Boundary-sensitive Pre-training for Temporal Localization in Videos – Supplementary Material –

Mengmeng Xu[1,2*] Juan-Manuel Pérez-Rúa[3] Victor Escorcia[1] Brais Martinez[1]
Xiatian Zhu[1] Li Zhang[4] Bernard Ghanem[2] Tao Xiang[5]
[1] Samsung AI Centre Cambridge, UK [2] King Abdullah University of Science and Technology, Saudi Arabia
[3] Facebook AI, UK [4] School of Data Science, Fudan University, China [5] University of Surrey, UK
mengmeng.xu@kaust.edu.sa, jmpr@fb.com, v.castillo@samsung.com, brais.mart@gmail.com,
eddy.zhuxt@gmail.com, lizhangfd@fudan.edu.cn, bernard.ghanem@kaust.edu.sa, t.xiang@surrey.ac.uk

1. Comparison on other TAL datasets

To further validate our method, we conduct experiments on two more datasets with different sizes for temporal action localization. On both datasets, we firstly use a Kinetics pre-trained TSM-50 model to extract feature and then run G-TAD on top of it. Then, for fair comparison, we use our BSP feature extractor for the same process. Both experiments are tested with the same environment and same hyper-parameters.

1.1. Evaluation on larger dataset

Human Action Clips and Segments (**HACS-1.1**) [5] is a recent large-scale temporal action localization dataset. It contains 140K complete segments on 50K videos in 200 action categories. Compared with ActivityNet holding 20K videos, HACS dataset is larger and more challenging.

We show the performance of BSP on HACS-1.1 dataset against an RGB-only baseline in Tab. 1. We observe a gain of $+0.78\%$ on Average mAP, shown by the grey column of the table.

Table 1. **Evaluations on HACS-1.1 dataset on temporal action localization.** "*" indicates RGB-only Kinetics pre-trained TSM feature without fine-tuning.

Method	0.5	0.75	0.95	Average
G-TAD*	37.50	23.80	7.10	24.25
G-TAD* +BSP	38.10	24.73	7.76	25.03
Gain	+0.60	+0.93	+0.66	+0.78

1.2. Evaluation on smaller dataset

THUMOS-14 [2] dataset contains 413 temporally annotated untrimmed videos with 20 action categories. We use the 200 videos in the validation set for training and evaluate on the 213 videos in the testing set. Different from ActivityNet, THUMOS-14 has in average 16 action instances per video for the test set. Thus, the sensitivity to action boundaries plays a more important role for its temporal action localization.

Table 2. **Evaluation of temporal action localization on THUMOS-14 dataset.** "*" indicates RGB-only Kinetics pre-trained TSM feature without fine-tuning.

Method	0.3	0.4	0.5	0.6	0.7
G-TAD*	46.61	39.46	30.14	20.13	12.15
G-TAD* +BSP	52.34	46.28	39.80	30.81	21.14
Gain	+5.73	+6.82	+9.66	+10.68	+8.99

The performance of our method on THUMOS-14 dataset against an RGB-only baseline is shown in Tab. 1. As shown by the grey column of the table, a significant gain of $+9.66\%$ on mAP at IoU=0.5 further demonstrates the effectiveness of our proposed approach.

2. Comparison on other TAL methods

To test the robustness of BSP to different TAL methods, we also conduct experiments on Boundary-Matching Network (BMN) [3] based on a publicly available re-implementation[1]. BMN is an effective, efficient and end-to-end trainable proposal generation method, which generates proposals with precise temporal boundaries and reliable confidence scores simultaneously. BMN proposed the Boundary-Matching (BM) mechanism to evaluate confidence scores of densely distributed temporal action proposals. Please note that BMN targets the problem of temporal action proposal generation, which is a sub-task of temporal action localization. In the proposal generation task, models

*Work done during an internship at Samsung AI Centre.

[1] https://github.com/JJBOY/BMN-Boundary-Matching-Network

only predict actions segments, but do not assign action labels to them. Thus, the evaluation metric of BMN is different than for TAL. The proposal generation task is evaluated by the top-k average recall of predictions (AR@k), where $k \in \{1, 5, 10, 100\}$. The area under the recall curve (AUC) is used to evaluate the overall performance.

It can be observed from Tab. 3 that BSP brings consistent improvements over all the evaluation metrics on BMN, showing that our method can generalize to different methods. Particularly, our method can get 67.61 AUC which is close to the value, 67.7, reported by the code repository. However, no optical flow is used in BSP method.

Table 3. **Evaluation of temporal action proposal generation on ActivityNet.** "*" indicates RGB-only Kinetics pre-trained TSM feature without fine-tuning.

Method	AR@1	AR@5	AR@10	AR@100	AUC
BMN*	33.45	49.41	56.55	75.17	67.26
BMN* +BSP	33.69	50.12	57.35	75.50	67.61
Gain	+0.24	+0.71	+0.80	+0.33	+0.35

3. Comparison to other pre-training methods

We compare BSP to two other popular video self-supervised methods in Tab. 4: Arrow of Time [4] and SpeedNet [1]. Furthermore, we also include the results when ensembling two classification-based pre-trained models and for a randomly-initialized network. All comparisons use the two-steam integration with the same classification-based pre-trained TSM-18 video encoder. A random model does not give extra information, while Speednet [1] and Arrow of Time [4] can enrich the original features. Since they are not sensitive to the action boundaries, these methods are experimentally equivalent to training an ensemble of two vanilla video encoder (last row). Their performance is thus clearly inferior to that of our BSP method.

Table 4. **Comparison to other pre-training methods.** We compare BSP to other pre-training strategies. The results show TAL performance of G-TAD on THUMOS-14 dataset.

Method	0.3	0.4	0.5	0.6	0.7
baseline	41.48	34.21	25.88	17.82	11.28
Arrow [4]	46.52	40.13	31.30	22.24	14.39
Speed [1]	46.76	39.49	31.92	23.21	15.18
BSP (ours)	**48.71**	**42.78**	**34.92**	**27.44**	**17.60**
ensemble	45.67	37.70	28.66	19.55	11.32

4. Robustness to model backbone and capacity

We investigate the impact of different model backbones and capacity in Tab. 5. We use G-TAD for the TAL task on ActivityNet 1.3, with TSM-18, TSM-50, and R(2+1)D-34 as backbone choices. It is clear that a deeper model produces better BSP features, and that the net contribution when concatenated to classification-based BSP features is consistently and similarly positive in all cases. This verifies the general efficacy of our method on different backbones with varying capacity.

Table 5. **Model backbone and capacity.** We compare performance when using TSM-18, TSM-50, and R(2+1)d-34 for G-TAD on ActivityNet 1.3 dataset.

Backbone	BSP	0.5	0.75	0.95	Average
TSM-18	✗	49.64	34.16	7.68	33.59
TSM-18	✓	50.09	34.66	7.95	33.96
TSM-50	✗	50.32	35.07	8.02	34.26
TSM-50	✓	50.94	35.61	7.98	34.75
R(2+1)d-34	✗	49.57	34.92	8.43	34.05
R(2+1)d-34	✓	50.28	35.65	8.06	34.53

References

[1] 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 *CVPR*, 2020. 2

[2] YG 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, 2014. 1

[3] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 1

[4] D. Wei, J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *CVPR*, 2018. 2

[5] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. HACS: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint*, 2019. 1