MAC: Mask-Augmentation for Motion-Aware Video Representation Learning

Arif Akar[1,2]
arifakar@gmail.com
Ufuk Umut Senturk[1,2]
ufukumutsenturk@gmail.com
Nazli Ikizler-Cinbis[1]
nazli@cs.hacettepe.edu.tr

[1] Department of Computer Engineering
Hacettepe University
Ankara, Turkey

[2] ASELSAN MGEO, Inc.
Ankara, Turkey

Abstract

We present MAC, a lightweight, efficient, and novel **M**ask-**A**ugmentation te**C**hnique and pretext task for self-supervised video representation learning. Most recent and successful methods leverage the instance discrimination approach that requires heavy computation and often leads to inefficient and exhaustive pretraining. We apply MAC augmentation on videos by blending foreground motion using frame-difference-based masks and set up a pretext task to recognize applied transformation. While we incorporate a game of predicting the correct blending multiplier at the pretraining stage, our model is enforced to encode motion-based features which are then successfully transferred to action recognition and video retrieval downstream tasks. Furthermore, we demonstrate the extension of the proposed approach step-by-step to improve representation capabilities in a joint contrastive framework. The proposed method achieves superior performance on UCF-101, HMDB51, and Diving-48 datasets at low resource settings and competitive results with instance discrimination methods at costly computation settings[1].

1 Introduction

Considering the abundance of unannotated raw video data, video representation learning is one of the domains that can potentially favor most from the self-supervised learning paradigm. The spirit of self-supervised learning lies in the heart of finding useful information from the data itself. Different from the image domain, temporal dimension of videos enables an additional direction to explore useful and supervisory consistencies in data [33, 52, 55]. There has been a tendency in early works to use only simple transformations to extract useful signals from data. Recent works, on the other hand, enforce models to become invariant and/or equivariant to applied transformations. Some works aim for spatial-invariance [12, 47] and others aim for invariance/equivariance to temporal transformations [26, 37] to improve representation capabilities of models.

Videos naturally contain redundant information[15] that consists of background for the most part. Models tend to learn clues from static background to solve a particular objective

[1] Codes are available at https://github.com/ufukpage/MAC_SSL.

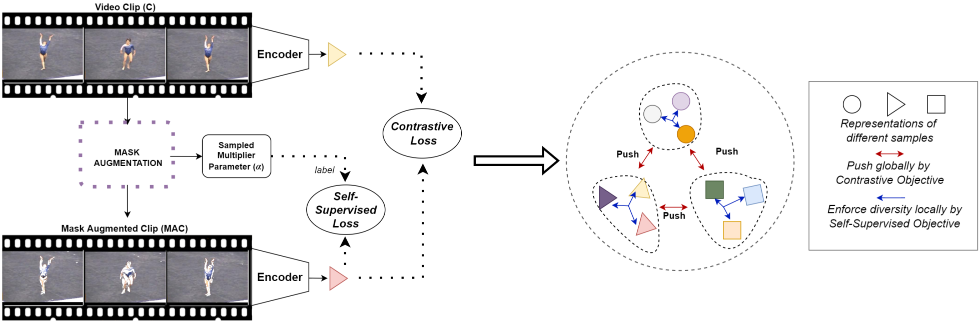
Video Clip (C)
Encoder
Contrastive Loss
MASK AUGMENTATION
Sampled Multiplier Parameter (α)
label
Self-Supervised Loss
Mask Augmented Clip (MAC)
Encoder
Push
Push
Push
Representations of different samples
Push globally by Contrastive Objective
Enforce diversity locally by Self-Supervised Objective

Figure 1: Clip (C) and Mask Augmented Clip (MAC) are preprocessed with different spatial and identical temporal augmentations. During pretraining, we train encoder via self-supervised and contrastive losses, both utilize novel mask augmentation technique MAC. Contrastive objective controls representation space globally by pulling distinct variants of the same video samples together and pushing instance representation of different video samples apart. This creates local clusters that are diversified by the self-supervised objective. The self-supervised objective enables the model to learn the internal structure of the clusters as the contrastive objective helps the discriminating potential of the model. Purple dotted area represents mask extraction and blending part.

that is not always semantically correct for that objective. Recent works have demonstrated that existing methods suffer from background bias problem [10, 32] and several works try to address this problem by decomposition of motion and scene elements [25, 42, 43, 46, 47, 48].

To mitigate this problem efficiently, we think that focusing on the regions of motion can be beneficial. To extract motion information, we take a voluntary choice of using a very simple and noisy form of dynamic foreground that can be computed as time-derivative of videos, i.e., frame differences. We verify that although noisy, foreground masks still possess strong potential as pseudo ground-truth supervision.

To this end, we extract binary motion masks by simple frame difference with a momentum structure. We then augment each frame by blending the extracted motion masks using a multiplier that is further utilized as a supervisory signal in a self-supervised objective. We call this approach MAC: **M**ask-**A**ugmentation te**C**hnique. We believe that our self-supervised objective enables any model to embed background-invariant representations of a video instance that are locally distinctive in representation space. Consequently, since MAC is a temporal transformation based on temporal derivatives and has a transformation-recognition objective, it enforces the model to become equivariant against spatio-temporal transformations. Additionally, we propose a contrastive objective for improving representation space. In our contrastive framework, we sample two views of the same instance and apply basic spatial and temporal data augmentations without changing the frame sequence. However, we apply MAC only on query view so that moving parts are further spatially diversified, but temporal relations are kept unchanged. Since we define these views as positive pairs, the strongest consistency between them remains as temporal motion patterns. Thus, the model focuses on pulling their representations globally by encoding motion and extracting background-invariant representations thanks to the proposed mask augmentation. By linking similarity relation over motion, contrastive objective works as a global force between different instances while self-supervised objective works instance-wise, i.e., locally. Figure

1 presents how self-supervised and contrastive objectives presumably operate in latent space.

We examine the performance of our main method and its variants for action recognition and video retrieval. We extensively conduct ablative studies corresponding to particular design choices. Our experiments demonstrate that the proposed MAC method performs on par with state-of-the-art methods in low resource settings, and competitive against state-of-the-art instance-discrimination methods with large-scale setups on mentioned downstream tasks. In summary, our contributions are as follows:

- We propose a simple yet effective mask augmentation technique that utilize regions of motion via foreground masks computed by frame differences.
- We propose a novel self-supervised objective, denoted as MAC-S, based on predicting the largely imperfect foreground masks. Moreover, we demonstrate how our foreground based augmentation can be used in combination with our novel contrastive objective, denoted as MAC-C.
- To the best of our knowledge, MAC is the first work that aims both background-invariance and spatio-temporal equivariance by exploiting transformation-recognition paradigm and utilizing motion cues as consistency signal.

2 Related Work

2.1 Contrastive Video Representation Learning

Representation learning using contrastive objective aims at extracting rich and useful features by discriminating instances based on positive and negative pair relations. After promising results in image domain[4, 7, 17, 21, 22], contrastive learning have been extensively applied for video data that can leverage an additional temporal dimension [2, 14, 19, 35, 36, 37]. In the context of video, there has been a wide pool of ideas to define positive and negative pairs. [46] proposes a spatial disturbance transformation to define positive pairs that could change local statistics but doesn't change the motion, [51] finds positive pairs by using a soft nearest neighbor search, [53], links positive pairs within the same video when they have correspondences of parts and objects among frames. Similar to ours, [12, 47] perform a blending operation for enhancing foreground on query view to create positive pairs.

2.2 Pretext-task Based Video Representation Learning

Pretext tasks based on spatio-temporal transformations. The initial attempts of pretext-based video representation learning include learning basic spatial transformations including cropping, rotating, warping or prediction-as-a-game [1, 29, 30, 34, 56]. Recent works include pretext tasks based on temporal order of frames or clips seeking to hypothesize over consistency and coherency of temporal features [27, 33, 49, 52]. Hu *et al.* [23] propose a simple task to enforce network to guess whether an augmented clip is preceded or followed by other augmented clips, [27, 33] proposes multi-objective training seeking the complementary success of both temporal and spatial transformations and show that extraction of both spatial and temporal information indeed leads to better results. Pretext tasks based on speed/frame rate have been largely studied in self-supervised learning [3, 6, 9, 45, 54, 55].

Pretext tasks based on motion. Motion is one of the strongest signals that can be utilized in video representation learning. Although speed-based self-supervision indirectly makes use of motion cues, there are works that directly focus on motion. Diba *et al.* [11] combine

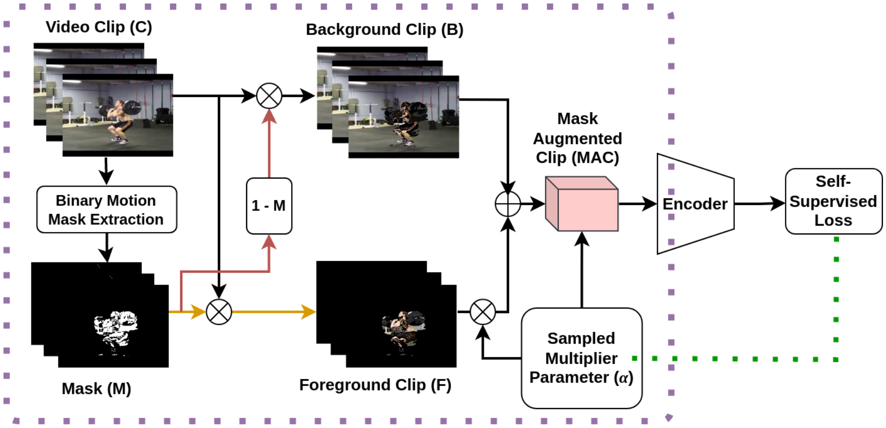
Video Clip (C)
Background Clip (B)
Binary Motion Mask Extraction
1 - M
Mask Augmented Clip (MAC)
Encoder
Self-Supervised Loss
Mask (M)
Foreground Clip (F)
Sampled Multiplier Parameter (α)

Figure 2: MAC pretraining framework. Foreground pixels of a video clip C are designated by binary foreground masks. FG region is multiplied with blending parameter α and merged with untouched background pixels to form mask-augmented clip. α is utilized as pseudo-ground truth to provide supervisory signal in self-supervised objective. During pretraining, we force our model to predict randomly sampled α by focusing on moving foreground that leads to learn motion-aware features. Contrastive-objective is not shown for simplicity.

visual appearance and motion learning in a dynamic motion representation layer, and [44] accommodate patches with pre-set trajectories into frames that constrains network to learn motion representations with pseudo-tracking.

Hybrid Approaches The combination of both pretext-based and contrastive learning methods have also caught attention in recent works [13, 40, 45]. [45] utilizes speed similarity to define positive pairs, Tao *et al.* [40] show that pretext task can be combined with a contrastive objective and both objectives provide a combination of local and global views in representation space.

3 Proposed Approach

The proposed MAC framework is illustrated in Figure 2. The core idea is to first extract the motion masks over the consecutive frames and then to utilize a motion-aware self-supervised objective to guide representation learning. Below, we first describe the foreground mask extraction procedure and then present our pretext method in detail. Lastly, we introduce the possible directions to create strong variants of our method.

3.1 Extracting Foreground Masks

The binary foreground masks are computed for each frame of a sampled clip on the fly. The idea is to extract motion information from frames in the form of foreground masks and use it as supervisory signal in pretraining. Since the time difference between frames is relatively small, static background can be subtracted to give a reasonable representation of active foreground and corresponding change in the scene.

A simple idea is to use a momentum structure to keep the track of background history similar to (but simpler than) that of [50]. Instead of taking a direct difference with current and previous frame, background is modeled using a moving average of recent frames. Accounting a brief history of the previous frames is an intuitively simple solution against illumination changes, ghosting effects or relative background motion. We provide steps of mask extraction as pseudo-code in Supplementary Work.

3.2 MAC Augmentation and Pretext Task MAC-S

Given a dataset D that consists of N video clips, we randomly sample clips $X_i = \{x_{ij}, j \in 1 \dots F\}$ of length F, using varying temporal strides T where T is sampled once from $\{1,2,3,4\}$ for each clip X. T provides different motion patterns of the same video by changing sampling rate, hence can be seen as temporal augmentation. The resulting input is designated as $X_i \in \mathbb{R}^{C \times F \times H \times W}$, where C is the number of input channels, H and W are spatial dimensions of frames. We extract binary foreground masks for each frame in clip X_i and denote them as $M_i = \{m_{ij}, j \in 1 \dots F\}$. With a blending function, we multiply each clip X's foreground region with a scalar α and merge with the background to construct the clip $\widehat{X}$ such that

$$\widehat{X} = \alpha \times (M \odot X) + ((1 - M) \odot X) \tag{1}$$

where $(1 - M)$ is the inverse of M and $\odot$ is an element-wise product. Foreground multiplier α in range (0,1] is then used as a pseudo-label for self-supervision. Finally, we apply a set of basic augmentations on each frame to conclude preprocessing. The final augmented clip $\widehat{X}$ is fed through a 3D-CNN network and model is trained with the self-supervised loss to predict the foreground multiplier α. We pose self-supervised objective as a classification task and utilize cross-entropy loss. In the rest of the paper, we will use MAC to denote augmentation process and MAC-S to refer to the corresponding self-supervised pretext task for predicting α's in pretraining.

Degenerate Solutions. Since MAC is a transformation at pixel level, there is a possibility that any enough-capacity network could potentially reverse engineer the prediction game without learning any motion features at all. In fact, we observe that without any standard augmentation, model tends to find shortcuts and converge to degenerate representations that do not transfer to downstream tasks well. Hence, we adapt strong augmentations to make sure that model does not receive any identical MAC-augmented frames of the same video instance. For instance, we apply color jittering on each frame to hinder the network from taking advantage of the overall color histogram as mentioned in [28]. In this way, the model learns to focus on motion in order to correctly predict the multiplier.

3.3 Variants of MAC

Using Multiple α's. To increase the complexity of α-prediction, we can extend the task to predict distinct α values for distinct subclips. We denote this pretext task as MAC-S-n, where each clip is equally divided into n subclips, each having F/n frames and we randomly sample a distinct α value for each subclip. We use same α value for all frames in a subclip and set up the pretext task to find the correct α for each subclip. It should be noted that model should find correct predictions for each subclip to satisfy the classification objective.

Let k indicate the number of different values that α can take. In this case, there will be a total of k^n possible cases for the classification objective. For example, for the task with 4 subclips denoted as MAC-S-4, there are 4^4 possible classes which makes it a harder classification task than MAC-S-2 model which tries to solve for 4^2 classes. In the former case, the network is more constrained to embed each subclip representations distinctively to solve for each α. We conjecture that this will in return potentially increase spatio-temporal variance in the representation space and our experimental results support this claim.

Combination with Contrastive Objective. MAC aims to extract video representations based on dynamic foreground regions. It is expected that the representations of the videos of the same class should also be similar as long as motion pattern is similar. In order to leverage this property and increase the regularization of pretraining phase, we combine MAC with a contrastive objective. Our contrastive objective is based on InfoNCE [18] loss for which we apply MAC on query samples along with standard augmentations previously mentioned. The key samples are preprocessed only with standard augmentations. We observed that MAC-C with contrastive objective brings strong improvements due to reasons that are detailed in the discussion in Introduction regarding contrastive objective.

In practice, MAC-C-n is optimized with InfoNCE loss using the similar setting in [21]. Query clip is denoted as q and key clip with only basic augmentations applied is denoted as k_+. Following the previous works, negative samples are coming from other videos that have been added to queue and the contrastive loss $\mathcal{L}_{\text{Contrastive}}$ is defined as

$$\mathcal{L}_{\text{Contrastive}} = -log\frac{\exp(q.k_+/\tau)}{\sum_{i=0}^{K}\exp(q.k_i/\tau)} \tag{2}$$

where τ is the temperature parameter that controls similarity/distance strength of embeddings in representation space.

Consequently, we introduce MAC-SC-n, that is a joint-optimization framework including both contrastive and self-supervised objectives. Specifically, we pretrain our models with both self-supervised and contrastive losses as follows:

$$\mathcal{L}_{\text{Total}} = \lambda * \mathcal{L}_{\text{Self-Supervised}} + \beta * \mathcal{L}_{\text{Contrastive}} \tag{3}$$

Here, β is defined as a weight parameter that controls the contribution of contrastive loss and $\mathcal{L}_{\text{Self-Supervised}}$ is defined as cross-entropy loss which is controlled by λ. Empirically, λ and β are set to 1 and 0.5 respectively in a default setting.

4 Experiments

Firstly, we inspect the contribution of each element through ablation studies. Then we present our experiments on downstream tasks including action recognition and video retrieval. We also examine the learning behavior both quantitatively and qualitatively. More evaluation with ablation experiments, details on experimental settings and failure case analysis with visualizations are available in Supplementary Material.

For self-supervised pretraining and action recognition downstream tasks, we utilize various benchmark datasets. UCF101 [39] is a widely used human action recognition dataset that contains more than 13k videos of 101 actions, of which 9.5k videos used for training and 3.5k videos used for test. HMDB51 dataset [31] contains nearly 7k videos of 51 human actions. Lastly, Kinetics400 [5] is a large video dataset that contains nearly 240k videos of 400 human actions. For fair comparison, we follow the same experimental setup in [44]. We also experiment with less-biased Diving48 dataset that has 18k videos of 48 fine-grained classes.

Method	Acc
Contrastive Obj. (w/o MAC aug.)	69.5
MAC-S-2 (SSL Obj., β=0)	81.5 ($\uparrow$ 12%)
MAC-C-2 (Contrastive Obj., λ=0)	81.5 ($\uparrow$ 11.9%)
MAC-SC-2	**84.8** ($\uparrow$ 15.3%)

Table 1: Contribution of each learning objective; self-supervised loss vs contrastive loss vs joint optimization (MAC-SC-2). First row shows contrastive objective when only standard augmentations are applied. Complementary performance of SSL and contrastive objective results in an improved representation learning.

# of Multiplier	Acc
MAC-SC-1 (4-way classification)	82
MAC-SC-2 (2x 4-way classification)	83.5 ($\uparrow$ 1.5%)
MAC-SC-2 (16-way classification)	84.8 ($\uparrow$ 2.8%)
MAC-SC-4 (256-way classification)	84.2 ($\uparrow$ 2.2%)
MAC-SC-4* (256-way classification)	**87.8** ($\uparrow$ 6.8%)

Table 2: Impact of multiple αs. Increasing number of α improves performance as long as there are enough frames in the subclip. MAC-SC-4* is pretrained with 64 frames (16 frames per α) and performs significantly better while 16-frame version suffers due to having 4 frames per α.

4.1 Implementation Details

Pretraining. We use a 3D CNN backbone and a single fully connected layer for self-supervised pretraining. For a fair comparison with previous works, we choose common backbones that have been extensively used in video representation learning. We utilize R(2+1)D-18 ([41]) for all ablation experiments and we also report downstream task results for R3D-18 ([20]). We apply standard augmentations such as *resize, crop, horizontal flip, temporal jitter, color jitter, grayscale* randomly. The number of different values that α can take $k = 4$ is set for all experiments. All models are pretrained for 300 epochs on UCF101 and 80 epochs on K400. We use SGD optimizer with momentum of 0.9 and weight decay of 1e-4. The initial learning rate is set as 0.01 which is decayed by factor 0.1 after each 100 epochs for UCF101, and after each 30 epochs for K400.

Downstream Tasks. For the finetuning stage, we transfer the weights of pretrained MAC models and use a randomly initalized fully connected layer for action recognition classification. We utilize R(2+1)D-18 and R3D backbones and report results for both. For video retrieval, we directly use pretrained models without any finetuning. Following practices in [52], we report Recall at k (R@k) results. If the top-k nearest neighbours include a video belonging to the same class of the query video, it is counted as a correct retrieval. All finetuning experiments are trained for 150 epochs. The initial learning rate is set as 0.01 which is decayed by factor 0.1 at 60th and 120th epoch.

Evaluation on Action Recognition. Following the common practice [52], we randomly sample 10 clips from test videos and average of these results are used for final evaluation.

4.2 Ablation Studies

In order to analyze the design choices of our proposed approach, we present various ablation studies. Unless otherwise stated, we use UCF101 dataset and sample clips of 16 frames with 112x112 resolution, report top-1 accuracy and utilize R(2+1)D-18 [41] backbone for all ablation experiments.

Contribution of Learning Objectives. As defined in Section 3.3, we combine self-supervised objective (MAC-S) with proposed contrastive objective (MAC-C). The Top-1 accuracy results on UCF-101 for optimizing each objective separately and jointly are given in Table 1.

Method	Evaluation	Pretrain	Backbone	Res.	UCF101	HMDB51
MoCo [8]	Linear	K400	R(2+1)D-18	112x16	67.4	39.8
FAME[12]	Linear	K400	R(2+1)D-18	112x16	72.2	42.2
MAC-SC-2	Linear	K400	R(2+1)D-18	112x16	73.6	43.1
Random Init.	Finetune	-	R(2+1)D-18	112x16	71.2	35.6
Supervised	Finetune	K400	R(2+1)D-18	112x16	89.9	63.5
TT[27]	Finetune	UCF101	R(2+1)D-18	128x16	81.6	46.4
FAME[12]	Finetune	UCF101	I3D-22	112x16	81.2	52.6
CtP[44]	Finetune	UCF101	R(2+1)D-18	112x16	86.2	57.1
MAC-SC-4	Finetune	UCF101	R(2+1)D-18	112x64	**87.8**	55.3
ASC-Net [24]	Finetune	K400	R3D-18	112x16	80.5	52.3
FAME[12]	Finetune	K400	R(2+1)D-18	112x16	84.8	53.5
CoCLR[19]	Finetune	K400	S3D	128x32	87.9	54.6
CtP[44]	Finetune	K400	R(2+1)D-18	112x16	88.4	61.7
MAC-SC-2	Finetune	K400	R(2+1)D-18	112x16	**87.1**	57.0
MAC-SC-4	Finetune	K400	R(2+1)D-18	112x64	**90.8**	58.5
TransRank[13]	Finetune	K400	R(2+1)D-18	112x64	90.7	**64.2**
BE [47]	Finetune	UCF101	R3D-34	224x16	83.4	53.7
STOR [56]	Finetune	K400	R2+1D-18	224x64	87.6	56.4
MotionFit[16]	Finetune	K400	S3D-G	224x64	90.1	50.6
ASC-Net [24]	Finetune	K400	S3D-G	224x64	90.8	60.5
CVRL [36]	Finetune	K400	R3D-50	224x32	92.2	66.7
BraVe [37]	Finetune	K400	R3D-50	224x64	93.7	72.0
ρBYOL [14]	Finetune	K400	R3D-50	224x16	95.5	73.6

Table 3: Comparison to prior and state-of-the-art methods for action recognition accuracy on UCF101 and HMDB51 datasets. We report both finetuning (No freeze for encoder) and linear probe (freeze encoder) results.

We also provide results of contrastive objective with only standard augmentations applied on videos at top row as a baseline. We observe that using MAC augmentation together with either self-supervised or contrastive objectives have gains nearly 12% over this baseline. More importantly, when our model is optimized with both objectives (Equation 3), i.e. MAC-SC-2, performance gain in accuracy increases another 3.3 points (from 81.5% to 84.8%). We conclude that both objectives operate cooperatively to complement each other in representation space.

Impact of using multiple αs per clip. In MAC-n variant, we assign different multipliers for distinct subclips of each clip as defined in Section 3.3. The results are shown in Table 2. We observe that the increase in the number of αs also increases the resulting action recognition performance in reasonable margin as long as there are enough frames for encoding. Key takeaway is that using multiple αs per clip enforces model to encode each subclip separately, increasing variance of temporal representations.

To analyze this behavior further, we investigate whether this performance gain results from a more challenging classification objective or treating the subclips separately. To understand this, we setup MAC-S-2 with $k = 4$ α values and assign two separate prediction heads for each α instead of a single prediction head. This reduces the possible number of α predictions from 16 to 4; a 2×4-way classification instead of 16-way classification. It can be seen from Table 2 that only %1.3 increase between MAC-SC-1 and MAC-SC-2 comes from classification complexity. Embedding 2 subclips of 8 frames with same classification complexity (MAC-SC-2 with 2 class heads) has %1.5 better than embedding 16 frames with only a single α (MAC-SC-1). This implies that model learns to embed each half subclips

Method	Pretrain Dataset	Accuracy
BE[47]	Diving-48	58.3
TE[26]	Diving-48	71.2
MAC-SC-2	Diving-48	72.3
BE[47]	UCF-101	58.8
FAME[12]	UCF-101	67.8
MAC-SC-2	UCF-101	69.2

Table 4: Comparison with state-of-the-art methods for Diving-48 dataset. All methods use I3D on same input size, only [26] use R2+1D-18.

	UCF101		HMDB51	
Method	Top-1	Top-5	Top-1	Top-5
VCP[33]	18.6	33.6	7.6	24.4
PRP[63]	22.8	38.5	8.2	25.8
PaceP[45]	19.9	36.2	8.2	24.2
BE[47]	-	-	11.9	31.3
CtP[44]	23.4	40.9	11.4	30.3
MAC-SC-2	**32.0**	**52.9**	**12.6**	**30.7**

Table 5: Comparison with state-of-the-art methods for video retrieval task on UCF101 and HMDB51 datasets.

separately to predict their corresponding blending multipliers.

4.3 Comparison To Prior Works on Downstream Tasks

Action Recognition on UCF101 and HMDB51. Table 3 compares MAC with previous state-of-the-art methods on action recognition task. We report top-1 accuracy results on UCF101 and HMDB51 datasets using common R2+1D-18 backbone. For linear probe, we add three fully connected layers and train only these layers while keeping the encoder frozen. We try to include best results of prior works in most comparable settings.

MAC performs superior for both datasets; especially on UCF101 it is on par with TransRank [13], state-of-the-art method for transformation-recognition based methods. Although both MAC and TransRank are successful at low resolution settings, TransRank uses less frames than ours. However, TransRank utilizes a two-stream network for ensembling RGB and RGBDiff modalities and it has to sample more clips per video with more than two times longer pretraining (80 vs 200 epoch).

We also include results of recent state-of-the-art instance discrimination methods with large scale experimental setups (shown in grey in Table 3) in terms of backbone, batch size, resolution and sampled clips per input. Our MAC performs not far from CVRL [36](%1.5 on UCF101) and comparable to BraVe [37] and ρBYOL [14] while being a much more simple method. As an example, CVRL use large batches (CVRL:1024, ours:16), heavy backbones (CVRL:50/152, ours:18 layers), more epochs (CVRL:800, ours:80) and higher resolution (CVRL:224, ours:112).

Action Recognition on Diving48-v2 dataset. Diving48 [32] is a challenging dataset that is specifically collected to be less biased against scene and appearance. Therefore, it favors models that encode motion patterns instead of relying on static scene context. We report Top-1 accuracy results in Table 4. Strong results on such a challenging dataset show that proposed MAC framework indeed can learn and generalize to motion patterns.

Video Retrieval. We evaluate our method on video retrieval benchmarks of UCF101 and HMDB51 in Table 5. We use R3D as backbone and UCF101 as pretraining dataset for the video retrieval experiments. Although our learning paradigm prioritizes motion information over scene information and the latter might be significantly useful in retrieval task, MAC still outperforms some recent successful methods.

4.4 Do we really focus on motion?

To show whether our learning actually focus on foreground motion, we perform an experiment by preparing a version of UCF101, that is more static and contains less temporal ac-

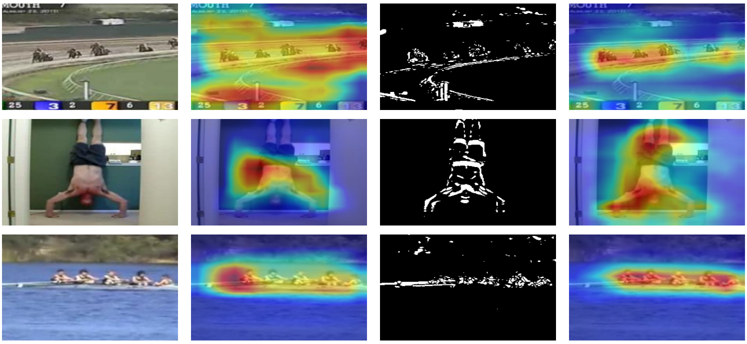
(a) Raw Frame (b) Supervised (c) MAC Mask (d) Ours

Figure 4: GradCAM [38] of models at finetuning stage.

tivity, dubbed as Still-UCF. We use only one third of each video in UCF101 to decrease the temporal variance, i.e. motion information. To make the sampled clip more static, we sample only 4 frames and use each frame four times repetitively. As shown in Figure 3, we observe that action recognition performances of K400 supervised model (↓ 4.1%) and MoCo pretrained model (↓ 3.2%) do not decrease as much as our MAC-SC-2 model (↓ 12.8%), which suggests that our model is more dependent on motion features.

4.5 Visual Analysis.

To provide qualitative analysis and reinforce the claims on motion-based learning, we present GradCAM activation maps of our pretrained encoder together with raw frames and MAC masks from three example videos in Figure 4.

We observe that our method is more focused on action regions than the supervised ((b) in Figure 4) pretrained model. Notably, at the first row, our method can attend running horses while supervised method picks up clues from different regions such as racetrack. Please refer to Supplementary Material for more detailed visual and failure case analysis.

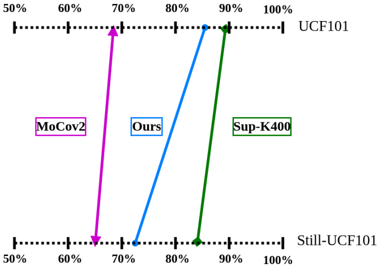

Figure 3: Comparison of finetuning results on UCF and Still-UCF datasets.

5 Conclusion

This paper introduces a novel mask-augmentation technique with corresponding self-supervised and contrastive objectives for learning rich representations from videos. Our work is motivated by background bias problem that might be mitigated by focusing more closely on motion information. In particular, our method aims to learn invariance to background while extracting distinctive features for temporal variants with the help of proposed mask-augmentation tecnique and corresponding learning objectives. Experimental results show that our approach achieves the state-of-the-art results in low resource settings and is comparable with methods that require large-scale computing resources.

6 Acknowledgement

The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

References

[1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[2] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9253, 2021.

[3] 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.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

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

[6] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI Conference on Artificial Intelligence*, volume 1, 2021.

[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.

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

[9] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised visual learning by variable playback speeds prediction of a video. *IEEE Access*, 2021.

[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.

[11] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. *arXiv*, pages 6192–6201, 2019. ISSN 23318422.

[12] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, and Jue Wang. Motion-aware self-supervised video representation learning via foreground-background merging. *arXiv preprint arXiv:2109.15130*, 2021.

[13] Haodong Duan, Nanxuan Zhao, Kai Chen, and Dahua Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3000–3010, 2022.

[14] 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.

[15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *ArXiv*, abs/2205.09113, 2022.

[16] Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees GM Snoek. Motion-augmented self-training for video recognition at smaller scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10429–10438, 2021.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[18] 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.

[19] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020.

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[21] 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.

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[23] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021.

[24] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8096–8105, 2021.

[25] 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.

[26] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9970–9980, 2021.

[27] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 425–442. Springer, 2020.

[28] Li Jing, Jiachen Zhu, and Yann LeCun. Masked siamese convnets. *ArXiv*, abs/2206.07700, 2022.

[29] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv: Computer Vision and Pattern Recognition*, 2018.

[30] Dahun Kim, Donghyeon Cho, and In-So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019.

[31] 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.

[32] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[33] 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.

[34] 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.

[35] 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.

[36] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[37] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[39] 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.

[40] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Pretext-contrastive learning: Toward good practices in self-supervised video representation leaning. *arXiv preprint arXiv:2010.15464*, 2020.

[41] 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.

[42] Sergey Tulyakov, Ming Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. *arXiv*, 2017. ISSN 23318422.

[43] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv*, pages 1–22, 2017. ISSN 23318422.

[44] Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, and Zhiwei Xiong. Unsupervised visual representation learning by tracking patches in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2563–2572, 2021.

[45] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020.

[46] Jinpeng Wang, Yuting Gao, Ke Li, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. In *AAAI*, 2021.

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11804–11813, June 2021.

[48] 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.

[49] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[50] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. Pfinder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997. doi: 10.1109/34.598236.

[51] Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. *arXiv preprint arXiv:2105.06463*, 2021.

[52] 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.

[53] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. *arXiv preprint arXiv:2103.17263*, 2021.

[54] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020.

[55] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6547–6556, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00658.

[56] Yujia Zhang, Lai-Man Po, Xuyuan Xu, Mengyang Liu, Yexin Wang, Weifeng Ou, Yuzhi Zhao, and Wing-Yin Yu. Contrastive spatio-temporal pretext learning for self-supervised video representation. *arXiv preprint arXiv:2112.08913*, 2021.