Supplementary Material
Learning from Noisy Pseudo Labels for Semi-Supervised Temporal Action Localization

Method	mAP (%)					
	0.3	0.4	0.5	0.6	0.7	Avg.
w/o VBA	71.9	65.4	55.7	40.9	23.4	51.5
w/ VBA	**73.1**	**68.4**	**59.9**	**47.6**	**31.3**	**56.1**

Table 1. Ablation study on our video background augmentation with 40% labeled THUMOS14.

1. Background Augmentation

An empirical observation [5, 9] in semi-supervised learning (SSL) is that training a model on a larger dataset could produce better performance. To this end, some SSL work [11, 4] devotes to diving into different data augmentation strategies to improve the generalization ability of the model. Recent advances [3, 7, 2] in semi-supervised temporal action localization (SS-TAL) provide insights into video perturbations for data augmentation. They aim to improve the consistency between the teacher and student networks' predictions when video features are augmented by different temporal perturbations, *e.g.*, time warping/masking [3], temporal feature shift/flip [7], or spatio-temporal feature crossover [2].

This paper introduces a novel data augmentation strategy, dubbed video background augmentation, to perform SS-TAL for improved performance. More specifically, given a labeled video, we first extract the snippet features outside the annotated action instances as video background segments. Then, we combine a background segment with another labeled video to construct a new training sample through linear interpolation:

$$\begin{aligned} \tilde{\mathbf{x}} &= \lambda \mathbf{x}_i + (1-\lambda)\mathbf{b}_j, \\ \tilde{y} &= y_i, \end{aligned} \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{b}_j$ are the raw video feature of the i-th labeled video and the background feature of the j-th labeled video, respectively. ($\mathbf{x}_i$, y_i) is a labeled sample drawn randomly from our training data, and y_i is the label of $\mathbf{x_i}$. $\lambda \in [0, 1]$. ($\tilde{\mathbf{x}}$, $\tilde{y}$) is a virtual training sample.

To verify the effectiveness of our video background augmentation (VBA) approach, we conduct an ablation study on random 40% labeled THUMOS14, where we set $\lambda = 0.5$. From Table 1, we can observe that our data augmentation strategy significantly outperforms the baseline by a large margin. In a nutshell, video background augmentation could not only alleviate the confusion of the model to the actions sharing similar context, but also improve the generalization ability of the model.

Method	mAP (%)			
	0.3	0.5	0.7	Avg.
CNN	49.9	36.6	15.9	34.7
+NR	50.6	38.1	17.5	36.0
+NF	52.2	39.4	18.7	37.3
+NL	**52.9**	**40.5**	**20.1**	**38.1**

Table 2. Ablation study on the effectiveness of each component of the proposed method with a CNN baseline, using 40% labeled videos on THUMOS14. + means training by the proposed method.

2. Effectiveness on CNN

Recent fully-supervised TAL methods [1, 10] achieve impressive results benefiting from the much more powerful representation ability of self-attention-based Transformer. To further demonstrate the effectiveness of our three key designs, *i.e.*, Noisy Label Ranking (NR), Noisy Label Filtering (NF) and Noisy Label Learning (NL), we conduct an ablation study with a CNN model using 40% labeled THUMOS14, where we replace the transformer with the CNN. Table 2 shows that the performance can be significantly and consistently improved over the CNN baseline. Therefore, our model can be deployed on different baselines with significant performance gains.

3. Explanation of Noisy Label Filtering

To further explain what the proposed noisy label filtering does, we visualize its process in Figure 1. It depicts pseudo labels in descending order of confidence scores, where different symbols represent different categories and the blue or green colors denote selections. Our method, *i.e.*, Eq. (6), samples pseudo labels from each category separately to alle-

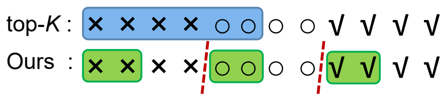

Figure 1. Difference between our noisy label filtering and conventional top-*K*

Method	mAP (%)					
	0.3	0.4	0.5	0.6	0.7	Avg.
w/o weighting	**71.9**	65.4	**55.7**	**40.9**	**23.4**	**51.5**
w/ weighting	71.8	**65.8**	55.6	40.8	**23.4**	**51.5**

Table 3. Ablation study on the sample re-weighting strategy with 40% labeled THUMOS14.

viate the class-imbalance problem of the conventional top-*K*. So, our sampling method is based on the confidence scores rather than random.

4. Sample Re-weighting Strategy

Deep learning-based detection models are susceptible to noisy labels, which degrades the model training [8]. In addition to the proposed Noisy Pseudo-Label Learning framework, we also make efforts to tackle the label noise on unlabeled data. Loss adjustment is a popular solution for reducing the negative impact of noisy labels by adjusting the loss of all training samples before updating the model [6]. Therefore, we attempt to use the foreground scores of all pseudo labels to weigh their loss values for loss adjustment. However, the results in Table 3 demonstrate that the sample re-weighting strategy fails to improve the performance. In our view, the reason behind this failure is that clean samples from labeled videos and noisy samples from unlabeled videos share similar semantic information, *e.g.*, motion patterns or background scenes. It results in similar loss values between the clean and noisy samples.

References

[1] Feng Cheng and Gedas Bertasius. TALLFormer: Temporal action localization with long-memory transformer. In *ECCV*, pages 503–521, 2022. 1

[2] Xinpeng Ding, Nannan Wang, Xinbo Gao, Jie Li, Xiaoyu Wang, and Tongliang Liu. KFC: An efficient framework for semi-supervised temporal action localization. *IEEE T-IP*, 30:6869–6878, 2021. 1

[3] Jingwei Ji, Kaidi Cao, and Juan Carlos Niebles. Learning temporal action proposals with fewer labels. In *ICCV*, pages 7073–7082, 2019. 1

[4] JongMok Kim, Jooyoung Jang, Seunghyeon Seo, Jisoo Jeong, Jongkeun Na, and Nojun Kwak. MUM: Mix image tiles and unmix feature tiles for semi-supervised object detection. In *CVPR*, pages 14512–14521, 2022. 1

[5] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, 2020. 1

[6] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE T-NNLS*, 2022. 2

[7] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. In *CVPR*, pages 1905–1914, 2021. 1

[8] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *CVPR*, pages 4568–4577, 2021. 2

[9] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 1

[10] Chenlin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510, 2022. 1

[11] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1