Enabling Detailed Action Recognition Evaluation Through Video Dataset Augmentation: Supplementary Material

1 Ablation Study

1.1 Selection of Segmentation Model

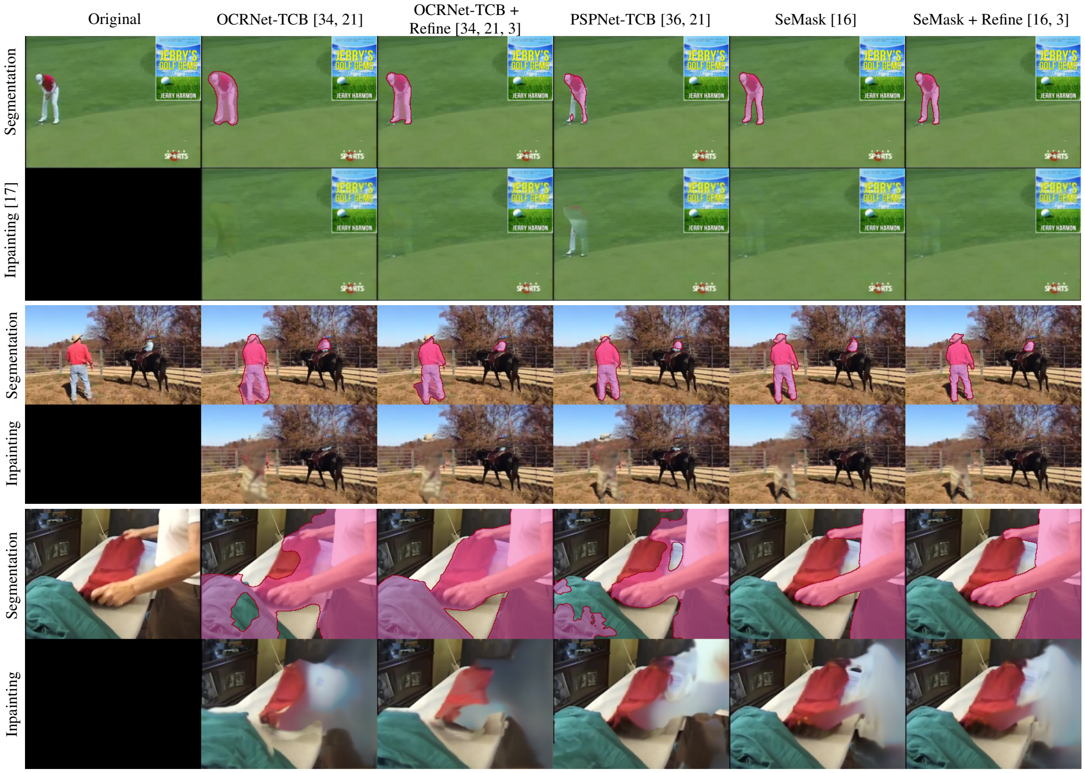

Figure 1: Segmentation results of different models. Check attached files for the video version of the figure.

While main contribution of our work is the toolkit framework, where segmentation model and the video inpainting model can be replaced to a better performing model, we show why we decided with the current choice for this paper. We qualitatively show the difference between segmentation models and how they perform differently in our toolkit. We have tested TCB [21][1] using PSPNet [36] or OCRNet [34] as a backbone, and SeMask [16]. We have also tested using a segmentation refinement method [3]. Figure 1 shows some of the examples using different segmentation models and the inpainting results. TCB methods show inaccurate boundary, often times with larger segmentation that the actual human. While using segmentation refinement model [3] can help, but when if the initial segmentation is bad (e.g., the last row), segmentation refinement only gets the accurate boundary of the unrelated objects. SeMask [16] already shows a good boundary so the refinement does not help. Moreover, we see that SeMask has better temporal robustness, where we do not see large change of segmentation for different frames. For instance, we see that on the last row, first three segmentation

---

[1] We contacted the original authors for the trained weights.

models clearly show the brown pants of the person when inpainted, as that part of the segmentation has been missing in the neighbouring frames (See attached video for better understanding). We decide to use SeMask without any additional refinement.

1.2 Need of Video Inpainting Model

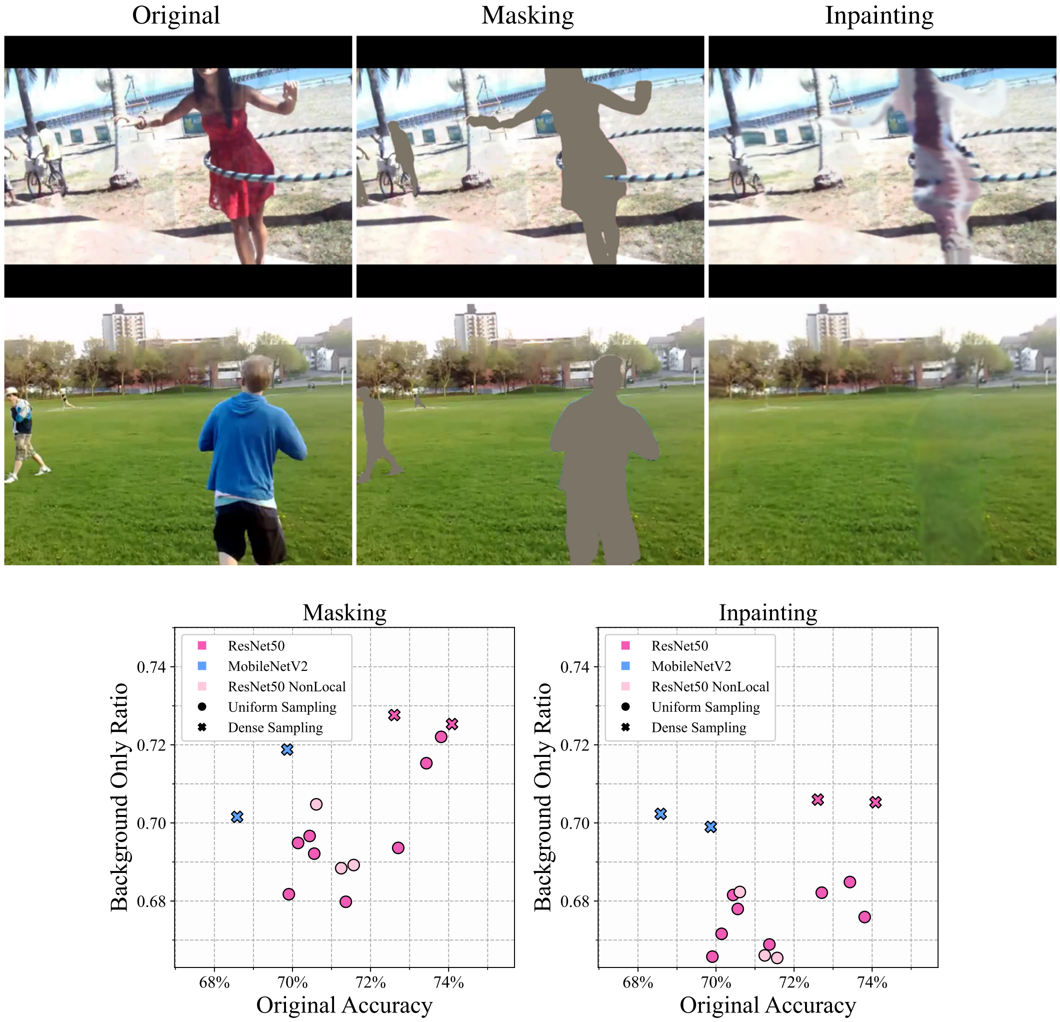

Figure 2: We compare Background Only Ratio using naïve masking and VINet inpainting. **Top:** Example frames using masking and inpainting. **Bottom:** BOR of different methods on TSM [19]

Here, we compare the performance between TSM models using VINet [17] or naïve inpainting method of filling the dataset average color. Figure 2 visualizes the method samples and the experiment results. Overall, using using masking shows higher background only accuracy. We expect this to be the model being able to see the shape of the human body in the masking method, thus the body information is not well obfuscated. This revisits the idea that the modern inpainting tool is crucial for evaluating human understanding of action recognition models. However, revealing of the body shape can be also seen in the inpainting method in some cases (e.g., top row of Figure 2). We believe having better inpainting model can resolve this, which we explain in Section 1.3.

1.3 Video Inpainting Model Comparisons

In this toolkit we used Deep Video Inpainting [17] for our inpainting module. However, a recent video inpainting module E2FGVI [18] was published in CVPR2022 and it shows better result than

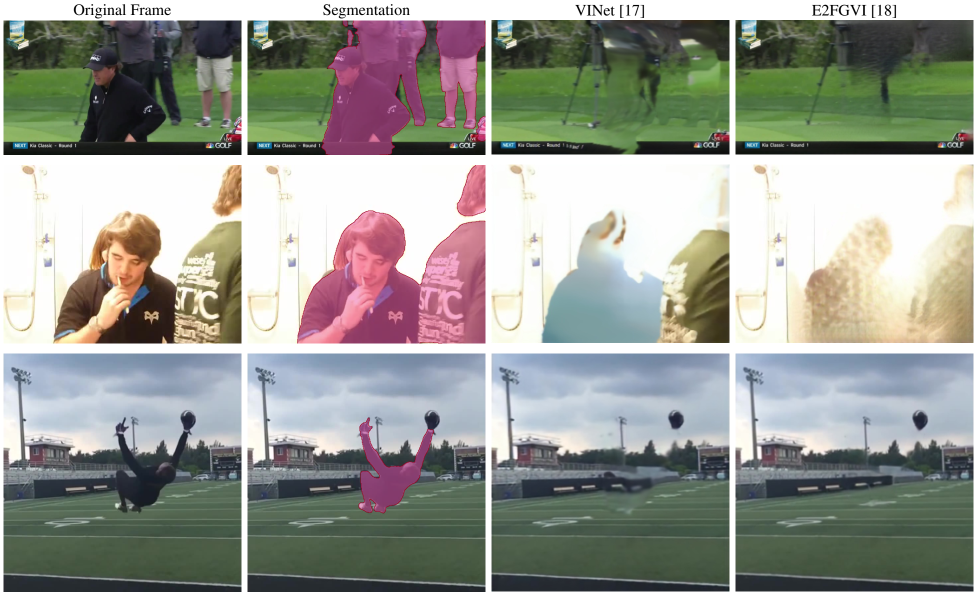

Figure 3: Inpainting results of VINet and E2FGVI. Although E2FGVI is more accurate, the textural artifacts could have worse effect to classification models. Check attached files for the video version of the figure.

our current inpainting module. We compare VINet [17] and E2FGVI [18] on Figure 3. While E2FGVI shows 'accurate' results, for example showing accurate black border on the first row of Figure 3 or non-distorted background panel in the third row. However, on many cases, E2FGVI shows textural artifacts, e.g., texture on the first and the second example. In similar cases, VINet tend to use inaccurate inpainting (e.g., grey blob) but the texture is more realistic. We believe such issues of E2FGVI can cause an issue as classification models tend to rely more on the textural information more than the shape information [12]. As the codes for E2FGVI were available recently, we did not have enough time to experiment with E2FGVI, and we plan to update our toolkit with more accurate video inpainting module in the future.

2 Additional Related Work

Out-of-context The context of an image refers to the relationship between the natural scenes and the object. It has been long known that the context is an important cue to the object recognition system [31, 35]. There is nothing intrinsically wrong of using contextual cues, as it has been known that human as well uses contextual information [22]. However, over-reliance [11] on the context can lead to inaccurate predictions on out-of-context objects [4]. This is mostly severe to machine learning systems where the model learns the context from the dataset distribution, which can be heavily biased due to its collection strategy.

Similarly, there is a well recognized belief that the deep-learning models tend to look at the background when predicting a human action. This is especially believed to be true for datasets that cover wide range of classes as in-the-wild videos, as the collected videos are heavily correlated with the scene and the background objects (e.g., videos of dribbling a basketball shows basketball hoop on the background). The belief is backed up by series of researches. [5, 24] have shown that the performance of a machine learning model when feeding a single frame shows very similar performance with the ones that are fed multiple frames in UCF101 [26] and ActivityNet [8]. Another study [15] shows that the performance of a model trained/tested with a video with a person masked out is on par with a model trained without any mask.

3 Additional Experiment

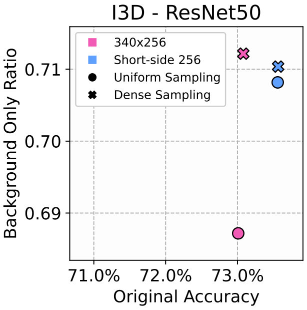

Figure 4: BOR on I3D

Extension of Section 4.2. Figure 3 shows effect of dense sampling strategy on I3D [2]. Colors show resolution used for training and testing. Similar to as TSM [19], dense sampling strategy shows that I3D models tend to rely lot more on video background. It also shows that the original accuracy cannot distinguish the difference between dense sampling strategy and uniform sampling strategy.

Table 1: SHAcc and SBErr per category. Human Action Category denotes the class label of the human action in the swapped video. Background Category denote the class label of the video where the background is from. **Top:** Categories with high SHAcc, **Bottom:** Categories with high SBErr.

Human Action Category			Background Category		
Category	SHAcc	SBErr	Category	SHAcc	SBErr
`belly dancing`	75.36	14.49	`exercising arm`	46.34	7.32
`carrying baby`	74.67	17.33	`getting a tattoo`	42.11	2.63
`capoeira`	69.57	17.39	`answering questions`	40.62	0.00
`riding mule`	0.00	75.00	`playing chess`	4.65	93.02
`bouncing on trampoline`	0.00	60.00	`cutting pineapple`	2.38	90.48
`scuba diving`	5.56	52.78	`parasailing`	2.04	79.59

Category-wise Accuracy. To understand how a human action recognizer performs on different categories of Action Swap Dataset, we break down the performance into each category. Here, we chose SlowFast [10] with ResNet50 [14] as the recognizer and Random Swap for the swap type.

Table 1 tabulates Human Action/Background categories and their SHAcc and SBErr. On the left, we tabulate the Human Action category, e.g., Among the swapped video where the human is from the class `belly dancing`, 75.36% of such videos, the model predict `belly dancing`, while 14.49% of videos were predicted the background. Similarly we tabulate analysis using background category on the right.

The model tends to predict the class of the action when the action only involves the human body and is often performed in various backgrounds. Meanwhile, actions that use specific objects (`riding mule`) or happen in a specific background (`scuba diving`) can not be well recognized by the model after the random swap. In such actions, models turn to predict from the background instead.

As for background categories, when the action does not have a specific background (e.g., `exercising arm` often happens under diverse settings in Kinetics-400), the model can predict using the pasted human body relatively well. On the other hand, if the action uses a very specific object (e.g., pineapple), has a distinctive object in the background (chessboard), or happens in a specific background (sky), the model leans towards using the background for its prediction.

3.1 Analysing Human-Only Videos

Table 2 tabulates the result from using human-only videos.

Table 2: Experiments over Human Only Videos. HAcc denotes accuracy on Human Only Videos. Check supplementary material for the full experiment results.

Model	Backbone	Pre-trained	OAcc (%)	HAcc (%)	$\frac{\text{HAcc}}{\text{OAcc}}$
Normal-scale dataset					
TSM [19]	MNetV2 [23]	ImageNet	69.87	20.27	0.2902
R(2+1)D [27]	ResNet34	-	74.22	20.95	0.2822
TSN [29]	ResNet50	ImageNet	71.75	17.92	0.2498
TIN [25]	ResNet50	TSM-Kinetics400	70.89	20.40	0.2877
TSM [19]	ResNet50	ImageNet	74.09	24.25	0.3273
I3D [2]	ResNet50	ImageNet	73.57	23.89	0.3247
NL-TSM [30]	ResNet50	ImageNet	71.57	19.75	0.2759
NL-I3D [30]	ResNet50	ImageNet	74.91	19.42	0.2592
NL-SlowOnly [30]	ResNet50	ImageNet	75.78	17.99	0.2374
CSN [28]	ResNet50	-	73.22	24.91	0.3403
TPN [33]	ResNet50	ImageNet	76.16	25.26	0.3316
SlowOnly [10]	ResNet50	ImageNet	75.35	22.30	0.2959
SlowFast [10]	ResNet50	-	76.61	25.23	0.3294
SlowOnly [10]	ResNet101	-	76.26	26.90	0.3528
SlowFast [10]	ResNet101+50	-	76.55	24.23	0.3166
SlowFast [10]	ResNet101	-	**78.10**	26.90	0.3445
CSN [28]	ResNet152	-	77.62	**28.33**	**0.3650**
SlowFast [10]	ResNet152+50	-	77.24	27.20	0.3521
X3D [9]	X3D_S	-	72.67	19.90	0.2738
X3D [9]	X3D_M	-	75.55	22.27	0.2948
TANet [20]	TANet	ImageNet	76.10	23.64	0.3107
Large-scale dataset					
TSN [29]	ResNet50	IG-1B [32]	70.96	15.84	0.2232
Omni-TSN [7]	ResNet50	IG-1B [32]	74.70	20.38	0.2728
Omni-SlowOnly [7]	ResNet50	-	76.49	25.18	0.3292
CSN [28]	ResNet50	IG65M [13]	79.09	29.48	0.3727
Omni-SlowOnly [7]	ResNet101	-	80.00	31.81	0.3976
CSN [28]	ResNet152	IG65M [13]	**82.38**	**33.58**	**0.4076**
TimeSFormer [1]	TimeSformer	ImageNet-21K [6]	77.97	22.27	0.2856

4 Qualitative Results on Random Action Swap

5 Experiments on UCF101 dataset

Here, we show the toolkit used on top of the UCF101 [26] dataset. Figure 6 visualizes example frames of modified UCF101.

5.1 Results

Table 3: Background Only Video and Human Only Video results on UCF101. C×S×N denotes Clip Length × Stride × Number of Clips.

#	Type	Backbone	Pre-trained	C×S×N	OAcc	BAcc	BOR	HAcc	HOR	Model Details	
1	TSN	ResNet50	ImageNet	1x1x3	81.31	64.34	0.7913	10.84	0.1333	ckpt	py
2	SlowOnly	ResNet50	Kinetics-400	8x4x1	92.60	70.74	0.7639	36.11	0.3900	ckpt	py
3	SlowOnly	ResNet50	ImageNet	8x4x1	71.08	54.06	0.7605	13.27	0.1867	ckpt	py
4	C3D*	C3D	Sports1M	16x1x1	73.83	49.33	0.6681	16.15	0.2188	ckpt	py
5	TSM	ResNet50	Kinetics-400	1x1x16	94.32	70.53	0.7478	41.85	0.4437	ckpt	py
6	TSM	ResNet50	Kinetics-400	1x1x8	94.16	69.92	0.7426	38.59	0.4099	ckpt	py

*: We were not able to get the reported accuracy of C3D. This is a known issue in MMAction2 implementation of C3D.

5.2 Discussion

The results show similar conclusion as what we see from the modified Kinetics-400. Notable discovery is that TSM show high SHAcc on Table 4, higher than SBErr. We see this to be coming

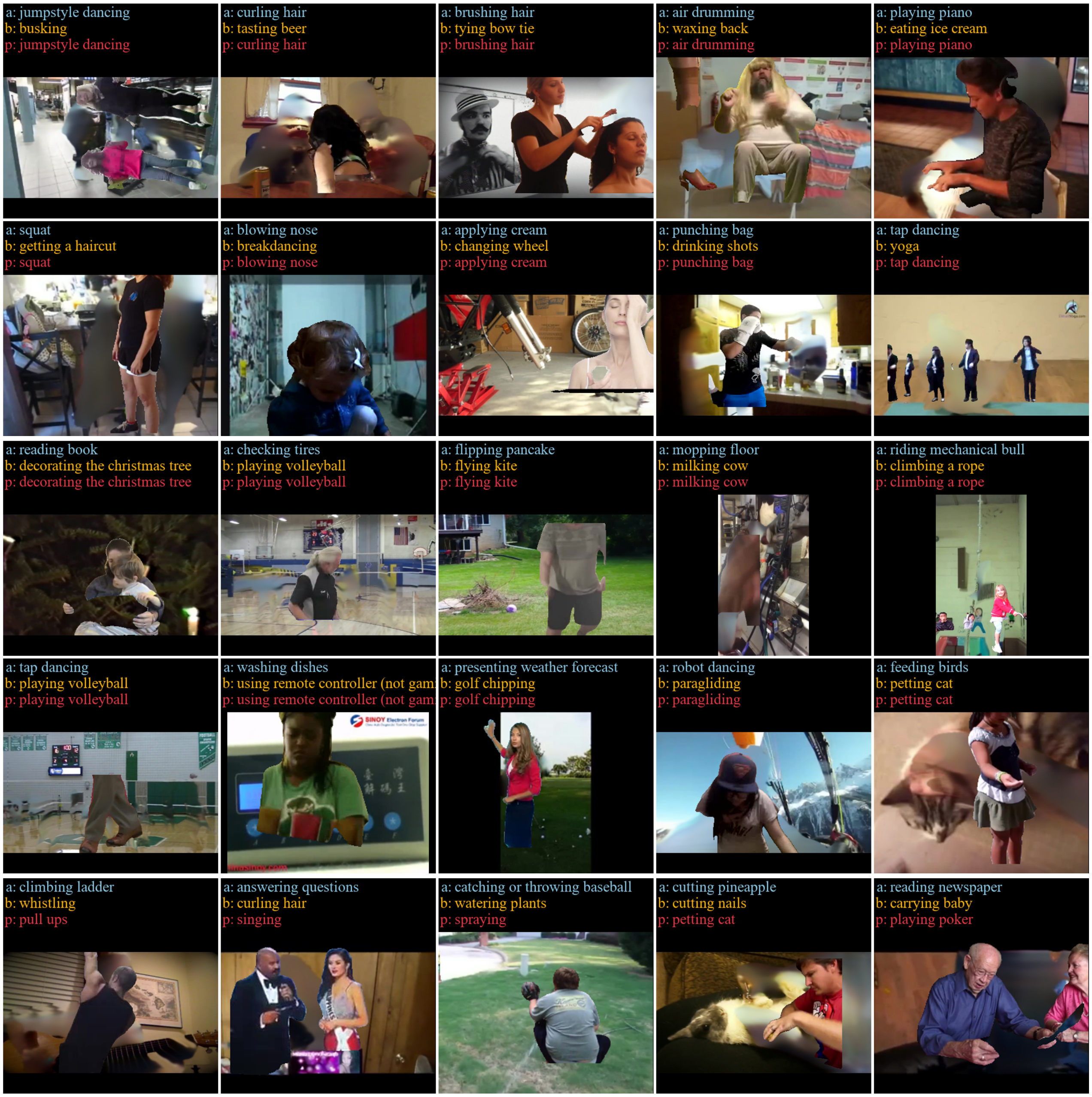

Figure 5: Random Swap Qualitative Results on ip-CSN [28]. a,b, and p denote action category, background category and prediction, respectively. Row 1-2: Samples where the prediction matches the action. Row 3-4: Samples where the prediction matches the background. Row 5: Prediction does not match any.

from Kinetics-400 pre-training, i.e., a model pre-trained with Kinetics-400 is heavily benefiting the downstream task of the smaller dataset. This can be seen from comparing #2 and #3 where a model pre-trained with Kinetics-400 is showing far superior performance than the same model pre-trained with ImageNet, under both original accuracy or the metrics we introduced.

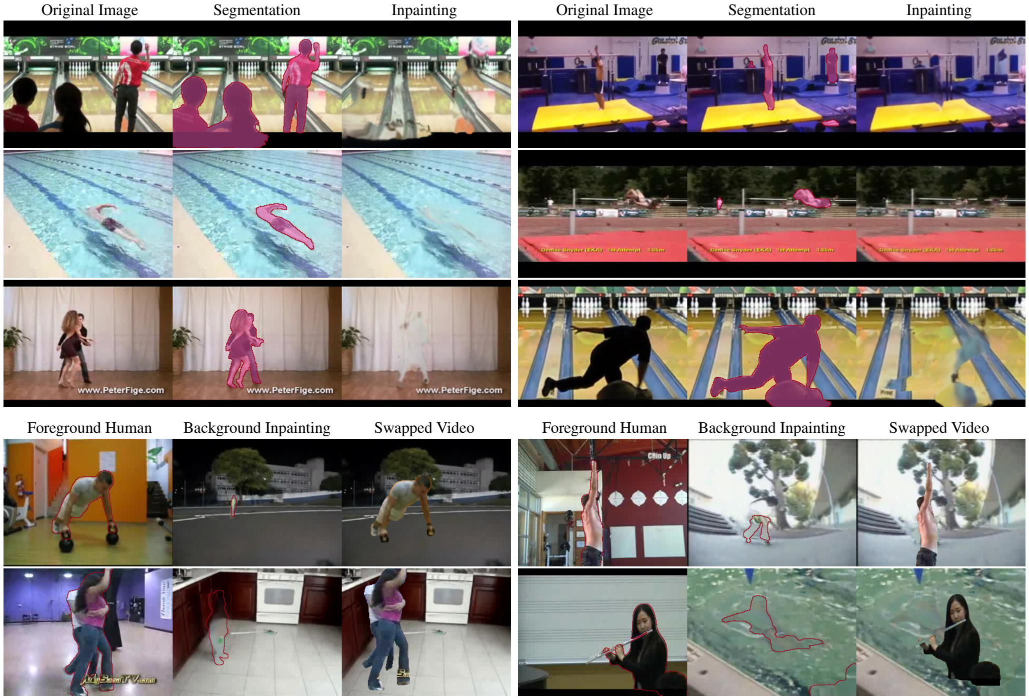

Figure 6: Human-centric Action Analysis on UCF101. First three rows show segmentation and inpainting results. The last two rows show Action Swap videos on UCF101. We draw human segmentation boundary for visibility.

Table 4: Action Swap Video results on UCF101. The numbers are averaged from 3 different random seeds.

#	Type	Backbone	Pre-trained	C×S×N	Random Swap SHAcc	Random Swap SBErr	Close Swap SHAcc	Close Swap SBErr	Far Swap SHAcc	Far Swap SBErr	Same Swap SHAcc	Model Details	
1	TSN	ResNet50	ImageNet	1x1x3	$13.12_{\pm 0.27}$	$31.88_{\pm 1.42}$	$21.53_{\pm 0.4}$	$31.54_{\pm 0.32}$	$9.82_{\pm 0.73}$	$33.06_{\pm 0.71}$	$67.93_{\pm 0.53}$	ckpt	py
2	SlowOnly	ResNet50	Kinetics-400	8x4x1	$33.71_{\pm 0.59}$	$39.22_{\pm 0.65}$	$45.91_{\pm 0.83}$	$32.86_{\pm 1.24}$	$26.96_{\pm 0.43}$	$47.52_{\pm 0.26}$	$90.91_{\pm 0.29}$	ckpt	py
3	SlowOnly	ResNet50	ImageNet	8x4x1	$9.74_{\pm 0.25}$	$35.28_{\pm 1.04}$	$16.37_{\pm 0.89}$	$29.12_{\pm 0.31}$	$7.34_{\pm 0.06}$	$43.03_{\pm 0.88}$	$59.43_{\pm 1.21}$	ckpt	py
4	C3D	C3D	Sports1M	16x1x1	$10.61_{\pm 0.19}$	$22.22_{\pm 0.39}$	$17.73_{\pm 0.27}$	$21.27_{\pm 0.24}$	$8.34_{\pm 0.14}$	$26.76_{\pm 1.82}$	$53.42_{\pm 0.26}$	ckpt	py
5	TSM	ResNet50	Kinetics-400	1x1x16	$40.42_{\pm 0.55}$	$26.84_{\pm 1.33}$	$49.11_{\pm 0.71}$	$26.51_{\pm 0.58}$	$36.07_{\pm 0.89}$	$32.15_{\pm 1.16}$	$89.87_{\pm 0.42}$	ckpt	py
6	TSM	ResNet50	Kinetics-400	1x1x8	$39.81_{\pm 1.11}$	$26.43_{\pm 1.18}$	$49.21_{\pm 1.16}$	$25.88_{\pm 0.53}$	$36.52_{\pm 0.68}$	$29.67_{\pm 0.9}$	$88.88_{\pm 0.53}$	ckpt	py

6 Downloads

We have uploaded the modified Kinetics-400 on Google Drive[2]. The link offers extracted frames of original Kinetics-400, segmentation mask, and inpainted version. As we draw Action Swap online, we did not upload any files for such. Instead, we offer visualizing code of Action Swap and action swap pair separately.

7 Details of the models we tested

We detail all 74 models we trained on Table 5. Since every model has some specific configurations, we cannot list every detail in this paper. We list only some of the characteristics we used in the paper. Rows that look like a duplicate have unlisted features that differentiate each other. E.g., #16 is trained for 50 epochs and #17 is trained for 100 epochs. #62 - #64 is three different implementations of TimeSFormer [1] listed in the original paper. The details of the settings of each row can be viewed by clicking the 'config' link.

---

[2] `https://drive.google.com/drive/folders/1IiLGMykqUjQp0IFHB-Yl7Hk1Yzzswrgx?usp=sharing`

8 Full experiment

Here we list the all the experiment in the main paper. The order of the models we tested are the same as Table 5.

Background Only Videos Table 6 tabulates full experiment results for Background Only Videos.

Human Only Videos Table 7 tabulates full experiment results for Human Only Videos.

Action Swap Videos Table 8 tabulates full experiment results for Action Swap Videos.

9 Details of Action-Swap

Here, we detail the method of our action-swap. Note that our method is only dependent on videos, but not on sampling strategy or image resizing pre-process.

Temporal Alignment As two videos can have different length when performing action-swap, for the generated video, we follow the length of the foreground action. For each of the action frames, corresponding new background is taken from the same relative temporal position. For example, assume the foreground action video has 100 frames, and the background video has 200 frames. The resulting action-swap video will take frame index of:

$$[[0,0],[1,2],\ldots,[\text{index of foreground video frames, index of background video frames},\ldots,[99,198]]]$$

Please check `https://github.com/princetonvisualai/HAT/blob/main/mmaction2/mmaction/datasets/actionswap_dataset.py#L258` for the code implementation.

Spatial Alignment When merging foreground and background when given two frames, we first use the original resolution for both frames. As both frames can have different resolution, we follow the resolution of the background. We use inpainted background frame as the base, and paste the foreground action on top of the background. For the first frame of the Action-Swap, we perform paste so that the foreground and the background has the same weight-center of the human segmentation. We keep this alignment and use the same alignment for the rest of the frames. This is done to keep the human body movement relative to the camera window. E.g., if a person moves from left to right in the camera window in the original video, we want to keep the same movement when the background is swapped.

Table 5: Details of all the models we have tested.

#	Dataset	Model Structure	Backbone	Pre-trained	Resolution	Dense Sampling	Clip Length	Stride	Number of Clips	Trained Weights	Config
1	Normal Scale	TSN	ResNet50	ImageNet	340x256	Uniform	1	1	3	ckpt	config
2			ResNet50	ImageNet	short-side 256	Uniform	1	1	3	ckpt	config
3			ResNet50	ImageNet	340x256	Dense	1	1	5	ckpt	config
4			ResNet50	ImageNet	short-side 320	Uniform	1	1	3	ckpt	config
5			ResNet50	ImageNet	short-side 256	Uniform	1	1	8	ckpt	config
6			ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
7			ResNet50	ImageNet	340x256	Dense	1	1	8	ckpt	config
8		R(2+1)D	ResNet34	None	short-side 256	Uniform	8	8	1	ckpt	config
9			ResNet34	None	short-side 320	Uniform	8	8	1	ckpt	config
10			ResNet34	None	short-side 320	Uniform	32	2	1	ckpt	config
11		TIN	ResNet50	TSM-Kinetics400	short-side 256	Uniform	1	1	8	ckpt	config
12		TSM	MobileNetV2	ImageNet	short-side 320	Dense	1	1	8	ckpt	config
13			MobileNetV2	ImageNet	short-side 320	Dense	1	1	8	ckpt	config
14			ResNet50	ImageNet	340x256	Uniform	1	1	8	ckpt	config
15			ResNet50	ImageNet	short-side 256	Uniform	1	1	8	ckpt	config
16			ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
17			ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
18			ResNet50	ImageNet	short-side 256	Uniform	1	1	8	ckpt	config
19			ResNet50	ImageNet	short-side 320	Dense	1	1	8	ckpt	config
20			ResNet50	ImageNet	short-side 320	Dense	1	1	8	ckpt	config
21			ResNet50	ImageNet	340x256	Uniform	1	1	16	ckpt	config
22			ResNet50	ImageNet	short-side 256	Uniform	1	1	16	ckpt	config
23			ResNet50	ImageNet	short-side 320	Uniform	1	1	16	ckpt	config
24		I3D	ResNet50	ImageNet	340x256	Uniform	32	2	1	ckpt	config
25			ResNet50	ImageNet	short-side 256	Uniform	32	2	1	ckpt	config
26			ResNet50	ImageNet	340x256	Dense	32	2	1	ckpt	config
27			ResNet50	ImageNet	short-side 256	Dense	32	2	1	ckpt	config
28		NL-TSM	ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
29			ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
30			ResNet50	ImageNet	short-side 320	Uniform	1	1	8	ckpt	config
31		NL-I3D	ResNet50	ImageNet	short-side 256p	Uniform	32	2	1	ckpt	config
32			ResNet50	ImageNet	short-side 256p	Uniform	32	2	1	ckpt	config
33			ResNet50	ImageNet	short-side 256p	Uniform	32	2	1	ckpt	config
34		NL-SlowOnly	ResNet50	ImageNet	short-side 320	Uniform	4	16	1	ckpt	config
35			ResNet50	ImageNet	short-side 320	Uniform	8	8	1	ckpt	config
36		CSN	ResNet50	None	short-side 320	Uniform	32	2	1	ckpt	config
37			ResNet152	None	short-side 320	Uniform	32	2	1	ckpt	config
38			ResNet152	None	short-side 320	Uniform	32	2	1	ckpt	config
39		TPN	ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
40			ResNet50	ImageNet	short-side 320	Uniform	8	8	1	ckpt	config
41		X3D	X3D_S	None	short-side 320	Uniform	13	6	1	ckpt	config
42			X3D_M	None	short-side 320	Uniform	16	5	1	ckpt	config
43		TANet	TANet	ImageNet	short-side 320	Dense	1	1	8	ckpt	config
44		SlowOnly	ResNet50	None	short-side 256	Uniform	4	16	1	ckpt	config
45			ResNet50	None	short-side 256	Uniform	8	8	1	ckpt	config
46			ResNet50	None	short-side 320	Uniform	4	16	1	ckpt	config
47			ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
48			ResNet50	ImageNet	short-side 320	Uniform	4	16	1	ckpt	config
49			ResNet50	ImageNet	short-side 320	Uniform	8	8	1	ckpt	config
50			ResNet101	None	short-side 320	Uniform	8	8	1	ckpt	config
51		SlowFast	ResNet50	None	short-side 256	Uniform	4	16	1	ckpt	config
52			ResNet50	None	short-side 320	Uniform	4	16	1	ckpt	config
53			ResNet50	None	short-side 320	Uniform	4	16	1	ckpt	config
54			ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
55			ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
56			ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
57			ResNet50	None	short-side 320	Uniform	8	8	1	ckpt	config
58			ResNet101 + ResNet50	None	short-side 256	Uniform	4	16	1	ckpt	config
59			ResNet152 + ResNet50	None	short-side 256	Uniform	4	16	1	ckpt	config
60			ResNet101	None	short-side 256	Uniform	8	8	1	ckpt	config
61	Large Scale	TSN	ResNet50	IG-1B	short-side 320	Uniform	1	1	3	ckpt	config
62		TimeSFormer	TimeSformer	ImageNet-21K	short-side 320	Uniform	8	32	1	ckpt	config
63			TimeSformer	ImageNet-21K	short-side 320	Uniform	8	32	1	ckpt	config
64			TimeSformer	ImageNet-21K	short-side 320	Uniform	8	32	1	ckpt	config
65		Omni-TSN	ResNet50	ImageNet	340x256	Uniform	1	1	3	ckpt	config
66			ResNet50	IG-1B	short-side 320	Uniform	1	1	3	ckpt	config
67		Omni-SlowOnly	ResNet50	None	short-side 320	Uniform	4	16	1	ckpt	config
68			ResNet101	None	short-side 320	Uniform	8	8	1	ckpt	config
69		CSN	ResNet50	IG65M	short-side 320	Uniform	32	2	1	ckpt	config
70			ResNet152	Sports1M	short-side 320	Uniform	32	2	1	ckpt	config
71			ResNet152	Sports1M	short-side 320	Uniform	32	2	1	ckpt	config
72			ResNet152	IG65M	short-side 320	Uniform	32	2	1	ckpt	config
73			ResNet152	IG65M	short-side 320	Uniform	32	2	1	ckpt	config
74			ResNet152	IG65M	short-side 320	Uniform	32	2	1	ckpt	config

Table 6: Full experiment result for Background Only videos.

#	Dataset	Model Structure	Backbone	Pre-trained	Sampling	Original Accuracy	Background Accuracy	Ratio
1	Normal Scale	TSN	ResNet50	ImageNet	Uniform	68.93	47.82	0.6937
2			ResNet50	ImageNet	Uniform	69.43	48.78	0.7025
3			ResNet50	ImageNet	Dense	68.37	46.72	0.6833
4			ResNet50	ImageNet	Uniform	70.06	48.43	0.6913
5			ResNet50	ImageNet	Uniform	71.12	48.00	0.6748
6			ResNet50	ImageNet	Uniform	71.75	49.02	0.6833
7			ResNet50	ImageNet	Dense	68.98	46.81	0.6787
8		R(2+1)D	ResNet34	None	Uniform	67.35	48.20	0.7157
9			ResNet34	None	Uniform	69.13	49.36	0.7140
10			ResNet34	None	Uniform	74.22	52.99	0.7140
11		TIN	ResNet50	TSM-Kinetics400	Uniform	70.89	48.32	0.6816
12		TSM	MobileNetV2	ImageNet	Dense	68.58	48.17	0.7023
13			MobileNetV2	ImageNet	Dense	69.87	48.84	0.6990
14			ResNet50	ImageNet	Uniform	69.91	46.54	0.6658
15			ResNet50	ImageNet	Uniform	70.56	47.84	0.6780
16			ResNet50	ImageNet	Uniform	70.14	47.11	0.6716
17			ResNet50	ImageNet	Uniform	71.37	47.74	0.6689
18			ResNet50	ImageNet	Uniform	70.44	48.01	0.6816
19			ResNet50	ImageNet	Dense	72.61	51.26	0.7060
20			ResNet50	ImageNet	Dense	74.09	52.25	0.7053
21			ResNet50	ImageNet	Uniform	72.71	49.60	0.6821
22			ResNet50	ImageNet	Uniform	73.43	50.29	0.6849
23			ResNet50	ImageNet	Uniform	73.81	49.89	0.6759
24		I3D	ResNet50	ImageNet	Uniform	73.01	50.17	0.6872
25			ResNet50	ImageNet	Uniform	73.56	52.09	0.7082
26			ResNet50	ImageNet	Dense	73.08	52.05	0.7122
27			ResNet50	ImageNet	Dense	73.57	52.26	0.7104
28		NL-TSM	ResNet50	ImageNet	Uniform	71.57	47.62	0.6654
29			ResNet50	ImageNet	Uniform	70.61	48.18	0.6823
30			ResNet50	ImageNet	Uniform	71.25	47.46	0.6661
31		NL-I3D	ResNet50	ImageNet	Uniform	74.91	52.84	0.7054
32			ResNet50	ImageNet	Uniform	73.60	52.35	0.7114
33			ResNet50	ImageNet	Uniform	74.00	52.96	0.7156
34		NL-SlowOnly	ResNet50	ImageNet	Uniform	74.52	53.17	0.7135
35			ResNet50	ImageNet	Uniform	75.78	53.51	0.7062
36		CSN	ResNet50	None	Uniform	73.22	51.97	0.7098
37			ResNet152	None	Uniform	76.30	53.70	0.7038
38			ResNet152	None	Uniform	77.62	54.33	0.6999
39		TPN	ResNet50	None	Uniform	73.52	51.76	0.7040
40			ResNet50	ImageNet	Uniform	76.16	54.40	0.7143
41		X3D	X3D_S	None	Uniform	72.67	50.61	0.6964
42			X3D_M	None	Uniform	75.55	52.47	0.6944
43		TANet	TANet	ImageNet	Dense	76.10	53.71	0.7059
44		SlowOnly	ResNet50	None	Uniform	72.79	51.67	0.7099
45			ResNet50	None	Uniform	74.60	53.15	0.7125
46			ResNet50	None	Uniform	72.72	51.60	0.7096
47			ResNet50	None	Uniform	72.55	50.82	0.7004
48			ResNet50	ImageNet	Uniform	73.33	52.58	0.7171
49			ResNet50	ImageNet	Uniform	75.35	53.97	0.7163
50			ResNet101	None	Uniform	76.26	54.38	0.7131
51		SlowFast	ResNet50	None	Uniform	74.78	53.13	0.7105
52			ResNet50	None	Uniform	75.81	53.33	0.7035
53			ResNet50	None	Uniform	76.02	54.26	0.7138
54			ResNet50	None	Uniform	76.61	53.46	0.6978
55			ResNet50	None	Uniform	76.19	53.44	0.7014
56			ResNet50	None	Uniform	75.56	53.77	0.7116
57			ResNet50	None	Uniform	76.46	54.21	0.7090
58			ResNet101 + ResNet50	None	Uniform	76.55	55.19	0.7210
59			ResNet152 + ResNet50	None	Uniform	77.24	55.46	0.7179
60			ResNet101	None	Uniform	78.10	56.14	0.7189
61	Large Scale	TSN	ResNet50	IG-1B	Uniform	70.96	49.05	0.6912
62		TimeSFormer	TimeSformer	ImageNet-21K	Uniform	77.97	53.88	0.6910
63			TimeSformer	ImageNet-21K	Uniform	76.97	54.25	0.7047
64			TimeSformer	ImageNet-21K	Uniform	76.85	52.06	0.6774
65		Omni-TSN	ResNet50	ImageNet	Uniform	72.45	50.98	0.7037
66			ResNet50	IG-1B	Uniform	74.70	52.09	0.6973
67		Omni-SlowOnly	ResNet50	None	Uniform	76.49	55.00	0.7190
68			ResNet101	None	Uniform	80.00	58.05	0.7255
69		CSN	ResNet50	IG65M	Uniform	79.09	55.83	0.7059
70			ResNet152	Sports1M	Uniform	78.12	55.78	0.7140
71			ResNet152	Sports1M	Uniform	78.34	55.07	0.7030
72			ResNet152	IG65M	Uniform	82.20	59.22	0.7204
73			ResNet152	IG65M	Uniform	82.38	58.97	0.7159
74			ResNet152	IG65M	Uniform	80.39	59.35	0.7382

Table 7: Full experiment result for Human Only videos.

#	Dataset	Model Structure	Backbone	Pre-trained	Sampling	Original Accuracy	Human Accuracy	Ratio
1	Normal Scale	TSN	ResNet50	ImageNet	Uniform	68.93	13.69	0.1986
2			ResNet50	ImageNet	Uniform	69.43	12.29	0.1769
3			ResNet50	ImageNet	Dense	68.37	15.85	0.2318
4			ResNet50	ImageNet	Uniform	70.06	15.71	0.2242
5			ResNet50	ImageNet	Uniform	71.12	16.71	0.2349
6			ResNet50	ImageNet	Uniform	71.75	17.92	0.2498
7			ResNet50	ImageNet	Dense	68.98	16.89	0.2448
8		R(2+1)D	ResNet34	None	Uniform	67.35	18.85	0.2798
9			ResNet34	None	Uniform	69.13	15.87	0.2296
10			ResNet34	None	Uniform	74.22	20.95	0.2822
11		TIN	ResNet50	TSM-Kinetics400	Uniform	70.89	20.40	0.2877
12		TSM	MobileNetV2	ImageNet	Dense	68.58	19.23	0.2803
13			MobileNetV2	ImageNet	Dense	69.87	20.27	0.2902
14			ResNet50	ImageNet	Uniform	69.91	19.49	0.2788
15			ResNet50	ImageNet	Uniform	70.56	20.05	0.2841
16			ResNet50	ImageNet	Uniform	70.14	20.48	0.2920
17			ResNet50	ImageNet	Uniform	71.37	20.67	0.2896
18			ResNet50	ImageNet	Uniform	70.44	19.85	0.2819
19			ResNet50	ImageNet	Dense	72.61	22.53	0.3102
20			ResNet50	ImageNet	Dense	74.09	24.25	0.3273
21			ResNet50	ImageNet	Uniform	72.71	22.34	0.3072
22			ResNet50	ImageNet	Uniform	73.43	22.72	0.3094
23			ResNet50	ImageNet	Uniform	73.81	23.28	0.3154
24		I3D	ResNet50	ImageNet	Uniform	73.01	22.45	0.3075
25			ResNet50	ImageNet	Uniform	73.56	22.77	0.3096
26			ResNet50	ImageNet	Dense	73.08	23.70	0.3243
27			ResNet50	ImageNet	Dense	73.57	23.89	0.3247
28		NL-TSM	ResNet50	ImageNet	Uniform	71.57	19.75	0.2759
29			ResNet50	ImageNet	Uniform	70.61	21.21	0.3004
30			ResNet50	ImageNet	Uniform	71.25	19.73	0.2769
31		NL-I3D	ResNet50	ImageNet	Uniform	74.91	19.42	0.2592
32			ResNet50	ImageNet	Uniform	73.60	23.80	0.3234
33			ResNet50	ImageNet	Uniform	74.00	20.04	0.2708
34		NL-SlowOnly	ResNet50	ImageNet	Uniform	74.52	19.20	0.2577
35			ResNet50	ImageNet	Uniform	75.78	17.99	0.2374
36		CSN	ResNet50	None	Uniform	73.22	24.91	0.3403
37			ResNet152	None	Uniform	76.30	27.20	0.3564
38			ResNet152	None	Uniform	77.62	28.33	0.3650
39		TPN	ResNet50	None	Uniform	73.52	26.40	0.3591
40			ResNet50	ImageNet	Uniform	76.16	25.26	0.3316
41		X3D	X3D_S	None	Uniform	72.67	19.90	0.2738
42			X3D_M	None	Uniform	75.55	22.27	0.2948
43		TANet	TANet	ImageNet	Dense	76.10	23.64	0.3107
44		SlowOnly	ResNet50	None	Uniform	72.79	20.12	0.2764
45			ResNet50	None	Uniform	74.60	22.08	0.2959
46			ResNet50	None	Uniform	72.72	20.53	0.2824
47			ResNet50	None	Uniform	72.55	19.50	0.2688
48			ResNet50	ImageNet	Uniform	73.33	20.39	0.2781
49			ResNet50	ImageNet	Uniform	75.35	22.30	0.2959
50			ResNet101	None	Uniform	76.26	26.90	0.3528
51		SlowFast	ResNet50	None	Uniform	74.78	22.90	0.3063
52			ResNet50	None	Uniform	75.81	26.08	0.3440
53			ResNet50	None	Uniform	76.02	24.07	0.3166
54			ResNet50	None	Uniform	76.61	25.23	0.3294
55			ResNet50	None	Uniform	76.19	25.00	0.3281
56			ResNet50	None	Uniform	75.56	22.96	0.3038
57			ResNet50	None	Uniform	76.46	25.17	0.3293
58			ResNet101 + ResNet50	None	Uniform	76.55	24.23	0.3166
59			ResNet152 + ResNet50	None	Uniform	77.24	27.20	0.3521
60			ResNet101	None	Uniform	78.10	26.90	0.3445
61	Large Scale	TSN	ResNet50	IG-1B	Uniform	70.96	15.84	0.2232
62		TimeSFormer	TimeSformer	ImageNet-21K	Uniform	77.97	22.27	0.2856
63			TimeSformer	ImageNet-21K	Uniform	76.97	22.70	0.2950
64			TimeSformer	ImageNet-21K	Uniform	76.85	21.11	0.2746
65		Omni-TSN	ResNet50	ImageNet	Uniform	72.45	16.50	0.2278
66			ResNet50	IG-1B	Uniform	74.70	20.38	0.2728
67		Omni-SlowOnly	ResNet50	None	Uniform	76.49	25.18	0.3292
68			ResNet101	None	Uniform	80.00	31.81	0.3976
69		CSN	ResNet50	IG65M	Uniform	79.09	29.48	0.3727
70			ResNet152	Sports1M	Uniform	78.12	27.17	0.3478
71			ResNet152	Sports1M	Uniform	78.34	26.46	0.3378
72			ResNet152	IG65M	Uniform	82.20	34.29	0.4171
73			ResNet152	IG65M	Uniform	82.38	33.58	0.4076
74			ResNet152	IG65M	Uniform	80.39	34.62	0.4307

Table 8: Full experiment result for Action Swap Videos. We list three numbers for each of the random seed. Original Accuracy (OAcc) is different from Table 6, 7 since Action Swap uses subset of validation set. We omit SBErr for Same Swap since the background class is as same as foreground class.

				Random Swap		Far Swap		Close Swap		Same Swap
#	Dataset	Model Structure	OAcc	SHAcc	SBErr	SHAcc	SBErr	SHAcc	SBErr	SHAcc
1	Normal Scale	TSN	71.44	10.66, 10.89, 11.28	26.85, 26.78, 26.87	09.45, 08.70, 09.15	31.22, 31.24, 31.86	20.14, 19.80, 19.91	24.92, 24.97, 25.36	55.44, 55.00, 55.55
2		TSN	71.69	09.59, 09.45, 10.12	26.76, 27.60, 26.90	08.61, 07.83, 08.15	32.62, 32.29, 32.85	18.15, 18.29, 18.38	25.29, 25.22, 25.87	54.34, 53.58, 53.97
3		TSN	70.49	12.47, 12.15, 12.89	21.01, 21.10, 21.26	11.35, 10.73, 11.05	24.29, 25.06, 25.02	20.87, 20.46, 20.83	20.88, 20.76, 21.38	50.83, 50.13, 50.56
4		TSN	72.55	11.17, 11.53, 11.31	28.40, 28.66, 28.79	10.37, 10.10, 09.80	32.92, 32.77, 33.58	21.72, 21.26, 21.43	26.48, 26.69, 27.19	59.14, 58.89, 59.08
5		TSN	72.97	12.40, 12.64, 12.63	25.41, 25.61, 25.63	11.24, 10.85, 10.82	30.08, 30.44, 30.74	21.49, 20.88, 21.51	24.19, 24.33, 25.20	56.10, 55.83, 55.76
6		TSN	73.54	13.35, 13.21, 13.32	28.20, 28.27, 27.76	12.02, 11.72, 11.95	32.48, 32.57, 32.94	23.69, 23.23, 23.18	26.37, 26.73, 27.10	59.95, 60.18, 60.49
7		TSN	70.93	13.69, 13.78, 13.67	21.19, 20.88, 21.42	12.59, 12.43, 11.88	24.37, 24.99, 25.43	21.79, 21.77, 22.02	21.68, 20.69, 21.51	51.91, 51.80, 52.69
8		R(2+1)D	70.41	13.80, 13.46, 14.21	29.36, 30.12, 29.05	11.08, 11.12, 10.76	34.72, 35.54, 35.86	23.46, 23.76, 24.56	26.55, 25.89, 26.18	61.36, 61.18, 61.39
9		R(2+1)D	71.86	12.50, 12.68, 12.73	29.00, 29.39, 28.98	10.09, 10.03, 10.12	33.88, 33.56, 34.43	22.89, 22.02, 23.25	26.69, 26.02, 26.35	60.70, 60.77, 61.18
10		R(2+1)D	76.19	15.81, 15.38, 16.07	30.74, 30.62, 29.59	12.89, 13.07, 12.96	35.75, 35.61, 35.50	26.21, 26.85, 26.71	27.69, 26.85, 26.73	64.64, 64.77, 63.99
11		TIN	72.95	18.08, 18.27, 18.56	21.03, 20.74, 20.71	16.52, 16.55, 16.68	23.16, 23.39, 23.94	27.44, 27.08, 26.89	21.24, 20.80, 20.92	58.53, 58.73, 58.48
12		TSM	70.87	13.75, 14.31, 13.96	27.74, 27.61, 26.71	11.81, 11.79, 11.90	32.91, 32.69, 32.39	24.68, 24.54, 24.67	24.47, 24.95, 25.54	61.45, 61.73, 61.64
13		TSM	72.22	13.91, 13.78, 14.01	29.91, 29.91, 29.53	11.21, 11.33, 11.06	35.23, 35.70, 35.55	24.03, 24.40, 24.68	27.01, 25.96, 26.23	62.33, 61.69, 62.44
14		TSM	72.19	16.55, 17.26, 17.37	20.71, 20.46, 20.74	15.70, 15.17, 15.38	22.66, 23.73, 23.89	26.02, 26.11, 26.02	20.99, 20.26, 20.25	57.72, 56.70, 56.72
15		TSM	72.72	16.78, 16.57, 17.39	22.02, 21.75, 21.91	15.06, 14.85, 15.52	24.74, 24.47, 25.20	25.80, 26.07, 26.18	22.06, 22.27, 21.91	58.55, 58.37, 58.14
16		TSM	72.89	16.46, 16.53, 16.64	23.18, 22.54, 22.20	14.63, 14.90, 14.70	25.50, 26.27, 26.25	25.54, 25.47, 26.09	22.54, 22.52, 22.66	58.46, 58.18, 58.23
17		TSM	73.81	16.80, 17.39, 17.14	23.32, 22.73, 22.87	15.72, 15.66, 16.02	25.91, 25.96, 26.73	26.58, 26.28, 26.82	22.52, 22.38, 22.77	60.22, 60.42, 59.19
18		TSM	72.68	16.98, 17.31, 17.55	22.62, 22.02, 21.99	15.27, 15.31, 15.01	25.13, 25.41, 25.86	25.71, 26.02, 26.80	22.25, 21.88, 21.79	58.91, 58.96, 58.34
19		TSM	74.84	16.41, 16.43, 16.36	32.48, 32.00, 31.77	13.59, 14.01, 13.44	38.70, 38.79, 39.09	26.27, 26.64, 26.62	28.82, 27.72, 28.40	66.10, 65.55, 65.69
20		TSM	76.08	17.19, 16.73, 17.83	34.17, 33.92, 33.00	14.69, 14.08, 14.14	40.10, 40.33, 40.65	27.69, 27.70, 27.93	29.23, 28.84, 29.43	67.04, 66.08, 66.79
21		TSM	74.51	17.58, 17.49, 18.03	27.85, 27.03, 26.58	16.11, 15.56, 15.81	31.29, 32.21, 32.52	27.05, 26.64, 26.89	24.97, 24.83, 25.34	62.30, 62.12, 61.71
22		TSM	75.35	18.04, 17.46, 18.70	29.02, 28.75, 28.65	15.66, 15.54, 15.98	33.60, 33.44, 34.10	27.51, 27.58, 27.79	25.75, 25.68, 26.11	64.50, 64.29, 63.65
23		TSM	75.49	17.74, 17.97, 18.90	29.41, 29.27, 28.73	15.86, 15.86, 15.79	33.92, 34.79, 35.11	28.11, 28.09, 28.34	26.69, 25.73, 26.57	65.37, 64.43, 63.97
24		I3D	75.58	17.07, 16.84, 17.21	29.94, 29.99, 29.68	15.43, 14.35, 14.65	34.35, 35.52, 34.63	27.31, 27.83, 27.05	26.23, 26.37, 27.26	64.84, 64.36, 65.37
25		I3D	75.33	17.17, 16.76, 17.23	31.81, 32.20, 31.33	14.81, 14.08, 14.37	37.36, 37.61, 37.31	27.61, 27.67, 27.65	28.34, 28.18, 28.33	66.86, 66.47, 66.63
26		I3D	75.05	17.08, 16.52, 16.94	32.52, 32.23, 31.68	14.42, 14.22, 14.10	38.11, 37.68, 38.13	27.83, 27.97, 27.54	28.79, 28.17, 28.89	67.55, 67.22, 67.59
27		I3D	75.26	17.30, 17.03, 17.60	32.73, 32.23, 31.93	15.08, 14.38, 14.54	38.43, 39.02, 39.12	28.15, 28.45, 28.20	29.09, 27.95, 28.40	67.41, 67.63, 68.09
28		NL-TSM	74.06	16.30, 16.60, 16.71	22.04, 21.51, 21.77	15.11, 14.65, 15.31	24.84, 25.18, 24.97	25.54, 25.40, 26.66	21.74, 22.04, 21.45	58.52, 59.12, 58.09
29		NL-TSM	72.61	17.60, 17.42, 17.55	20.03, 20.17, 20.01	16.04, 15.97, 16.14	22.66, 22.78, 23.12	25.59, 25.80, 26.89	21.10, 21.31, 20.67	57.15, 57.38, 56.76
30		NL-TSM	73.52	17.00, 17.71, 17.12	21.03, 21.12, 20.60	15.56, 15.13, 16.14	23.32, 23.35, 24.28	26.14, 26.30, 26.53	21.20, 21.35, 21.43	58.23, 58.02, 57.40
31		NL-I3D	76.90	16.16, 16.00, 16.52	30.47, 30.12, 29.53	13.69, 13.55, 13.07	35.14, 35.96, 35.82	27.01, 27.15, 26.78	26.67, 26.43, 26.60	64.43, 65.30, 65.07
32		NL-I3D	75.96	16.92, 16.98, 17.76	31.59, 30.83, 30.40	14.88, 14.37, 14.62	36.58, 36.74, 36.42	27.86, 28.11, 27.70	27.51, 27.49, 27.40	66.38, 66.84, 66.38
33		NL-I3D	76.17	15.25, 15.61, 16.02	30.67, 29.68, 30.07	12.96, 12.70, 12.72	35.84, 35.73, 35.22	26.66, 26.73, 26.60	27.03, 26.35, 27.12	64.41, 64.80, 64.55
34		NL-SlowOnly	76.10	15.04, 15.27, 16.05	30.33, 30.53, 29.76	13.39, 12.82, 12.77	35.25, 36.16, 36.33	25.77, 25.47, 26.03	26.96, 27.03, 27.24	64.45, 64.68, 63.97
35		NL-SlowOnly	77.74	17.42, 17.39, 17.81	28.66, 29.02, 27.76	15.54, 14.42, 14.56	33.56, 34.13, 34.43	27.10, 27.15, 26.78	25.86, 25.20, 25.68	63.77, 63.68, 63.97
36		CSN	75.51	17.76, 18.13, 17.74	31.66, 31.74, 31.34	15.43, 14.99, 15.29	36.97, 36.62, 37.70	27.92, 28.47, 28.17	28.25, 27.19, 27.28	65.76, 65.73, 66.33
37		CSN	78.08	19.39, 19.46, 19.66	30.56, 31.08, 30.63	16.39, 16.43, 16.59	36.49, 36.81, 36.21	29.76, 29.60, 29.07	27.72, 27.24, 26.90	66.79, 66.79, 67.43
38		CSN	79.26	20.05, 19.98, 21.08	30.26, 30.42, 29.64	17.39, 17.44, 18.04	35.23, 35.29, 35.20	30.62, 31.08, 30.78	26.53, 25.87, 26.55	67.29, 68.26, 67.86
39		TPN	76.04	17.55, 17.95, 18.58	32.43, 31.89, 31.08	15.11, 14.79, 14.79	37.65, 38.43, 38.38	28.86, 28.73, 28.98	27.70, 27.69, 27.92	67.77, 67.45, 67.39
40		TPN	78.63	18.84, 18.49, 19.04	33.48, 33.62, 32.53	16.11, 15.66, 15.75	38.38, 39.41, 38.93	28.77, 29.37, 28.96	29.53, 28.13, 29.48	69.54, 69.21, 69.22
41		X3D	74.76	13.66, 13.73, 14.30	26.69, 27.49, 25.82	10.90, 11.03, 10.97	32.04, 31.63, 32.27	24.37, 24.26, 23.97	25.00, 24.77, 24.37	60.42, 61.13, 60.91
42		X3D	77.34	15.52, 15.47, 15.86	27.38, 27.24, 27.33	12.84, 12.84, 12.86	31.98, 33.17, 33.17	27.08, 26.16, 26.12	25.64, 25.29, 25.45	63.88, 64.71, 64.38
43		TANet	78.35	18.27, 17.92, 18.59	30.85, 30.74, 29.87	15.40, 15.41, 15.61	36.46, 36.17, 37.13	28.31, 28.84, 28.43	27.33, 26.64, 27.06	66.83, 67.04, 67.47
44		SlowOnly	74.80	15.11, 15.88, 15.72	33.72, 33.12, 32.27	13.00, 12.63, 12.63	38.63, 39.48, 39.67	25.98, 26.09, 27.01	28.79, 29.30, 28.63	66.58, 66.28, 66.44
45		SlowOnly	76.73	16.28, 15.98, 16.82	33.85, 33.49, 32.55	13.71, 13.80, 13.59	38.93, 39.50, 39.58	26.98, 26.89, 26.78	28.77, 29.44, 29.28	66.90, 66.26, 66.67
46		SlowOnly	75.28	14.44, 14.51, 15.31	33.88, 33.33, 32.61	12.29, 11.77, 12.00	39.53, 39.80, 40.15	25.61, 24.88, 25.45	29.02, 28.22, 28.56	65.02, 64.50, 65.07
47		SlowOnly	75.18	15.82, 16.18, 16.36	31.88, 32.37, 30.23	13.19, 13.09, 13.48	37.24, 37.68, 38.13	27.14, 26.57, 26.71	27.70, 27.58, 27.83	65.49, 65.53, 65.37
48		SlowOnly	75.87	15.04, 15.45, 15.65	33.37, 32.98, 31.84	13.19, 12.75, 12.40	38.29, 39.09, 39.30	25.87, 25.71, 26.02	28.98, 28.77, 29.18	66.61, 66.06, 66.33
49		SlowOnly	77.85	17.19, 17.67, 17.63	33.37, 33.03, 32.13	15.27, 14.47, 14.69	38.32, 38.79, 39.32	27.65, 28.04, 28.70	29.05, 28.70, 28.45	68.32, 68.02, 68.19
50		SlowOnly	78.27	19.78, 19.37, 20.10	31.79, 31.33, 30.30	17.23, 16.94, 16.78	36.96, 36.49, 37.40	31.06, 31.06, 30.90	28.08, 27.63, 28.45	69.28, 69.92, 69.01
51		SlowFast	76.99	16.50, 16.27, 17.10	34.10, 33.81, 32.73	13.51, 13.59, 13.27	40.29, 39.78, 40.17	27.21, 27.60, 27.40	29.30, 29.28, 29.27	67.18, 67.50, 67.11
52		SlowFast	78.21	16.84, 16.87, 18.36	34.68, 33.95, 33.40	14.62, 14.37, 14.21	40.12, 40.60, 40.63	28.77, 28.66, 28.52	29.12, 29.46, 29.11	68.30, 67.77, 68.03
53		SlowFast	78.25	17.30, 17.39, 18.08	34.83, 34.68, 33.64	14.72, 14.60, 14.70	40.72, 40.85, 40.72	28.45, 28.79, 28.43	30.72, 30.39, 30.12	69.03, 68.90, 68.82
54		SlowFast	78.60	17.85, 17.24, 18.56	33.81, 33.85, 32.82	14.95, 14.97, 14.97	39.23, 39.53, 39.73	29.12, 28.68, 28.43	29.62, 29.11, 29.37	67.06, 67.95, 67.96
55		SlowFast	78.71	17.62, 17.40, 18.50	33.74, 33.37, 32.37	15.31, 15.10, 15.08	39.23, 39.35, 39.39	28.43, 28.27, 28.45	29.21, 28.98, 29.04	67.47, 67.77, 67.91
56		SlowFast	77.68	16.59, 16.21, 16.94	35.55, 35.66, 33.90	13.41, 13.11, 13.18	40.93, 41.50, 41.98	26.80, 27.01, 27.10	30.39, 29.91, 29.39	67.00, 67.47, 67.71
57		SlowFast	78.84	17.92, 17.79, 18.38	34.40, 33.76, 32.82	15.10, 14.70, 15.22	39.73, 40.24, 40.10	28.77, 28.52, 28.98	30.21, 29.71, 29.09	68.09, 68.34, 68.87
58		SlowFast	79.05	17.49, 17.14, 17.99	32.27, 32.16, 31.33	15.24, 14.95, 15.04	36.92, 38.02, 38.16	28.31, 28.04, 28.73	29.30, 28.73, 29.09	68.25, 67.68, 67.87
59		SlowFast	79.30	20.07, 19.75, 21.06	32.41, 32.46, 30.94	17.69, 17.69, 17.26	36.78, 36.80, 37.20	31.04, 30.97, 30.85	28.77, 28.15, 28.54	70.02, 68.89, 69.05
60		SlowFast	79.80	18.06, 18.13, 18.52	34.29, 33.64, 32.84	15.36, 15.54, 15.31	39.53, 39.83, 40.74	29.05, 29.62, 29.07	29.80, 29.20, 29.12	69.63, 69.97, 69.14
61	Large Scale	TSN	73.39	11.60, 11.86, 12.43	27.28, 27.85, 27.22	10.44, 09.93, 09.89	32.02, 31.98, 32.45	21.12, 21.24, 21.83	25.70, 25.82, 25.59	58.23, 57.04, 57.95
62		TimeSFormer	79.09	15.24, 15.70, 15.88	28.68, 28.88, 28.73	13.41, 12.66, 13.00	32.68, 33.80, 33.21	25.70, 25.77, 25.86	27.56, 27.08, 27.67	64.98, 65.71, 65.09
63		TimeSFormer	78.33	14.53, 14.99, 14.78	31.65, 31.81, 31.17	12.64, 12.11, 12.40	36.05, 36.85, 36.92	25.96, 25.47, 25.96	28.47, 28.77, 28.96	66.54, 66.61, 66.60
64		TimeSFormer	77.89	14.70, 14.97, 14.92	28.70, 28.47, 28.20	13.27, 12.54, 12.80	32.59, 33.12, 32.91	24.84, 24.17, 25.18	26.92, 27.14, 27.69	63.10, 62.81, 62.99
65		Omni-TSN	74.19	13.23, 13.09, 13.66	28.95, 28.88, 29.25	11.45, 11.19, 11.67	35.09, 34.98, 35.25	23.30, 22.54, 22.78	26.71, 26.92, 26.69	60.47, 59.76, 60.06
66		Omni-TSN	76.33	13.27, 13.89, 14.21	30.49, 30.78, 30.76	11.95, 11.54, 11.86	36.69, 36.19, 37.58	24.35, 24.49, 24.45	28.15, 27.63, 27.85	64.07, 63.10, 64.57
67		Omni-SlowOnly	78.44	18.42, 17.19, 18.34	34.33, 35.00, 33.78	15.11, 14.72, 15.04	40.54, 40.81, 40.99	29.12, 28.86, 29.28	29.57, 29.80, 29.94	69.79, 69.06, 69.63
68		Omni-SlowOnly	81.63	22.46, 22.36, 22.84	33.94, 33.65, 33.03	19.37, 19.11, 19.61	38.80, 39.25, 39.62	32.77, 33.87, 33.48	30.53, 29.36, 30.33	73.75, 72.88, 73.17
69		CSN	81.46	22.04, 21.45, 22.66	32.75, 32.84, 32.45	18.82, 18.97, 18.63	38.59, 38.57, 39.00	32.80, 32.57, 31.82	29.14, 28.31, 29.23	70.01, 70.59, 70.70
70		CSN	78.98	20.39, 20.37, 20.81	31.75, 31.17, 30.51	17.08, 16.92, 17.30	37.88, 37.56, 37.93	31.58, 30.60, 31.38	27.83, 26.99, 26.32	68.07, 68.51, 67.96
71		CSN	79.38	20.19, 20.05, 20.87	31.84, 32.50, 31.84	16.55, 16.78, 16.73	38.59, 38.79, 39.02	30.28, 30.74, 30.28	27.58, 26.98, 27.79	68.16, 67.86, 68.03
72		CSN	83.17	25.20, 24.84, 25.71	32.37, 32.30, 31.52	21.65, 21.88, 21.75	38.04, 39.02, 37.97	35.57, 35.32, 34.79	28.33, 27.77, 28.45	72.46, 72.88, 72.40
73		CSN	83.92	25.06, 24.77, 25.75	32.16, 32.73, 31.59	21.70, 21.95, 22.54	37.67, 37.97, 38.38	35.54, 36.01, 35.27	28.88, 27.56, 28.72	73.01, 72.74, 72.94
74		CSN	82.58	25.87, 25.71, 26.60	32.13, 32.18, 31.59	22.73, 22.66, 23.21	36.74, 37.93, 37.83	35.61, 36.60, 35.61	28.59, 27.63, 28.11	71.98, 72.72, 72.19

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[2] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[3] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, 2020.

[4] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 2012.

[5] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2021.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, 2020.

[8] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[9] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv*, 2018.

[13] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. Human action recognition without human. In *ECCV Workshops*, 2016.

[16] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masking transformer backbones for effective semantic segmentation. *arXiv*, 2021.

[17] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, 2019.

[18] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *CVPR*, 2022.

[19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

[20] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *ICCV*, 2021.

[21] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021.

[22] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 2007.

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[24] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020.

[25] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *AAAI*, 2020.

[26] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012.

[27] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[28] Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *ICCV*, 2019.

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[31] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *ICLR*, 2021.

[32] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv*, 2019.

[33] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, 2020.

[34] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020.

[35] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.

[36] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.