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

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, Hongkai Xiong*

SHANGHAI JIAO TONG UNIVERSITY

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Abstract:

- **Task:** Video Representation Learning
- **Problem:** Background Bias. As seen in Fig 1, when naively pulling two augmented views of a video closer, the model tends to learn the common static background as a shortcut but fails to capture the motion information.
- **Our solution:** We propose **F**oreground-ba**c**kground **Me**rging (FAME) to deliberately compose the moving foreground region of the selected video onto the static background of others.

Method (FAME):

Background Bias in Contrastive Learning

The vanilla learning optimize the InfoNCE Loss:

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

The vanilla contrastive learning in the video domain cannot fully utilize the dynamic motion and tends to discriminate different instances according to the background cues. We show this phenomenon by the learned weights in Fig3.

Our Video Representation Learning Pipeline

As seen in Fig.2, the overall pipeline can be described within 4 steps:

- We randomly sample two clips from different timestamps.
- We use our FAME to compound the foreground of one clip with the background from other videos in the same mini-batch.
- 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.
- We minimize the InfoNCE loss to pretrain the 3D encoder.

The propose FAME

- We differentiate adjacent frames iteratively and sum up the magnitude of the difference along channel and timespan to generate the seed region S:

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

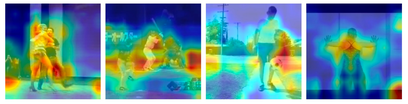
- We binarize the mask from the unsupervised foreground discovery method [1] for seed propagation:

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

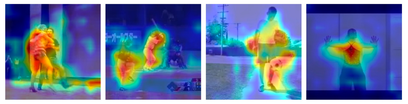
- Having foreground mask $\widetilde{M}$, we then fill the rest with a random background. Denoting X, Y as foreground and background source clips, the synthetic clip are generated by:

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

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



(a) Results of vanilla contrastive learning.



(b) Results of our approach FAME.

Fig1: Class-agnostic activation map visualization of important areas.

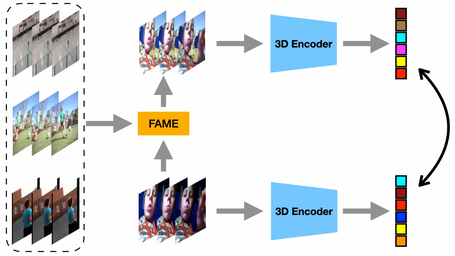


Fig2: The contrastive learning framework with the proposed FAME.

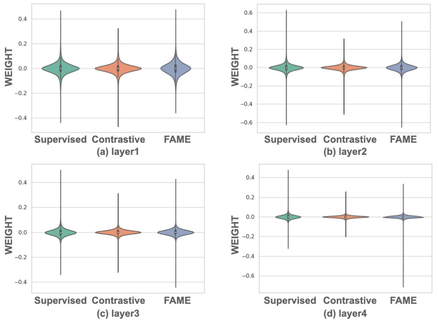


Fig3: The statistics of temporal kernel weights at all layers of R(2+1)D. The learned kernel weights in the supervised/contrastive/FAME manner are violin-plotted from left to right.

Experiments:

- We conduct ablation study on the range of β. The performance is reported in Table 1.
- We explore whether the performance would change dramatically using the background in the same video instead of other videos in Table 3.
- To verify the effect of moving foreground, we devise three variants of ground mask: (i) Gauss, (ii) Seed, and (iii) Grid. The results are in Table 4 and Fig 4.
- We report Top-1 accuracy on UCF101 and HMDB51 in Table 5.
- We finetune and test our FAME on a more challenging fine-grained dataset Diving48 and report the results in Table 2.
- We report the performance comparison on the video retrieval task in Table 6.

	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.

Method	Pretrain Dataset	Diving48
Random Init.	✗	57.4
BE [53]	UCF101	58.8
FAME(ours)	UCF101	67.8
BE [53]	Kinectics-400	62.4
FAME(ours)	Kinectics-400	72.9

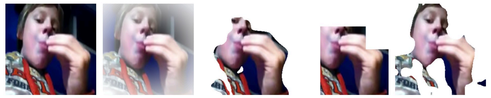
Table 2: Top-1 accuracy on Diving48 according to updated labels (V2).

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.



Frame Gauss Seed Grid FAME

Fig4: The illustration about FAME and three variants.

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF101	HMDB51
CBT [48]	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 [63]	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 [54]	C3D	Kinetics-400	16	112	✗	61.2	33.4
Pace Prediction [55]	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 [68]	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 [52]	I3D	Kinetics-400	16	224	✗	74.8	52.5
BE [53]	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.

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.