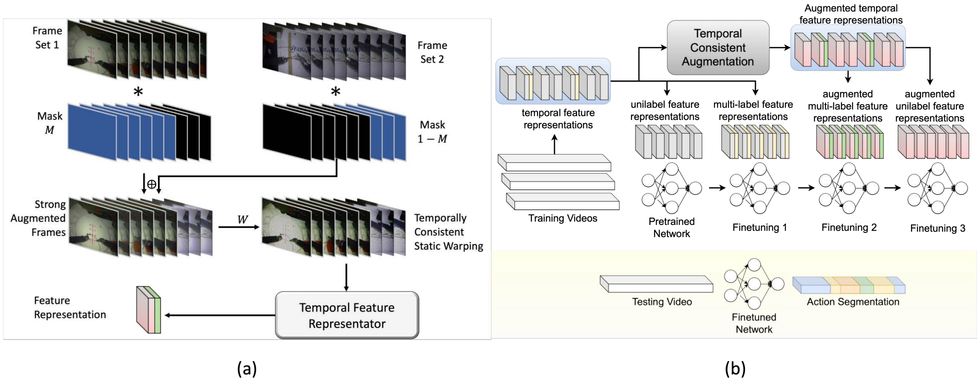

Frame Set 1
Frame Set 2
*
*
Mask M
Mask $1-M$
⊕
Strong Augmented Frames
W
Temporally Consistent Static Warping
Feature Representation
Temporal Feature Representator
(a)
Temporal Consistent Augmentation
Augmented temporal feature representations
temporal feature representations
unilabel feature representations
multi-label feature representations
augmented multi-label feature representations
augmented unilabel feature representations
Training Videos
Pretrained Network
Finetuning 1
Finetuning 2
Finetuning 3
Testing Video
Finetuned Network
Action Segmentation
(b)

Fig. 3. Proposed architecture. (a) Proposed video-specific data augmentation. Here, W refers to one of the temporally consistent, static warping transformations, which corrupts (e.g. blur) a frame independently, but consistently, across a video. (b) Proposed curriculum learning framework.

it to generate a *Temporal Feature Representation* (TFR), $F_i := g(f_{i-\delta}, \ldots, f_{i+\delta})$. Here, g is a temporal feature map extractor, such as I3D [3] or VideoMAE [20]. The feature maps F are further fine-tuned typically using a smaller neural network h to generate a frame level prediction, $\widehat{y_i} :== h(F_i)$. Note that $\widehat{y_i}$ is a k dimensional probability vector generated after applying the Softmax function on the last layer logits, and k is the number of action classes. We use $\widehat{y_i^j}$ to denote the predicted probability of j^{th} action class.

2.1 VideoCutMix: Proposed data augmentation technique

Overview. We propose to modify the boundary frames in a temporal window used for generating the TFRs to synthesize an augmented sample. For a temporal window, $T_i = \{f_{i-\delta}, \ldots, f_{i+\delta}\}$, used to generate TFR for a frame i, we propose to replace the first β frames, or the last β frames, with consecutive β frames from another temporal location. These β frames are chosen from a random location of the same or another video sample. Augmented TFR is generated using these modified set of frames. Replacing the first or last frames ensures that the flow is consistent inside the sample and merely a pseudo-action-boundary is created. Besides nudging the network to focus on the real flow far from the boundary, the proposed augmentation also helps avoid overfitting on action-label-sequence. In a small dataset, when the number of action boundaries are small, a DNN model may potentially overfit frequent action label ordering (e.g. *Pick* precedes *Move* most of the time in NETS dataset). Adding frames from another location, and potentially different label, help creates new action boundaries which do not follow such pattern. We train the SOTA action segmentation architectures with the augmented TFRs and observe improvement in the performance. Fig. 3 visually describes the proposed video-specific data augmentation technique.

Mask Generation. Consider two original video samples i, and k, and two frames f_{ij} and f_{kl} respectively. Note that i may or may not be equal to k, i.e., frames f_{ij} and f_{kl} may or may not come from a same video. However, if $i == k$, we ensure non-overlap in the temporal window, by constraining $|j - l| \geq 2\delta$. Recall that, the temporal window for extracting `TFR` is of length $(2\delta + 1)$. We generate a boolean mask vector M of size $(2\delta + 1)$ as follows:

$$M = \alpha \times \mathbb{1}_{\beta} \quad \oplus \quad \mathbb{0}_{(2\delta+1-2\beta)} \quad \oplus \quad (1-\alpha) \times \mathbb{1}_{\beta}. \tag{1}$$

Here $\mathbb{1}_d$ and $\mathbb{0}_d$ denote a d-dimensional vector of all ones and all zeros respectively. Further, $\alpha \in \{0, 1\}$ is the hyper-parameter deciding the position of the augmentation, in the beginning for $\alpha = 1$, and in the end for $\alpha = 0$. Hyper-parameter β is the augmentation factor, typically set as an integer 3 or 4 in our experiments. Further, $\oplus$ indicates the concatenation of the vectors. The value of α is chosen randomly for every sample in the input. Therefore, for $\beta = 3$, if α is chosen to be 1, we get $M = \{1, 1, 1, 0, 0, 0, \ldots, 0\}_{(2\delta+1)}$ and if we choose $\alpha = 0$, we get $M = \{0, 0, 0, \ldots, 0, 1, 1, 1\}_{(2\delta+1)}$.

Augmented sample generation. We now obtain augmented samples as:

$$\{\widehat{f}_{i-\delta}, \ldots, \widehat{f}_{i+\delta}\} = M \times \{f_{kl-\delta}, \ldots, f_{kl+\delta}\} + (1 - M) \times \{f_{ij-\delta}, \ldots, f_{ij+\delta}\}. \tag{2}$$

That is, we replace the first or last β frames of the set $\{f_{ij-\delta}, \ldots, f_{ij+\delta}\}$ from the corresponding frames in the set, $\{f_{kl-\delta}, \ldots, f_{kl+\delta}\}$ to generate the augmented sample. The augmented sample contains the original optical flow and adds at most one pseudo action boundary.

Static warping, and updating target probability vector. The augmented samples undergo temporal consistent static warping, i.e. applying the same set of weak augmentations (e.g., random rotation, random flip) to each frame. Warped augmented frames are used to generate the `TFRs`, $\widetilde{F}$. Instead of using the one-hot vector based on the ground-truth label, we create a k dimensional Softmax vector, such that the probability of each action class is now as per the actual proportion of the number of frames belonging to that class in the temporal window. We use $\widetilde{F}$ and the modified target vector to train the action segmentation model.

2.2 Proposed curriculum learning technique

The proposed curriculum learning mechanism is demonstrated in Fig. 3(b). Consider a video V and a frame $f_i \in V$, with its temporal window, T_i, and feature representation, F_i. If every frame in the temporal window has the same label, we call F_i as the *Unilabel Feature Representation* (`UFR`). Otherwise, we call it the *Multi-label Feature Representation* (`MFR`). Similarly, post-augmentation, a `TFR` $\widetilde{F}_i$ is said to be *Augmented Unilabel Feature Representation* (`AUFR`) if every frame responsible for $\widetilde{F}_i$ has the same label. Otherwise, it's called the *Augmented Multi-label Feature Representation* (`AMFR`).

We argue that `UFRs` are the easiest samples to learn; therefore, we first fine-tune the pre-trained network using these training samples. Next, `MFRs` are used

Table 1. The performance of our VideoCutMix algorithm on 5 different datasets, using VideoMAE feature extractor. Here, **"H-"** refers to HASR architecture [1]

Datasets		**NETS**		**JNP** [7]		**JS** [7]		**JKT** [7]		**C-T50** [14]	
Venue	**Model**	**Edit**	**F1**	**Edit**	**F1**	**Edit**	**F1**	**Edit**	**F1**	**Edit**	**F1**
CVPR	MSTCN [6]	95.02	95.18	77.53	78.02	84.04	88.36	78.45	84.67	32.99	38.25
2019	+Proposed	**95.31**	**95.43**	**84.86**	**84.71**	**86.46**	**91.26**	**81.73**	**88.43**	**38.99**	**42.81**
WACV	ASRF [8]	63.50	75.19	79.52	77.91	86.13	89.31	85.16	88.89	20.90	26.55
2021	+Proposed	**92.8**	**94.83**	**83.93**	**84.38**	**88.34**	**92.32**	**90.75**	**93.46**	**26.43**	**34.09**
ICCV	mGRU [1]	**95.14**	93.56	56.63	62.36	74.98	82.43	70.33	77.65	39.87	40.81
2021	+Proposed	94.72	**93.95**	**68.59**	**73.99**	**81.98**	**88.01**	**79.47**	**87.00**	**42.94**	**43.05**
ICCV	H-ASRF [1]	92.87	**94.62**	89.94	91.33	85.86	90.00	90.68	93.44	23.29	32.93
2021	+Proposed	**93.24**	94.14	**95.05**	**96.76**	**91.85**	**95.43**	**91.51**	**95.45**	**29.00**	**39.99**
ICCV	H-MSTCN [1]	91.83	94.07	89.94	91.33	86.8	90.67	90.45	93.54	31.26	41.36
2021	+Proposed	**96.13**	**95.91**	**95.10**	**96.89**	**91.72**	**95.38**	**91.51**	**95.45**	**37.98**	**50.72**
ICCV	H-SSTDA [1]	83.43	89.13	89.75	91.14	86.14	90.42	90.79	93.52	30.11	39.82
2021	+Proposed	**96.22**	**95.83**	**95.16**	**96.88**	**91.66**	**95.35**	**91.48**	**95.32**	**36.99**	**49.75**
BMVC	ASFormer [22]	96.55	93.50	**84.22**	81.60	81.92	87.77	82.68	87.94	34.36	38.99
2022	+Proposed	**96.64**	**95.52**	83.74	**84.03**	**88.30**	**91.79**	**83.28**	**89.80**	**38.64**	**42.76**
ECCV	UVAST [2]	85.81	90.28	36.38	47.38	58.46	70.82	44.86	58.83	49.31	43.19
2023	+Proposed	**89.72**	**90.95**	**45.03**	**56.16**	**66.25**	**77.04**	**56.01**	**68.15**	**49.96**	**43.47**
MM	CETNet [19]	95.20	94.51	77.25	83.00	68.40	70.27	79.73	86.65	37.02	41.17
2023	+Proposed	**95.29**	**95.64**	**80.89**	**87.33**	**78.32**	**80.03**	**84.76**	**90.54**	**40.14**	**42.31**
Average Gain		**4.39**	**3.05**	**6.33**	**6.91**	**5.24**	**4.54**	**4.00**	**4.33**	**4.74**	**5.04**

to fine-tune the network further, followed by AMFRs and AUMRs. Note that in the proposed setting, we first train the model with the *Augmented Multi-label Feature Representation* and then with the *Augmented Uni-label Feature Representations*. This is because, in the proposed architecture, we replace β frames from the base set of frames used to generate TFR and hence, we expect one shift in the optical flow at the boundary. Thus, AMFRs act as pseudo boundaries between the actions and are easier to learn, when compared to AUMRs. To avoid catastrophic forgetting, we use 100% of the samples from the previous sets while fine-tuning the network with the next set. The fine-tuned model is then used to perform temporal action segmentation on an unseen video.

3 Results and Discussions

Datasets. We perform our experiments on the proposed NETS dataset and the publicly available JIGSAWS and Cholec T-50 datasets. JIGSAWS is an endoscopic skills assessment dataset of 176 videos, split into three parts, namely Suturing (JS), knot tying (JKT) and needle passing (JNP), approximately 58 mins for whole dataset. Cholec T-50 (C-50), on the other hand, is a 90-minute dataset of 50 laparoscopic videos of cholecystectomy. It involves 10 actions such as grasp, dissect, etc.

Evaluation Metric. We evaluate the performance of action segmentation using *Segmental Edit Score* [8], and Frame-wise F1 score at 0.1 IOU. For any frame f_i, the Edit score is the Levenshtein Distance [12] between the ground truth and the predicted labels for a set of $(2\delta + 1)$ frames, $[f_{i-\delta}, \dots, f_{i+\delta}]$.

Table 2. The comparison of our VideoCutMix technique against current SOTA augmentation on the JIGSAWS-Knot-tying dataset using I3D features

		UVAST [2]		mGRU [1]		HASR-ASRF [1]	
Algorithm	**Venue**	**Edit**	**F1@10**	**Edit**	**F1@10**	**Edit**	**F1@10**
CutMix [23]	ICCV19	51.05	62.99	63.09	66.45	63.28	62.28
Cutout [5]	ArXiv17	49.59	61.52	63.37	72.86	82.83	80.68
Mixup [25]	ICLR18	50.25	59.5	51.87	55.53	66.42	67.53
RandMix [21]	CVPR23	43.95	51.91	51.91	52.77	75.59	75.82
TubeMix [21]	CVPR23	51.83	59.17	51.79	51.97	77.91	75.05
FrameMix [21]	CVPR23	51.63	61.98	60.71	65.75	80.46	77.63
Dynaugment [9]	ICLR23	44.51	56.51	51.56	52.97	86.42	85.17
Proposed Technique		**54.49**	**65.32**	**74.29**	**81.86**	**89.11**	**93.45**

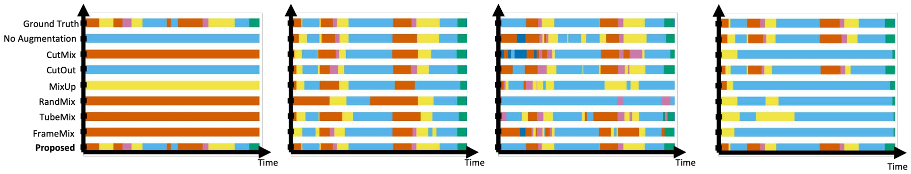


Fig. 4. Visualization of the performance of the proposed data augmentation technique against SOTA methods using UVAST [2] architecture on JIGSAWS-Knot-tying dataset

Implementation Details. We use a batch size of 1. The value of δ is typically set to 8. The initial learning rate is set to 0.0001, which drops by 0.1 after every 20 epochs. We set 0.9 as the Nesterov momentum coefficient. We train the network for 25 epochs on a server with 8 NVidia A100, 40GB GPUs. Other hyper-parameter details can be found in Table S1 in the supplementary.

Action Segmentation Results. The results of the various SOTA temporal action segmentation architectures on NETS, JS, JKT, JNP, and Cholec T-50 datasets using VideoMAE [20] features is shown in Table 1. Similar results using I3D [3] features is given in Table S2 of the supplementary material. All the reported baseline and proposed models were trained using the same set of weak augmentations and hyperparameters.

Effect of dataset size. One observes in Table S2 that as the datasets become smaller, the improvement gained using our method also increases. Though this is expected with any data augmentation strategy, we confirm this hypothesis by testing on a subset of the Breakfast [11] (containing natural images) dataset, with 50%, 10%, and 5% data. Results are given in Table S3 in the supplementary.

Ablation Analysis - Effect of every component in the proposed architecture. Table S4 in the supplementary material demonstrates the importance of each of the components in the proposed architecture.

Comparison with other augmentation methods. Table 2 compares proposed augmentation strategy against the SOTA techniques on the three action segmentation architectures (UVAST [2], Asformer [22] and HASR-ASRF [1]) using I3D features on JIGSAWS-Knot-tying dataset. It can be observed that the proposed architecture outperforms the current SOTA augmentation technique at least by 1.8% in edit score. Table S5 in the supplementary shows comparison for other architectures.

Comparison with other augmentation methods. A few sample outputs are visualized in Fig. 4. One observes that after augmentation using the proposed technique, the UVAST model is able to correctly detect the boundaries between the actions, accurately detect small actions, and reduce the misclassification rate.

4 Conclusion

We proposed a video-specific, flow-consistent, data augmentation technique for temporal action segmentation in surgical video analysis. The technique relies on the thesis that optical flow is an important cue for action segmentation and must not be disturbed during augmentation. The proposed augmentation technique, when coupled with the proposed curriculum learning, achieved significant performance gain. Though we focused only on temporal action segmentation, we believe that this work can be extended to other video analysis tasks as well.

Acknowledgments. We thank Mr. Sushil Kumar Vemuri for his assistance. We acknowledge and thank the funding support from the Department of Biotechnology (grant number BT/CRS01557/PACE-27/22, and BT/PR33193/AI/133/5/2019), and Ministry of Education, Government of India, Central Project Management Unit, IIT Jammu with sanction number IITJMU/CPMU-AI/2024/0002 for the India grant for the Centre of Excellence in AI (CoEAI) for healthcare, first phase. We also thank ICMR, Govt of India, for the funding under the project 5/3/8/1/2022/MDMS(CARE).

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

References

1. Ahn, H., Lee, D.: Refining action segmentation with hierarchical video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16302–16310 (2021) 7, 8
2. Behrmann, N., Golestaneh, S.A., Kolter, Z., Gall, J., Noroozi, M.: Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: European Conference on Computer Vision. pp. 52–68. Springer (2022) 7, 8
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 5, 8
4. Cauli, N., Reforgiato Recupero, D.: Survey on videos data augmentation for deep learning models. Future Internet **14**(3), 93 (2022) 3
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 8
6. 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) 7

7. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI workshop: M2cai. vol. 3 (2014) 2, 3, 7
8. Ishikawa, Y., Kasai, S., Aoki, Y., Kataoka, H.: Alleviating over-segmentation errors by detecting action boundaries. In: Proceedings of the IEEE/CVF Winter Applications of Computer Vision Conference. pp. 2322–2331 (2021) 7
9. Kim, T., Kim, J., Shim, M., Yun, S., Kang, M., Wee, D., Lee, S.: Exploring temporally dynamic data augmentation for video recognition. International Conference on Learning Representations (ICLR) (2023) 8
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012) 2
11. Kuehne, H., Gall, J., Serre, T.: An end-to-end generative framework for video segmentation and recognition. In: Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16). Lake Placid (Mar 2016) 2, 8
12. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710. Soviet Union (1966) 7
13. Liu, D., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 476–484. Springer (2019) 2
14. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 364–374. Springer (2020) 3, 7
15. Paulius, D., Sun, Y.: A survey of knowledge representation in service robotics. Robotics and Autonomous Systems **118**, 13–30 (2019) 2
16. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021) 3
17. Singh, G.K., Shukla, V., Patil, S., Shah, P.: Automatic detection of abnormal event using smart video surveillance system in a nuclear power plant. In: 55th Annual Meeting of the Institute of Nuclear Materials Management–Atlanta, USA: Institute for Nuclear Materials and Management. vol. 1, pp. 3139–3146 (2014) 2
18. Singh, R., Baby, B., Damodaran, N., Srivastav, V., Suri, A., Banerjee, S., Kumar, S., Kalra, P., Prasad, S., Paul, K., et al.: Design and validation of an open-source, partial task trainer for endonasal neuro-endoscopic skills development: Indian experience. World neurosurgery **86**, 259–269 (2016) 2, 4
19. Wang, J., Wang, Z., Zhuang, S., Hao, Y., Wang, H.: Cross-enhancement transformer for action segmentation. Multimedia Tools and Applications pp. 1–14 (2023) 7
20. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14560 (June 2023) 5, 8

21. Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., Jiang, Y.G.: Svformer: Semi-supervised video transformer for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18816–18826 (2023) 3, 8
22. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. British Machine Vision Conference (BMVC) (2021) 7, 8
23. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019) 2, 3, 8
24. Yun, S., Oh, S.J., Heo, B., Han, D., Kim, J.: Videomix: Rethinking data augmentation for video classification. arXiv preprint arXiv:2012.03457 (2020) 3
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 2, 8