Mitigating and Evaluating Static Bias of Action Representations in the Background and the Foreground

Haoxin Li[1], Yuan Liu[2], Hanwang Zhang[1], Boyang Li[1]
[1]Nanyang Technological University [2]Guangzhou University
{haoxin003, hanwangzhang, boyang.li}@ntu.edu.sg, yuanliu@gzhu.edu.cn

Abstract

In video action recognition, shortcut static features can interfere with the learning of motion features, resulting in poor out-of-distribution (OOD) generalization. The video background is clearly a source of static bias, but the video foreground, such as the clothing of the actor, can also provide static bias. In this paper, we empirically verify the existence of foreground static bias by creating test videos with conflicting signals from the static and moving portions of the video. To tackle this issue, we propose a simple yet effective technique, StillMix, to learn robust action representations. Specifically, StillMix identifies bias-inducing video frames using a 2D reference network and mixes them with videos for training, serving as effective bias suppression even when we cannot explicitly extract the source of bias within each video frame or enumerate types of bias. Finally, to precisely evaluate static bias, we synthesize two new benchmarks, SCUBA for static cues in the background, and SCUFO for static cues in the foreground. With extensive experiments, we demonstrate that StillMix mitigates both types of static bias and improves video representations for downstream applications. Code is available at https://github.com/lihaoxin05/StillMix.

1. Introduction

Traditional computer vision techniques perform well on independent and identically distributed (IID) test data, but often lack out-of-distribution (OOD) generalization [9, 32, 12]. This is intimately tied to the learning of shortcut features [27, 16, 17], which are easy to learn and correlate strongly with IID labels but cause poor OOD generalization [53, 62, 49, 22]. In video action recognition, shortcut features often manifest as static cues. For example, a network may classify a video as *golf swinging* based on its background, a golf course, even if the motion patterns indicate another action such as *walking*. While static cues can provide valuable information [74, 11, 77], they often outcompete motion features [23, 40, 41, 52, 69] and result in low OOD performance [41, 63, 26]. In contrast to the rich literature on mitigating background static bias (*e.g.*, golf courses for *golf swinging*) [5, 63, 73, 10, 6], foreground static bias has been underexplored. Examples of foreground bias include swimsuits for *swimming* and guitars for *guitar playing* — people can swim without swimsuits or show guitars in the video without playing them.

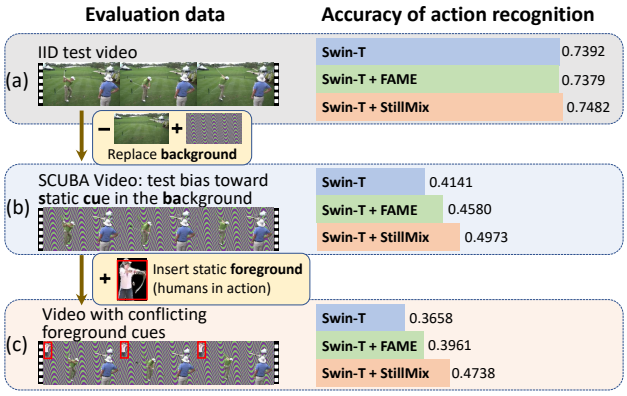


Figure 1: Evaluation of background and foreground static bias. (a) Testing on IID HMDB51 [36] test videos. (b) Testing on SCUBA videos, constructed by replacing the video background with a synthetic sinusoidal stripe image. (c) Testing on videos with conflicting foreground cues, constructed by inserting a random static foreground into the SCUBA video.

The first question we ask is if foreground static bias exists and if it is captured by the representations learned by neural networks. Our investigation technique is to create test videos with conflicting action cues from the moving part and the static part of the video. In the first step, shown in Figure 1(b), we replace the backgrounds of IID HMDB51 [36] test videos by sinusoidal stripe images. These videos have no meaningful backgrounds, so the action information must come from the foreground. Therefore, models overly reliant on background static cues

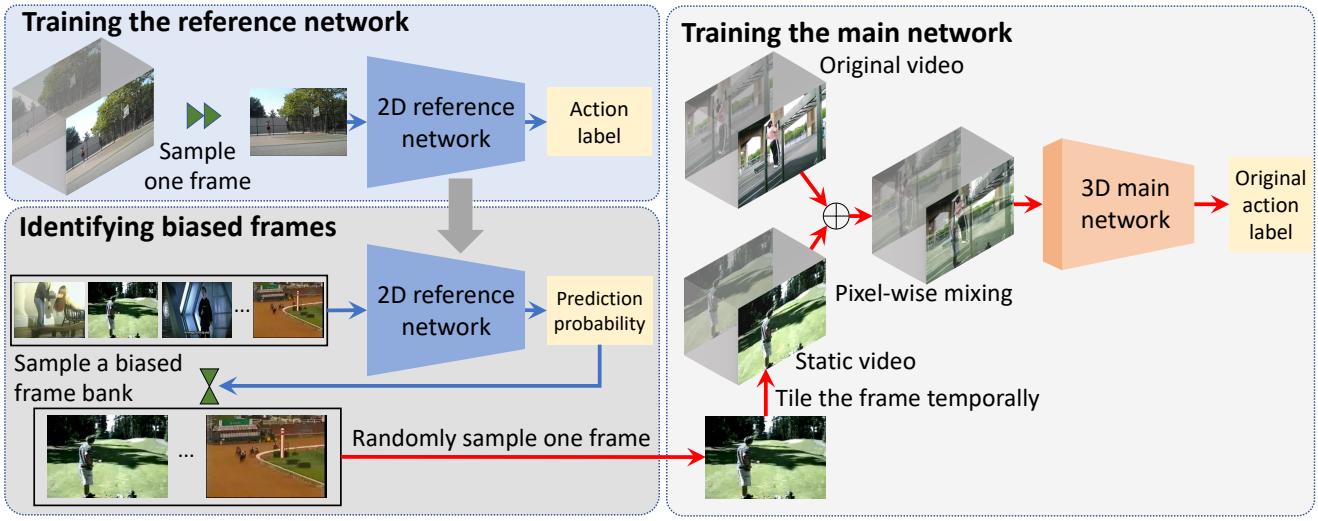


Figure 2: An illustration of StillMix. We train a 2D reference network that classifies still frames into actions to capture static bias. With the reference network, we sample frames inducing static bias to construct a biased frame bank. We mix the frames from the bank with a given video to generate an augmented video, which is used to train a 3D main network to mitigate static bias.

should perform poorly. A background debiasing technique, FAME [10], coupled with a tiny Video Swin Transformer (Swin-T) [46], works relatively well on this test.

In the second step, shown in Figure 1(c), from a single frame of a random video, we extract its foreground (mainly human actors), and insert the static foreground into all the frames of the current SCUBA video. The resultant video contains only two action features: a static foreground that indicates one action label and a moving foreground that indicates another action label. Predictions made using the static foreground would be wrong. This design allows the quantification of foreground static bias. More details can be found in Sec. S1 of the Supplementary Material.

The results clearly show the existence of foreground static bias and its negative effects. On the second test set, both Swin-T and Swin-T+FAME suffer similar degradation and perform 5% worse than SCUBA videos. FAME works by procedurally isolating the foreground regions from each frame and use those for training. However, it is hard to separate the foreground motion from the static foreground (*e.g.*, clothing, equipment, or other people attributes [40]) in the training videos, since both types of features are strongly tied to the human actors.

We propose StillMix, a technique that mitigates static bias in both the background and the foreground, without the need to explicitly isolate (or even enumerate [5]) the bias-inducing content within a frame. StillMix identifies bias-inducing frames using a reference network and mixes them with training videos without affecting motion features. The process is illustrated in Figure 2. Unlike FAME, StillMix could suppress static bias anywhere in a frame, including the background and the foreground. In Figure 1, StillMix outperforms FAME and suffers only 2% accuracy drop on the second benchmark, highlighting its resilience.

Evaluating OOD action recognition is challenging as test videos with OOD foregrounds, such as *swimming* without swimsuits or *cycling* while carrying a guitar, are rare. To pinpoint the static bias in either the background or the foreground, we create new synthetic sets of OOD benchmarks by altering the static features in IID test videos, as illustrated in Figure 3. Specifically, we retain the foregrounds of actions and replace the backgrounds with diverse natural and synthetic images. This procedure yields a test set that quantifies representation bias toward static cues in the background (SCUBA). Second, we create videos that repeat a single random frame from SCUBA, producing a test set that quantifies representation bias toward static cues in the foreground (SCUFO). As these videos disassociate the backgrounds from the action and contain no motion, their actions can be recognized by only static foreground features. Thus, high accuracy on SCUFO indicates strong foreground static bias.

With the synthetic OOD benchmarks, we extensively evaluate several mainstream action recognition methods and make the following observations. First, all examined methods exhibit static bias. Second, existing debiasing methods like ActorCutMix [78] and FAME [10] demonstrate resistance to background static bias, but remain vulnerable to foreground static bias. In contrast, the proposed StillMix consistently boosts performance of action recognition mod-

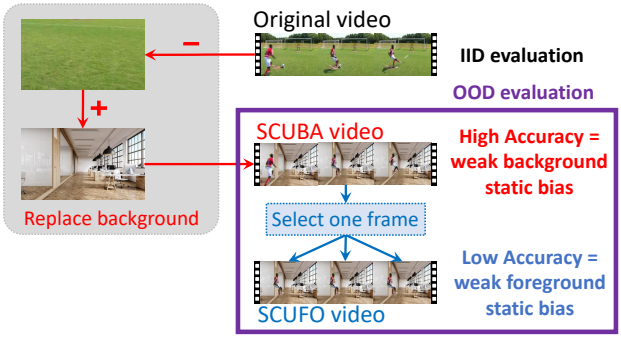


Figure 3: An illustration of OOD benchmark construction. To quantify static cues in the background, we reserve the foreground actions and replace the backgrounds with other images to synthesize SCUBA videos. To quantify static cues in the foreground, we randomly select one frame in the SCUBA video and stack it into a single-frame video without motion, named SCUFO videos.

els and compares favorably with the other debiasing techniques on both background and foreground static bias. In addition, StillMix improves the performance of transfer learning and downstream weakly supervised action localization.

The paper makes the following contributions:

- Through quantitative experiments, we highlight the importance to address foreground static bias in learning robust action representations.
- We propose StillMix, a video data augmentation technique to mitigate static bias in not only the background but also the foreground.
- We create new benchmarks to quantitatively evaluate static bias of action representations and pinpoint the source of static bias (backgrounds or foregrounds).
- We compare action recognition methods on the created benchmarks to reveal their characteristics and validate the effectiveness of StillMix.

2. Related Work

Bias Evaluation. Biases are surface features that are easily learned by neural networks and strongly influence their predictions. Such features perform well on IID data [60, 29] but do not generalize to OOD data [53, 49]. In action recognition, models easily capture static bias [40, 41, 5, 63]. The following methods are used for bias evaluation: *(1) Visualization techniques* [15, 47] visualize the regions that models focus on to interpret the static bias qualitatively. *(2) Proxy data or tasks.* Synthetic videos with altered backgrounds [6], videos with white-noise textures [26], dynamic texture videos [21, 3] are used to reveal the bias toward backgrounds or texture. Proxy tasks evaluating temporal asymmetry, continuity, and causality are designed to show the static bias in video representations [18]. *(3) Mutual information.* [33] quantifies the static bias using mutual information between representations of different types of videos. Although these works evaluate the static bias in the whole video, they do not specify the source of static bias. In this paper, we create new benchmarks to pinpoint the source of static bias as the background and the foreground.

Bias Mitigation. Prevalent techniques of mitigating bias in action representations can be broadly classified into four categories. *(1) Attribute supervision.* [5] uses scene pseudo-labels and human masks to discourage models from predicting scenes and recognizing actions without human, but it needs extra attribute labels. *(2) Re-weighting.* [40, 41] identify videos containing bias and downweight them in training, but [65] suggests merely weight adjustment is insufficient. *(3) Context separation.* [66] learns to separate action and contexts by collecting samples with similar contexts but different actions. *(4) Data augmentation.* Similar to the proposed StillMix, a few works utilize augmented videos. BE [63] mixes a frame from a video with other frames in the same video. ActorCutMix [78], FAME [10], ObjectMix [31] and FreqAug [30] carefully carve out the foreground (human actors or regions of motion), and replace the background with other images to create augmented training data. SSVC [73] and MCL [38] focus the models to the dynamic regions. However, these methods have not addressed static cues in the foreground.

A particular advantage of StillMix is that it does not require specially designed procedures to carve out the bias-inducing pixels within the frames like ActorCutMix [78] and FAME [10], or even to enumerate the source of bias like [5]. Rather, it automatically identifies bias-inducing frames using a reference network. Consequently, StillMix can suppress static bias in both the background and the foreground.

StillMix is also similar to two debiasing techniques designed for image recognition and text classification [48, 44], which use a reference network to identify bias-inducing data instances. However, StillMix exploits the special property of videos that they can be decomposed into individual frames. StillMix identifies bias-inducing components (frames) using 2D networks rather than whole data points as in [48, 44].

Action Recognition. 3D convolution or decomposed 3D convolutions [28, 58, 4, 61, 59, 42] are popular choices for action recognition. Two-stream architectures employ two modalities to classify actions, such as both RGB frames and optical flow [54, 64], or videos with two different frame rates and resolutions [14]. Multi-scale temporal convolutions or feature fusion are designed for fine-grained actions with strong temporal structures [75, 24, 39, 67]. Transformer networks are proposed to capture the long-range dependencies [1, 2, 46]. However, our understanding of the

representations learned by these models remains limited. In this paper, we create benchmarks to evaluate what features are captured by action models and propose a simple data augmentation method that effectively improves the robustness of action models.

3. The StillMix Technique

In order to learn robust and generalizable action representations that are invariant to static cues, we propose a simple but effective video data augmentation technique, StillMix. Instead of using manually designed rules to identify and remove biased data from the training set, as in ActorCutMix [78] and FAME [10], StillMix learns to identify still frames that induce biased representation using a neural network and mitigate static bias through mixing the identified frames with videos. As a result, StillMix offers a flexible bias suppression technique that works for both the background and the foreground.

We begin with some notations. We denote the i^{th} video in the training set as tensor $\boldsymbol{x}_i \in \mathbb{R}^{C\times T\times H\times W}$, where C, T, H and W are the number of channels, number of frames, height and width of the video, respectively. The associated ground-truth action label is y_i. The video $\boldsymbol{x}_i$ contains a sequence of frames $\langle \boldsymbol{z}_{i,j}\rangle_{j=1}^{T}$, $\boldsymbol{z}_{i,j} \in \mathbb{R}^{C\times H\times W}$. The training set contains N training video samples and is written as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. The goal of StillMix is to augment a given training sample $(\boldsymbol{x}_i, y_i)$ into a transformed sample $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$. The procedures of StillMix are illustrated in Figure 2 and introduced as follows.

Step 1: Training the Reference Network. We identify bias-inducing frames using a 2D reference network that predicts the action label from a still frame of a video. As the still frame contains no motion, we expect the network to rely on static features to make the predictions.

Specifically, at every epoch we randomly sample a frame $\boldsymbol{z}_{i,j} \in \mathbb{R}^{C\times H\times W}$ from each video $\boldsymbol{x}_i$, and train the reference network $\mathcal{R}(\cdot)$ to predict the label y_i. The loss is

$$L_{ref} = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathcal{R}(\boldsymbol{z}_{i,j}), y_i), \tag{1}$$

where $\ell(\cdot)$ can be any classification loss, such as the cross-entropy. After training, the reference network $\mathcal{R}(\cdot)$ encodes the correlations between static cues within the frames and the action classes.

Step 2: Identifying Biased Frames. The output of reference network $\mathcal{R}(\boldsymbol{z}_{i,j})$ is a categorical distribution over action classes. We take the probability of the predicted class $p_{i,j} = \max_k P(y = k|\boldsymbol{z}_{i,j})$. A high $p_{i,j}$ indicates strong correlation between $\boldsymbol{z}_{i,j}$ and the action class, which means $\boldsymbol{z}_{i,j}$ can induce static bias. Therefore, we select frames with high $p_{i,j}$ to construct the biased frame bank S:

$$S = \{\boldsymbol{z}_{i,j} | p_{i,j} \geq p_\tau\}, \tag{2}$$

where p_τ is the τ-th percentile value of $p_{i,j}$. In practice, we perform another round of uniformly random selection to control the size of the biased frame bank.

Step 3: Mixing Video and Biased Frames. To break the strong correlation between the biased frame and the action class, we mix a video of any action class with different biased frames identified above. Specifically, in each epoch, given a video sample $(\boldsymbol{x}_i, y_i)$, we sample a frame $\boldsymbol{z}^{\text{biased}}$ from the biased frame bank S and tile it T times along the temporal dimension, yielding a static video with T identical frames. We denote this operation as $\text{Tile}(\boldsymbol{z}^{\text{biased}}, T)$. The augmented video sample $\tilde{\boldsymbol{x}}_i$ is generated by the pixel-wise interpolation of $\boldsymbol{x}_i$ and the static video. The augmented video label $\tilde{y}_i$ is the same as the original action label y_i.

$$\tilde{\boldsymbol{x}}_i = \lambda \boldsymbol{x}_i + (1-\lambda)\text{Tile}(\boldsymbol{z}^{\text{biased}}, T), \;\; \tilde{y}_i = y_i, \tag{3}$$

where the scalar λ is sampled from a Beta distribution $Beta(\alpha, \beta)$.

The rationale for keeping the video label unchanged after augmentation is that the static video contains no motion and does not affects the motion patterns in the mixed video, thus it should not contribute to the action label. This setting of StillMix can be intuitively understood as randomly permuting the labels of the static video, so that the network is forced to ignore the correlations between the static cues in the biased frames and actions.

Training with Augmented Videos. We apply StillMix to each video with a predefined probability P_{aug}.

$$(\boldsymbol{x}_i^*, y_i^*) = \begin{cases} (\boldsymbol{x}_i, y_i) & a_i = 0 \\ (\tilde{\boldsymbol{x}}_i, \tilde{y}_i) & a_i = 1 \end{cases}, a_i \sim Ber(P_{\text{aug}}), \tag{4}$$

where a is a scalar sampled from a Bernoulli distribution $Ber(P_{\text{aug}})$. The samples $\{(\boldsymbol{x}_i^*, y_i^*)\}_{i=1}^{N}$ are used to train the main network $\mathcal{F}(\cdot)$ using the following loss function:

$$L = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathcal{F}(\boldsymbol{x}_i^*), y_i^*), \tag{5}$$

where $\ell(\cdot)$ could be any classification loss.

Discussion. StillMix aims to learn robust action representations that generalize to OOD data. One popular formulation of OOD generalization [68, 34, 51, 35, 50] considers shortcut features as features that work under a specific environment but not others. For example, a classifier that excels in well-lit environments may perform terribly in dim environments. To learn robust classifiers, we ought to discover invariant features that work equally well in all environments.

More formally, the optimal predictor $\mathcal{F}^*$ can be found with the bi-level optimization

$$\mathcal{F}^* = \underset{\mathcal{F}}{\operatorname{argmin}} \max_{e} \mathop{\mathbb{E}}_{\boldsymbol{x}^e, y^e} [\ell(\mathcal{F}(\boldsymbol{x}^e), y^e)], \tag{6}$$

where the feature-label pair, $(\boldsymbol{x}^e, y^e)$, are drawn from the data distribution $P(\boldsymbol{x}, y|e)$ of environment e and ℓ is the per-sample loss. $\boldsymbol{x}^e$ contains both class features and environment features; a good predictor $\mathcal{F}$ is sensitive to the former and ignores the latter. The optimization encourages this because if $\mathcal{F}$ utilizes features that work for environment e_1 but not e_2, the loss will increase as the $\max_e$ operation will select e_2.

However, directly optimizing Eq. (6), such as in [51], requires sampling data from all environments, which is impractical due to skewed environment distributions. For example, videos of people *playing soccer* in tuxedos on beaches are exceedingly rare. Maximizing over all environments is also challenging.

The mixing operation in StillMix may be understood in the same framework. A static frame $\boldsymbol{z}^{\text{biased}}$ can be considered as coming from an environment e' which biases predictions toward certain action labels. Mixing $\boldsymbol{z}^{\text{biased}}$ with $\boldsymbol{x}_i$ simulates sampling $\boldsymbol{x}^{e'}$ from the environment e'. StillMix may be considered to optimize the following loss,

$$\mathcal{F}^* = \underset{\mathcal{F}}{\operatorname{argmin}} \mathop{\mathbb{E}}_{e} \Big[\mathop{\mathbb{E}}_{\boldsymbol{x}^e, y^e} [\ell(\mathcal{F}(\boldsymbol{x}^e), y^e)] \Big], \tag{7}$$

which replaces the maximization over environments in Eq. (6) with an expectation over environments due to the random sampling of $\boldsymbol{z}^{\text{biased}}$.

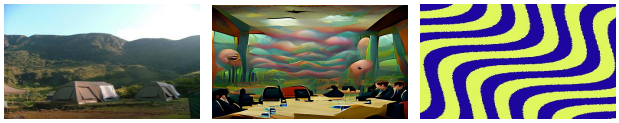
4. SCUBA and SCUFO: OOD Benchmarks

To quantitatively evaluate static bias in the background and the foreground, we create OOD benchmarks based on three commonly used video datasets, *i.e.*, HMDB51 [36], UCF101 [55] and Kinetics-400 [4], as detailed below.

4.1. Foreground Masks and Background Images

Foreground Masks. To extract the foreground area of actions, we use available human-annotated masks of people for UCF101 and HMDB51. There are totally 910 videos in the UCF101 test set and 256 videos in the HMDB51 test set having foreground annotations. Since there is no human-annotated masks for Kinetics-400, we use video segmentation models [57, 56] to generate foreground masks. After filtering out the videos with small foreground masks (likely to be wrong), we obtain totally 10,190 videos in the Kinetics-400 validation set to construct the benchmark.

Background Images. In order to synthesize diverse test videos, we collect background images from three different image sources: 1) the test set of Place365 [76]. 2) images generated by VQGAN-CLIP [8] from a random scene category of Place365 and a random artistic style. 3) randomly generated images with S-shaped stripes defined by sinusoidal functions. For each image source, we construct a background image pool. In Figure 4, we show three example background images from the three sources. More details are described in Sec. S2 of the Supplementary Material.



(a) Place365 (b) VQGAN-CLIP (c) Sinusoid

Figure 4: Background images from different sources. (a) An image from Place365. (b) An image generated by VQGAN-CLIP from the query "A painting of a conference room in the style of surreal art". (c) An image of randomly generated sinusoidal stripes.

4.2. Test Video Synthesis

Testing for Background Static Cues. Given a video $\boldsymbol{x}$ with T frames $\{\boldsymbol{x}_t\}_{t=1}^T$, we create a synthetic video $\hat{\boldsymbol{x}}$ by combining the foreground of $\boldsymbol{x}$ and a background image sampled from a background image pool.

$$\hat{\boldsymbol{x}}_t = \boldsymbol{m}_t \odot \boldsymbol{x}_t + (1 - \boldsymbol{m}_t) \odot \text{Tile}(\boldsymbol{z}^{\text{bg}}, T), \tag{8}$$

where $\boldsymbol{m}_t$ is the foreground mask, $\odot$ denotes pixel-wise multiplication, $\boldsymbol{z}^{\text{bg}}$ is a background image sampled from the image pool. $\text{Tile}(\boldsymbol{z}^{\text{bg}}, T)$ repeats $\boldsymbol{z}^{\text{bg}}$ T times along the temporal dimension. For each video with foreground masks, we pair it with m randomly selected background images from each of the 3 background image pools to synthesize $3m$ videos. We set $m = 10, 5, 1$ for HMDB51, UCF101 and Kinetics-400, respectively, since HMDB51 and UCF101 have fewer videos with foreground masks and we would like to increase the diversity of the synthetic videos.

The generated videos retain the original action foreground, including the human actors and their motion, on new random backgrounds. They are designed to test bias toward static cues from the background, and are named SCUBA videos. We expect models invariant to static backgrounds to obtain high classification accuracy on SCUBA.

Testing for Foreground Static Cues. In addition, we create another set of videos to test the amount of foreground static bias in the learned representations. Foreground static cues include people and object attributes, such as bicycle helmets for *cycling* and bows for *archery* — people can ride a bicycle without helmets or hold bows when not performing archery. As the SCUBA videos contain most foreground elements in the original videos, they cannot distinguish whether models rely on foreground static cues.

To this end, we create videos that contain only a single frame. Specifically, from each SCUBA video, we ran-

Table 1: Statistics of the created benchmarks.

Video Source	# Original Videos	Background Source	# Synthetic Videos	# Domain Gap of SCUBA	# Domain Gap of SCUFO
HMDB51	256	Place365	2,560	5.646±0.276	5.745±0.290
		VQGAN-CLIP	2,560	8.178±0.685	8.307±0.533
		Sinusoid	2,560	11.739±0.444	11.998±0.932
UCF101	910	Place365	4,550	20.493±2.199	20.829±2.093
		VQGAN-CLIP	4,550	51.320±5.790	55.202±9.477
		Sinusoid	4,550	52.249±3.522	52.534±5.930
Kinetics-400	10,190	Place365	10,190	6.094±0.208	6.455±0.224
		VQGAN-CLIP	10,190	7.504±0.296	8.052±0.273
		Sinusoid	10,190	7.211±0.311	7.766±0.148

domly select one frame and repeat it temporally to create a video with zero motion. As these videos quantify the representation bias toward static cues in the foreground, we name them SCUFO videos. In SCUFO videos, the foreground static features are identical to the corresponding SCUBA videos, but the motion information is totally removed. Therefore, a model invariant to foreground static features should obtain low classification accuracy on them.

We summarize the dataset statistics in Table 1. SCUBA and SCUFO have the same number of videos for each pair of video source and background source. We also report the domain gap between original videos to show their OOD characteristics, as explained in the next section.

4.3. Quality Assessment

We empirically verify that the SCUBA datasets retain the motion features of the original videos but replace background static features using the following two tests.

Human Assessment. To test if SCUBA preserves the motion information sufficiently for action recognition, we carry out an experiment on Amazon Mechanical Turk (AMT) to verify if human workers can recognize the actions in SCUBA videos.

From the same original video, we randomly sampled one synthetic video and asked the AMT workers if the moving parts in the video show the labeled action. The workers are given three options: yes, no, and can't tell. We also create control questions with original videos to detect random clicking and design control groups to prevent the workers from always answering yes to synthetic videos. The final answer for each video is obtained by majority voting of three workers. Workers who do not reach at least 75% accuracy on the control questions are rejected. More details are described in Sec. S2 of the Supplementary Material.

Collectively, the AMT workers were able to correctly recognize the actions in 96.15% of UCF101-SCUBA, 86.33% of HMDB51-SCUBA and 85.19% of Kinetics400-SCUBA videos. We conclude that SCUBA videos preserve sufficient action information for humans to recognize.

Domain Gaps of the Static Features. To verify if SCUBA and SCUFO have successfully replaced the background static features and qualify as OOD test sets, we test if a classifier based on purely static features trained on IID videos can generalize to SCUBA and SCUFO.

Using a variation of scene representation bias [5], we define the domain gap G_{scene} as

$$G_{scene} = Acc(D_{\text{ori}}, \Phi_{scene})/Acc(D_{\text{syn}}, \Phi_{scene}). \quad (9)$$

Here Φ_{scene} is the average frame feature extracted from a ResNet-50 pretrained on Place365 [76]. Thus, the extracted feature captures static scene information, mostly from the background. We train a linear classifier on the original video training set and apply it to the original test set D_{ori}, obtaining the accuracy $Acc(D_{\text{ori}}, \Phi_{scene})$. After that, we apply the same classifier to the synthetic dataset D_{syn}, obtaining the accuracy $Acc(D_{\text{syn}}, \Phi_{scene})$. A higher ratio indicates greater domain gap with respect to static features.

In Table 1, we show the means and standard deviations computed from three random repeats of video synthesis. We observe large domain gaps, ranging from 5.6-fold to 52-fold decrease in accuracy on the synthetic test set. This demonstrates the static features of synthetic videos differ substantially from the original videos and the synthetic videos can serve as OOD tests. Moreover, the low standard deviations show that the effects of random sampling are marginal. In later experiments, we use the dataset from one random seed.

5. Experiments

In this section, we compare the performance of several mainstream action recognition methods on IID and OOD test data and validate the effectiveness of StillMix.

5.1. Comparing Methods

Action Recognition Models. (1) TSM [43], a temporal shift module learning spatiotemporal features with 2D CNN. (2) SlowFast [14], a two-branch 3D CNN learning spatiotemporal signals under two frame rates. (3) Video Swin Transformer [46], an adapted Swin Transformer [45] for videos. We use the tiny version, denoted as Swin-T.

Video Data Augmentation and Debiasing Methods. We compare the debiasing performance of several video data augmentation and debiasing methods by adapting them to supervised action recognition. (1) Mixup [72] and VideoMix [70]. (2) SDN [5]. (3) BE [63], ActorCutMix[78] and FAME [10]. We adapt these three self-supervised debiasing methods as data augmentations, which carve out the foreground and replace the background as in the original papers. All the data augmentation techniques are applied stochastically as in [19]. More implementation details are described in Sec. S3 of the Supplementary Material.

Table 2: IID and OOD test accuracy (%) of augmentation and debiasing methods on Kinetics-400. † indicates adaptation from self-supervised debiasing methods. Confl-FG denotes synthetic videos with conflicting foreground cues. All models are pretrained on ImageNet.

Model	Augmentation or Debiasing	IID	OOD				
			Avg SCUBA ↑	Avg SCUFO ↓	Contra. Acc. ↑	Confl-FG ↑	ARAS ↑
TSM	No	71.13	37.39	17.22	22.80	20.15	57.86
	Mixup	71.33	40.81	17.53	25.98	23.48	58.05
	VideoMix	**71.35**	38.87	17.25	24.57	23.43	56.61
	SDN	69.99	36.95	16.55	22.38	20.29	55.06
	BE†	71.30	37.89	16.08	24.35	20.11	57.47
	ActorCutMix†	71.07	40.42	16.29	26.52	21.41	57.09
	FAME†	71.13	**40.91**	18.34	25.63	24.41	57.47
	StillMix (Ours)	71.28	40.48	**5.23**	**36.07**	**25.73**	**59.69**
Swin-T	No	73.95	41.74	18.17	25.93	25.25	60.17
	Mixup	73.91	43.95	17.92	28.24	27.64	59.59
	VideoMix	73.80	43.17	19.26	26.40	29.37	60.95
	SDN	72.23	42.34	21.46	24.46	27.14	60.26
	BE†	73.93	43.40	19.56	26.28	26.67	59.79
	ActorCutMix†	**73.97**	45.70	19.39	28.64	29.02	61.23
	FAME†	73.81	**48.79**	21.27	30.03	29.50	60.37
	StillMix (Ours)	73.86	44.10	**5.51**	**39.41**	**30.77**	**62.49**

5.2. Evaluation Metrics

We conduct the following experiments on Kinetics-400, UCF101 and HMDB51. First, we perform IID tests on the original test sets and use the top-1 accuracy as metrics. After that, we perform OOD tests on SCUBA and SCUFO and report the average top-1 accuracy across background image sources. Note that higher accuracy on SCUBA is better (low background static bias), while lower accuracy on SCUFO is better (low foreground static bias).

To show the performance of utilizing pure foreground motion information, we propose another performance metric called contrasted accuracy (Contra. Acc.). As one SCUFO video is derived from a SCUBA video, we count one correct prediction if the model is correct on the SCUBA but incorrect on the associated SCUFO video.

We further evaluate on the synthetic videos with conflicting foreground cues (Figure 1). Finally, we also evaluate on ARAS [13], a real-world OOD dataset with rare scenes, to show the performance of scene bias reduction.

5.3. Results on IID and OOD Benchmarks

Table 2, 3 and 4 compare the IID and OOD performance of different video data augmentation and debiasing methods on Kinetics-400, HMDB51 and UCF101. Given limited computational resources, we ran experiments on Kinetics-400 using a single seed. However, on the smaller HMDB51 and UCF101, we repeated experiments with three seeds. In Sec. S1 of the Supplementary Material, we provide more detailed results (*e.g.*, tests on videos with conflicting foreground cues and ARAS [13]).

Table 3: IID and OOD test accuracy (%) of augmentation and debiasing methods on HMDB51. All models are pretrained on Kinetics-400.

Model	Augmentation or Debiasing	IID	OOD			
			Avg SCUBA ↑	Avg SCUFO ↓	Contra. Acc. ↑	Confl-FG ↑
TSM	No	70.39±0.51	38.03±1.39	19.23±1.30	22.02±0.64	25.44±1.31
	Mixup	**72.00±0.47**	39.76±1.72	19.08±1.37	23.76±0.84	26.94±1.23
	VideoMix	70.72±0.12	35.71±1.57	17.48±1.11	21.03±0.55	22.19±1.47
	SDN	69.51±0.30	37.05±0.73	17.60±0.37	23.74±0.95	28.38±0.87
	BE	71.22±0.24	38.48±1.42	19.45±1.06	22.39±0.67	25.21±1.35
	ActorCutMix	70.52±0.82	38.40±0.53	19.61±0.56	21.94±0.40	26.16±0.36
	FAME	70.39±0.88	47.19±1.52	22.33±0.91	28.21±0.89	33.98±2.09
	StillMix	71.52±0.38	**48.23±0.96**	**8.43±0.88**	**42.05±0.99**	**36.89±1.09**
Swin-T	No	73.92±0.74	43.93±0.78	20.46±0.71	27.84±1.28	36.58±1.65
	Mixup	74.58±0.43	43.10±1.13	21.17±0.66	26.09±1.05	36.62±2.98
	VideoMix	73.31±0.53	39.39±0.71	20.44±0.73	23.13±0.54	32.68±1.04
	SDN	74.66±0.82	40.02±1.48	20.22±1.24	22.88±1.05	34.87±2.43
	BE	74.31±0.41	43.56±1.38	19.96±0.71	27.84±1.32	35.99±0.67
	ActorCutMix	74.05±0.60	46.79±1.38	22.07±0.36	28.12±1.27	36.97±1.63
	FAME	73.79±0.29	51.40±1.54	26.92±0.71	29.66±2.11	39.61±1.87
	StillMix	**74.82±0.43**	**51.81±1.78**	**13.39±0.71**	**40.28±1.61**	**47.38±1.50**

Table 4: IID and OOD test accuracy (%) of augmentation and debiasing methods on UCF101. All models are pretrained on Kinetics-400.

Model	Augmentation or Debiasing	IID	OOD			
			Avg SCUBA ↑	Avg SCUFO ↓	Contra. Acc. ↑	Confl-FG ↑
TSM	No	94.62±0.08	25.60±1.36	4.21±0.84	21.83±1.48	27.68±1.35
	Mixup	**94.71±0.14**	27.80±0.95	4.04±0.81	24.17±1.00	30.31±1.10
	VideoMix	94.50±0.19	31.55±1.68	5.77±0.74	26.69±1.38	30.69±1.79
	SDN	93.84±0.27	19.91±0.61	3.10±0.19	17.19±0.51	20.89±0.36
	BE	94.49±0.14	25.91±1.37	4.62±0.84	21.82±1.38	28.06±1.32
	ActorCutMix	94.47±0.15	**38.11±1.48**	4.56±0.16	33.90±1.51	38.12±2.12
	FAME	93.72±0.09	35.72±1.15	3.67±0.52	32.28±1.28	34.58±0.93
	StillMix	94.30±0.14	37.18±1.29	**0.79±0.12**	**36.47±1.24**	**40.59±0.80**
Swin-T	No	**96.21±0.19**	42.31±2.24	5.78±0.68	36.82±2.12	44.65±2.10
	Mixup	96.17±0.14	46.16±1.74	5.93±0.43	40.46±1.96	47.16±2.82
	VideoMix	96.00±0.02	41.40±1.11	13.27±0.85	29.37±0.91	42.59±1.51
	SDN	95.76±0.11	39.25±2.32	2.98±0.88	36.42±1.74	48.47±2.06
	BE	96.06±0.11	43.98±0.80	5.54±0.94	38.62±1.13	46.62±0.96
	ActorCutMix	95.87±0.19	**58.61±0.48**	11.92±0.25	46.87±0.45	56.88±0.39
	FAME	95.81±0.15	40.90±1.57	6.36±0.71	35.14±1.66	28.21±1.83
	StillMix	96.02±0.08	58.22±0.41	**3.44±0.51**	**54.90±0.77**	**57.30±0.60**

OOD data cause performance degradation. Comparing the performance of TSM and Swin-T on IID and OOD tests, we observe that they perform much worse (more than 20%) on SCUBA than IID videos. Given that human workers can recognize the action in more than 85% of SCUBA videos, the results indicate that the models are not robust to the domain shifts, probably due to the reliance of static background features; when the backgrounds are replaced, performance deterioration ensues.

IID tests do not fully reveal representation quality. Comparing the performance of different augmentation and debi-

asing methods, we observe that all methods obtain similar accuracies on IID tests, which fall within a 2% band. However, they show vastly different performance on SCUBA and SCUFO — the maximum difference is larger than 15%. Therefore, we argue that IID tests alone may not be good indicators of the robustness of action representations.

In particular, VideoMix, SDN and BE provide little debiasing effects. Mixup leads to inconsistent performance gains. ActorCutMix and FAME consistently improve performance on SCUBA. Nevertheless, they decrease performance (increase accuracy) on SCUFO, which suggests that they improve performance on SCUBA partially by increasing reliance on foreground static features. The action features learned with ActorCutMix and FAME are likely still vulnerable to foreground static bias.

StillMix alleviates foreground and background static bias. StillMix boosts the performance on both SCUBA and SCUFO videos and obtains the best contrasted accuracy (Contra. Acc.). The significant improvements on SCUFO indicate that StillMix can suppress static bias from the entire video and not only the background. In addition, StillMix outperforms other methods on videos with conflicting foreground cues as well as ARAS. Overall, these results demonstrate the ability of StillMix to reduce static bias that is difficult to exhaustively name or pixel-wise cut out.

5.4. StillMix Improves Representation Learning

We further investigate the effects of StillMix on improving representation learning through the following tests.

Transferring action features across datasets. We evaluate the representations learned with different augmentation and debiasing methods by their capability to transfer to different datasets. We adopt the linear probing protocol, which trains a linear classifier on the target dataset on top of the backbone network trained on the source dataset. Table 5 shows the results of TSM, where StillMix obtains the best performance, especially in transferring across small datasets.

Downstream weakly supervised action localization. We evaluate the representations learned with StillMix by their ability to improve downstream weakly supervised action localization. We pretrain TSM on Kinetics-400 with StillMix. After that, we extract RGB features for each video segments on THUMOS14 [25] and use the extracted features to train weakly supervised action localization models BaSNet [37] and CoLA [71]. StillMix improves the performance by more than 1.0% of average mAP for BaSNet and more than 0.5% of average mAP for CoLA.

5.5. Ablation Study

We conduct ablation study on UCF101 and HMDB51 to examine design choices of StillMix.

Table 5: Action recognition accuracy (%) of transferring features across Kinetics-400, UCF101, and HMDB51.

Augmentation or Debiasing	Source→Target			
	K400→UCF	K400→HMDB	HMDB→UCF	UCF→HMDB
No	92.52	66.67	61.64	44.95
Mixup	93.07	68.69	63.58	46.60
VideoMix	93.55	69.22	61.49	40.33
SDN	92.81	63.79	61.12	41.90
BE	93.10	67.45	62.71	45.88
ActorCutMix	92.73	67.39	61.67	42.92
FAME	93.87	67.84	58.87	44.99
StillMix	**93.89**	**70.07**	**65.69**	**47.99**

Table 6: Weakly supervised action localization performance of features learned by StillMix.

Method	Feature	Debiasing	Avg mAP@IoU=[0.1:0.9]
BaSNet	TSM (RGB)	No	0.1810
	TSM (RGB)	StillMix	**0.1935**
CoLA	TSM (RGB)	No	0.2380
	TSM (RGB)	StillMix	**0.2436**

Debiasing works the best when the reference network and the main network share the same architecture. We compare the results of StillMix with different network structures in Table 7. When the structures of the reference network and the main network are identical, the OOD performance is the best and the IID performance is very close to the best, indicating good bias mitigation. We hypothesize that networks with same architecture tend to learn the same bias. As a result, using a reference network with the same architecture as the main network could be the most effective at identifying bias-inducing frames.

Sampling biased frames improves debiasing. We compare three frame sampling strategies when constructing the biased frame bank: (1) *No RefNet*: the frame bank is uniformly sampled from the whole dataset; (2) *RefNet*: as in StillMix, we sample frames with high prediction probabilities from the reference network according to Eq. (2); (3) *RefNet Inversed*: contrary to StillMix, we sample frames with low prediction probabilities from the reference network, $S = \{\boldsymbol{z}_{i,j} | p_{i,j} < p_\tau\}$. Table 8 shows results of ImageNet pretrained TSM and Swin-T. The reference network (RefNet) approach achieves the best OOD performance, whereas RefNet Inversed performs the worst.

We observe the difference between RefNet and No RefNet is small on UCF101 but is large on HMDB51. We attribute this to the prevalence of bias-inducing frames in UCF101. MMAction2 [7] trained TSN [64] using only three frames per video on UCF101 and achieved 83.03% classification accuracy but achieved only 48.95% with 8

Table 7: Action recognition accuracy (%) of StillMix with different reference network structures. All networks are pretrained on ImageNet.

Main Network	Reference Network	UCF101		HMDB51	
		IID	Contra. Acc.	IID	Contra. Acc.
TSM	ResNet50-2D	87.29	**24.60**	54.66	**33.14**
	SlowFast-2D	87.44	22.20	55.03	30.51
	Swin-T-2D	86.72	23.08	55.05	31.62
SlowFast	ResNet50-2D	84.85	18.86	50.74	20.89
	SlowFast-2D	84.96	**19.76**	51.53	**21.21**
	Swin-T-2D	85.16	19.18	51.85	20.28
Swin-T	ResNet50-2D	88.59	31.09	56.10	18.44
	SlowFast-2D	88.60	29.34	54.43	19.25
	Swin-T-2D	88.92	**32.14**	55.36	**21.40**

Table 8: Action recognition accuracy (%) of StillMix with different frame sampling strategies.

Main Network	Sampling Strategy	UCF101		HMDB51	
		IID	Contra. Acc.	IID	Contra. Acc.
TSM	No RefNet	87.39	24.49	54.07	31.21
	RefNet	87.29	**24.60**	54.66	**33.14**
	RefNet Inversed	87.38	23.53	54.79	29.17
SlowFast	No RefNet	85.03	18.98	51.79	20.94
	RefNet	84.96	**19.76**	51.53	**21.21**
	RefNet Inversed	84.33	18.77	50.94	18.61
Swin-T	No RefNet	88.37	31.24	55.62	18.89
	RefNet	88.92	**32.14**	55.36	**21.40**
	RefNet Inversed	88.59	30.51	56.34	18.18

frames on HMDB51[1]. This shows many frames in UCF101 contain static cues correlated with the class labels. Random sampling can yield many bias-inducing frames on UCF101 but cannot do so on HMDB51, where the strength of RefNet becomes apparent.

In Sec. S1 of the Supplementary Material, we provide more ablation studies showing that mixing action labels in StillMix decreases performance and sufficient mixing strength (*i.e.*, small values of λ in Eq. (3)) is necessary for debiasing.

5.6. Performance on Something-Something-V2

To validate the effectiveness of different debiasing methods on recognizing fine-grained actions with strong temporal structures, we perform tests on Something-Something-V2 [20]. In Table 9, we show the performance of different debiasing methods with TSM as the base model. Since SDN and ActorCutMix require bounding boxes of human, which are time-consuming to extract, we did not include the results of these two methods. The results show that StillMix outperforms other data augmentation methods, illustrating its effectiveness on fine-grained action videos.

[1] https://github.com/open-mmlab/mmaction2/blob/02a06bb3180e951b00ccceb48dab055f95acd1a7/configs/recognition/tsn/README.md

Table 9: Action recognition accuracy (%) of different debiasing methods on Something-Something-V2.

Debiasing	Accuracy
No	57.49
Mixup	57.86
VideoMix	58.23
BE	57.68
FAME	58.10
StillMix (Ours)	**58.68**

6. Conclusion and Discussion

To learn robust and generalizable action representations, we explore techniques that mitigate static bias in both the background and the foreground. We propose a simple yet effective video data augmentation method, StillMix, and create two new sets of OOD benchmarks, SCUBA and SCUFO, to quantify static bias in the background and the foreground. Through extensive evaluation, we conclude that StillMix mitigates static bias in the background and the foreground and improves the performance of transferring learning and downstream tasks. In contrast, existing debiasing methods remain vulnerable to foreground static bias despite their robustness to background static bias.

Despite the strengths of StillMix on mitigating static bias in the background and the foreground, it has the following limitations: (1) additional computational overhead in training the reference network (about 8% of the training time of the main network); and (2) little improvement (and little degradation) on IID tests.

For future work, we believe that evaluating static bias in large pretrained models with the created benchmarks and adapting StillMix to mitigate static bias in such models would be promising directions.

Acknowledgments

This work has been supported by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *IEEE International Conference on Computer Vision*, pages 6836–6846, 2021. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 3

[3] Sofia Broomé, Ernest Pokropek, Boyu Li, and Hedvig Kjellström. Recur, attend or convolve? on whether temporal modeling matters for cross-domain robustness in action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4199–4209, 2023. 3

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733. IEEE, 2017. 3, 5

[5] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 3, 6

[6] Jihoon Chung, Yu Wu, and Olga Russakovsky. Enabling detailed action recognition evaluation through video dataset augmentation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 3

[7] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 8

[8] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022. 5

[9] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint 2011.03395*, 2020. 1

[10] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9726, June 2022. 1, 2, 3, 4, 6

[11] Shuangrui Ding, Rui Qian, and Hongkai Xiong. Dual contrastive learning for spatio-temporal representation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5649–5658, 2022. 1

[12] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16458–16468, 2021. 1

[13] Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, and Dahua Lin. Mitigating representation bias in action recognition: Algorithms and benchmarks. *arXiv preprint arXiv:2209.09393*, 2022. 7

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE international conference on computer vision*, pages 6202–6211, 2019. 3, 6

[15] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes, and Andrew Zisserman. Deep insights into convolutional networks for video recognition. *International Journal of Computer Vision*, 128(2):420–437, 2020. 3

[16] 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*, 2(11):665–673, nov 2020. 1

[17] 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 preprint 1811.12231*, 2018. 1

[18] Amir Ghodrati, Efstratios Gavves, and Cees GM Snoek. Video time: Properties, encoders and evaluation. *arXiv preprint arXiv:1807.06980*, 2018. 3

[19] Raphael Gontijo-Lopes, Sylvia Smullin, Ekin Dogus Cubuk, and Ethan Dyer. Tradeoffs in data augmentation: An empirical study. In *International Conference on Learning Representations*, 2020. 6

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision*, pages 5843–5851. IEEE, 2017. 9

[21] Isma Hadji and Richard P Wildes. A new large scale dynamic texture dataset with application to convnet understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 320–335, 2018. 3

[22] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020. 1

[23] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 7366–7375, 2018. 1

[24] Noureldien Hussein, Efstratios Gavves, and Arnold W.M. Smeulders. Timeception for complex action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 3

[25] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 8

[26] Filip Ilic, Thomas Pock, and Richard P Wildes. Is appearance free action recognition possible? In *European Conference on Computer Vision*, pages 156–173. Springer, 2022. 1, 3

[27] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[28] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 3

[29] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019. 3

[30] Jinhyung Kim, Taeoh Kim, Minho Shim, Dongyoon Han, Dongyoon Wee, and Junmo Kim. Spatiotemporal augmentation on selective frequencies for video representation learning. *arXiv preprint arXiv:2204.03865*, 2022. 3

[31] Jun Kimata, Tomoya Nitta, and Toru Tamaki. Objectmix: Data augmentation by copy-pasting objects in videos for action recognition. *arXiv preprint arXiv:2204.00239*, 2022. 3

[32] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. 1

[33] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13999–14009, 2022. 3

[34] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020. 4

[35] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 4

[36] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1, 5

[37] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11320–11327, 2020. 8

[38] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *IEEE International Conference on Computer Vision*, pages 2105–2114, 2021. 3

[39] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang. Tea: Temporal excitation and aggregation for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 906–915, 2020. 3

[40] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *European Conference on Computer Vision*, pages 513–528, 2018. 1, 2, 3

[41] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019. 1, 3

[42] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 3

[43] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *IEEE International Conference on Computer Vision*, pages 7082–7092. IEEE, 2021. 6

[44] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 3

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[46] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2, 3, 6

[47] Joonatan Manttari, Sofia Broomé, John Folkesson, and Hedvig Kjellstrom. Interpreting video features: A comparison of 3d convolutional networks and convolutional lstm networks. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[48] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 3

[49] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021. 1, 3

[50] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. 4

[51] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 4, 5

[52] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 535–544, 2021. 1

[53] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020. 1, 3

[54] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. 3

[55] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[56] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *arXiv preprint arXiv:2203.04708*, 2022. 5

[57] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3126–3137, 2022. 5

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 3

[59] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[60] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018. 3

[61] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, 2018. 3

[62] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[63] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11804–11813, 2021. 1, 3, 6

[64] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 3, 8

[65] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019. 3

[66] Yang Wang and Minh Hoai. Pulling actions out of context: Explicit separation for effective combination. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2018. 3

[67] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2020. 3

[68] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021. 4

[69] Sukmin Yun, Jaehyung Kim, Dongyoon Han, Hwanjun Song, Jung-Woo Ha, and Jinwoo Shin. Time is matter: Temporal self-supervision for video transformers. *arXiv preprint arXiv:2207.09067*, 2022. 1

[70] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 6

[71] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 8

[72] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[73] Manlin Zhang, Jinpeng Wang, and Andy J Ma. Suppressing static visual cues via normalizing flows for self-supervised video representation learning. *arXiv preprint arXiv:2112.03803*, 2021. 1, 3

[74] Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video rep-

resentation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3235–3245, 2022. 1

[75] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, pages 831–846, 2018. 3

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 5, 6

[77] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022. 1

[78] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *arXiv preprint arXiv:2103.16565*, 2021. 2, 3, 4, 6