Motion-aware Contrastive Video Representation Learning via Foreground-background Merging

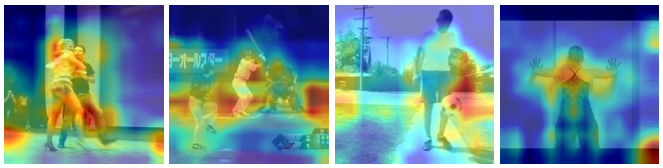
Shuangrui Ding[1*] Maomao Li[2] Tianyu Yang[2] Rui Qian[3]
Haohang Xu[1] Qingyi Chen[4] Jue Wang[2] Hongkai Xiong[1†]
[1]Shanghai Jiao Tong University [2]Tencent AI Lab
[3]The Chinese University of Hong Kong [4]University of Michigan
{dsr1212, xuhaohang, xionghongkai}@sjtu.edu.cn tianyu-yang@outlook.com
{limaomao07, arphid}@gmail.com qr021@ie.cuhk.edu.hk chenqy@umich.edu

Abstract

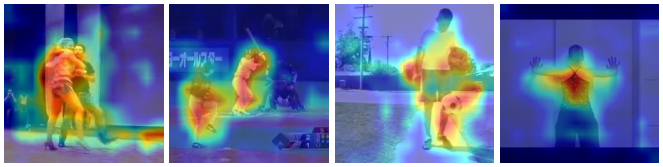
*In light of the success of contrastive learning in the image domain, current self-supervised video representation learning methods usually employ contrastive loss to facilitate video representation learning. When naively pulling two augmented views of a video closer, the model however tends to learn the common static background as a shortcut but fails to capture the motion information, a phenomenon dubbed as background bias. Such bias makes the model suffer from weak generalization ability, leading to worse performance on downstream tasks such as action recognition. To alleviate such bias, we propose **F**oreground-b**a**ckground **Me**rging (FAME) to deliberately compose the moving foreground region of the selected video onto the static background of others. Specifically, without any off-the-shelf detector, we extract the moving foreground out of background regions via the frame difference and color statistics, and shuffle the background regions among the videos. By leveraging the semantic consistency between the original clips and the fused ones, the model focuses more on the motion patterns and is debiased from the background shortcut. Extensive experiments demonstrate that FAME can effectively resist background cheating and thus achieve the state-of-the-art performance on downstream tasks across UCF101, HMDB51, and Diving48 datasets. The code and configurations are released at https://github.com/Mark12Ding/FAME.*

1. Introduction

The recent development of deep learning has promoted a series of applications in videos, such as video recognition [14, 51, 57], video retrieval [16, 67], and video object segmentation [11, 28, 70]. While various large-scale benchmarks [1, 5, 18] are the key to those successes, the costly manual annotation involved in fully-supervised methods excludes the potential utilization of millions of uncurated videos on the Internet. To further advance the video-related research, learning video representation in an unsupervised manner is of great significance and emerges as a general trend in the computer vision community.



(a) Results of vanilla contrastive learning.



(b) Results of our approach FAME.

Figure 1. Class-agnostic activation map [3] visualization of important areas. The heatmap indicates how much the pretrained model attends to the region. Compared to the conventional approach, our method mitigates the background bias significantly.

Recently, unsupervised learning in images [9, 40, 50, 59] has achieved competitive performances compared to their supervised counterparts, especially with the contrastive self-supervised learning formulation [7, 24]. The common idea of contrastive learning is to pull 'positive' pairs together in the embedding space and push apart the anchor from 'negative' samples. Due to inaccessibility to the label, a positive pair is usually formed by data augmentations of the anchor sample while the negative samples come from other samples. Inspired by these successes, various attempts

*Work done during an internship at Tencent AI Lab.
†Corresponding author. Email: xionghongkai@sjtu.edu.cn.

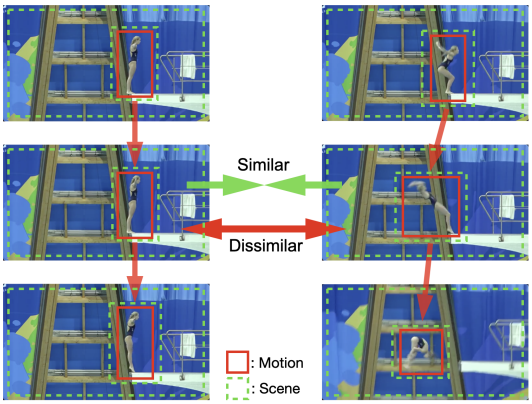


Figure 2. Illustration on a diving sequence. The green dashed box represents the scene and the red box means motion area. The two clips have the same background but distinct motions. Drawing such positive pairs closer inclines the model towards static bias.

have also been made in self-supervised video representation learning [15, 44]. However, we find applying vanilla contrastive learning on the video domain directly will lead the model to attending to the static area. As illustrated in Fig. 1, vanilla contrastive learning does not concentrate on the moving actors or objects but lays much emphasis on background areas. There might be two possible reasons: 1) the background usually covers much more area than the moving objects in the whole video, so that the model is more likely to focus on the background. 2) when sampling two different clips of the video, the static contexts are almost the same but there tend to be subtle differences in motion patterns. We show an example in Fig. 2. There are two clips sampled from one diving video. The green region is the background, which occupies over 3/4 area. And the red box, a small area, contains the moving diver. In addition, the backgrounds of the two clips are almost identical while two motions appear somewhat different, i.e., one is standing on the springboard, the other is taking off. That is to say, when we follow conventional visual augmentation techniques to form the positive pair and employ the multi-view constraint as self-supervision, it is intuitive for the model to pull static features close but attend less to motions. Hence, to let the contrastive learning pipeline be more motion-aware, we need to construct positive pairs in the way where the motions are more similar than backgrounds.

Here comes to a question on how to fabricate the motion-aware positive pair that makes contrastive learning prioritize the motion patterns. This paper explores this feasibility and presents a new augmentation technique named **F**oreground-b**a**ckground **Me**rging (FAME). Our motivation is to keep motion areas (foreground) unchanged as much as possible and replace static areas (background) with irrelevant content. Specifically, we first circle out the edge region of the moving object as the seed region via frame difference. Then, we use color statistics to extrapolate the entire moving object from the seed region. This efficient foreground discovery method extracts dynamic areas on which we expect the model to put emphasis. Then, we fuse the extracted foreground regions of each video with random backgrounds from other videos to form new action samples. In this way, when we force the model to learn the consistent representation between original clips and distracting clips, the model has to learn representations that are sensitive to motion patterns and overcome the background cheating. We evaluate the proposed FAME on three action recognition benchmarks. The superior experimental performance verifies that FAME enables self-supervised contrastive video representation learning to generalize better and distill the motion-aware representations. In short, we summarize our contributions as follows:

- We demonstrate the background bias caused by vanilla contrastive learning and propose a simple yet effective augmentation method FAME to help the model break background shortcuts and learn motion-aware representations.
- Our method enhances the conventional contrastive learning without whistle and bell and achieves the start-of-the-art performance among UCF101, HMDB51, and Diving48 datasets.

2. Related Work

Contrastive Visual Representation Learning. Recently, contrastive learning has greatly facilitated self-supervised visual representation learning [7, 24, 40, 50, 59]. It performs instance discrimination in a fully self-supervised manner to pull the representations of the same instance close and push those of different instances far away. Following this idea, [59] proposes to formulate the instance discrimination as a non-parametric classification problem. [40] mathematically proves that we could estimate mutual information with InfoNCE loss [20], which can be easily used for optimization. Later, MoCo [24] proposes to make use of key representations calculated in previous iterations as negative samples to facilitate contrastive learning. SimCLR [7] employs a large batch size instead of the memory bank to expand the negative pool for more robust visual representation. Considering that SimCLR requires tremendous computational resources, we adopt the MoCo framework as a strong baseline for self-supervised pretraining in our work.

Self-supervised Video Representation Learning. In video representation learning, there has been a line of works that employ diverse pretext tasks for self-supervised representation learning [35, 39, 62]. The most prevalent approaches include temporal order prediction [39, 62], video colorization [52], spatio-temporal puzzling [31] and speed prediction [4]. These methods generally employ manually

designed tasks to seek the spatio-temporal cues in video data, but the performance is limited. Then, for further improvement, some works apply contrastive learning formulation into video representation learning [15, 44]. Han et al. use InfoNCE loss to guide dense predictive coding in videos [21,22]. Based on the contrastive formulation, [6,26] jointly learn appearance and speed of videos, and [69] simultaneously encodes inter- as well as intra-variance in videos. [2, 23, 45] propose to leverage the consistency between different modalities to enhance video representation. Our method focuses only on the single modality, i.e., raw RGB video, to explicitly construct positive samples with the same motions but different backgrounds for self-supervised contrastive video representation learning.

Video Background Bias Mitigation. How to mitigate the background bias has been a long-standing topic [10, 25, 37, 58] for action recognition. In the supervised scenario, [10] uses an off-the-shelf human detector to mask out the human regions and train the model in an adversarial manner. [37] proposes a procedure to reassemble existing datasets that alleviates static representation bias. Later, to make the self-supervised video representations more robust to the background bias, a line of works employ other natural supervision [27, 36, 60] to guide the model to capture motion information explicitly. However, these methods require more than one backbone to pretrain multi-modality data, resulting in an undesired computational cost. To better utilize the implicit motion information in videos, DSM [53] aims to decouple the motion and context by deliberately constructing the positive/negative samples through spatial and temporal disturbance. BE [54] proposes to add a static frame as background noise for static bias mitigation. However, these two methods would erode the moving objects and impair the motion patterns. In contrast, our method solves this shortcoming by meticulously extracting dynamic foreground regions and preserving high-quality motion patterns.

Copy-paste Augmentation. Copy-paste augmentation [12, 13, 17] is a simple way to combine information from various instances and has been proved to be a good match for object-aware learning. In addition, Mixup [68] and CutMix [65] share a similar idea to increase the robustness against input corruptions. Inspired by these successes in supervised learning, MixCo [32] applied Mixup into visual contrastive learning and construct the semi-positive image from the mix-up of positive and negative images. Besides, InsLoc [63] proposes to copy image instances and paste them onto background images at diverse locations and scales, which advances self-supervised pretraining for object detection. FAME also copy and paste foreground content onto another video, a bit like CutMix. But one key difference in our work compared to CutMix [65] is that we leverage motion inductive bias to guide the extraction of the foreground region. Therefore, we can guarantee synthesized sample contains motion information rather than a random patch like CutMix.

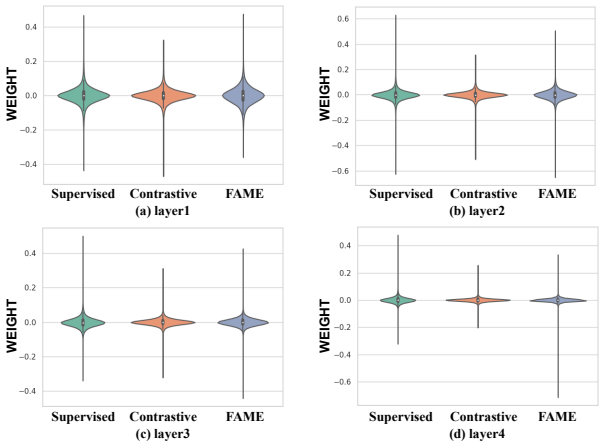


Figure 3. The statistics of temporal kernel weights at all layers of R(2+1)D. The learned kernel weights in the supervised/contrastive/FAME manner are violin-plotted from left to right.

3. Approach

In this section, we introduce our Foreground-background Merging (FAME) method. In section 3.1, we first revisit the vanilla contrastive learning framework based on instance discrimination [59] and shed light on the background bias when vanilla contrastive learning is transferred into the video domain. In section 3.2, we elaborate on how to separate foreground regions using our method. To clarify the notation, we denote the video clips as $X \in R^{C\times T\times H\times W}$, where C, T, H, W represents the dimension of the channel, timespan, height, width, respectively.

3.1. Background Bias in Contrastive Learning

The vanilla contrastive learning approach employs instance discrimination to learn the feature representation in a fully self-supervised manner [7,19,24]. Generally, it aims to maximize the similarity between the query sample q and its positive keys k^+, and minimize the similarity between q and negative keys k^-. We empirically use InfoNCE loss [20] for optimization:

$$\mathcal{L}_{nce} = -\log \frac{\sum_{k\in\{k^+\}} \exp(\text{sim}(q,k)/\tau)}{\sum_{k\in\{k^+,k^-\}} \exp(\text{sim}(q,k)/\tau)}, \quad (1)$$

where τ is the temperature hyper-parameter controlling the concentration level of the distribution, and $\text{sim}(q,k)$ measures the cosine similarity between the latent embeddings, i.e., $\text{sim}(q,k) = q^T k/(\|q\|_2 \|k\|_2)$. In most existing works [15], k^+ is the set of clip embeddings extracted from the same video as q, and k^- is the set from other videos.

However, this vanilla contrastive learning formulation in the video domain cannot fully utilize the dynamic motion information and tends to discriminate different instances according to the background cues [54]. To demonstrate this phenomenon, we plot the 1D convolution layers' kernel weights of R(2+1)D [51] trained by both supervised manner [1] and contrastive manner. As depicted in Fig. 3, the weights learned via contrastive formulation are more compact and clustered at all layers than the weights learned under supervision. It reveals that the supervised model allows more flexible temporal modeling, while the contrastive-based counterpart presents less temporal diversity and prefers static cues to temporal dynamics. Moreover, to consolidate our findings, we adopt the class-agnostic activation map (CAAM) [3] to measure the spatial attention in that CAAM can fairly evaluate pretrained representations without additional training. As shown in Fig. 1(a), the model trained by the traditional contrastive task cannot capture the moving object correctly and is distracted by the static background. This phenomenon further indicates that there exists the static background bias in the positive pair formulation. As mentioned in Sec. 1, two temporally different clips usually own similar static backgrounds but distinct motion patterns. Thus, when simply pulling two augmented clips closer, the model leans to prioritize the background alignment and give up grasping the dynamic motion. To deal with it, we carefully design FAME as an augmentation technique. Our idea is simple. We erase the static areas on purpose and retain the dynamic areas to construct the positive pairs. By doing so, the model has to align the motion area firstly and break the static shortcut. We show the contrastive learning framework with the proposed FAME in Fig. 4. In detail, we randomly sample two clips from different timestamps. Before applying the basic augmentation, we use our proposed FAME method to compound the foreground of one clip with the background from other videos in the same mini-batch. After that, the two clips are more similar in motions than backgrounds. Then, we feed these two clips into the 3D encoder and treat them as the positive keys while the rest of the clips serve as negative keys. Finally, we minimize the InfoNCE loss to pretrain the 3D encoder. By constructing the positive pair with the same foreground but diverse backgrounds, we guide the model to focus on temporal cues and suppress the impact of the background.

3.2. Foreground-background Merging

Motivated by mitigating background bias in self-supervised video representation learning, we intend to retain the foreground regions in original videos and shuffle the background areas among various videos. To achieve this goal, we propose the Foreground-background Merging method to augment the clips with minimal computation overhead. Concretely, FAME firstly separates the dynamic region of the static area and then composes the foreground on the other backgrounds.

[1]The supervised pretrained R(2+1)D is from torchvision library.

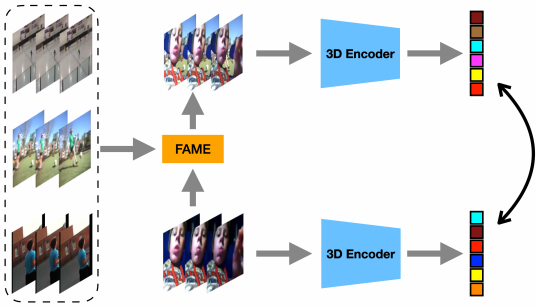


Figure 4. The contrastive learning framework with the proposed FAME. We first randomly sample two clips from a video and use FAME to generate new clips by composing the original foreground onto various backgrounds from other videos. Then, we feed the augmented clips into the existing contrastive learning scheme and perform self-supervised pretraining.

We first differentiate adjacent frames iteratively and then sum up the magnitude of the difference along channel and timespan dimensions to generate the seed region S. We formulate $S \in \mathbb{R}^{H \times W}$ as

$$S = \frac{1}{T-1} \sum_{c=1}^{C} \sum_{t=1}^{T-1} \left\| X_{c,t+1} - X_{c,t} \right\|_1 . \tag{2}$$

Intuitively, frame difference delivers natural dynamic motion that moving foreground objects tend to possess a great magnitude, while the static backgrounds are minor in this metric. Apart from frame difference, we also consider other methods to convey motion information like optical flow [66]. But we find the extraction of the dense optical flow of each frame is time-consuming and the frame difference can be an ideal substitute for reducing the computational cost. In practice, we find that the large values of the seed region S usually correspond to the moving objects' edge region. To expand the edge of the foreground objects, we take inspiration from the unsupervised foreground discovery [48] for seed propagation. Specifically, we leverage the color distributions to estimate the entire object. Denoting $N^{(F)}$ as the total number of pixels in the foreground region and $N_x^{(F)}$ as the number of the given color x appearing in the foreground region, the probability of a given color x appearing in the foreground region can be estimated as $P(x \mid F) = N_x^{(F)}/N^{(F)}$. Similarly, the probability of x belonging to the background region is $P(x \mid B) = N_x^{(B)}/N^{(B)}$. In practice, we sample the foreground color distribution in the top 50% of seed region S and the background color distribution in the last 10% of seed region S. Namely, in our setting, $N^{(F)} = [0.5 \times H \times W]$ and $N^{(B)} = [0.1 \times H \times W]$.

Given the above two distributions for the color x and the assumption that all pixels with the same color have the same probability of being the foreground and background, we approximate the foreground likelihood for a given color x as $P(F \mid x) = P(x \mid F)/[P(x \mid F) + P(x \mid B)]$. Therefore, the soft segmentation mask $M \in \mathbb{R}^{H \times W}$ can be calculated based on the color of each pixel. We formulate it as $[M]_{ij} = P(F \mid x_{ij})$, where x_{ij} is the color at pixel (i, j). To better filter out the background region, we binarize the mask as follows:

$$\left[\widetilde{M}\right]_{ij} = \begin{cases} 1, \text{ if } [M]_{ij} \text{ is among Top-}[\beta HW] \text{ of } M, \\ 0, \text{ otherwise,} \end{cases} \tag{3}$$

where $\beta \in [0, 1]$ is a hyper-parameter to describe the portion of the foreground. For the sake of computational efficiency, the mask we generate is constant with respect of timespan T. We view video clips as "image" when counting the color statistics, i.e., $\widetilde{X} = \sum_{t=1}^{T} X_t/T$. Having foreground mask $\widetilde{M}$, we then fill the rest with a random background. Denoting X, Y as foreground and background source clips, the synthetic clip

$$X_{\text{merge}} = X \otimes \widetilde{M} + Y \otimes (1 - \widetilde{M}), \tag{4}$$

where $\otimes$ is the element-wise multiplication. Noted that the background area we blend into the foreground video may not be the actual background and might contain unrelated motions. Those motions are necessary for robust motion-pattern learning. If all the background is filled with static pixels, the model will be collapsed to learn whether the region contains dynamic pixels as a shortcut.

Besides, we have tried three variants to obtain the foreground mask $\widetilde{M}$. Though we admit the quality of our generated mask cannot be comparable to the (semi-)supervised foreground discovery methods [11, 61], we find all variants of FAME consistently enhance the representation ability shown in Table 4 and FAME works best among them. In addition, we test the real-time performance of FAME (**480** fps) on a single 8G NVIDIA T4 GPU at 16×224×224 pixels, which is negligible in terms of whole pretraining.

4. Experiments

In this section, we first introduce datasets and the implementation details for our experiments. Then, we conduct a set of ablation studies to analyze and validate our FAME method quantitatively. As the evaluation, we report our results on downstream tasks: action recognition and video retrieval. Finally, we investigate and make sense of what the model learns with FAME qualitatively.

4.1. Datasets

We evaluate our method on four standard video benchmarks. Kinetics-400 [5] is a large-scale and high-quality dataset for action recognition, which consists of around 240K video clips with 400 human action classes. We use the training set of Kinetics-400 to pretrain our model in a self-supervised manner. UCF101 [47] and HMDB51 [34] are two smaller human action datasets, where the former contains over 13k clips covering 101 action classes and the latter annotates around 7,000 manually annotated clips with 51 action categories. Following previous methods [38,41,56,62], we use split 1 of UCF101 and HMDB51 in our experiments for downstream tasks. Also, we adopt UCF101 split 1 to conduct pre-training of our model for ablation experiments. Finally, we consider a more challenging dataset Diving48 [37] for evaluation, which involves around 18k trimmed video clips of 48 dive categories. It is noted that the different diving sequences in Diving48 often occur in a similar background and primarily differ in the fine-grained motion pattern.

4.2. Implementation Details

Self-supervised Pretraining. In the stage of self-supervised training, we adopt the MoCo framework [8, 24] as the representative of vanilla contrastive methods and apply our FAME method on MoCo framework. We select two common backbone choices, R(2+1)D-18 [51] and I3D-22 [5], as the 3D encoder. First, we randomly sample two different temporal clips in the same video as positive pair. Each clip consists of 16 frames with a temporal stride of 2. We spatially crop a random portion of clips and resize it to the size of 224×224 or 112×112. Then, we employ FAME to distract one out of the positive pairs. Notice that the background videos are from the clips in the same mini-batch. Next, following the prior work [15], we perform the basic augmentation containing random grayscale, color jittering, random horizontal flip, and random Gaussian blur. All these augmentations are temporally consistent according to [44]. We pretrain the model for 200 epochs with a batch size of 64 on 8 Tesla V100 GPUs during the training phase. The SGD optimizer is adopted with the initial learning rate of 10^{-2} and weight decay of 10^{-4}. We show more implementation details in the Supplementary Material.

Action Recognition. After pretraining, we initialize the backbone with the pretrained parameters except for the last fully connected layer. There are two popular protocols of action recognition to validate the self-supervised representations. One is *linear probe*. The encoder is frozen, and we only train the last fully connected layer. The second one is *finetune*, where we train the whole network in a supervised fashion. During the inference phase, we take the standard evaluation protocol [41, 56, 62]. We uniformly sample ten 16-frame video clips with a temporal stride of 2 from each testing video, then crop and resize them to 224×224 or 112×112. We average the prediction of each testing video clip and report Top-1 accuracy to measure the performance.

	UCF101		HMDB51	
β	single	both	single	both
1.0(baseline)	75.8		45.5	
0.7	80.3	79.6	49.6	50.8
0.5	81.2	81.2	52.6	51.4
0.3	82.0	81.1	51.6	53.1

Table 1. Top-1 accuracy with β on UCF101 and HMDB51. We denote the operating FAME on single branch (default setting) as *single* and the operating FAME on both branches as *both*.

Method	Pretrain Dataset	Diving48
Random Init.	✗	57.4
BE [54]	UCF101	58.8
FAME(ours)	UCF101	67.8
BE [54]	Kinectics-400	62.4
FAME(ours)	Kinectics-400	72.9

Table 2. Top-1 accuracy on Diving48 according to updated labels (**V2**). Both methods use I3D and 16×224×224 pixels.

Background	UCF101	HMDB51
none	75.8	45.5
intra-video	77.4(1.6↑)	47.6(2.1↑)
inter-video	81.2(5.4↑)	52.6(7.1↑)

Table 3. Top-1 accuracy on UCF101 and HMDB51 in terms of intra-/inter-video background.

Method	UCF101	HMDB51
baseline	75.8	45.5
Gauss	77.9	46.4
Seed	80.4	51.3
Grid	81.5	51.5
FAME	81.2	52.6
Grid†	86.5	58.7
FAME†	88.6	61.1

Table 4. Top-1 accuracy of various foreground-background separation methods on UCF101 and HMDB51. † indicates the pretrain dataset is Kinetics-400. FAME performs best.

Video Retrieval. Without further training, we directly leverage the representation from the pretrained encoder for evaluation. Following [38, 62], we take video clips in the test set to query k nearest neighbors in the training set. Specifically, we average ten uniformly sampled clips to obtain the global representation. If the category of the testing clip appears in the k nearest neighbors, it counts as a hit. We report Top-k recall R@k for evaluation.

4.3. Ablation Study

To analyze how our FAME improves self-supervised video representation learning, we conduct the following ablation studies. We choose split 1 of UCF101 as the pretrain dataset and I3D as the backbone for computational efficiency. All of the Top-1 accuracy in our ablation study is measured under the protocol of finetune.

Area Ratio of Foreground Region. To inspect how the area ratio of the foreground region contributes to the representation quality, we ablate β (i.e., the portion of the foreground) in the range of $\{1, 0.7, 0.5, 0.3\}$. We report the performance comparison in Table 1. Note that $\beta = 1$ reverts to the baseline method without applying FAME. It can be observed that the results of $\beta = 0.3$ and 0.5 vastly outperform baseline by $\sim 6\%$ on both UCF101 and HMDB51. The improvement of $\beta = 0.7$ is also considerable, though slightly inferior to the smaller value of β due to insufficient background replacement. It validates our idea that replacing the static area guides the model to distill motion-aware representations and thus enhance the downstream performance.

Stronger background debiasing. To explore whether FAME is sufficiently strong to reduce the background bias in contrastive learning, we design a stronger contrastive objective. That is, we apply FAME on both branches of MoCo and neither of the two processed video clips contains initial background information. We report the results in Table 1. The neglectable difference in performance between both settings proves that our default setting is strong enough to learn the scene-debiased representations.

Background source. Besides the foreground ratio, we also investigate how the source of background affects the representation ability to capture the motion. Specifically, we aim to explore whether the performance would change dramatically using the background in the same video instead of other videos. We perform an experiment where we merge the foreground of one video with the background sampled at different timestamps of the video itself. As shown in Table 3, we find that using the background from intra-video slightly boosts the baseline with 1.6% and 2.1% improvement on UCF101 and HMDB51 and the introduction of other videos' backgrounds brings further improvement, i.e., 5.4% and 7.1% gain on UCF101 and HMDB51. In general, the intra-video background is almost the same as the original one, while the inter-video background is quite distinct. Thus, it is consistent with our intuition that the modification from the intra-video is not adequate to mitigate background bias while replacing the background with diverse scenes better strengthens motion pattern learning.

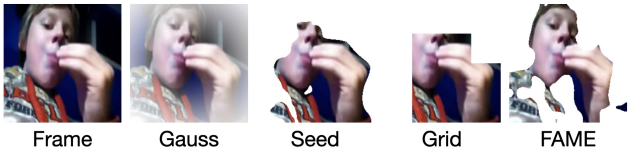


Figure 5. The illustration about FAME and three variants.

Variants of Foreground Mask. To verify that emphasizing moving foreground advances the motion understanding

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF101	HMDB51
CBT [49]	S3D	Kinetics-600	16	112	✓	54.0	29.5
CCL [33]	R3D-18	Kinetics-400	16	112	✓	52.1	27.8
MemDPC [22]	R3D-34	Kinetics-400	40	224	✓	54.1	30.5
RSPNet [6]	R3D-18	Kinetics-400	16	112	✓	61.8	**42.8**
MLRep [43]	R3D-18	Kinetics-400	16	112	✓	63.2	33.4
FAME (Ours)	R(2+1)D	Kinetics-400	16	112	✓	**72.2**	42.2
VCP [38]	R(2+1)D	UCF101	16	112	✗	66.3	32.2
PRP [64]	R(2+1)D	UCF101	16	112	✗	72.1	35.0
TempTrans [29]	R(2+1)D	UCF101	16	112	✗	81.6	46.4
3DRotNet [30]	R3D-18	Kinetics-400	16	112	✗	62.9	33.7
Spatio-Temp [55]	C3D	Kinetics-400	16	112	✗	61.2	33.4
Pace Prediction [56]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
SpeedNet [4]	S3D-G	Kinetics-400	64	224	✗	81.1	48.8
VideoMoCo [41]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
RSPNet [6]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
MLRep [43]	R3D-18	Kinetics-400	16	112	✗	79.1	47.6
ASCNet [26]	R3D-18	Kinetics-400	16	112	✗	80.5	52.3
SRTC [69]	R(2+1)D	Kinetics-400	16	112	✗	82.0	51.2
FAME (ours)	R(2+1)D	Kinetics-400	16	112	✗	**84.8**	**53.5**
DSM [53]	I3D	Kinetics-400	16	224	✗	74.8	52.5
BE [54]	I3D	Kinetics-400	16	224	✗	86.8	55.4
FAME (ours)	I3D	Kinetics-400	16	224	✗	**88.6**	**61.1**

Table 5. Comparison with the existing self-supervised video representation learning methods for action recognition on UCF101 and HMDB51. To compare fairly, we list each work's setting, including backbone architecture used, pretrain dataset and spatial-temporal resolution. Freeze (tick) indicates linear probe, and no freeze (cross) means finetune.

Method	Backbone	R@k				
		R@1	R@5	R@10	R@20	R@50
SpeedNet [4]	S3D-G	13.0	28.1	37.5	49.5	65.0
TempTrans [29]	R3D-18	26.1	48.5	59.1	69.6	82.8
MLRep [43]	R3D-18	41.5	60.0	71.2	80.1	-
GDT [42]	R(2+1)D	57.4	73.4	80.8	88.1	92.9
ASCNet [26]	R3D-18	58.9	76.3	82.2	87.5	93.4
FAME (ours)	R(2+1)D	64.6	77.7	82.9	87.6	94.2

Table 6. Comparison with the existing self-supervised video representation learning methods for video retrieval. All methods are pretrained on Kinetics-400. We report the Top-k recall R@k when k=1, 5, 10, 20, 50 on UCF101.

in contrastive framework, we devise three variants of foreground mask: (i) Gauss: we adopt a 2D Gaussian kernel matrix as the foreground mask. It derives from the assumption that videos are shot in the object-centric form. (ii) Seed: we just take the seed region S to characterize the foreground. (iii) Grid: the video is split into 4×4 grids spatially. We count the sum of S in each grid and take the greatest eight grids as the foreground area. A brief illustration is displayed in Fig. 5. We compare FAME with these three variants in Table 4. First, we note that all variants improve the baseline by a large margin, demonstrating the benefit of the introduction of different backgrounds. Furthermore, refining the foreground mask from Gauss, Seed, Grid to FAME continually increases the action recognition performance. Interestingly, we notice that Grid outperforms FAME a little on UCF101. We conjecture that since both the pretrain dataset and downstream dataset are UCF101, similar backgrounds that occurred might be leveraged as a shortcut. To delve into this phenomenon, we carry out an extra experiment on another pretrain dataset Kinetics-400. Top-1 accuracy of Grid variant is over 2% lower than FAME on both UCF101 and HMDB51. It indicates that a meticulous segmentation mask instead of a rough grid box is more effective in facilitating generalization ability when transferring the motion-aware representations to different downstream benchmarks.

4.4. Evaluation on Downstream Tasks

Action Recognition on UCF101 and HMDB51. To verify the effectiveness of the proposed method, we compare our method with the prior arts. In Table 5, we report Top-1 accuracy on UCF101 and HMDB51. For a fair comparison, we do not report methods with a deeper backbone or non-single modality, e.g., optical flow, audio, and text.

Our method obtains the best result on UCF101 and a comparable result on HMDB51 in the linear probe setting. FAME beats MLRep [43] by a large margin, i.e., about 9.0% gain on both UCF101 and HMDB51, where MLRep carefully designs the multi-level feature optimization and temporal modeling. The outstanding performance verifies that our method captures the moving foreground patterns without further finetune.

In the finetune protocol, FAME with R(2+1)D backbone achieves the best result on UCF101 and HMDB51. It shows that FAME learns the scene-debiased and motion-aware representations on the Kinetics-400 dataset, which would generalize better to downstream datasets. Remarkably, our simple formulation outperforms SRTC [69] by 2.8% and 2.3% with the same backbone R(2+1)D, despite its two additional sub-loss terms regularizing the self-supervised pretraining. Notably, we share similar motivation with BE [54], which directly adds a static frame to every other frame and regards this distracting video as the positive pair to the original video. Such subtle disturbance cannot sufficiently mitigate the static background bias, confirmed by the experimental results. When using the same backbone I3D, our FAME outperforms BE by 1.8% and 5.7% on UCF101 and HMDB51, respectively. It proves that our method can better highlight the motion patterns.

Video Retrieval on UCF101 and HMDB51. We report the performance comparison on the video retrieval task in Table 6. Our method achieves significant performance gain from R@1 to R@50. Remarkably, though ASCNet [26] devises two particular tasks to learn the appearance and speed consistency, we still gain 6.7% improvement on Top-1 retrieval accuracy only through fabricating motion-aware positive pairs, which demonstrates our methods can recognize the action semantics more precisely.

Evaluation on Diving48. Besides common action recognition benchmarks, we finetune and test our FAME on a more challenging fine-grained dataset Diving48 and report the results in Table 2. In Diving48, since the static backgrounds are not strongly related to the fine-grained diving label, our motion-aware representations can strongly benefit the action recognition. FAME can boost a randomly initialized model by 15.5% when pretraining on Kinetics-400. In comparison, no matter whether the pretrain dataset is UCF101 or Kinetics-400, BE is far less effective than FAME. This is because BE does not construct the motion-aware positive pairs where background features are still more similar in contrast to the motion counterparts. The result on Diving48 indicates FAME can indeed make the model perceive the long-term motion patterns and hinder the scene bias.

4.5. Visualization Analysis

To better demonstrate the effectiveness of FAME, we visualize the CAAM [3] in Fig. 1. By comparison, the model learned by FAME enhances the activation on the moving foreground area and suppresses the background area. For example, in the second column of Fig. 1, FAME precisely captures two baseball players in the court, while the vanilla contrastive method displays a dispersed highlight map and fails to attend to the motion area.

Moreover, compared to vanilla contrastive method, the distribution of temporal kernel learned by FAME is more scattered with larger variance as shown in Fig. 3. Surprisingly, the shape of FAME's temporal kernel weights is similar to supervised learning, showing that via FAME, contrastive learning can well grasp action semantics. In light of the aforementioned evidence, we safely conclude that guided by the strong motion inductive augmentation like FAME, contrastive learning can also prevent background cheating and pay attention to the motion patterns.

5. Conclusion

In this work, we propose a new Foreground-background Merging (FAME) method to alleviate the background bias in self-supervised video representation learning. Via Foreground-background Merging, we augment the original video by fusing the original foreground with other videos' backgrounds. When forcing the backbone model to learn semantically consistent representation between the original video and the fused video, the model can learn the scene-debiased and motion-aware representations of videos. Experimental results on a bunch of downstream tasks manifest the effectiveness of our method.

While our work shows some promising results, there are still some limitations. One is that the quality of foreground extraction is not stable, especially when foreground and background have no significant differences in the color distribution or the camera is dynamically moving. Besides, the foreground area ratio is now fixed by hyper-parameter β. It would be better to set an adaptive foreground area ratio.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61932022, Grant 61720106001, Grant 61971285, Grant 61831018, Grant 61871267, Grant T2122024, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 20511100100.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 3

[3] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 07, pages 10451–10459, 2020. 1, 4, 8

[4] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2, 7

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 5

[6] Peihao Chen et al. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI*, 2021. 3, 7, 13

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5, 12

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 1

[10] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 3

[11] Ioana Croitoru, Simion-Vlad Bogolin, and Marius Leordeanu. Unsupervised learning from video to detect foreground objects in single images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 5

[12] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 3

[13] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017. 3

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 1

[15] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2, 3, 5

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020. 1

[17] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2928, 2021. 3

[18] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 13

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[20] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2, 3

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE international conference on computer vision Workshops*, pages 0–0, 2019. 3

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Proceedings of the European conference on computer vision*, pages 312–329. Springer, 2020. 3, 7

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 3

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 3, 5

[25] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. Human action recognition without human. In *European Conference on Computer Vision*, pages 11–17. Springer, 2016. 3

[26] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. *arXiv preprint arXiv:2106.02342*, 2021. 3, 7, 8

[27] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2021. 3

[28] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggregation network and dynamic template matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[29] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *Proceedings of the European conference on computer vision*, pages 425–442. Springer, 2020. 7

[30] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 7

[31] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 2

[32] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020. 3

[33] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810*, 2020. 7

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

[35] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017. 2

[36] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2105–2114, October 2021. 3

[37] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European conference on computer vision*, pages 513–528, 2018. 3, 5

[38] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11701–11708, 2020. 5, 6, 7

[39] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[41] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 5, 7

[42] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 7

[43] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. *arXiv preprint arXiv:2108.02183*, 2021. 7, 8

[44] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020. 2, 3, 5

[45] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2103.16559*, 2021. 3

[46] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 12

[47] 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

[48] Otilia Stretcu and Marius Leordeanu. Multiple frames matching for object discovery in video. In *BMVC*, volume 1, page 3, 2015. 4

[49] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 7

[50] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1, 4, 5

[52] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision*, pages 391–408, 2018. 2

[53] Jinpeng Wang et al. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI21*, 2021. 3, 7

[54] Jinpeng Wang et al. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *CVPR*, 2021. 3, 4, 6, 7, 8

[55] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019. 7

[56] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *Proceedings of the European conference on computer vision*, 2020. 5, 7

[57] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 1

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

[59] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 1, 2, 3

[60] Fanyi Xiao, Joseph Tighe, and Davide Modolo. Modist: Motion distillation for self-supervised video representation learning. *arXiv preprint arXiv:2106.09703*, 2021. 3

[61] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[62] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2, 5, 6

[63] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. 3

[64] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020. 7

[65] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3

[66] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 4

[67] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390, 2018. 1

[68] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3

[69] Lin Zhang, Qi She, Zhengyang Shen, and Changhu Wang. How incomplete is contrastive learning? an inter-intra variant dual representation method for self-supervised video recognition. *arXiv preprint arXiv:2107.01194*, 2021. 3, 7, 8

[70] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE international conference on computer vision*, pages 2921–2929, 2016. 12

A. More Implementation Details

A.1. Self-supervised Pretraining Details.

In the pretraining stage, we adopt the SGD optimizer with the initial learning rate of 0.01 and weight decay of 10^{-4}, and we decay the learning rate by 0.1 at epoch 120 and 180. For the implementation of MoCo, we closely follow the parameter setting in [8]. The number of the negative queue is set to 65536 for Kinetics-400, and 2048 for UCF101, respectively. We also swap the key/queue samples so that each sample can generate the gradient for optimization. The momentum of updating the key encoder is 0.999, and the temperature hyper-parameter τ is 0.1. We use a 2-layer MLP projection head.

A.2. Augmentation Details.

We perform data augmentation using Kornia package [46]. In the pretraining and finetune phase, we crop 224×224 or 112×112 pixels from a video with RandomResizedCrop, which randomly resizes the input area between a lower bound and upper bound. We set the bound as [0.2, 1]. Then, the basic augmentation set consists of RandomGrayscale (probability 0.2), ColorJitter (probability 0.8, {brightness, contrast, saturation, hue} = {0.4, 0.4, 0.4, 0.1}), RandomHorizontalFlip (probability 0.5) and RandomGaussianBlur (probability 0.5, the kernel with radius 23 and standard deviation $\in$ [0.1, 2.0]). In the linear probe stage, we take a simpler augmentation setting instead. We only apply RandomResizedCrop with the bound [0.2, 1] and RandomHorizontalFlip (probability 0.5).

A.3. More Details on Action Recognition.

In the finetune stage, the SGD optimizer is adopted with the initial learning rate of 0.025 and weight decay of 10^{-4}. We finetune the model for 150 epochs with a batch size of 128 on 4 Tesla V100 GPUs. We decay the learning rate by 0.1 at epoch 60 and 120. Besides, we add the dropout layer before the last fully connected layer. We set dropout rate 0.7 for UCF101 and 0.5 for HMDB51, respectively.

We train the last fully connected layer in the linear probe with the initial learning rate of 5 and weight decay of 0. We finetune the model for 100 epochs with a batch size of 128 on 4 Tesla V100 GPUs. We decay the learning rate by 0.1 at epoch 60 and 80. Besides, We $\mathcal{L}_2$ normalize the embeddings before the last fully connected layer.

B. More Visualization of FAME

In Figure 6, we show more foreground masks obtained from FAME. We show that FAME can discover most regions of the foreground objects and remove the monotonous backgrounds.

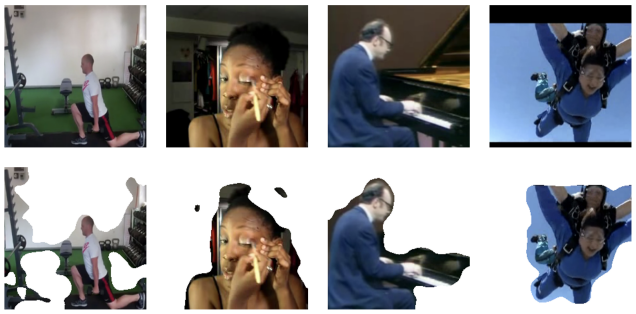


Figure 6. Illustration of FAME visualization. The first row is the video frame while the second row is the foreground mask FAME generates.

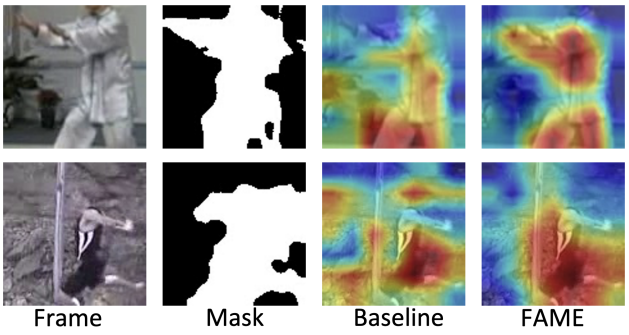


Figure 7. Class activation maps (CAM) visualization. Red areas indicate the important areas for the model to predict the action class. Comparing to baseline, FAME resists the impact of background and highlights the motion areas.

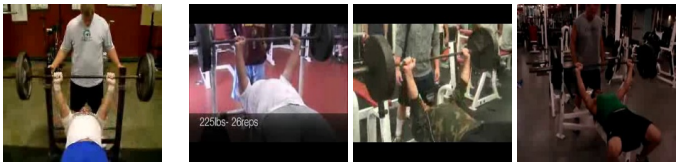
C. CAM Visualization

Besides CAAM visualization, we provide the CAM [71] visualization in Figure 7. With that, we can spot the contribution of each area and find crucial regions for discriminating the specific action class. We find that when integrated with FAME, the model can focus on moving foreground area rather than background context. For example, in the first row of Figure 7, FAME precisely captures the moving upper and lower body when the man is practicing TaiChi, while the baseline displays a dispersed highlight map and fails to attend to the motion area. In addition, we illustrate that the CAM activation map can almost overlap with the foreground mask generated by FAME. It testifies that our strong motion inductive augmentation guides the model to perceive the motion patterns and hinder the background bias.

D. Visualization of Video Retrieval

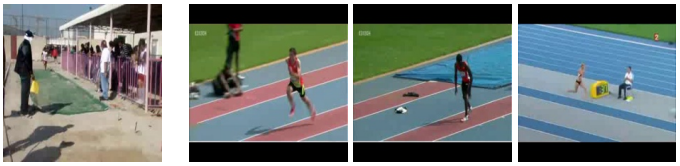
In Figure 8, we demonstrate the results of video retrieval. After pretraining the model on Kinetics-400, we conduct the video retrieval experiment on UCF101. The results show that our model can retrieve diverse video samples that

(a) BandMarching



(b) BenchPress

(c) HorseRace



(d) LongJump

Figure 8. Visualization results of video retrieval.The first column is the video frame of query instances. The rightmost three columns are Top-3 nearest retrieval results.

share the same action semantics with the query, regardless of the background context. For example, in Fig 8d, the query sample contains the action in the sandpit, and our model could retrieve the long jump samples in the standard stadium. Though the backgrounds in the query and retrieved videos are quite different, our model achieves accurate retrieval by attending to the dynamic motions and understanding the true action semantics.

E. More Results on Something-something V2

We finetune our pretained model on Something-Something V2 [18]. We obtain 53.3% Top-1 accuracy with R(2+1)D, which beats RSPNet [6] under same resolution.