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

Arif Akar, Ufuk Umut Senturk, Nazli Ikizler-Cinbis

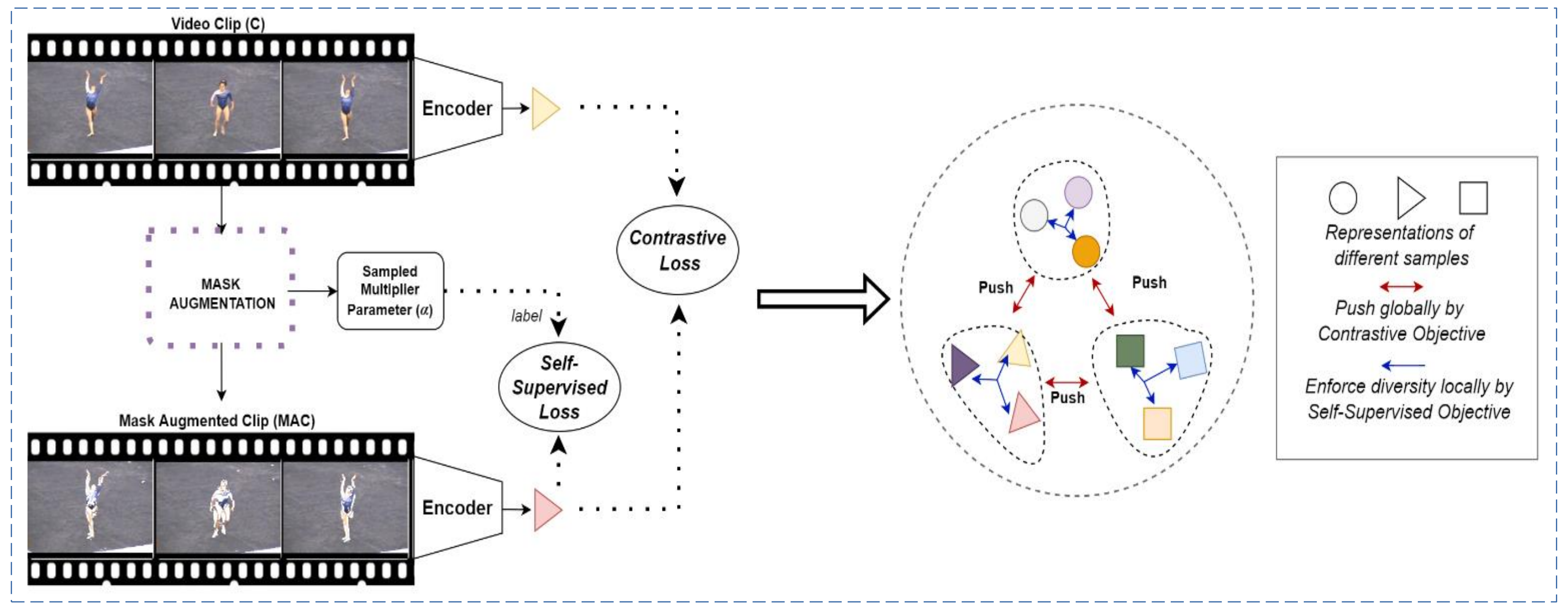
Problem Description

- Learning video representations based on actor/motion is critical for motion-related downstream tasks.
- Video redundancy might lead to background-bias.

Motivation:

- Utilize motion information as a form of data augmentation step, which potentially removes background reliance.
- Leverage pretext-based self-supervised learning with a *transformation-recognition* approach.

Approach


Video Clip (C)
Encoder
MASK AUGMENTATION
Sampled Multiplier Parameter (α)
label
Contrastive Loss
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

- **Self-Supervised (Pretext) Objective:** Predict applied mask Augmentation
- **Contrastive Objective:** Apply MAC on query to have similarity relation over moving foreground

Mask Extraction

- A momentum structure used to keep the track of background history, instead of taking direct frame differences.
- A very simple background modelling: moving average of recent frames.

Algorithm 1 Foreground Mask Extraction

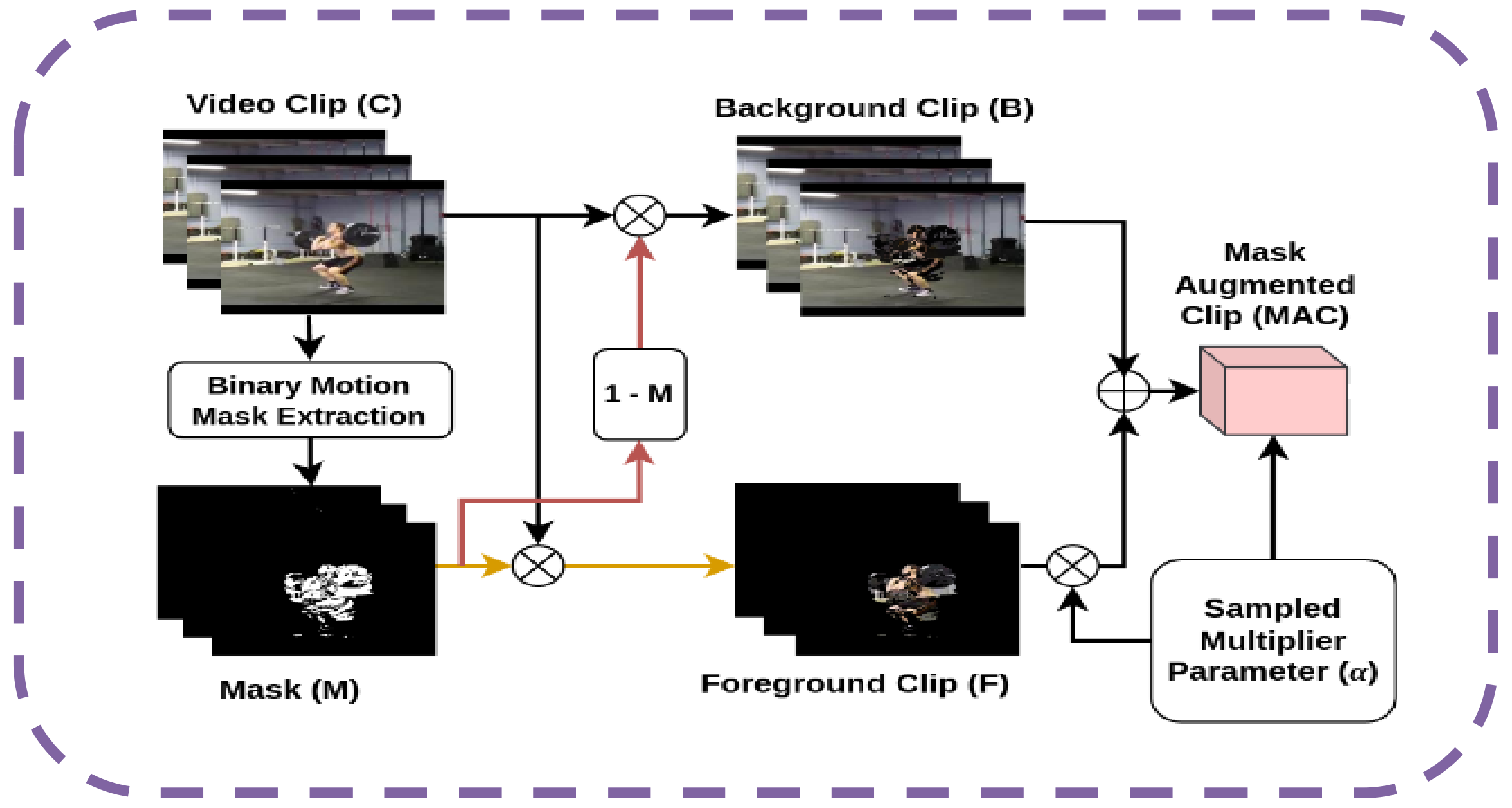
```
for each frame F of clip C:
  I(t) = F;
  diff = abs[BG(t-1) - I(t)]
  FG_Mask(t) = threshold(diff, lambda)
  BG(t) = (1-m) * I(t) + m * BG(t-1)
end
```

Method

Foreground and background regions blended using foreground masks

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

where (1 − M) is the inverse of M and ⊙ is an element-wise product


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

$$L_{\text{Self−Supervised}} = \sum_{i=1}^{N} \alpha_i \log(\widehat{\alpha}_i)$$

α is the index of randomly sampled multiplier parameter in range (0, 1]

$$L_{\text{Contrastive}} = -\log \frac{\exp(\frac{q\,.\,k^{+}}{\tau})}{\sum_{i=0}^{K} \exp(\frac{q\,.\,k^{-}}{\tau})}$$

Query clip is q and key clip with only basic augmentations applied is denoted as k^{+}. Negative samples(k^{-}) are coming from other videos that have been added to queue.

$$L_{\text{Total}} = \lambda * L_{\text{Self−Supervised}} + \beta * L_{\text{Contrastive}}$$

Experimental Results

Action Recognition Results

Method	Pretrain	Backbone	Res.	UCF101	HMDB51
TT [27]	UCF101	R(2+1)D-18	128x16	81.6	46.4
CtP [44]	UCF101	R(2+1)D-18	112x16	86.2	**57.1**
Ours	UCF101	R(2+1)D-18	112x64	**87.8**	55.3
FAME [12]	K400	R(2+1)D-18	112x16	84.8	52.3
CtP [44]	K400	R(2+1)D-18	112x16	88.4	61.7
CoCLR [19]	K400	S3D	128x38	87.9	54.6
Ours	K400	R(2+1)D-18	112x16	87.1	57.0
Ours	K400	R(2+1)D-18	112x64	**90.8**	58.5
TransRank [13]	K400	R(2+1)D-18	112x64	90.7	**64.2**
Brave [37]	K400	R3D-50	224x64	93.7	72.0
BYOL [14]	K400	R3D-50	224x16	95.5	73.6

Video Retrieval Results

	UCF101		HMDB51	
Method	Top-1	Top-5	Top-1	Top-5
VCP [33]	18.6	33.6	7.6	24.4
PRP [55]	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
Ours	**32**	**52.9**	**12.6**	**30.7**

Action Recognition results on Diving-48 v2 dataset

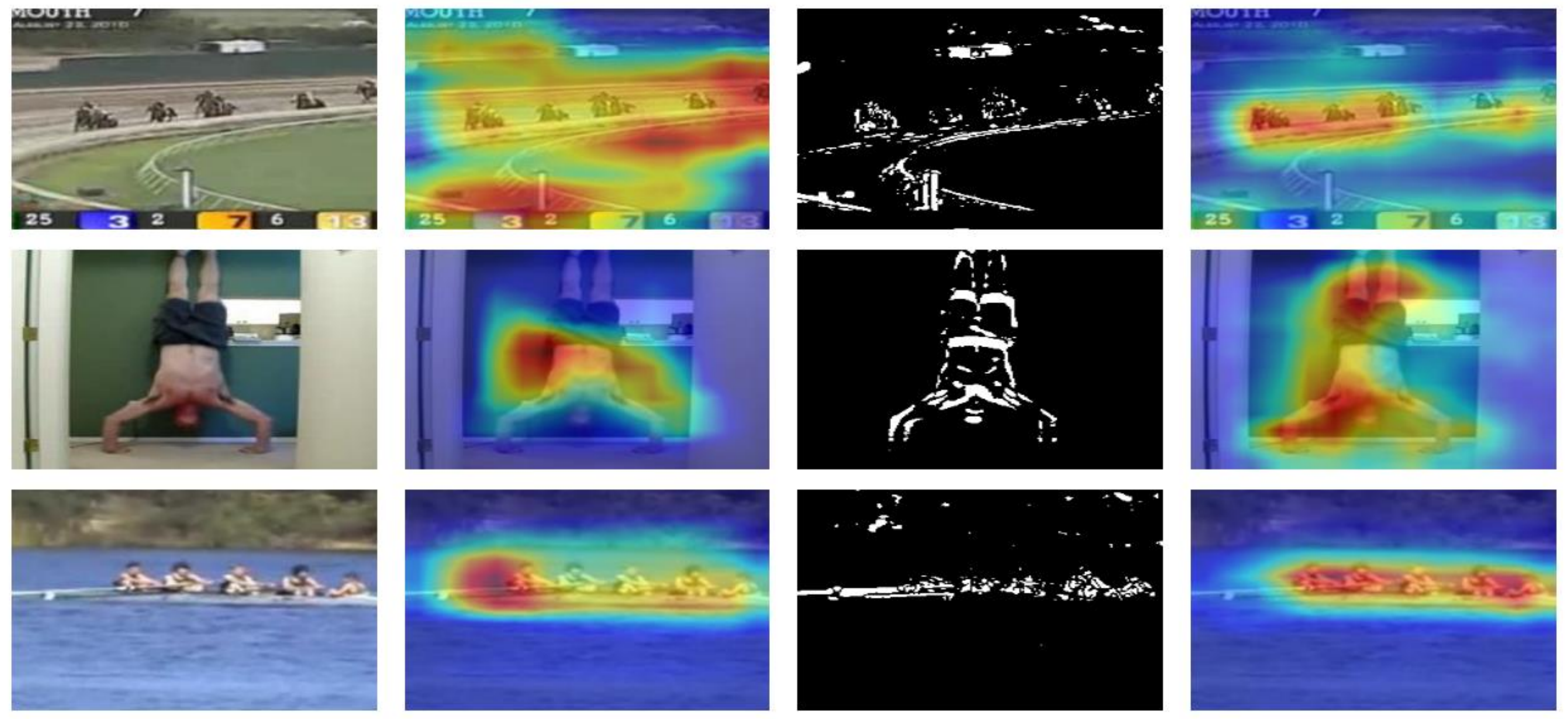
Method	Accuracy
BE [47]	58.8
FAME [12]	67.8
Ours	**69.2**

Experimental Results

Results for different choice of number of multipliers. 2 means input clip is split into two subclips and different multipliers are sampled and predicted for each subclip.

# of multiplier	UCF101 Acc.
MAC-SC-1 (4-way clas.)	82
MAC-SC-2 (2 x 4-way clas.)	83.5
MAC-SC-2 (16-way clas.)	84.8
MAC-SC-4 (256-way clas.)	84.2
MAC-SC-4* (256-way clas.)	**87.8**

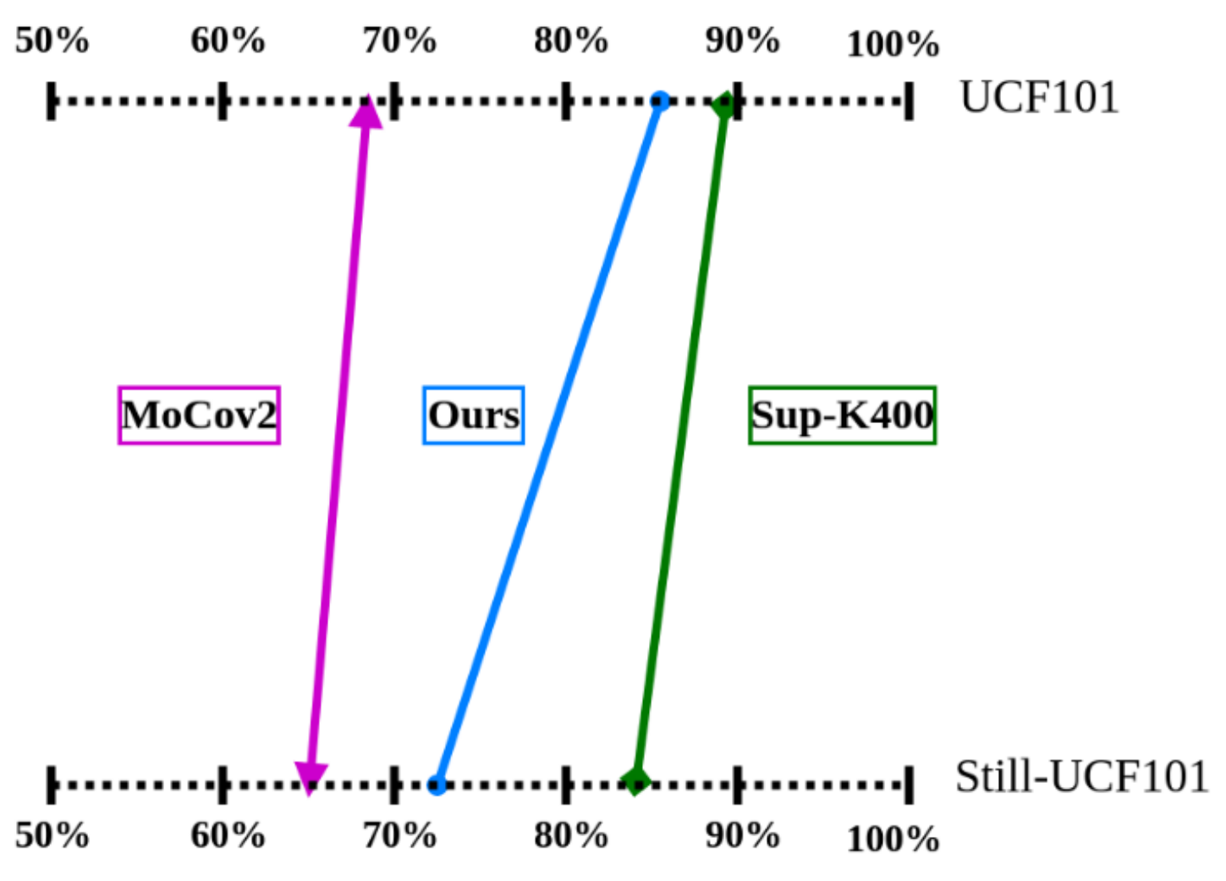
GradCAM Visualizations



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

Mask Extraction

- A more static version of UCF101 with less temporal activity to compare MAC with pretrained Supervised K400
- Only one third of each video used
- Sampled only 4 frames and use each frame four times repetitively



Contributions

- Simple and effective mask augmentation technique (MAC) based on frame differences.
- A novel self-supervised objective, denoted as MAC-S, based on predicting the largely imperfect foreground masks.
- A novel contrastive objective, denoted as MAC-C, describing positive pairs via MAC augmentation.
- Learning video representations with background-invariance and spatio-temporal equivariance by exploiting transformation-recognition paradigm.