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

Kun Xia[1,3†] Le Wang[1*] Sanping Zhou[1] Gang Hua[2] Wei Tang[3]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Wormpex AI Research [3]University of Illinois Chicago

Abstract

Semi-Supervised Temporal Action Localization (SS-TAL) aims to improve the generalization ability of action detectors with large-scale unlabeled videos. Albeit the recent advancement, one of the major challenges still remains: noisy pseudo labels hinder efficient learning on abundant unlabeled videos, embodied as location biases and category errors. In this paper, we dive deep into such an important but understudied dilemma. To this end, we propose a unified framework, termed Noisy Pseudo-Label Learning, to handle both location biases and category errors. Specifically, our method is featured with (1) Noisy Label Ranking to rank pseudo labels based on the semantic confidence and boundary reliability, (2) Noisy Label Filtering to address the class-imbalance problem of pseudo labels caused by category errors, (3) Noisy Label Learning to penalize inconsistent boundary predictions to achieve noise-tolerant learning for heavy location biases. As a result, our method could effectively handle the label noise problem and improve the utilization of a large amount of unlabeled videos. Extensive experiments on THUMOS14 and ActivityNet v1.3 demonstrate the effectiveness of our method. The code is available at github.com/kunnxia/NPL.

1. Introduction

Temporal Action Localization (TAL) aims at detecting action instances of interest in an untrimmed video by locating their temporal boundaries and recognizing their action categories. Most existing TAL methods rely on dense temporal annotations for the training videos. However, labeling human actions is very tedious and time-consuming. As a remedy, Semi-Supervised TAL (SS-TAL) requires only a few labeled videos in conjunction with a large amount of unlabeled videos. It has attracted growing attention in academia and industry.

Figure 1. Illustration of **label noise**. Two examples demonstrate that pseudo labels may contain location biases and category errors. The fewer labeled videos are available, the heavier the noise of the pseudo labels will be.

Existing SS-TAL methods [12, 37, 28] are based on consistency regularization or self-training. Consistency-based methods [12, 37] aim to generate consistent action proposals for the same video subject to different augmentations, *e.g.*, time warping [12] and temporal feature shift [37]. In contrast, the self-training-based method [28] achieves new state-of-the-art performance by alternating between pseudo-labeling and re-training. It focuses on designing a proposal-free framework to address proposal error propagation from [12, 37] but neglects the important role of pseudo labels. Albeit its advancement, *label noise* still remains a core challenge, hindering efficient learning on abundant un-

*Corresponding author
† Part of this work was done while Kun Xia was a visiting scholar at University of Illinois Chicago.

labeled videos. From Figure 1, we can observe that label noise commonly leads to two intractable issues, *i.e.*, *location bias* and *category error*, which become worse as the amount of labeled videos decreases. As a result, noisy pseudo labels will significantly degrade the performance of SS-TAL.

In this paper, we propose a **N**oisy **P**seudo-Label **L**earning (NPL) framework tailored for SS-TAL to combat detrimental label noise. It follows the self-training paradigm that alternates between pseudo-labeling and model training. But unlike all previous self-training methods, our NPL includes three novel components, termed *Noisy Label Ranking*, *Noisy Label Filtering*, and *Noisy Label Learning*, respectively, which alleviate the negative effects caused by location biases and category errors in a unified framework.

First, Noisy Label Ranking aims to rank and select high-quality pseudo labels based on both semantic confidence and boundary reliability. Classification scores have been widely used in self-training to measure the quality of pseudo labels, but they only reflect semantic confidence and fail to account for localization reliability. To close this gap, we explicitly model the localization reliability of a detected action instance in an unlabeled video as the variance of the boundary predictions from dense snippets within the action. We then introduce a new integrated metric of semantic confidence and boundary reliability to rank the pseudo labels.

Second, Noisy Label Filtering aims at addressing the class-imbalance problem in noisy pseudo labels. Owing to the category error, the class-imbalance problem occurs regardless of whether pseudo labels are sampled based on a confidence threshold or the number of samples. The model will be dominated by redundant noisy pseudo labels in training, especially for the ones of category errors, and further harm its generalization ability. To address this issue, we introduce an *adaptive filtering strategy* to regularize the distribution of pseudo labels and adaptively assign class-balanced pseudo labels to unlabeled videos.

Last, Noisy Label Learning aims to improve the robustness of training to location bias. While noisy label ranking and filtering can improve the quality of sampled pseudo labels, location bias will not be removed completely. Training on biased boundary labels hinders the convergence of the model and further impedes accurate action localization. To this end, we propose a noise-tolerant training algorithm based on an unsupervised *temporal consistency* loss, which penalizes inconsistent predictions from adjacent action frames.

The main contributions are summarized as follows:

- This paper introduces a Noisy Pseudo-Label Learning (NPL) framework tailored for SS-TAL, which handles both the location bias and category error in a unified framework. Extensive experiments conducted on THUMOS14 [13] and ActivityNet v1.3 [2] demonstrate the effectiveness of the proposed method.
- We propose a Noisy Label Ranking method to rank and select high-quality pseudo labels based on a new integrated metric of semantic confidence and boundary reliability.
- We propose a Noisy Label Filtering method to tackle the largely ignored class-imbalance problem in pseudo labels based on a new adaptive filtering strategy.
- We introduce a Noisy Label Learning method, which adopts an unsupervised temporal consistency loss to penalize inconsistent predictions from adjacent frames for noise-tolerant learning.

2. Related Work

Temporal Action Localization. The recent significant advancement of TAL arguably owes to the availability of large-scale and well-annotated datasets. We can roughly categorize existing TAL methods into three groups. *Anchor-based methods* [43, 5, 45, 52, 36] employ multi-scale anchors that may contain an action and refine them via a boundary regression head. GTAN [25] dynamically optimizes the scale of each anchor via Gaussian kernels. G-TAD [45] improves temporal anchor representation with semantic and temporal context. *Anchor-free methods* [54, 29, 18, 27, 40, 39, 41] directly learn to predict action proposals by grouping possible start/end locations [20, 19] or regressing the distance to action boundaries [18, 49]. *Transformer-based methods* [31, 24, 49, 8] adopt the Transformer architecture [33, 3] for action localization and achieve remarkable performance on TAL benchmarks. In addition, *unsupervised temporal action localization* [11, 50] has made remarkable achievements without requiring any annotations.

Semi-Supervised Learning. Semi-Supervised Learning (SSL) methods mostly focus on image classification. They can be classified into two categories. *Consistency-based methods* [32, 26, 51, 1] enforce the model to produce consistent predictions across label-preserving image augmentations. *Self-training methods* [42, 15, 10] retrain the model with high-confidence pseudo labels of unlabeled raw data. Thereby, the final performance is largely limited by the quality of pseudo labels generated by an inaccurate model trained using a few labeled data. Much effort tries to remedy noisy pseudo labels [16, 23, 44, 55, 17, 46]. Li *et al.* [16] train a noise-tolerant model, which is encouraged to produce consistent predictions under a variety of noisy synthetic labels. DSL [7] directly ignores the gradient computation and propagation for ambiguous pseudo labels. Wang *et al.* [38] propose an iterative training method to identify and down-weigh noisy samples. Chen *et al.* [6] address severe confirmation bias during self-training and generate unbiased pseudo labels to drive student learning. In this paper, we focus on the challenging SS-TAL problem.

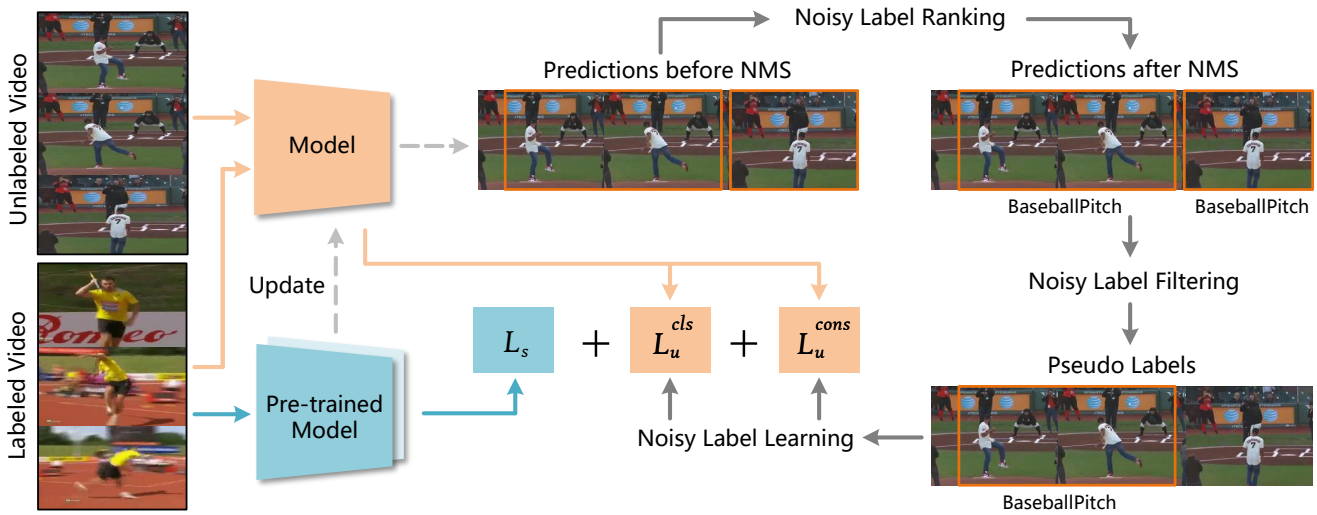

Figure 2. An overview of our proposed Noisy Pseudo-Label Learning framework. The training data contain both labeled and unlabeled videos. We first obtain a pre-trained model using a small amount of labeled videos and generate pseudo labels on unlabeled videos. To deal with noise in pseudo labels, Noisy Label Ranking first assesses the quality of pseudo labels through both semantic confidence and boundary reliability. Noisy Label Filtering then adaptively filters pseudo labels so that the model learning will not be dominated by pseudo labels with category errors. To make the training robust to label noise, we introduce dense temporal consistency training with L_u^{cons}. The model can be optimized by the final loss, which is the sum of L_s, L_u^{cls}, and L_u^{cons}.

Semi-Supervised Temporal Action Localization. SS-TAL is a worthwhile but under-explored field. Existing works [12, 37, 9, 28] are intrinsically driven by SSL. Some works [12, 37, 9] aim to improve the consistent predictions of the teacher-student network when video features are augmented by different temporal perturbations, *e.g.*, time warping/masking [12], temporal feature shift/flip [37] or spatio-temporal feature crossover [9]. The other one [28] focuses on addressing proposal error propagation, which alternates between predicting and applying pseudo labels. However, all these approaches fail to dive deep into the major challenge of SS-TAL, *i.e.*, noisy pseudo labels hindering efficient learning on unlabeled videos. We attribute the label noise problem of pseudo labels from unlabeled videos to the semantic uncertainty and boundary ambiguity of labeled actions. This paper tackles both the location bias and category error of pseudo labels into a unified framework for efficient learning from unlabeled videos.

3. Method

3.1. Preliminary

Problem Setting. Semi-supervised temporal action localization (SS-TAL) aims to perform TAL from a small amount of labeled videos $\{X_i|_{i=1}^{N_l}\}$ and a large amount of unlabeled videos $\{U_i|_{i=1}^{N_u}\}$, where N_l and N_u are the numbers of labeled and unlabeled videos, respectively. In a labeled video, its annotation is a set of action instances, and each action instance could be represented as (t_s, t_e, c), where t_s, t_e, and c denote the start time, end time, and action class, respectively.

Feature Encoding. Following conventions [45, 47], we sample a video snippet from every few consecutive frames. Then, we adopt a fine-tuned two-stream network to extract RGB and optical flow features at each video snippet.

Overview. Recent anchor-free TAL methods [18, 49, 27] are more attractive and practical since they are powerful and portable to be deployed in many real-world applications. We take the anchor-free detector [49] as our baseline, where each video snippet is directly supervised by the corresponding labels, *i.e.*, distances to boundaries and the action category.

Our SS-TAL framework follows the self-training paradigm [30], which can be decomposed into two steps. In the first step, we obtain a pre-trained model based on labeled videos with a supervised loss and then generate pseudo labels on unlabeled videos. In the second step, we update the model parameters with a semi-supervised training objective, including a supervised loss and an unsupervised loss:

$$\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_u, \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote the supervised loss of labeled videos and the unsupervised loss of unlabeled videos, respectively, and α controls the contribution of the unsupervised loss. Both of them are normalized by the corresponding numbers of training samples as follows:

$$\mathcal{L}_s = \frac{1}{N_l} \sum_{i=1}^{N_l} \left(\mathcal{L}_s^{\text{cls}}(X_i) + \mathcal{L}_s^{\text{reg}}(X_i) \right), \tag{2}$$

$$\mathcal{L}_u = \frac{1}{N_u}\sum_{i=1}^{N_u}\left(\mathcal{L}_u^{\text{cls}}\left(U_i\right) + \mathcal{L}_u^{\text{reg}}\left(U_i\right)\right), \tag{3}$$

where $\mathcal{L}^{\text{cls}}$ is a classification loss, $\mathcal{L}^{\text{reg}}$ is a regression loss, and N_l and N_u denote the number of training samples for labeled and unlabeled videos, respectively.

We claim that the major obstacle that hinders self-training performance lies in *label noise*, which leads to two issues: location biases and category errors. In this paper, we propose a Noisy Pseudo-Label Learning framework (NPL) to alleviate the negative effects caused by location biases and category errors for SS-TAL. The pipeline of our method is illustrated in Figure 2. It includes three novel components: Noisy Label Ranking, Noisy Label Filtering, and Noisy Label Learning, described in the following sections.

3.2. Noisy Label Ranking

In SS-TAL, the quality of pseudo labels can be measured from two perspectives, *i.e.*, semantic confidence and boundary reliability. It is desirable to have high-quality pseudo labels for semi-supervised learning.

The majority of TAL frameworks [20, 53, 52, 36] rely on the foreground score of an anchor or a proposal to measure its quality. However, the foreground score only reflects semantic confidence and cannot account for boundary reliability. From Figure 3 (a), we can observe that foreground scores may not strongly correlate with the localization quality (*i.e.*, IoU w.r.t. ground truth). Thus, it is necessary to measure the quality of pseudo labels from both aspects of action classification and boundary localization.

For labeled videos, anchors or proposals corresponding to the same action instance are supervised by the same ground truth during training. Thus, they are gradually regressed to the similar temporal boundary locations. This motivates us to use the consistency of the boundary predictions to measure their location reliability. Specifically, taking our anchor-free framework as an example, we can obtain dense predictions of action instances before the post-processing phase. Given a prediction $p_j = (\hat{t}_{s,j}, \hat{t}_{e,j}, c_j)$ where $\hat{t}_{s,j}$ and $\hat{t}_{e,j}$ are the predicted start and end boundaries of the j-th video snippet respectively and c_j is its foreground score, we can obtain a set of predictions $\{p_n\}_{n=1}^{N_a}$ from its N_a adjacent snippets. Then, we formulate the boundary ambiguity of p_j as follows:

$$\begin{aligned} \bar{\sigma}_j &= \hat{\sigma}_{s,j} + \hat{\sigma}_{e,j}, \\ \hat{\sigma}_{s,j} &= \frac{\sigma_{s,j}}{d\left(p_j\right)}, \\ \hat{\sigma}_{e,j} &= \frac{\sigma_{e,j}}{d\left(p_j\right)}, \end{aligned} \tag{4}$$

where $\sigma_{s,j}$ and $\sigma_{e,j}$ are the variances of the start and end boundaries of $\{p_n\}_{n=1}^{N_a}$, respectively. $\hat{\sigma}_{s,j}$ and $\hat{\sigma}_{e,j}$ are normalized by their duration $d\left(p_j\right)$. A smaller boundary variance $\bar{\sigma}_j$ indicates lower boundary ambiguity of the prediction p_j, thus leading to higher localization reliability. In Figure 3 (b), we experimentally illustrate the correlation between the boundary variance and IoU w.r.t ground truth. We can observe that boundary variances could better measure the localization quality of pseudo labels than the classification scores. Particularly, this motivates us to define the confidence score of p_j as:

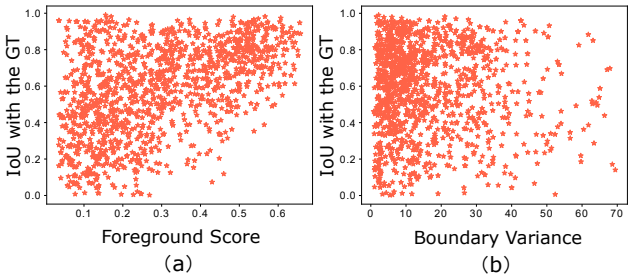

Figure 3. Each star represents a prediction result (*i.e.*, a pseudo label) on 60% unlabeled videos based on the model trained with 40% labeled videos on THUMOS14. (a) depicts the correlation between the foreground score and the IoU with ground truth. (b) depicts the correlation between the boundary variance and the IoU with ground truth. It can be observed that the boundary variance highly correlates with the localization quality.

$$s_j = \frac{c_j}{\bar{\sigma}_j}. \tag{5}$$

Eq. (5) provides a comprehensive metric to measure the quality of pseudo labels by their semantic confidence (foreground score c) and boundary ambiguity (boundary variance $\bar{\sigma}$). As a result, our Noisy Label Ranking could improve the quality of pseudo labels only from intrinsic video snippets without requiring additional learnable networks or hyperparameters.

3.3. Noisy Label Filtering

Training on a few labeled videos inevitably produces noisy pseudo labels with category errors on unlabeled videos. As a result, the frequencies of some action categories in the selected pseudo labels will be much larger or smaller than their natural frequencies in the videos, regardless of whether pseudo labels are sampled based on a confidence threshold or the number of samples. We call this phenomenon the class-imbalance problem of pseudo labels in SS-TAL, illustrated in Figure 4. For example, the frequency of "Golf Swing" in pseudo labels is much higher than that in ground truth labels, while the frequency of "Throw Discus" is much lower in the pseudo labels.

Under this class-imbalance problem, overwhelming pseudo labels with category errors will dominate the model learning and further aggravate the semantic uncertainty of actions of category errors, resulting in limited gains.

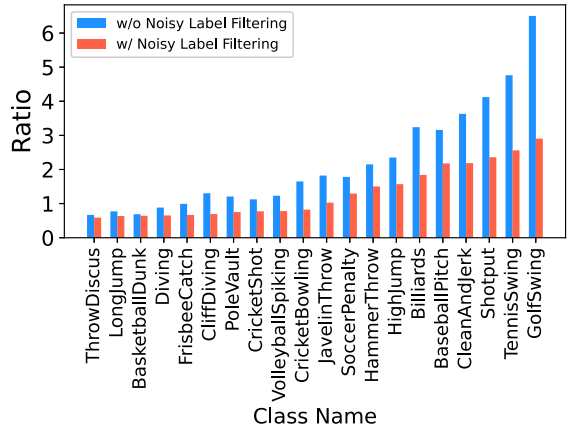

Figure 4. Illustration of the class-imbalance problem of pseudo labels obtained on 60% unlabeled THUMOS14. The horizontal axis of the histogram represents the class name, and the vertical axis represents the ratio between the number of pseudo labels and the number of ground truth labels. It can be observed that our Noisy Label Filtering could alleviate the class-imbalance problem of pseudo labels.

To address this class-imbalance problem, we introduce an adaptive filtering strategy to regularize the distribution of pseudo labels for each category. Concretely, we arrange dense labels according to the category of actions and the number of pseudo labels in an unlabeled video, as follows:

$$K_{adp} = \frac{|\{p_n^c\}_{n=1}^{N_v}|}{N_v} \cdot N_{pos}, \tag{6}$$

where N_v is the total number of pseudo labels within a video, and N_{pos} is a hyper-parameter to control the number of positive predictions. Eq. (6) is used to adaptively select the number K_{adp} of class-specific pseudo labels for the category c in each unlabeled video. From Figure 4, we experimentally demonstrate that our adaptive filtering strategy could address the class-imbalance distribution of pseudo labels. As a result, our Noisy Label Filtering could not only assign accurate pseudo labels to improve the model learning but also prevent the pseudo labels from being dominated by overwhelming pseudo labels with category errors, while it also avoids tediously adjusting heuristic thresholds.

3.4. Noisy Label Learning

The boundary ambiguity of labeled actions gives rise to noisy pseudo labels with location biases. Subsequently, training on the biased boundaries results in the boundary ambiguity of unlabeled actions and impedes accurate action localization.

To remedy it, we propose a noise-tolerant training algorithm for unlabeled videos with heavy location biases. It aims to penalize inconsistency boundary predictions from adjacent feature locations within a video feature sequence. Our unsupervised temporal consistency loss is defined as follows:

$$\mathcal{L}_u^{\text{cons}} = \frac{1}{M_u(M_p-1)} \sum_{j=1}^{M_u} \sum_{i=1}^{M_p-1} |b_{i+1}^j - b_i^j|, \tag{7}$$

where M_u denotes the number of pseudo labels within an unlabeled video. M_p represents the number of the training snippets within an action. b_i represents the predicted start/end boundary locations from the i-th snippet. In this way, the model could learn to produce more reliable predictions. Consequently, our complete unsupervised loss is denoted below:

$$\mathcal{L}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \left(\mathcal{L}_u^{\text{cls}}(U_i) + \mathcal{L}_u^{\text{cons}}(U_i) \right). \tag{8}$$

The Noisy Label Learning makes the training robust to location biases in the pseudo labels. Finally, the model can be optimized by the sum of $\mathcal{L}_s$ and $\mathcal{L}_u$ on all videos.

4. Experiments

4.1. Datasets and Metrics

Evaluation Datasets. We evaluate our proposed method on two benchmark TAL datasets, *i.e.*, THUMOS14 [13] and ActivityNet v1.3 [2]. THUMOS14 [13] is a standard benchmark for TAL. It contains 200 validation videos and 213 testing videos, including 20 action categories. It is very challenging since each video has more than 15 action instances. Following the common setting [48], we use the validation set for training and evaluate on the testing set.

ActivityNet v1.3 [2] is a large-scale benchmark for video-based action localization. It contains 10k training videos and 5k validation videos corresponding to 200 different actions. Following the standard practice [22], we train our method on the training set and test it on the validation set.

Evaluation Metrics. We use the mean Average Precision (mAP) as the evaluation metric. The tIoU thresholds are $[0.3:0.1:0.7]$ for THUMOS14 and $[0.5:0.05:0.95]$ for ActivityNet v1.3. We report the average mAP of the IoU thresholds between 0.5 and 0.95 with the step of 0.05 on ActivityNet v1.3. Also, we present the average mAP of the tIoU thresholds from 0.3 to 0.7 on THUMOS14.

The goal of SS-TAL is to use a large amount of unlabeled data to improve a well-trained detector on a small amount of labeled data. We randomly sample 10%, 20%, 40%, and 60% of the training data as labeled data and treat the remainder as unlabeled data.

4.2. Implementation Details

We implement our semi-supervised action localization framework based on the recent prevalent anchor-free detector

Label	Method	Backbone	THUMOS14 (%)						ActivityNet v1.3 (%)			
			0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
10%	ActF* [49]	I3D	28.5	22.9	14.1	8.2	4.1	15.6	47.8	24.2	1.7	25.6
	ActF [49] + MixUp [51]	I3D	29.7	24.2	14.5	9.6	5.4	16.7	49.4	27.9	3.1	28.8
	NPL (ActF)	I3D	**32.8**	**29.6**	**20.1**	**11.7**	**7.2**	**20.3**	**51.9**	**33.4**	**3.6**	**32.5**
	SSP [12]	TSN	44.2	34.1	24.6	16.9	9.3	25.8	38.9	28.7	8.4	27.6
	SSTAP [37]	TSN	45.6	35.2	26.3	17.5	10.7	27.0	40.7	29.6	**9.0**	28.2
	SPOT [28]	TSN	49.4	40.4	31.5	22.9	12.4	31.3	49.9	31.1	8.3	32.1
	NPL (BMN)	TSN	**50.0**	**41.7**	**33.5**	**23.6**	**13.4**	**32.4**	**50.9**	**32.0**	7.9	**32.6**
20%	ActF* [49]	I3D	49.1	41.6	32.6	21.5	12.1	31.4	51.2	34.3	3.8	32.9
	ActF [49] + MixUp [51]	I3D	51.2	43.2	34.0	23.9	14.1	33.3	52.9	34.7	3.9	33.3
	NPL (ActF)	I3D	**54.5**	**47.1**	**39.3**	**29.7**	**18.5**	**37.8**	**53.1**	**35.8**	**3.9**	**33.8**
	SPOT [28]	TSN	52.6	43.9	34.1	25.2	16.2	34.4	51.7	32.0	6.9	32.3
	NPL (BMN)	TSN	**53.9**	**45.6**	**36.2**	**26.9**	**16.5**	**35.8**	**52.1**	**32.9**	**7.9**	**32.9**
40%	ActF* [49]	I3D	69.0	60.4	49.3	31.5	19.3	45.9	53.2	35.7	3.8	34.2
	ActF [49] + MixUp [51]	I3D	69.7	61.9	52.4	34.4	20.1	47.7	53.1	36.0	4.3	34.5
	NPL (ActF)	I3D	**71.9**	**65.4**	**55.7**	**40.9**	**23.4**	**51.5**	**53.6**	**36.5**	**4.6**	**35.3**
	SPOT [28]	TSN	54.4	45.8	37.2	29.7	19.4	37.3	53.3	33.0	6.6	33.2
	NPL (BMN)	TSN	**56.2**	**46.7**	**38.8**	**30.3**	**19.5**	**38.3**	**53.4**	**33.9**	**8.1**	**33.8**
60%	ActF* [49]	I3D	71.5	65.6	59.9	47.3	32.7	55.4	53.9	36.1	5.7	35.0
	ActF [49] + MixUp [51]	I3D	72.2	67.5	61.2	48.7	34.0	56.7	54.1	36.4	5.7	35.2
	NPL (ActF)	I3D	**74.5**	**69.9**	**62.8**	**51.1**	**36.6**	**59.0**	**54.3**	**36.7**	**6.5**	**35.8**
	SSP [12]	TSN	53.2	46.8	39.3	29.7	19.8	37.8	49.8	34.5	7.0	33.5
	SSTAP [37]	TSN	56.4	49.5	41.0	30.9	21.6	39.9	50.1	34.9	7.4	34.0
	SPOT [28]	TSN	58.9	50.1	42.3	33.5	22.9	41.5	52.8	35.0	8.1	35.2
	NPL (BMN)	TSN	**59.0**	**51.4**	**42.9**	**34.3**	**23.3**	**42.2**	**53.9**	**35.8**	**8.5**	**35.7**

Table 1. Performance comparison with state-of-the-art SS-TAL methods on THUMOS14 testing set and ActivityNet v1.3 validation set. Notably, SSP and SSTAP employ UntrimmedNet [34] trained with 100% class labels for proposal classification. ActF refers to ActionFormer [49]. * means using only labeled training videos.

ActionFormer [49], which is composed of a self-attention backbone [33], an FPN [21] neck and two parallel heads. Also, we combine the proposed semi-supervised method with BMN [19], which is a two-stage proposal-based detector. For video feature encoding, we used two popular backbones, *i.e.*, two-stream network [35] and I3D [4] pre-trained on Kinetics [14] to extract the video features, following [18, 49]. For THUMOS14, the initial learning rate is $1e$-4, and a cosine learning rate decay is used. The mini-batch size is 2, and a weight decay of $1e$-4 is used. For ActivityNet v1.3, the learning rate is $1e$-3, the mini-batch size is 32, and the weight decay is $1e$-4.

For semi-supervised configurations, an initial model is trained on available labeled videos, where labeling ratios are set to 10%, 20%, 40%, and 60%. Further using this well-trained model generates pseudo labels. For the re-training phase, we pre-train the model on the labeled data and then compute the supervised loss and unsupervised loss for 40 epochs (THUMOS14) and 15 epochs (ActivityNet v1.3). We set $\alpha = 1$ and $N_{pos} = 15$. We apply background data augmentation to 10% of unlabeled videos for performance improvements. Additional implementation details and more attempts for SS-TAL refer to the supplementary material.

4.3. Comparison with State-of-the-art Methods

THUMOS14. We compare the proposed method NPL with existing SS-TAL methods in Table 1, where we combine two different frameworks, ActionFormer [49] and BMN [19] with our NPL. We report mAP at different tIoU thresholds as well as average mAP between 0.3 and 0.7 with the step of 0.1. Also, we reproduce the SPOT [28] results of 20% and 40% labeling ratios from the source code for comparison. When labeled data is scarce (*i.e.*, 10% and 20% labeling ratios), our method achieves the competitive performance compared to existing methods. Especially, our method outperforms the baseline, semi-supervised ActF [49], by a large margin with 10% labeled THUMOS14. When more labeled data is accessible (*i.e.,* 40% and 60% labeling ratios), our method outperforms all the competing methods, demonstrating the effectiveness and superiority of our method.

ActivityNet v1.3. We also conduct experiments on the more challenging benchmark ActivityNet v1.3. As depicted in Table 1, the proposed method also achieves significant performance gains over all the compared methods. When

Label	Method	mAP (%)			
		0.3	0.5	0.7	Avg.
10%	baseline	29.5	15.8	5.0	16.8
	+NR	30.7	16.9	6.3	18.0
	+NF	32.0	18.5	6.9	19.1
	+NL	**32.8**	**20.1**	**7.2**	**20.3**
40%	baseline	70.1	49.5	20.2	47.1
	+NR	70.7	52.1	20.7	48.5
	+NF	71.6	54.4	21.3	50.3
	+NL	**71.9**	**55.7**	**23.4**	**51.5**

Table 2. Ablation study on effectiveness of each component of the proposed method on THUMOS14, using 10% and 40% labeled videos. + means training by the proposed method.

Label	Method	Class (%)	Avg. tIoU (%)
10%	baseline	21.0	27.4
	NPL	**25.9**	**36.2**
40%	baseline	73.0	54.2
	NPL	**74.4**	**59.0**

Table 3. Ablation study on the quality of pseudo labels in terms of action classification accuracy (Class) and average tIoU (Avg. tIoU), using 10% and 40% labeled videos.

there are more unlabeled data introduced, our method also can improve the average mAP, which mainly comes from the high-quality pseudo labels for effective training.

4.4. Ablation Study

To better understand how the proposed method works, we conduct a series of ablation studies, where we use ActionFormer [49] as the localization framework with I3D [4].

Effectiveness of individual component. To validate our key designs, we ablate the effects of the proposed Noisy Label Ranking (NR), Noisy Label Filtering (NF), and Noisy Label Learning (NL). The results are shown in Table 2. Under the 40% labeling ratio, one can see that our model equipped with NR improves performance by 1.4%. Further applying our adaptive filtering strategy and the temporal consistency learning, the performance reaches an average mAP of 51.5%. In summary, our designs significantly improve the utilization of unlabeled data for effectively semi-supervised learning.

Quality improvements of pseudo labels. The quality of pseudo labels is the important indicator for the SS-TAL performance improvement. We further study the quality of psudo labels from their average temporal IoU (tIoU) w.r.t ground truth and classification accuracy (Acc.). We report the comparison results in Table 3. It can be observed that the significant improvements attribute to our Noisy Label Ranking to rank and select high-quality pseudo labels based on the comprehensive metric, *i.e.*, semantic confidence and localization reliability, as well as our Noisy Label Filtering to

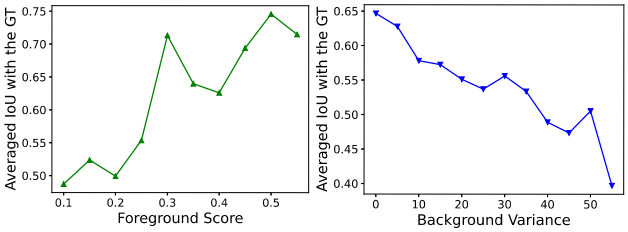

Figure 5. Curves of averaged IoU vs. foreground score or boundary variance, where each averaged IoU is calculated by the foreground scores or boundary variances of the samples in different intervals.

Method	mAP (%)					
	0.3	0.4	0.5	0.6	0.7	Avg.
$\theta_s = 0.3$	70.1	63.5	52.9	37.4	22.2	49.2
top-10	71.0	63.3	53.4	38.1	22.8	49.7
NF	**71.9**	**65.4**	**55.7**	**40.9**	**23.4**	**51.5**

Table 4. Comparison of pseudo-label filtering strategies on THUMOS14 with 40% labeled videos and 60% unlabeled videos.

adaptively assign class-balanced pseudo labels for training.

Visualization of noisy label ranking. To further support our motivation of the noisy label ranking, we add a curve of averaged IoU (between pseudo labels and ground truth) versus foreground score or boundary variance, where each averaged IoU corresponds to the foreground scores or boundary variances of the samples in different intervals, as shown in Figure 5. Particularly, the Pearson correlation coefficient between IoU and boundary variance is 0.496, while that between IoU and foreground score is 0.449.

Superiority of adaptive filtering strategy. Our Noisy Label Filtering (NF) could alleviate the class-imbalance problem of pseudo labels caused by the category error. In addition, it also could avoid tediously adjusting hard-crafted thresholds or numbers for pseudo-label filtering. To demonstrate this superiority, we experiment with two common approaches as a reference, *i.e.*, single thresholding and fixed number strategies. More specifically, we empirically set the foreground score threshold θ_s to 0.3, where instances are regarded as foreground if their scores are above the threshold and background otherwise. Besides, we also empirically select top-10 confidence predictions for each unlabeled video. As shown in Table 4, these approaches cannot achieve satisfactory performance. In contrast, our adaptive filtering strategy could adaptively sample class-specific pseudo labels, showing its effectiveness and importance.

Analysis of temporal consistency. Our Noisy Label Learning introduces a temporal consistency loss to tackle the location bias problem by penalizing inconsistent boundary predictions from each intrinsic video snippet. We conduct an ablation study on this key design, where we train the pre-trained model with and without the temporal consistency only on 60% unlabeled videos, as shown in Figure 6. Based

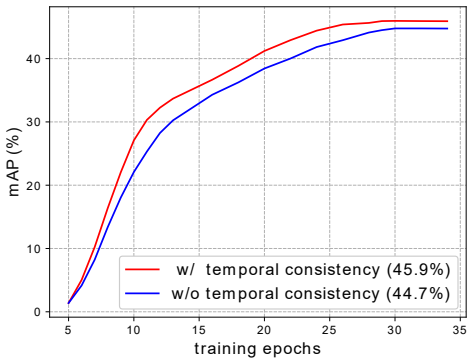

Figure 6. Comparison of model performance and convergence speed with and without the temporal consistency loss on THUMOS14.

N_{pos}	mAP (%)			
	0.3	0.5	0.7	Avg.
5	71.5	55.2	22.8	51.0
15	**71.9**	**55.7**	**23.4**	**51.5**
30	71.0	55.3	22.5	50.7

Table 5. Comparison of mAP with various values of N_{pos} using 40% labeled THUMOS14.

Label	100%	60%	40%	20%	10%
mAP@0.5 (%)	71.0	62.8	55.7	39.3	20.1
mAP@Avg (%)	66.8	59.0	51.5	37.8	20.3

Table 6. Comparison of mAP with various labeling ratios on THUMOS14, where the remainder is treated as unlabeled data.

on the temporal consistency, it obtains a 1.2% improvement on the average mAP. Apart from the performance gain, our contribution can also significantly boost the convergence speed. Both of these advances validate the effectiveness of the proposed Noisy Label Learning.

Choice of hyper-parameters. N_{pos} is a pre-defined value to handle the number of positive samples for unlabeled videos. We conduct an ablation study on the hyper-parameter N_{pos} under 40% labeling ratio. From Table 5, it can be observed that the performance peaks around $N_{pos} = 15$. A large value such as $N_{pos} = 30$ will introduce more low-quality pseudo labels and confuse the model in training, while a small value such as $N_{pos} = 5$ only provides a small number of samples for training so as to obtain limited gains.

Bottleneck of our NPL. The performance limitation of SS-TAL following self-training lies in the noise level of pseudo labels. To explore the performance bottleneck, we train a fully-supervised model with 100% labeled videos on THUMOS14. The experimental results in Table 6 show that our method still has great room for improvement. Also, it demonstrates that the negative impacts of the label noise problem cannot be ignored in semi-supervised learning.

Figure 7. Qualitative SS-TAL result comparison of our proposed method with SPOT [28] on two untrimmed videos from (a) ActivityNet v1.3 and (b) THUMOS14, respectively.

4.5. Visualization Analysis

To further validate the effectiveness of the proposed method, we provide some qualitative results by prior art SPOT [28] and our model with 10% and 40% labeled data on both ActivityNet v1.3 and THUMOS14. From the illustration in Figure 7, we can observe that our method can localize the target actions more accurately, demonstrating the effectiveness and superiority of our method.

5. Conclusion

In this paper, we propose a novel Noisy Pseudo-Label Learning framework tailored for SS-TAL to tackle the label noise problem. Specifically, we first present a new integrated metric of semantic confidence and boundary reliability for ranking high-quality pseudo labels. Then, an adaptive filtering strategy alleviates the class-imbalance distribution of pseudo labels caused by the category error. Finally, we introduce an unsupervised temporal consistency loss to tackle biased boundary labels for noise-tolerant learning. Experiments show that the effectiveness of our method on both THUMOS14 and ActivityNet v1.3.

Limitation. The number of positive snippets inside an action instance affects our method for quality ranking and noise-tolerant training, which indicates that tackling short action instances remains a challenging problem.

Acknowledgement

This work was supported partly by National Key R&D Program of China under Grant 2021YFB1714700, NSFC under Grants 62088102 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

[1] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *CVPR*, pages 6923–6932, 2021. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 2, 5

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6, 7

[5] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster R-CNN architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018. 2

[6] Binbin Chen, Weijie Chen, Shicai Yang, Yunyi Xuan, Jie Song, Di Xie, Shiliang Pu, Mingli Song, and Yueting Zhuang. Label matching semi-supervised object detection. In *CVPR*, pages 14381–14390, 2022. 2

[7] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *CVPR*, pages 4815–4824, 2022. 2

[8] Feng Cheng and Gedas Bertasius. TALLFormer: Temporal action localization with long-memory transformer. In *ECCV*, pages 503–521, 2022. 2

[9] 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. 3

[10] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, pages 1301–1310, 2017. 2

[11] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian. Learning temporal co-attention models for unsupervised video action localization. In *CVPR*, pages 9819–9828, 2020. 2

[12] Jingwei Ji, Kaidi Cao, and Juan Carlos Niebles. Learning temporal action proposals with fewer labels. In *ICCV*, pages 7073–7082, 2019. 1, 3, 6

[13] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014. 2, 5

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[15] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, page 896, 2013. 2

[16] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *CVPR*, pages 5051–5059, 2019. 2

[17] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training. In *ECCV*, pages 589–607, 2020. 2

[18] 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*, pages 3320–3329, 2021. 2, 3, 6

[19] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 2, 6

[20] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–19, 2018. 2, 4

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 6

[22] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. Multi-shot temporal event localization: a benchmark. In *CVPR*, pages 12596–12606, 2021. 5

[23] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *CVPR*, pages 14207–14216, 2022. 2

[24] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE T-IP*, 31:5427–5441, 2022. 2

[25] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, pages 344–353, 2019. 2

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE T-PAMI*, 41(8):1979–1993, 2018. 2

[27] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *ECCV*, pages 645–662, 2022. 2, 3

[28] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Semi-supervised temporal action detection with proposal-free masking. In *ECCV*, 2022. 1, 3, 6, 8

[29] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, pages 5734–5743, 2017. 2

[30] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 3

[31] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *ICCV*, pages 13526–13535, 2021. 2

[32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, pages 1196–1205, 2017. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2, 6

[34] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 6

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 6

[36] Qiang Wang, Yanhao Zhang, Yun Zheng, and Pan Pan. RCL: Recurrent continuous localization for temporal action detection. In *CVPR*, pages 13566–13575, 2022. 2, 4

[37] 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, 3, 6

[38] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, pages 8688–8696, 2018. 2

[39] Kun Xia, Le Wang, Yichao Shen, Sanpin Zhou, Gang Hua, and Wei Tang. Exploring action centers for temporal action localization. *IEEE T-MM*, 2023. 2

[40] Kun Xia, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. Dual relation network for temporal action localization. *PR*, 129:108725, 2022. 2

[41] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. Learning to refactor action and co-occurrence features for temporal action localization. In *CVPR*, pages 13874–13883, 2022. 2

[42] 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. 2

[43] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5783–5792, 2017. 2

[44] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, pages 3060–3069, 2021. 2

[45] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-graph localization for temporal action detection. In *CVPR*, pages 10156–10165, 2020. 2, 3

[46] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *CVPR*, pages 5941–5950, 2021. 2

[47] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7094–7103, 2019. 3

[48] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional module for temporal action localization in videos. *IEEE T-PAMI*, 44(10):6209–6223, 2022. 5

[49] Chenlin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510, 2022. 2, 3, 6, 7

[50] Can Zhang, Tianyu Yang, Junwu Weng, Meng Cao, Jue Wang, and Yuexian Zou. Unsupervised pre-training for temporal action localization tasks. In *CVPR*, pages 14031–14041, 2022. 2

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 6

[52] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *ICCV*, pages 13658–13667, 2021. 2, 4

[53] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, pages 539–555, 2020. 4

[54] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 2

[55] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *ECCV*, pages 35–50, 2022. 2