Supplementary Materials for Motion Aware Self-Supervision for Generic Event Boundary Detection

Ayush K. Rai, Tarun Krishna, Julia Dietlmeier, Kevin McGuinness*, Alan F. Smeaton*, Noel E. O'Connor*
Insight SFI Centre for Data Analytics, Dublin City University (DCU)
ayush.rai3@mail.dcu.ie

1. Additional Details

1.1. Implementation Details

Stage 2: Finetuning. Input to the encoder is based on the temporal window $W = 5$ and stride $m = 3$. (W, m) can be thought of as hyper-parameter, setting a larger value of each might introduce noise information when two different boundaries lie close to each other, a smaller value might be unable to capture the necessary context information for a boundary. Among the 5 annotations available for Kinetics-GEBD for every video, the ones with highest annotator F1 consistency score is used for fine-tuning.

1.2. Architectural Design Choice

Temporal Shift Module (TSM) [3] is inserted in every residual block of ResNet50 encoder. *MotionSqueeze* module is added after the `conv3_x` layer of the ResNet50 encoder.

Table 1. Modified ResNet50 Encoder

Layers	ResNet-50	Modified ResNet-50	Output size
conv1	$7 \times 7, 64$, stride 2		112×112
	3×3, max-pool, stride 2		
conv2_x	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{TSM} \\ 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	56×56
conv3_x	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{TSM} \\ 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 128 \end{bmatrix} \times 4$	28×28
MS Module	✗	✓	28×28
conv4_x	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{TSM} \\ 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	14×14
conv5_x	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{TSM} \\ 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	7×7

It should be noted that our encoder definition is consistent with the architecture design introduced in ResNet [1] and is different from the encoder in the work of *MotionSqueeze* in [2] as shown in Table 1.

1.3. Evaluation Protocol.

We conduct evaluation on two datasets Kinetics-GEBD [5] and TAPOS [4]. For evaluation, we follow the standard evaluation protocol explained in [5], which uses the F1 score as the measurement metric. *Rel. Dis* (Relative Distance) is used to decide whether a detected event boundary is correct (if detection probability ≥ 0.5) or otherwise incorrect. More formally, *Rel. Dis* is defined as the error between detected and ground-truth timestamps, divided by the length of the whole video. F1 score calculated at *Rel. Dis* threshold 0.05 was used as the evaluation metric for the GEBD challenge[1]. We compare our detection results with all annotations (5 annotations per video for Kinetics-GEBD and 1 annotation for TAPOS) in the same video and select the annotation with the highest F1 score.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *European Conference on Computer Vision*, pages 345–362. Springer, 2020.

[3] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[4] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

*Equal supervision

[1] LOVEU@CVPR2021, LOVEU@CVPR2022

[5] Mike Zheng Shou, Stan W Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. *arXiv preprint arXiv:2101.10511*, 2021.