Supplementary Material for Action-aware Masking Network with Group-based Attention for Temporal Action Localization

Tae-Kyung Kang[1], Gun-Hee Lee[2], Kyung-Min Jin[1], and Seong-Whan Lee[1]

[1]Dept. of Artificial Intelligence, Korea University, South Korea
[2]Dept. of Computer Science and Engineering, Korea University, South Korea
{tk_kang, gunhlee, km_jin, sw.lee}@korea.ac.kr

In the supplementary material, we conduct additional ablation studies (Section 1) and provide our qualitative results (Section 2), as shown below.

1. Additional Ablation Study

Experiment settings We use a single GPU (NVIDIA TITAN Xp) for all additional ablation studies and provide training and inference time for each experiment. Our overall experiments are conducted on THUMOS14 [2] dataset, and the other settings are the same as those in the main paper.

Feature embedding dimension Our method adopts the 2048-dimensional video features extracted by I3D [1], then projects the features to a specific embedding dimension. We conduct the ablation study to verify performance and the speeds (training and inference) according to each embedding dimension, as shown in Tab. 1. The performance with 128-dimensional embedding is the lowest because it is not enough to represent the video features. On the other hand, the 256-dimensional embedding shows competitive performance compared to the 512-dimensional embedding. However, the gap in speeds of training and inference between 256 and 512 dimensions is not much, so the 512 dimension is the most reasonable choice. When we trained our model with 1024-dimensional embedding features, out-of-memory occurred.

Encoder layer and multi-heads Tab. 2 shows the results of the different number of encoder layers and multi-heads where the encoder denotes self-attention and cross-attention. In our environment, out-of-memory occurs when the number of multi-heads exceeds 8. Therefore, we set the attention modules to two encoder layers with four multi-heads, maximizing the detection performance.

2. Qualitative Results

Visualization of action-aware attention process We provide the visualization of action-aware attention processes in Fig. 1. Here, the video features are not salient around human activities, weakening localization performance improvement. On the other hand, the masked features are more salient around human activities by applying action-aware attention, which help to utilize semantic temporal knowledge.

Comparison our results with AFSD To provide intuitive comparisons, we visualize our detection results on THUMOS14, as shown in Fig. 2. Here, we compare the results to an end-to-end TAL framework called AFSD [3]. As mentioned in the main paper, our method refines the offline extracted video features by applying action-aware and group-based attention. In contrast, AFSD extracts the video feature by fine-tuning the video encoder I3D with pre-trained weights. The videos in Fig. 2 consist of many similar frames, which makes it hard to localize the actions due to the ambiguity between consecutive frames. Despite that, these results show that our model predicts the temporal action boundaries and classes more precisely by refining the video features than AFSD trained by end-to-end learning.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[2] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[3] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, 2021.

Embedding dim.	THUMOS14				
	0.5	0.7	Avg.	Training time	Inference time
128	60.7	34.2	56.8	**54m 11s**	**29.38s**
256	66.2	40.6	62.2	60m 43s	32.88s
512	**66.8**	**42.7**	**63.3**	76m 2s	45.72s
1024	Out-of-memory				

Table 1. Ablation study of different feature embedding dimensions on THUMOS14 dataset. The results are measured by mAP (%) at different tIoU thresholds.

# Layers	# Heads	THUMOS14				
		0.5	0.7	Avg.	Training time	Inference time
1	2	66.2	42.2	63.2	**60m 43s**	**31.65s**
1	4	65.7	40.6	62.0	63m 41s	33.84s
2	4	**66.8**	**42.7**	**63.3**	64m 47s	34.73s
2	8	Out-of-memory				

Table 2. Ablation study of different the numbers of encoder layers and multi-heads on THUMOS14 dataset. The results are measured by mAP (%) at different tIoU thresholds.

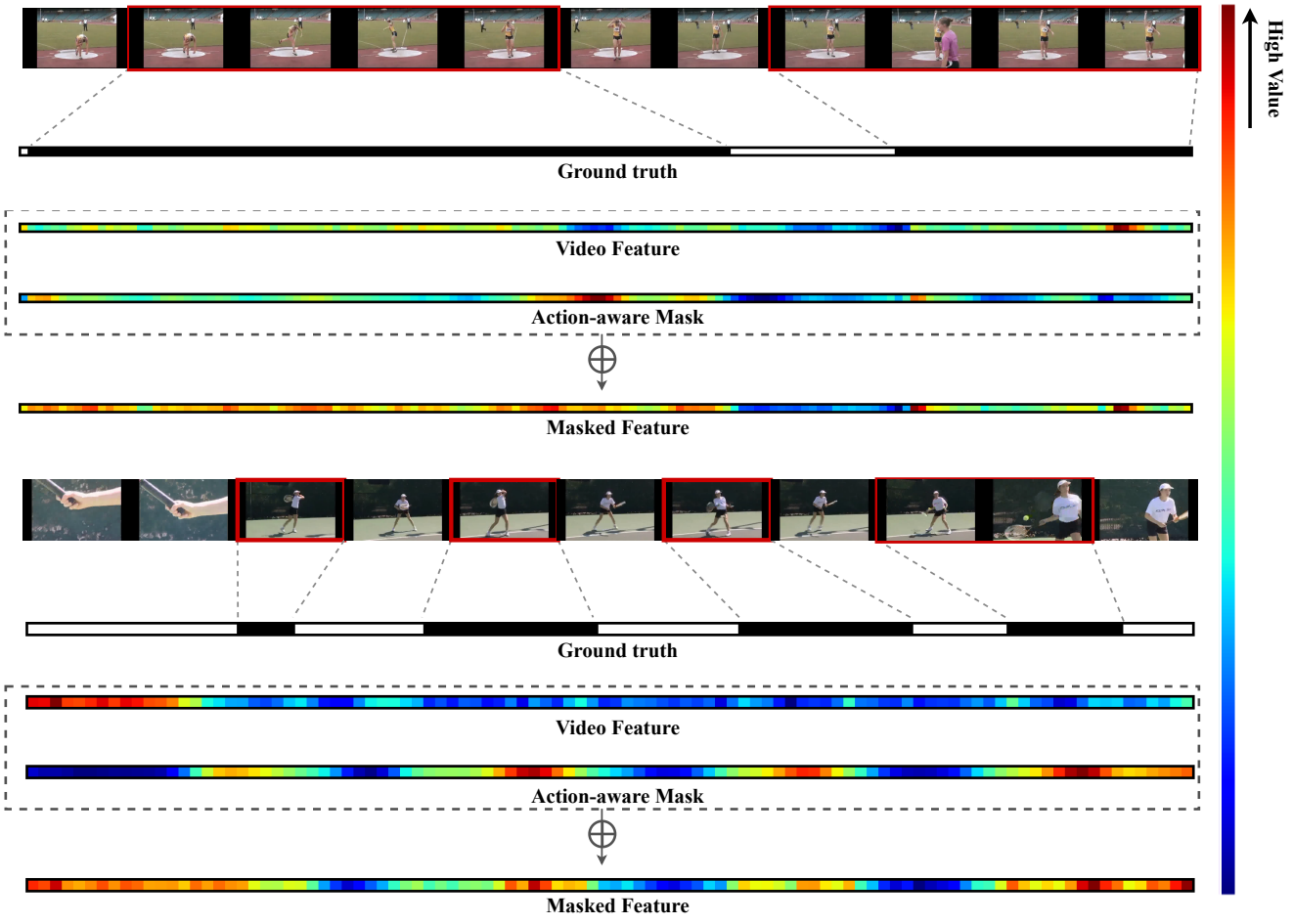

Figure 1. Visualization of the action-aware attention processes. In the figure, the red boxes denote ground truths, and each feature denotes a mean of channels. The action-aware masks make the video feature more salient around actions.

Figure 2. Qualitative results of our method on THUMOS14, compared to AFSD [3]. In this figure, our model utilizes the action frames (red box) as positive components and the background frames (blue box) as negative components in the Action Masking Encoder (AME).