Supplementary material for: 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

1 Overview

In this supplementary work, we provide additional ablation experiments and demonstrative qualitative results for in-depth analysis of the proposed framework. We also provide source codes that include all the models and configurations described in main paper to help future research and conform with reproducibility principles. All experiments are done on UCF101 dataset with MAC-SC-2 model, using R(2+1)D backbone over 112×112 spatial resolution and 16 input frames unless otherwise stated.

2 Mask Extraction Algorithm

A simple idea is to use a momentum structure to keep the track of background history. 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 in Algorithm 1.

Algorithm 1 Foreground Mask Extraction

```
for each frame F of clip C:
  I(t) = F;
  diff = abs[BG(t-1) - I(t)]
  FG_Mask(t) = threshold(diff, lambda)
  BG(t) = (1-m) * I(t) + m * BG(t-1)
end
```

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

3 Impact of Augmentations

To examine the effect of each data augmentation technique, we provide ablation results for color jitter, greyscale and MAC in Table 2. We present results for MAC version with only self-supervised objective (MAC-S-2) in order to observe possible degenerate solutions when lack of spatial augmentations. Setting up MAC augmentation without any spatial augmentation leads to degenarate representations that do not translate to downstream tasks well. Applying color jitter increase gains for an additional 0.9% and grey scale augmentation helps for another 2.3%. We conclude that all augmentation techniques improve model invariance for spatial transformations and prevent model from finding trivial solutions during pretraining.

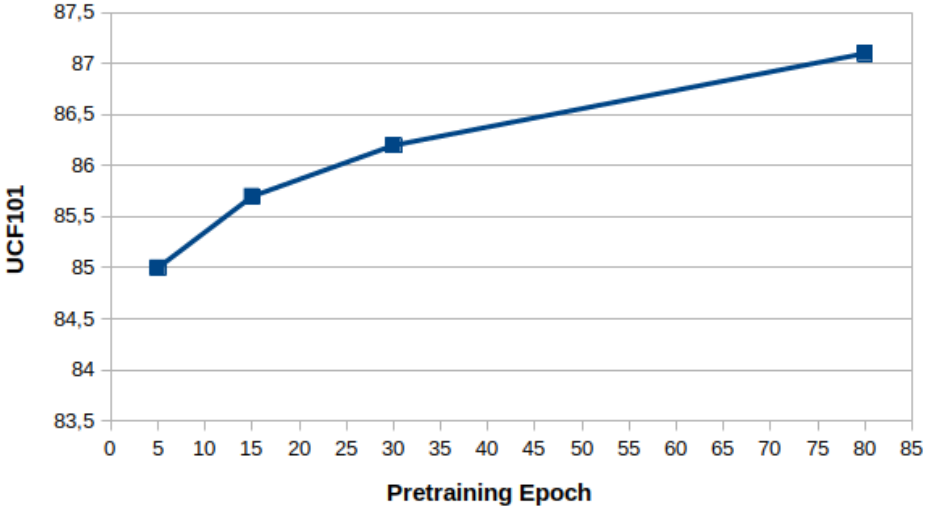


Figure 1: Pretraining Efficiency. Finetuning results on UCF-101 are given for each pretrained model. Models are picked at different epoch of self supervised pretraining stage. In this case, models at epoch 5, 15, 30, 80 of pretraining stage are selected for finetuning.

4 Efficiency

We observe that our model is able to converge quite fast, and only a few epochs of pretraining is enough for superior transfer learning performance. To demonstrate this, we pretrain our model on K400 dataset, and pick checkpoints from epoch 5, 15, 30 and 80. We then finetune these models on UCF101 dataset to evaluate on action recognition and report results in Figure 1. Our model which is pretrained for 5 epochs gives remarkable results, yielding a boost up to ↑ 14% from random initialized model. We think that this is also a valuable aspect and the fast convergence at even early-phase pretraining models can provide efficiency for downstream tasks. Given that ρBYOL achieves 84.6%[7] for UCF101 after 50 epochs with heavy training, ours outperforms by solely utilizing 5 epochs (85%). That implies that the proposed method might be a strong alternative against large-scale video representation

learning methods that mostly benefits longer pretraining.

Method	UCF101	HMDB51
PacePred [8]	66.3	32.2
VCOP[9]	72.4	30.9
VCP[5]	77.1	36.6
HDC[11]	76.8	40.0
MAC-SC-2	84.6	53.2

Table 1: R2+1D-10 [7] results which is far smaller backbone than R2+1D-18

We additionally demonstrate the effectiveness of our approach over a small-scale and shallower backbone in Table 1. The proposed model pre-trained on UCF-101 outperforms many existing models by a large margin for both datasets.

5 Impact of Source of Foreground Mask

In order to analyze the impact of foreground mask, we design a small-scale experiment. Here, we utilize a subset of MiniKinetics (13k training, 1.7k test videos of 100 classes of actions) dataset and compare the cases where the foreground mask is acquired via our simple frame-differences based framework vs. human detections introduced in [1] that are extracted with Mask R-CNN [4]. Namely, we pretrain our MAC model using foreground detections extracted from off-the-shelf person detector instead of frame difference based masks and compare it to our pretrained model. The results are given in Table 3. Surprisingly, frame-difference based imperfect masks outperforms an off-the-shelf person detector by a margin. We speculate that although the detector leverages prior knowledge, our instantly evolving masks captures the moving foreground behavior better leading to better video.

6 Effect of Training Objective

We also experiment with using MSE(Mean Squared Error) objective instead of cross entropy loss. In Table 2 MSE Regression refers to regressing preset α values, whereas MSE-n regression refers to regressing uniformly sampled α values from (0,1]. As can be seen from the results in Table 4, the performance of the model trained with MSE loss is rather inferior than the one with cross-entropy objective. This might be due to the observation stated in

Augmentation	Accuracy
MAC-S-2	76.6
MAC-S-2 + CJ (↑ 4.6%)	81.2
MAC-S-2 + CJ + GS (↑ 1.3%)	82.5

Table 2: The effect of additional augmentations on UCF101 dataset. CJ stands for Color Jitter, and GS is Grayscaling.

Source	Accuracy
B.Box ([4])	82.4
Ours	83.5

Table 3: Impact of Mask Source

Method	Backbone	Accuracy
MAC-SC-2	R3D	78
MAC-SC-2	R(2+1)D-18	84.8
MAC-SC-4	R3D	86.4
MAC-SC-4	R(2+1)D-18	87.8

Table 5: Impact of Backbone

# of classes	Accuracy
2	81.5
3	82.6
4	84.8
5	84.1
6	84.2

Table 6: Impact of # of α classes

m	Accuracy
0.1	82.6
0.5	84.8
1	83.4

Table 7: Impact of momentum parameter

λ	Accuracy
10	81.1
30	84.8
60	83.9

Table 8: Impact of threshold parameter

Colorful Image Colorization [10] that MSE loss is not robust to problems with multi-modal nature, i.e. temporal variance of pixel color space. In unconstrained conditions, the optimal case for MSE regression objective would reduce to mean of all possible values.

7 Effect of Mask Extraction Parameters

We demonstrate ablation experiments for mask extraction parameters; threshold parameter λ and momentum parameter m that is defined in Algorithm 1 for the edge cases. The corresponding evaluations are presented in Table 7 and Table 8. Recall that obtaining perfectly realistic foreground masks is not in the scope of this study. We observe the impact of setting both parameters at extreme values by simply comparing the performance of downstream action recognition task. It should be noted that $m = 0.1$ converges to pure frame difference while $m = 1$ corresponds to the case of calculating moving average of previous frame differences. Similarly, $\lambda = 10$ is an extreme low threshold with high response to differences and $\lambda = 60$ is a rather high threshold with low response to differences. Without doing a thorough search in parameter space, we pick average values for $\lambda = 30$ and $m = 0.5$ for the rest of the experiments.

Loss Function	Accuracy
MSE Regression Loss	81.5
MSE-n Regression Loss	82.0
Cross-Entropy Loss	84.8

Table 4: The impact of loss functions.

8 Impact of granularity of α

Here, we present the impact of granularity of α for cases of k from $\{2,3,4,5,6\}$. There are k different values that α can take and here we analyze the impact of this value of k. We analyze cases where α is sampled from the list of values $A = \{(a_i)_{k-1}, 1\}$, where (a_i) designates centers of clusters formed by splitting (0,1) into $k-1$ clusters uniformly, recalling

that $\alpha = 1$ is always retained to represent natural form of augmented clips. As can be seen from Table 6, best performance is achieved when $k = 4$ that is slightly superior to $k = 5$ case. When $k = 3$, the number of possible cases for classification objective becomes $k^n = 3^2$ which is a fairly easy task than for $k = 4$ ($k^n = 4^2$).

9 Impact of Backbone

The impact of using different backbone models is given in Table 5. We experiment with two different 3D CNN networks that are widely used in video representation learning. We observe that R(2+1)D-18 [7] model yields significantly superior results than R3D-18 [3] model. Both networks have the same number of parameters and depth, on the other hand, R(2+1)D-18 model decomposes 3D convolutions into 2D spatial and 1D temporal convolutions. Since our MAC relies on temporal information for motion augmentation and more responsive to temporal occurrences, we conclude that performance gap between two backbones might have resulted from R(2+1)D-18 model's handling temporal relations better than R3D-18 model.

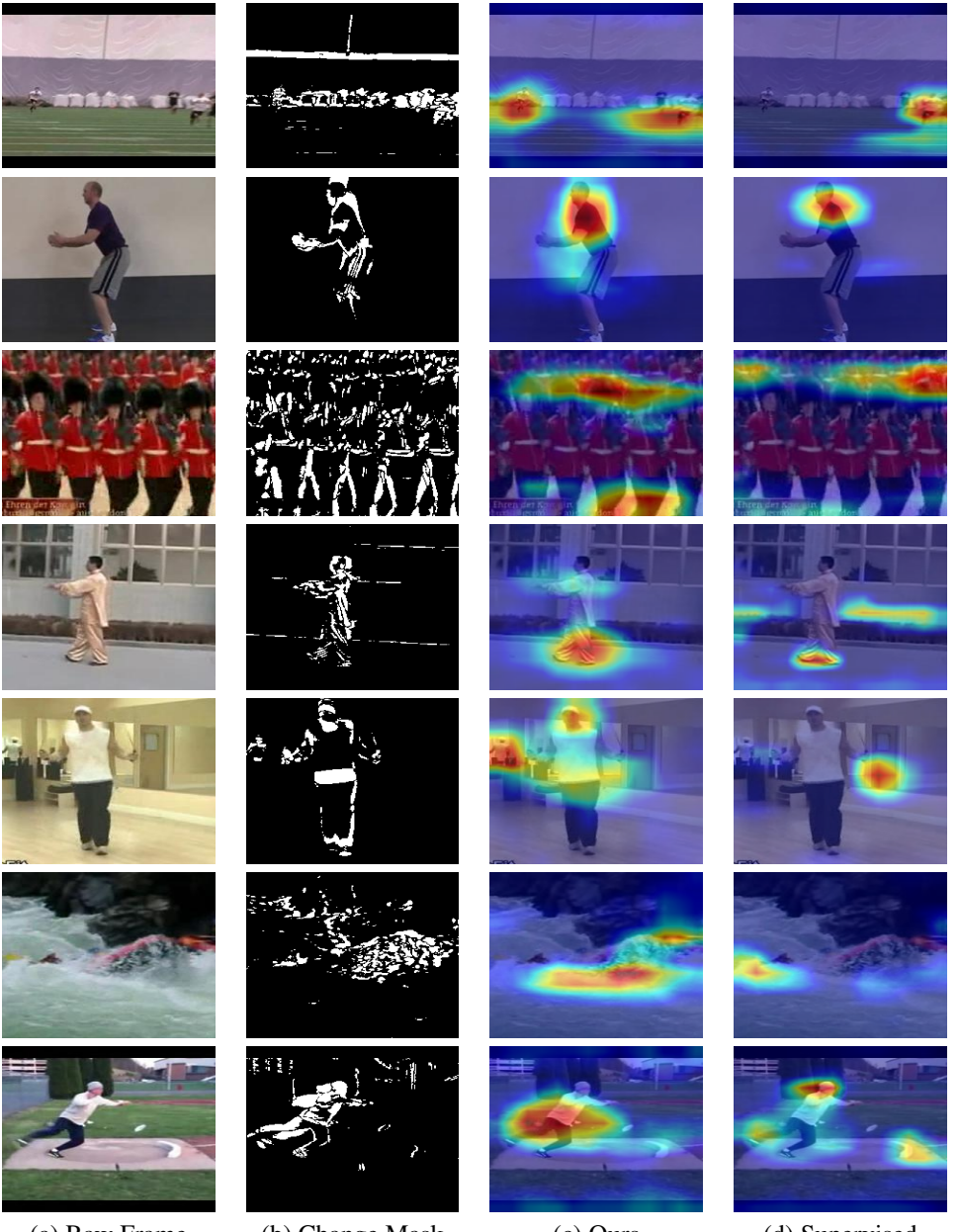
10 Qualitative Analysis

In this section, we demonstrate class activation maps (CAM) visualizations, corresponding change masks and raw frames for samples both K400 supervised model and MAC-SC-2 model incorrectly classify and for cases only one of them incorrectly classifies.

Samples correctly classified by MAC but failed by K400 supervised model is given in Figure 2. K400 supervised model apparently either loses focus from dynamic regions or pick false clues from background. Surprisingly, our model can even pick motion in the mirror in one particular case while K400 supervised model falsely attends background. These visualizations encourage that MAC is more robust to background clues while being more focused on motion.

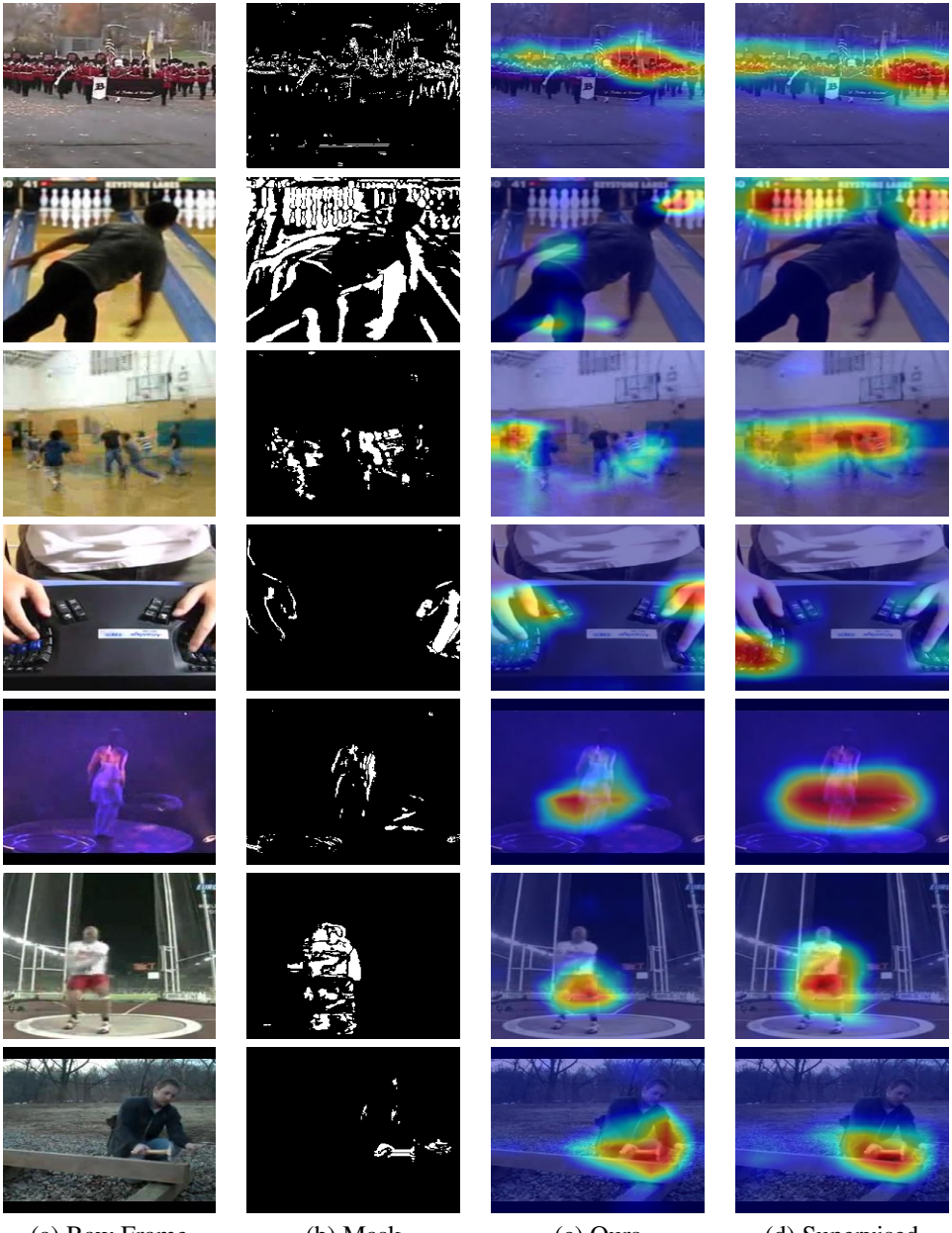
Samples correctly classified by K400 supervised model but failed by MAC is given in Figure 3. It can be observed that although MAC fails due to losing its focus on dynamic regions for some cases, it sometimes still fails even with a good focus on correct dynamic regions. It is not straight forward to evaluate these failure cases with only qualitative analysis but possible reasons might be related to inter-class variances between similar categories. On the other hand, K400 supervised model seems to attend on background more, picking up useful clues for some cases.

Failure cases both models incorrectly classify are demonstrated in Figure 4. Surprisingly, for some cases both models seem to focus on correct regions by either attending on dynamic parts or spatially related part of the scene but still fail. Although reasons of failure on both sides are not evident, there might be incorrect context information (such as nunchucks in front of a bridge at fourth row in Figure 4) or lack of distinguishing motion pattern (sixth row in Figure 4). For others, lack of visual appearance information due to occlusion, illumination variance and low resolution/bad quality imaging might be potential reasons of failure.



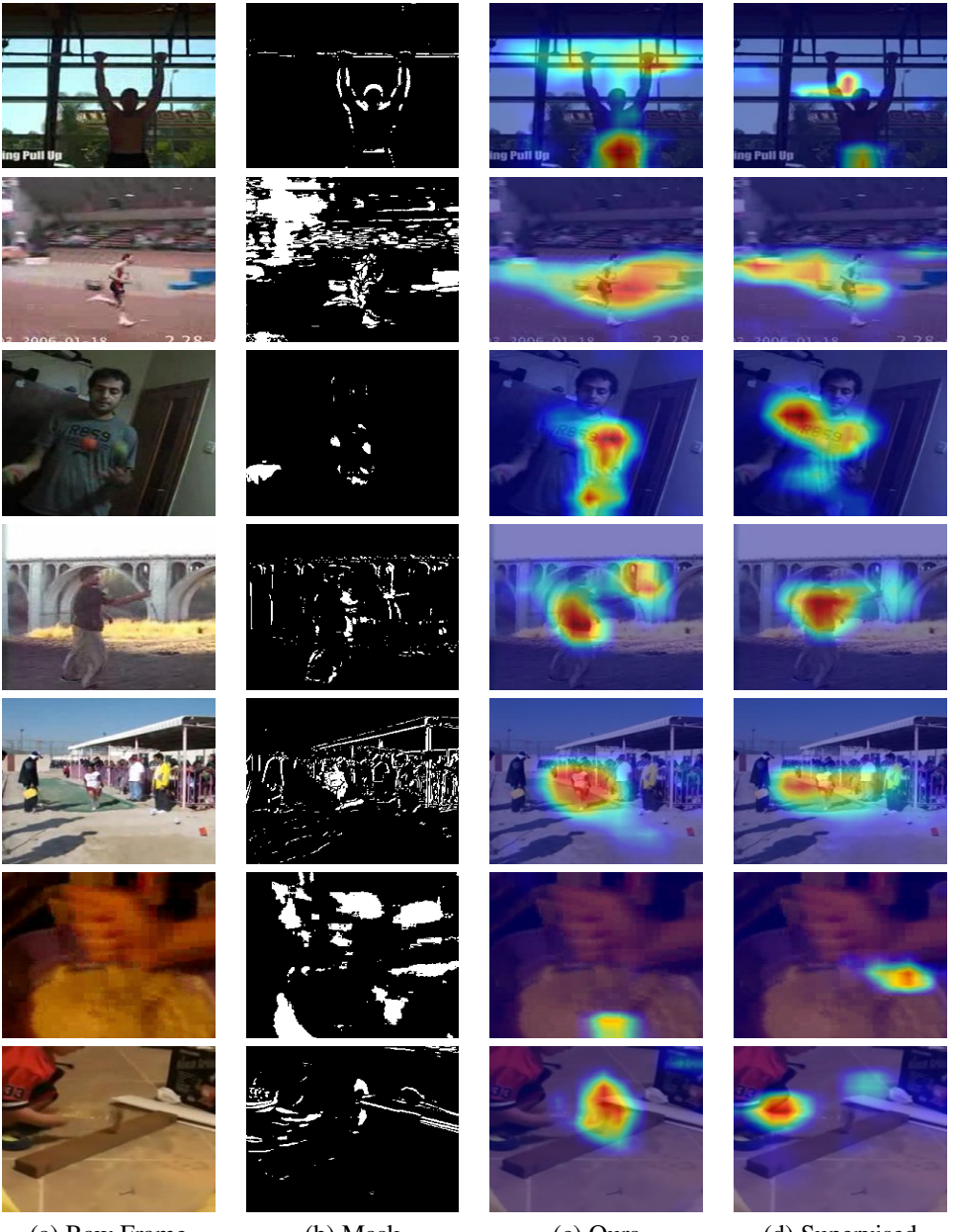
(a) Raw Frame (b) Change Mask (c) Ours (d) Supervised

Figure 2: Examples for correctly classified cases by MAC-SC-2 and incorrectly classified cases by Kinetics-400 supervised model from UCF101 dataset. GradCAM [6] activation maps of models at finetuning stage. (a) raw video frames, (b) change masks, (c) GradCAM maps of MAC-SC-2, and (d) GradCAM maps of K400 supervised pretrained model.



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

Figure 3: Examples for correctly classified cases by Kinetics-400 supervised model and incorrectly classified cases by MAC-SC-2 from UCF101 dataset. GradCAM [6] activation maps of models at finetuning stage. (a) raw video frames, (b) change masks, (c) GradCAM maps of MAC-SC-2, and (d) GradCAM maps of K400 supervised pretrained model.



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

Figure 4: Examples for incorrectly classified cases by Kinetics-400 supervised model and MAC-SC-2 from UCF101 dataset. GradCAM [6] activation maps of models at finetuning stage. (a) raw video frames, (b) change masks, (c) GradCAM maps of MAC-SC-2, and (d) GradCAM maps of K400 supervised pretrained model.

References

[1] 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.

[2] 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.

[3] 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.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[11] Zehua Zhang and David Crandall. Hierarchically decoupled spatial-temporal contrast for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3235–3245, 2022.