

# S3Aug: Segmentation, Sampling, and Shift for Action Recognition

Taiki Sugiura and Toru Tamaki<sup>a</sup>  
Nagoya Institute of Technology, Japan

**Keywords:** Action Recognition, Data Augmentation, Out-Of-Context, Segmentation, Image Translation.

**Abstract:** Action recognition is a well-established area of research in computer vision. In this paper, we propose S3Aug, a video data augmentation for action recognition. Unlike conventional video data augmentation methods that involve cutting and pasting regions from two videos, the proposed method generates new videos from a single training video through segmentation and label-to-image transformation. Furthermore, the proposed method modifies certain categories of label images by sampling to generate a variety of videos, and shifts intermediate features to enhance the temporal coherency between frames of the generated videos. Experimental results on the UCF101, HMDB51, and Mimetics datasets demonstrate the effectiveness of the proposed method, particularly for out-of-context videos of the Mimetics dataset.

## 1 INTRODUCTION

Action recognition is an active area of research in computer vision and is used in a variety of applications. A major difficulty in developing action recognition methods is the need for a large amount of training data. To address this, several large datasets have been proposed (Kuehne et al., 2011; Soomro et al., 2012; Kay et al., 2017; Goyal et al., 2017).

In certain tasks, it can be hard to collect a large number of videos. To address this issue, data augmentation has been employed (Cauli and Reforgiato Recupero, 2022). This technique involves virtually increasing the number of training samples by applying geometric transformations, such as vertical and horizontal flip, or image processing, such as cropping a part of one image and pasting it onto another.


Various data augmentation techniques have been proposed for both images (Shorten and Khoshgoftaar, 2019) and video tasks (Cauli and Reforgiato Recupero, 2022). These video data augmentation methods are based on cutmix (Yun et al., 2019) and copy paste (Ghiasi et al., 2021), which involve cutting (or copying) regions of two videos to create a new video. However, these approaches have two drawbacks. First, the spatio-temporal continuity of the actor in the video may be compromised. Unlike general image recognition tasks, the region of the ac-

tor is essential for action recognition, and thus simple extensions of cutmix or copy-paste are not suitable since the actor region may be cut off or obscured by augmentation. Second, action recognition datasets are known to have considerable dataset biases (Chung et al., 2022), therefore, simple augmentation does not address the issue of out-of-context (or out-of-distribution) samples.

Therefore, in this paper, we propose an alternative to cutmix-based data augmentation methods, called S3Aug (Segmentation, category Sampling, and feature Shift for video Augmentation). This method produces multiple videos from a single training video while maintaining the semantics of the regions by using panoptic segmentation and image translation. We evaluated the effectiveness of our proposed method using two well-known action recognition datasets, UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011). Furthermore, we evaluated its performance on out-of-context samples with the Mimetics data set (Weinzaepfel and Rogez, 2021).

## 2 RELATED WORK

Action recognition is a long-standing and significant area of study in computer vision (Hutchinson and Gadepally, 2021; Kong and Fu, 2022; Ulhaq et al., 2022), with a variety of models being proposed, in-

<sup>a</sup>  <https://orcid.org/0000-0001-9712-7777>

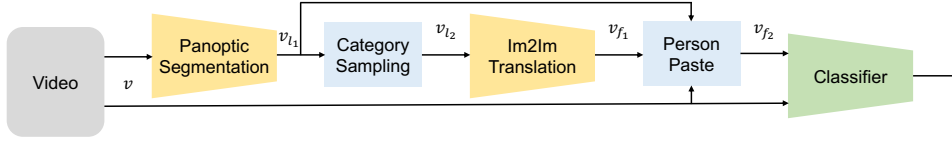


Figure 1: A schematic diagram of the proposed S3Aug. The green component is a classifier that is trained, while the yellow components are pre-trained segmentation and image translation components. The blue components are non-training processes.

cluding CNN-based (Feichtenhofer, 2020a; Feichtenhofer et al., 2019) and Transformer-based (Arnab et al., 2021; Bertasius et al., 2021).

For this data-demanding task, video data augmentation has been proposed (Cauli and Reforgiato Recupero, 2022). The main approach is cutmix (Yun et al., 2019) and copy-paste (Ghiasi et al., 2021), which cut (or copy) a random rectangle or actor region from frames of one video and paste it onto frames of the other video. This approach is used by VideoMix (Yun et al., 2020), ActorCutMix (Zou et al., 2022), and ObjectMix (Kimata et al., 2022), however, it has the issue of the spatial and temporal discontinuity of the actor regions. To address this issue, Learn2Augment (Gowda et al., 2022) and Action-Swap (Chung et al., 2022) generate a background image by utilizing inpainting to remove the extracted actor regions from one video frame, and then paste the actors extracted from the other video frame onto the background image.

Another issue is background bias (Chung et al., 2022; Weinzaepfel and Rogez, 2021; He et al., 2016), where models tend to heavily rely on cues in appearances of the scene (e.g., background or object) and fail to predict the actions of out-of-context samples. To address this, some simple methods generate various videos from samples in the given dataset. Action Data Augmentation Framework (Wu et al., 2019) stacks the generated still images, which does not produce a video with appropriate variations. Self-Paced Selection (Zhang et al., 2020) treats a video as a single “dynamic image”, resulting in the loss of temporal information. Our approach is similar in spirit but instead uses segmentation as a guide to generate video frames to maintain the semantics of the original source video.

Note that generating videos is still a difficult task despite advances in generative models such as GAN (Jabbar et al., 2021; Goodfellow et al., 2014; Yi et al., 2019) and diffusion models (Rombach et al., 2022; Ramesh et al., 2021). There have been some attempts to generate videos using diffusion models (Ho et al., 2022; Luo et al., 2023) and GPT (Yan et al., 2021), but they require specific prompts to control the content of the videos, which is an ongoing exploration. On the contrary, our approach produces video frames from

segmented label frames, similar to Vid2Vid (Wang et al., 2018a) and the more recent ControlNet (Zhang and Agrawala, 2023). However, these methods are computationally expensive and are not suitable for this study. Therefore, we use a GAN-based method (Park et al., 2019) as frame-wise image translation, which is a compromise between speed and computational cost. Frame-wise processes are known to produce temporally incoherent results, so we propose the shit feature, which was originally proposed for lightweight action recognition models (Zhang et al., 2021; Lin et al., 2019; Hashiguchi and Tamaki, 2022; Wang et al., 2022).

### 3 METHOD

This section begins by providing an overview of the proposed S3Aug (Fig.1), followed by a description of key components such as category sampling and feature shift.

An input video clip  $v \in \mathcal{R}^{T \times 3 \times H \times W}$  is a sequence of  $T$  frames  $v(t) \in \mathcal{R}^{3 \times H \times W}$ ,  $t = 1, \dots, T$ , where  $H, W$  are the height and width of the frame.  $y \in \{0, 1\}^{L_a}$  is a one-hot vector of the action label, with  $L_a$  being the number of action categories.

First, we apply panoptic segmentation to each frame  $v(t)$  to obtain the corresponding label image  $v_{l_1}(t) \in \{0, 1, \dots, L_s\}^{H \times W}$ , where  $L_s$  is the number of segmentation categories. In addition, an instance map  $I(t) \in \{0, 1, \dots, N(t)\}^{H \times W}$  is obtained, which assigns a unique value to each detected instance in the frame, with  $N(t)$  being the number of instances detected in the frame. To obtain another label image  $v_{l_2}(t)$ , some of the pixels in  $v_{l_1}(t)$  are replaced by the proposed category sampling (which will be discussed in sec.3.1).

Then the label image  $v_{l_2}(t)$  and the instance map  $I(t)$  are used to generate the image  $v_{f_1}(t) \in \mathcal{R}^{3 \times H \times W}$  using the image translation with feature shift (sec.3.2), and then the actors’ regions are pasted to generate the final frame  $v_{f_2}(t)$ .

#### 3.1 Category Sampling

In the image translation stage, the frame generated from a given frame  $v(t)$  would remain the same for

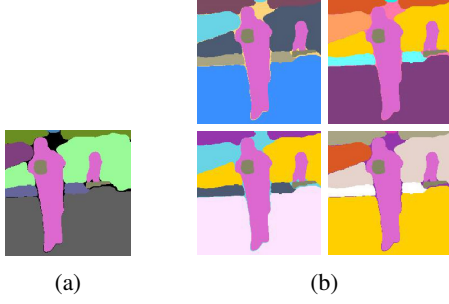


Figure 2: Category sampling. (a) A frame of a label video  $v_{l_1}$ , (b) corresponding several frames of  $v_{l_2}$ .

each epoch unless the method has a stochastic mechanism. Diffusion models have this, but we opted for a deterministic GAN for this step for the aforementioned reason. To introduce variability in the generated frames even with deterministic methods, we propose replacing the segmentation label category in the label images with a different category. We call this process *category sampling*.

This is similar to introducing noise into the latent variable (Zhu et al., 2017) to create a variety of images; however, it is difficult to maintain frame-to-frame temporal consistency. On the other hand, simply replacing the categories of labeled images can be done very quickly, and it is possible to maintain temporal continuity between frames when the categories are replaced in the same way for all frames.

The importance of which categories are replaced is a key factor in this work. We use the COCO dataset (Lin et al., 2014), which is the de fact standard for segmentation tasks. The idea is to maintain objects in the scene that are essential for understanding the actions and people-object interactions. Therefore, the COCO things (Lin et al., 2014) categories, including the person class, are kept as is, while the COCO stuff (Caesar et al., 2018) categories are replaced. In the following, we propose two methods, random sampling (for random categories) and semantic sampling (for semantically similar categories).

### 3.1.1 Random Category Sampling

We use a segmentation model pre-trained on the COCO panoptic segmentation (Kirillov et al., 2019). The category set of segmentation  $\{0, 1, \dots, L_s = 200\}$  is consists of the unlabeled class  $\{0\}$ , the things class set  $L_{\text{things}} = \{1, \dots, 91\}$ , the stuff class set  $L_{\text{stuff}} = \{92, \dots, 182\}$ , and the merged stuff class set  $L_{\text{mstuff}} = \{183, \dots, 200\}$ <sup>1</sup>.

For each video, we use a permutation  $\sigma$  that represents replacement sampling, randomly replacing cat-

egories of the COCO stuff and merged stuff to one of the categories of the COCO stuff.

$$\sigma = \begin{pmatrix} 92 & \dots & 200 \\ \sigma(92) & \dots & \sigma(200) \end{pmatrix}, \quad (1)$$

and each  $\sigma(c)$  is sampled by

$$\sigma(c) \sim \text{Unif}(L_{\text{stuff}}) \quad \forall c \in L_{\text{stuff}} \cup L_{\text{mstuff}}, \quad (2)$$

where Unif is a uniform distribution. Note that  $\{\sigma(c) \mid c \in L_{\text{stuff}} \cup L_{\text{mstuff}}\} \subseteq L_{\text{stuff}}$  holds due to the replacement. The same permutation is used for each video, and it is applied to all pixels in all frames to create a new label image;

$$v_{l_2}(x, y, t) = \sigma(v_{l_1}(x, y, t)), \quad (3)$$

where  $v(x, y, t)$  denotes the pixel values of the corresponding frame of the video  $v$ .

### 3.1.2 Semantic Category Sampling

Rather than randomly selecting categories, a category sampling that takes into account the similarity between them by using word embedding is expected to generate more realistic frames than simply substituting categories randomly. We call this sampling semantic category sampling.

First, the category name  $w_c$  of each stuff category  $c \in L_{\text{stuff}}$  is encoded into an embedding  $t_c$ . Then, we compute the cos similarity of the embedding  $t_c$  of category  $c$  to the embedding  $t_{c'}$  of other categories  $c'$ ,

$$p(c'|c) = \frac{\exp(t_c^T t_{c'})}{\sum_{i \in L_{\text{stuff}}} \exp(t_c^T t_i)}, \quad (4)$$

and sample a new category

$$c' \sim p(c'|c) \quad \forall c \in L_{\text{stuff}}. \quad (5)$$

Similarly to random category sampling, we fix  $\sigma(c) = c'$  for all frames of each video to obtain a new label image.

## 3.2 Feature Shift

It is known that frame-wise processing often results in temporal incoherency; the resulting video exhibit artifacts such as flickering between frames. In this study, we propose the use of feature shift, which has been proposed to give the ability of temporal modeling to frame-wise image recognition models (Lin et al., 2019; Zhang et al., 2021; Hashiguchi and Tamaki, 2022; Wang et al., 2022). This approach inserts feature shift modules inside or between layers of a 2D CNN or transformer model to swap parts of features between consecutive frames. We use feature shift to enhance coherency between frames.

<sup>1</sup><https://cocodataset.org/#panoptic-eval>

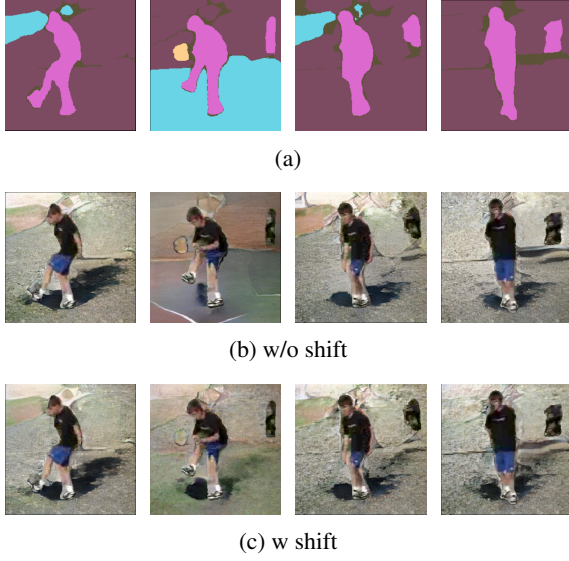


Figure 3: Effect of feature shift. (a) Label images  $V_{l_2}$ , and (b) corresponding generated frames without feature shift and (c) with feature shift.

A typical structure of image translation models consists of a combination of an encoder and a decoder, both of which are composed of multiple blocks. Assuming that there are no skip connections across blocks, we write the  $\ell$ -th decoder block as follows

$$z_\ell = D^\ell(z_{\ell-1}), \quad (6)$$

where  $z_\ell \in \mathbb{R}^{T \times C_\ell \times H_\ell \times W_\ell}$  are intermediate features and  $C_\ell, H_\ell, W_\ell$  are the number of channels, height and width. In this work, we insert a feature shift module between the decoder blocks as follows;

$$z'_\ell = D^\ell(z_{\ell-1}) \quad (7)$$

$$z_\ell = \text{shift}(z'_\ell). \quad (8)$$

Let  $z_\ell(t, c) \in \mathbb{R}^{H_\ell \times W_\ell}$  be the channel  $c$  of  $t$ -th frame of  $z_\ell$ , then the shift module can be represented as follows (Hashiguchi and Tamaki, 2022);

$$z_\ell(t, c) = \begin{cases} z'_\ell(t-1, c), & 1 < t \leq T, 1 \leq c < C_b \\ z'_\ell(t+1, c), & 1 \leq t < T, C_b \leq c < C_b + C_f \\ z'_\ell(t, c), & \forall t, C_b + C_f \leq c \leq C \end{cases}. \quad (9)$$

This means that the first  $C_b$  channels at time  $t$  are shifted backward to time  $t-1$ , and the next  $C_f$  channels are shifted forward to time  $t+1$ .

Note that we used a pre-trained image translation model (Park et al., 2019) in which shifting was not considered. However, it is expected to contribute to the reduction of artifacts between frames, as shown in Figure 3.



Figure 4: Example of person paste. (a) From the labeled moving image  $v_{l_2}$  of Figure 2(b) The generated video  $v_{f_1}$  and (b) Video image with person area pasted on it  $V_{F_2}$ .

### 3.3 Person Paste

A pre-trained image translation model might work in general; however, it does not guarantee to generate plausible actors that are import to action recognition. Therefore, we use the actor regions in the original video frame and paste them into the generated frame as shown in Fig. 4.

$$v_{f_2}(x, y, t) = \begin{cases} v(x, y, t) & v_{l_1}(x, y, t) = \text{“person”} \\ v_{f_1}(x, y, t) & \text{otherwise} \end{cases} \quad (10)$$

## 4 EXPERIMENTAL RESULTS

We evaluate the proposed S3Aug with two commonly used action recognition datasets and an out-of-context dataset. We also compare it with the conventional methods.

### 4.1 Settings

#### 4.1.1 Datasets

UCF101 (Soomro et al., 2012) is a motion recognition dataset of 101 categories of human actions, consisting of a training set of approximately 9500 videos and a validation set of approximately 3500 videos.

HMDB51 (Kuehne et al., 2011) consists of a training set of about 3600 videos and a validation set of about 1500 videos. HMDB51 is a motion recognition dataset of 51 categories of human motions.

Mimetics (Weinzaepfel and Rogez, 2021) is an evaluation-only dataset consisting of 713 videos with 50 categories, which is a subset of the category of Kinetics400 (Kay et al., 2017). Videos are out-of-context that does not align the usual context of action recognition, such as surfing in a room or bowling on a football pitch. After training on 50 categories of the Kinetics400 training set, of which videos are normal context, we evaluated on the 50 categories of the Mimetics dataset.



### 4.1.2 Model

Mask2Former (Cheng et al., 2022) pre-trained on the COCO Panoptic segmentation (Kirillov et al., 2019) (80 things, 36 stuff, 17 other, 1 unlabeled classes) was used for the segmentation of each frame of the video.

For the image translation model, we used SPADE (Park et al., 2019) pre-trained on the COCO stuff (Caesar et al., 2018) (80 things, 91 stuff, 1 unlabelled classes). The number of channels  $C_b, C_f$  to be shifted was set to  $C_\ell/8$  in every layer, where  $C_\ell$  is the number of channels of the feature  $z_\ell$ .

A pretrained BERT (Devlin et al., 2019) was used for the word embedding model. For action recognition, we used X3D-M (Feichtenhofer, 2020b) pre-trained on Kinetics400 (Kay et al., 2017).

In the experiment, only the action recognition model, X3D-M, was fine-trained, while the other models were pre-trained and fixed. For feature shift, a shift module was inserted into each decoder block of the pre-trained SPADE with the weights fixed.

### 4.1.3 Training and Evaluation

We followed a standard training setting. We randomly sampled 16 frames from a video to form a clip, randomly determined the short side of the frame in the [256, 320] pixel range, resized it to preserve the aspect ratio, and randomly cropped a  $224 \times 224$  patch. Unless otherwise noted, the number of training epochs was set at 10, batch size at 2, and learning rate at  $1e-4$  with Adam optimizer (Kingma and Ba, 2015).

In validation, we used the multiview test (Wang et al., 2018b) with 30 views; three different crops from 10 clips randomly sampled.

We applied the proposed method to each batch with probability  $0 \leq p \leq 1$ . In the experiment, the performance was evaluated from  $p = 0$  to  $p = 1$  in increments of 0.2. Note that  $p = 0$  is equivalent to no augmentation.

## 4.2 Effects of Components

Table 1 shows the effect of category sampling, feature shift (fs), and person paste (pp). Note that the results are identical for  $p = 0.0$ .

The first row shows results without any proposed modules. Performance decreases when  $p > 0$ , demonstrating that a simple image translation only does not work as a video data augmentation.

The second row shows the result of the person paste, showing that the person paste consistently improves performance for all  $p$  values. The performance decrease for large  $p$  is less significant than when the

Table 1: Evaluation of top-1 performance on the UCF101 validation set for random (r) and semantic (s) category sampling (cs), feature shift (fs), and person paste (pp).

cs	fs	pp	0.0	0.2	0.4	0.6	0.8	1.0
			93.68	92.71	93.38	90.88	89.20	74.41
		✓	93.68	94.15	94.10	91.66	90.60	85.28
r		✓	93.68	94.04	92.99	92.93	91.85	82.93
s		✓	93.68	93.99	93.63	92.96	89.83	81.71
r	✓	✓	93.68	<b>94.26</b>	93.85	92.93	91.86	82.93
s	✓	✓	93.68	<b>94.26</b>	93.54	92.77	90.49	83.73

person paste is used, indicating that the effect of the person paste is more pronounced.

Without feature shift, random sampling looks slightly better than semantic sampling as shown in the third and fourth rows. However, as shown in the last two rows, semantic category sampling shows better than or comparable performance with feature shift. The best performances of the random and semantic category sampling are the same at  $p = 0.2$ , while the semantic category sampling performs slightly better for other values of  $p$ .

Note that in all settings, performance decreases as  $p$  increases and, in particular, performance decreases significantly for  $p \geq 0.6$ , regardless of which setting was used. This indicates that the augmented samples clearly change the content of the frames and that too much augmentation does not help the model to be generalized.

## 4.3 Comparisons

The comparison with VideoMix (Yun et al., 2020) and ObjectMix (Kimata et al., 2022) on UCF101 and HMDB51 is shown in Table 2. The batch size was 16, which is the same as in the previous work (Kimata et al., 2022).

The results of the experiments vary depending on the randomness of the training and the augmentation applied. Therefore, we ran each setting three times for each method, and the results are presented in a single cell, along with the average performance of the cell.

The proposed S3Aug performs competitively on UCF101 and significantly better on HMDB51, with an average of 78.88% ( $p = 0.2$ ), which is 2 points higher than the best of VideoMix and ObjectMix. It is likely that a similar performance of the three methods is obtained on UCF101, as the data set is relatively easy to predict, and the state-of-the-art methods exceed 98% (Wang et al., 2023).

Generally, VideoMix and ObjectMix appear to be more effective when  $p$  is larger (around 0.6), while S3Aug is most successful when  $p$  is around 0.2. This discrepancy is due to the fact that the techniques generate videos in the same or different contexts.

Table 2: Performance comparison of the proposed S3Aug with two previous work; VideoMix and ObjectMix. The top one is on the validation set of UCF101 and the bottom is HMDB51.

method	0.0	0.2	0.4	0.6	0.8	1.0
VideoMix	93.40	93.18	93.49	93.60	92.96	92.66
	93.68	93.51	93.82	93.65	93.99	92.85
	94.06	94.51	94.15	93.96	94.20	93.23
avg	93.71	93.73	93.82	93.74	93.72	92.91
ObjectMix	93.40	94.07	94.10	93.68	94.10	92.74
	93.68	94.10	94.15	93.71	94.34	92.82
	94.06	94.20	94.37	94.76	94.48	93.76
avg	93.71	94.12	94.21	94.05	94.31	93.11
S3Aug	93.40	93.29	93.21	92.07	90.19	83.54
	93.68	94.04	93.54	92.77	90.49	83.73
	94.06	94.15	94.76	93.38	90.74	84.45
avg	93.71	93.83	93.84	92.74	90.47	83.91

method	0.0	0.2	0.4	0.6	0.8	1.0
VideoMix	74.22	75.11	74.33	76.67	74.56	74.42
	76.83	76.50	76.89	76.89	75.56	75.50
	78.22	77.83	78.39	77.67	76.39	75.89
avg	76.42	76.48	76.54	77.08	75.50	75.27
ObjectMix	74.22	76.33	75.56	75.33	72.17	74.72
	76.83	77.00	76.67	75.83	75.83	73.83
	78.22	77.33	78.39	77.67	76.39	75.89
avg	76.42	76.89	76.87	76.28	74.80	74.81
S3Aug	74.22	77.00	77.72	76.17	72.22	70.11
	76.83	79.81	77.89	76.50	75.83	73.94
	78.22	79.83	78.22	79.17	77.22	74.11
avg	76.42	78.88	77.94	77.28	75.09	72.72

VideoMix and ObjectMix generate new videos by utilizing two training videos, which share a similar context in terms of the background. On the other hand, S3Aug produces a video with a completely different background from the original video. We compare these methods in this paper, but our method is complementary to them, and thus a synergistic effect can be expected when they are used together.

#### 4.4 Performance on Out-of-Context Videos

One of the motivations of the proposed S3Aug is to address the issue of the background bias by generating various background while keeping the semantic layout of the action scene. Table 3 shows the performance comparisons of the proposed method and other two prior work. The top table shows performances on the same 50 categories of Kinetics validation set, which is in-context samples. Three methods are almost comparable while S3Aug is inferior due to the reason mentioned above.

The motivation behind S3Aug is to tackle the problem of background bias by creating a variety of backgrounds while preserving the semantic layout of the action scene. Table 3 compares the performance

Table 3: Evaluation of top-1 performance on 50 categories of the Kinetics (top) and Mimetics (bottom) validation sets.

method	0.0	0.2	0.4	0.6	0.8	1.0
VideoMix	81.99	79.88	78.75	79.60	79.72	78.38
ObjectMix	81.99	78.30	78.55	78.30	77.41	79.07
S3Aug	81.99	81.63	80.54	79.60	77.22	66.72

method	0.0	0.2	0.4	0.6	0.8	1.0
VideoMix	16.72	16.09	16.09	15.61	16.72	17.98
ObjectMix	16.72	15.68	15.77	16.24	13.88	17.35
S3Aug	16.72	19.30	22.37	<b>22.40</b>	19.08	16.45

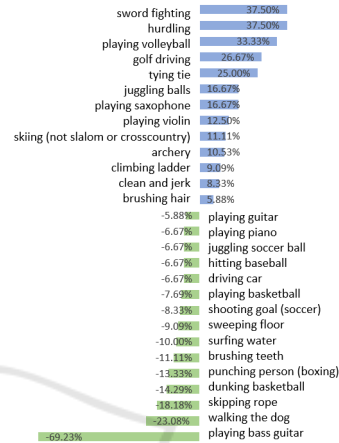


Figure 5: The score differences of 50 action categories of the Mimetics datasets when S3Aug is used and when it is not. Categories with no differences are not included in the comparison.

of the proposed method with two prior works. The top table displays the results on the same 50 categories of the Kinetics validation set, which are in-context samples. All three methods are quite similar, however S3Aug is slightly weaker due to the previously mentioned reason.

The bottom table shows the results for the Mimetics dataset, which clearly demonstrate the superiority of the proposed method. This is likely due to the various background generated by the proposed method. Figure 5 shows how scores of each category were improved or deteriorated when S3Aug is used relative to the case when it is not used ( $p = 0.0$ ). The top four categories are of sports, and training a model with generated various background may help to handle out-of-context videos of the Mimetics dataset. The worst categories look involving objects (e.g. guitar, leash for dogs, rope) that are not included as a category of the COCO dataset, or are too small to be detected by the segmentation model. This is a limitation of the proposed approach and using more sophisticated segmentation models or datasets with fine categories would be helpful.

## 5 CONCLUSION

In this study, we proposed S3Aug, a video data augmentation for action recognition using segmentation, category sampling, image generation, and feature shift. The proposed method is different from conventional data augmentation methods that cut and paste object regions from two videos in that it generates a label video from a single video by segmentation and creates a new video by image translation. Experiments using UCF101 and HMDB51 have confirmed that UCF101 is effective as a data augmentation method to suppress overfit during training.

## ACKNOWLEDGEMENTS

This work was supported in part by JSPS KAKENHI Grant Number JP22K12090.

## REFERENCES

- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846.
- Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.
- Caesar, H., Uijlings, J., and Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Cauli, N. and Reforgiato Recupero, D. (2022). Survey on videos data augmentation for deep learning models. *Future Internet*, 14(3).
- Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation.
- Chung, J., Wu, Y., and Russakovsky, O. (2022). Enabling detailed action recognition evaluation through video dataset augmentation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 39020–39033. Curran Associates, Inc.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Feichtenhofer, C. (2020a). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Feichtenhofer, C. (2020b). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.
- Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Gowda, S. N., Rohrbach, M., Keller, F., and Sevilla-Lara, L. (2022). Learn2augment: Learning to composite videos for data augmentation in action recognition. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 242–259, Cham. Springer Nature Switzerland.
- Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thureau, C., Bax, I., and Memisevic, R. (2017). The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Hashiguchi, R. and Tamaki, T. (2022). Temporal cross-attention for action recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*, pages 276–288.
- He, Y., Shirakabe, S., Satoh, Y., and Kataoka, H. (2016). Human action recognition without human. In Hua, G. and Jégou, H., editors, *Computer Vision – ECCV 2016 Workshops*, pages 11–17, Cham. Springer International Publishing.
- Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video diffusion models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc.
- Hutchinson, M. S. and Gadepally, V. N. (2021). Video action understanding. *IEEE Access*, 9:134611–134637.
- Jabbar, A., Li, X., and Omar, B. (2021). A survey on gener-

- ative adversarial networks: Variants, applications, and training. *ACM Computing Surveys*, 54(8).
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The kinetics human action video dataset. *CoRR*, abs/1705.06950.
- Kimata, J., Nitta, T., and Tamaki, T. (2022). Objectmix: Data augmentation by copy-pasting objects in videos for action recognition. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia, MMAsia '22*, New York, NY, USA. Association for Computing Machinery.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Kirillov, A., He, K., Girshick, R., Rother, C., and Dollar, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kong, Y. and Fu, Y. (2022). Human action recognition and prediction: A survey. *Int. J. Comput. Vis.*, 130(5):1366–1401.
- Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. A., and Serre, T. (2011). HMDB: A large video database for human motion recognition. In Metaxas, D. N., Quan, L., Sanfeliu, A., and Gool, L. V., editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2556–2563. IEEE Computer Society.
- Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., and Tan, T. (2023). Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218.
- Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60.
- Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402.
- Ulhaq, A., Akhtar, N., Pogrebna, G., and Mian, A. (2022). Vision transformers for action recognition: A survey. *CoRR*, abs/2209.05700.
- Wang, G., Zhao, Y., Tang, C., Luo, C., and Zeng, W. (2022). When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. *CoRR*, abs/2201.10801.
- Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. (2023). Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018a). Video-to-video synthesis. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018b). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Weinzaepfel, P. and Rogez, G. (2021). Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690.
- Wu, D., Chen, J., Sharma, N., Pan, S., Long, G., and Blumenstein, M. (2019). Adversarial action data augmentation for similar gesture action recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. (2021). Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157.
- Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552.
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Yun, S., Oh, S. J., Heo, B., Han, D., and Kim, J. (2020). Videomix: Rethinking data augmentation for video classification.



- Zhang, H., Hao, Y., and Ngo, C.-W. (2021). Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia, MM '21*, page 917–925, New York, NY, USA. Association for Computing Machinery.
- Zhang, L. and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *CoRR*, abs/2302.05543.
- Zhang, Y., Jia, G., Chen, L., Zhang, M., and Yong, J. (2020). Self-paced video data augmentation by generative adversarial networks with insufficient samples. In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, page 1652–1660, New York, NY, USA. Association for Computing Machinery.
- Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017). Toward multi-modal image-to-image translation.
- Zou, Y., Choi, J., Wang, Q., and Huang, J.-B. (2022). Learning representational invariances for data-efficient action recognition. *Computer Vision and Image Understanding*, page 103597.

