Rethinking Pseudo-Label Guided Learning for Weakly Supervised Temporal Action Localization from the Perspective of Noise Correction

Quan Zhang[1*], Yuxin Qi[2,3*], Xi Tang[1], Rui Yuan[1], Xi Lin[2,3], Ke Zhang[4†], Chun Yuan[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai, China
[4]School of Computer Science and Technology, Soochow University, Suzhou, China
{zhangqua22, tangx24, yuanr22}@mails.tsinghua.edu.cn, {qiyuxin98,linxi234}@sjtu.edu.cn, kzhang19@suda.edu.cn, yuanc@sz.tsinghua.edu.cn

Abstract

Pseudo-label learning methods have been widely applied in weakly-supervised temporal action localization. Existing works directly utilize weakly-supervised base model to generate instance-level pseudo-labels for training the fully-supervised detection head. We argue that the noise in pseudo-labels would interfere with the learning of fully-supervised detection head, leading to significant performance leakage. Issues with noisy labels include:(1) inaccurate boundary localization; (2) undetected short action clips; (3) multiple adjacent segments incorrectly detected as one segment. To target these issues, we introduce a two-stage noisy label learning strategy to harness every potential useful signal in noisy labels. First, we propose a frame-level pseudo-label generation model with a context-aware denoising algorithm to refine the boundaries. Second, we introduce an online-revised teacher-student framework with a missing instance compensation module and an ambiguous instance correction module to solve the short-action-missing and many-to-one problems. Besides, we apply a high-quality pseudo-label mining loss in our online-revised teacher-student framework to add different weights to the noisy labels to train more effectively. Our model outperforms the previous state-of-the-art method in detection accuracy and inference speed greatly upon the THUMOS14 and ActivityNet v1.2 benchmarks.

Introduction

Temporal action localization (TAL) (Chao et al. 2018) is a significant task in video comprehension, which aims to locate the start and end times of target action instances within untrimmed videos and identify their categories. It has shown great potential in tasks like intelligent surveillance, video summarization, etc. Inspired by the widespread application of deep learning methods across various fields(Chen et al. 2024; Zhang et al. 2024b; Lu et al. 2024b,a,c), temporal action localization has widely adopted deep learning-based approaches, which has sparked significant attention. Although the TAL problem has garnered much attention, current prominent methods often rely on costly, labor-intensive temporal boundary annotations, hindering the widespread applications of TAL.

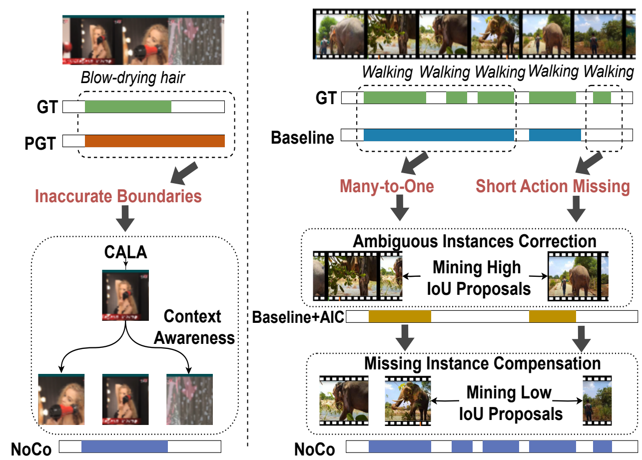

Figure 1: Noise correction modeling for WTAL. NoCo introduces pseudo-label noise correction modules to address the typical failures in existing WTAL methods.

To reduce reliance on instance-level labels, Weakly Supervised Temporal Action Localization (WTAL) solely based on video-level labels (Chen et al. 2022; Zhang and Qi 2024) has gained increasing attention. WTAL only requires knowledge of which actions occurred in the video, without the need for precise location information.

Existing WTAL works mainly follow per-category localization pipelines (Wang et al. 2017; Zhao et al. 2022), using video-level classifiers trained on class labels (Narayan et al. 2021) to filter each video clip. However, due to the absence of refinement mechanisms, these models attribute high confidence to incorrect segments, such as background elements, leading to imprecise localization. Many studies (Liu, Jiang, and Wang 2019; Liu et al. 2021a,b) attempt to address the disparity between classification and localization, with a promising solution being the generation and utilization of pseudo-labels which benefits from utilizing clip-level supervision instead of video-level labels. Existing works (Yang et al. 2021c; Luo et al. 2020; Pardo et al. 2021; Zhai

*These authors contributed equally.
†Corresponding author.

et al. 2020) have achieved remarkable results by introducing pseudo-labels on WTAL.

However, due to the sparsity of video-level label supervision information, there exists a large amount of noises in pseudo-labels. We argue that these noises can be classified into three types: 1) Inaccurate instance boundaries; 2) Missing action instances; 3) Adjacent instances aggregated into one instance, as shown in Figure 1.

To mitigate the impact of noisy pseudo-labels on WTAL, we reevaluate the pseudo-label learning process from the perspective of noise correction. We propose NoCo, a weakly-supervised temporal action localization method, which trains a weakly-supervised pseudo-label generator first, and then incorporate proposed teacher-student-based online noise learning framework over noisy pseudo-labels. In practice, there exists two issues in the above two stages. (1) How to generate accurate pseudo-label. (2) How to implement the online noise correction framework.

To generate accurate pseudo-labels, we argue that using Non-Maximum Suppression (NMS) (Neubeck and Van Gool 2006) loses the contextual information, leading to inaccurate instance boundaries. In the pseudo-label generation process, contextual information serves as a valuable supplementary source. Motivated by this, we introduce the Context-Aware Label Augmentation module (CALA), which adaptively perceives the contextual information for each pseudo-label and and corrects them through weighted averaging. Such augmentation not only yields more accurate action boundaries but also provides a more reasonable initial pseudo-label set for subsequent noise correction framework.

To realize online noise correction, we introduce a strategy that combines noise correction with teacher-student training. This approach leverages the potential of the teacher-student framework in noise learning, allows for the convenient addition of modules tailored for WTAL, and ensures high inference efficiency. Different from the previous teacher-student strategy, our approach introduces several key innovations. Firstly, to generate robust and stable pseudo-label compensation information, our teacher model employs weighted aggregation of multiple historical student models. Secondly, we propose a Missing Instance Compensation module (MIC) to solve the short clips missing problem by identifying compensation instances with low intersection-over-union (IoU) ratio from the teacher prediction information. Thirdly, the Ambiguous Instance Correction module (AIC) is introduced to address the issue of aggregating adjacent instances into one by mining the high IoU ratio instance from the teacher's online pseudo-labels. Lastly, a High-quality Pseudo-label Mining loss (HPM) is proposed to construct adaptive optimization weights, fully leveraging high-quality supervision information.

It should be noted that our noise correction framework is decoupled from the WTAL base model. The inference process of NoCo does not depend on the WTAL base model.In this paper, we build a student model based on the efficient TAL model TriDet (Shi et al. 2023), and during the inference phase, only the student model needs to be run.

In summary ,the main contributions are listed as follows.

- We revisit the WTAL process from the perspective of noise correction and propose NoCo, a noise learning framework aimed at gradually mitigating the pseudo-label noise effect in TAL. We observe three types of noise information present in the pseudo-label learning method.
- We propose a Context-aware Label Augmentation module aimed at rectifying inaccurate action boundaries. We present a teacher-student-based online noise correction framework to address short-action omissions and many-to-one problems, encompassing Missing Instance Compensation, Ambiguous Instance Correction , and High-quality Pseudo-label Mining.
- We conduct extensive experiments to demonstrate the effectiveness of NoCo. Our methods bring a significant improvement for WTAL baselines on two benchmarks and achieve state-of-the-art performance. Additionally, NoCo demonstrates fast inference speed.

Related Work

Weakly-supervised temporal action localization aims to determine the start and end times, as well as the category, of specified action instances in untrimmed videos using only video-level labels. Unlike traditional fully-supervised methods, it doesn't rely on costly instance-level labels, making it a widely studied approach(Paul, Roy, and Roy-Chowdhury 2018). Existing weakly-supervised methods fall into four categories: a) Multi-instance Learning: UntrimmedNet(Wang et al. 2017) was first applied to WTAL, generating action proposals which were then classified. However, such methods tend to focus on the most discriminative regions due to the difference between classification and localization tasks. b) Metric Learning: WTALC(Paul, Roy, and Roy-Chowdhury 2018) utilizes metric learning to make features of the same class similar. c) Attention-Based Learning: STPN(Nguyen et al. 2018) introduced a sparse loss for attention sequences, capturing key foreground segments. d) Pseudo-label Learning: TSCN(Zhai et al. 2020) generates segment-level pseudo-labels by fusing attention from RGB and FLOW modalities,while UGCT (Yang et al. 2021c) considers their complementarity and ASM-Loc (He et al. 2022) enhances features with action proposals. Previous pseudo-label learning methods have made great progress, but did not consider the existence of noise in pseudo-labels.We are the first to improve pseudo-label learning methods from the perspective of noise correction.

Pseudo Label Guided Training has been widely applied in visual tasks with weak or limited supervision. In weakly-supervised object detection, one common approach is self-training (Ren et al. 2020). This method involves training an teacher model first and using high-confidence predictions to generate instance-level pseudo-labels. These pseudo-labels are then utilized to train the final detector. Similarly, in semi-supervised learning (Weng et al. 2022) and domain adaptation (Das and Lee 2018), models are first trained on labeled/source datasets and then used to generate pseudo-labels for unlabeled/target datasets, guiding the training process. Similar to these works, our method utilizes segment-level pseudo-labels to guide training process for the WTAL

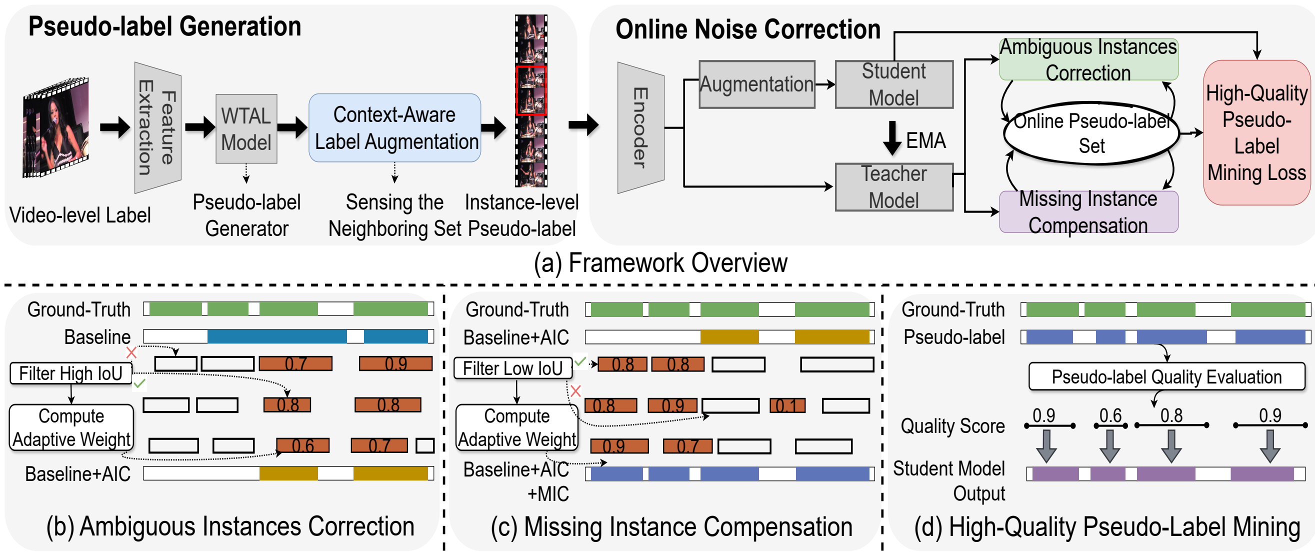

Figure 2: (a). Framework Overview. The dark gray modules indicate the base method (eg. WTAL model and augmentation), while the others are our noise correction modules. (b). Ambiguous Instances Correction is based on teacher predictions to accurate instance boundaries through mining high IoU samples and adaptively aggregating context. It is able to solve many-to-one and in-accurate boundary position problem. (c). Missing Instance Compensation focuses on adding missing instance based on teacher predictions. It aims to solve short action missing problem. (d). High-Quality Pseudo-label Mining Loss assigns adaptive weights for each noisy action instance to mine high-quality pseudo-label.

task. The difference is that we notice noises in the pseudo-label, and progressively correct these noises.

Learning with Noisy Labels has been widely applied in image classification tasks. To effectively leverage noise labels to guide the training process, DivideMix (Li, Socher, and Hoi 2020) uses two networks for sample selection via binary mixture modeling. Price et al. (Pleiss et al. 2020) introduced the marginal area statistic, measuring the average difference between a sample's assigned class logit and its highest unspecified class logit, to separate correctly from incorrectly labeled data. Liu et al. (Liu et al. 2020) found the model initially predicts ground-truth labels but eventually memorizes wrong ones, motivating exploiting the model's early output. Motivated by these works, we propose noisy label learning framework for TAL task.

Methods

In this section, we introduce the proposed NoCo framework in detail. First, we present the problem definition and describe the framework overview. Then we present the main components of NoCo in detail.

Problem Definition. Given an untrimmed training video V_i and its video-level label $y_i \in \{0,1\}^C$, where y_i is an C-dimension one-hot vector indicating category. The goal of the inference phase is to predict an action instance set $\mathcal{A} = \{(c_i, \theta_i, s_i, e_i)\}$, where c_i denotes the category of the instance $A_i \in \mathcal{A}$, θ_i is the confidence score, and s_i and e_i denote the start and end timestamp.

Framework Overview. As illustrated in Figure 2, our noise correction framework comprises two main stages. Firstly, the WTAL Base Model generates noisy pseudo-labels and then CALA is employed for context-aware enhancement of noisy pseudo-labels to achieve more accurate boundaries. Secondly, the enhanced pseudo-labels are input into the online noise correction framework as the initial pseudo-label set.

In the online correction framework, the teacher model is aggregated from various historical versions of student models, providing stable compensatory information for noise correction. The Missing Instance Compensation module (MIC) identifies action instances with low IoU with the pseudo-label, generating compensatory action instances. Simultaneously, the Ambiguous Instance Correction module (AIC) discovers action instances with high IoU with the pseudo-label, creating an associated action instance set. Then this associated set is used to rectify over-complete action instances in pseudo-labels, aligning them with corrected instances. Adjacent missing instances are compensated through MIC. The High-Quality Pseudo-Label Mining Loss (HPM), combined with the associated instance set, adaptively mines high-quality pseudo-labels, assigning greater optimization weights to these labels.

Noisy Label Generator

We first train the WTAL model based on video-level labels to obtain instance-level pseudo-labels. To refine the boundary of inaccurate instance of pseudo-labels, we propose CALA to enhance localization through context aug-

Algorithm 1: Context-Aware Label Augmentation (CALA)

Input: Weakly-supervised output $\mathcal{P}_w$, NMS threshold ρ, Confidence threshold ψ, IoU threshold η_0, Weight Aggregation Function $\mathcal{F}_{fuse}^{aug}(\cdot,\cdot)$

```
A^aug, W^aug = φ
P_filt^aug = {p_t | p_t ∈ P_w & p_t(θ_{p_t}) > ψ}
index_keep^aug = NMS(P_filt^aug, ρ)
P_keep^aug = P_filt^aug[index_keep^aug, :]
for every P_a ∈ P_keep^aug do
    P_{a,ass}, W_{a,ass} = φ
    for every P_w ∈ P_w do
        if P_a(c_a) = P_w(c_w) and IoU(P_a, P_w) > η_0 then
            # every associated action instance of P_a
            Calculate adaptive weights w_{a,w}
            P_{a,ass} = P_{a,ass} ∪ P_w, W_{a,ass} = W_{a,ass} ∪ w_{a,w}
        end if
    end for
    if P_{a,ass} is not φ then
        P_a^aug, w_a^aug = F_fuse^aug(P_{a,ass}, W_{a,ass})
    end if
    A^aug = A^aug ∪ P_a^aug, W^aug = W^aug ∪ w_a^aug
end for
return Augmented Pseudo Ground-Truth A^aug and Weight W^aug.
```

mentation. The detail is shown as follows.

WTAL Base Model Following the previous work (Pardo et al. 2021; He et al. 2022), for each video V_i, we use the features $F \in R^{2048}$ extracted by the I3D network (Carreira and Zisserman 2017), where the I3D network is pre-trained on the Kinetics400 dataset. Using video-level labels, we train a WTAL via which we can obtain a noisy pseudo-label for each video sample. In this paper, the WTAL we use is ASM-Loc (He et al. 2022) if not specified. Note that our method is decoupled from the WTAL model and can be easily cooperated to other weakly-supervised models. In the experiment section, we conduct an analysis of the generalizability for WTAL models.

Context-Aware Label Augmentation Due to the sparsity of video-level label information, there exists non-negligible noises in the obtained pseudo-labels.

We propose the Context-Aware Label Augmentation (CALA) algorithm to refine instance pseudo-labels and reduce noise by considering neighboring instances. The CALA process is outlined in Algorithm 1.

First, initialize the enhanced pseudo-labels and their corresponding weights. By setting a confidence threshold and deploying NMS processing, a set of high-confidence and low-overlapping action instances P_{keep}^{aug} is selected. The context enhancement process is executed for each action instance P_a in $\mathcal{P}_{keep}^{aug}$, including determining the associated action instance set $\mathcal{P}_{a,ass}$ and calculating the adaptive weights. The constraint for the associated action instance is: a) sharing the same category; b) having an IoU greater than a given threshold η_0. The formula for calculating the adaptive weight of P_a's associate instance P_r is as follows:

$$w_r = e^{\sqrt{\mathrm{IoU}(P_r,\mathcal{P}_{keep}^{aug})\cdot\min(\max(P_r(\theta_r),0),1)}}, \tag{1}$$

where $P_r(\theta_r)$ is the confidence score of $P_r \in \mathcal{P}_{a,ass}$. If $\mathcal{P}_{a,ass}$ is not empty, then the augmented P_a^{aug} of P_a is obtained via weighted aggregation of associated action instances set $\mathcal{P}_{a,ass}$, which is computed as follows:

$$\begin{aligned} P_a^{aug}(\zeta_a) &= \mathcal{F}_{fuse,stage}^{aug}(P_a,\mathcal{P}_{a,ass}) \\ &= \frac{\sum_{i=1}^{n} w_i \cdot P_i(\zeta_i), P_i \in \mathcal{P}_{a,ass}}{\sum_{i=1}^{n} w_i}, \end{aligned} \tag{2}$$

$$\text{where } \zeta_i = \begin{cases} s_i & stage = start \\ e_i & stage = end, \end{cases} \tag{3}$$

weight aggregation function $\mathcal{F}_{fuse}^{aug}(\cdot,\cdot)$ consists of two stage: start timestamp aggregation $\mathcal{F}_{fuse,start}^{aug}(\cdot,\cdot)$ and end timestamp aggregation $\mathcal{F}_{fuse,end}^{aug}(\cdot,\cdot)$. The Weight of P_a^{aug} is computed as $w_a^{aug} = \sqrt[n]{\prod_{i=1}^{n} w_i}$. Then add the P_a^{aug} to the augmented pseudo-label set $\mathcal{A}^{aug}$ and repeat this process till all action instance is scanned.

Online Noise Correction Framework

The online noise correction framework aims to progressively correct three types of noise: missing action instances, many-to-one and incomplete instance problem. Specifically, we introduce a teacher-student-based online noise correction framework. As illustrated in Figure 2, our online noise correction framework consists of several components: feature extraction encoder, data augmentation module, student model, teacher model, AIC, MIC, online pseudo-label set $\mathcal{A}^*$ and HPM.

Teacher Model Pre-training The online noise correction process begins with the pre-training of the teacher model, involving using the enhanced pseudo-labels generated by CALA for full supervised training. After completing pre-training, the teacher model plays a crucial role in the online generation of stable compensation information during the noise correction stage, aiding the processes of AIC and MIC. The online pseudo-label $\mathcal{A}^*$ and weight $\mathcal{W}^*$ is first initialized as $\mathcal{A}^{aug}$ and $\mathcal{W}^{aug}$.

Ambiguous Instances Correction Ambiguous Instances Correction (AIC) aims to leverage the pseudo-label information derived from teacher predictions to accurately position instance boundaries through adjacent context aggregation, thus enabling the online update of the pseudo-label set. For low-quality samples, the output from the teacher model is further corrected through the AIC module, effectively suppressing noise in the pseudo-labels. Compared with traditional semi-supervised learning, we possess not only the output of the teacher model but also maintain an online, realtime corrected pseudo-label set. We refer to **Supplementary** for the workflow of AIC.

The input to the AIC consists of teacher predictions $\mathcal{P}_w$ and the online pseudo-label set $\mathcal{A}^*$. For each instance A_r^* in the $\mathcal{A}^*$, we calculate its IoU with every action instance $P_t \in \mathcal{P}_w$. If the IoU is greater than the threshold η_1, and the confidence score of the action instance exceeds the threshold ψ, we include it in the corresponding associated set $\mathcal{A}_{r,ass}$ and compute its adaptive weight. The formula for calculating the adaptive weight w_t of A_r^*'s associated instance P_t is the same as the one proposed in the CALA algorithm, as described above in equation (1). By combining $\mathcal{A}_{r,ass}$ and A_r^*, we perform weighted aggregation to obtain the corrected action instance A_r^{cor} of A_r^* as follows:

$$\begin{aligned} A_r^{cor}(\zeta_r) &= \mathcal{F}_{fuse,stage}^{cor}(A_r^*, \mathcal{A}_{r,ass}) = \alpha \cdot w_r^* \cdot A_r^*(\zeta_r) \\ &+ (1-\alpha) \cdot \frac{\sum_{i=1}^{n} w_i \cdot A_i(\zeta_i), A_i \in \mathcal{A}_{r,ass}}{\sum_{i=1}^{n} w_i}, \end{aligned} \tag{4}$$

$$\text{where } \zeta_i = \begin{cases} s_i & stage = start \\ e_i & stage = end, \end{cases} \tag{5}$$

where α is hyper-parameter, w_r^* is the weight of instance A_r^*'s weight in $\mathcal{W}^*$, n is the size of $A_{r,ass}$.The corrected weight of A_r^* is computed as:

$$w_r^{cor} = \sqrt[n+\beta]{(\prod_{i=1}^{n} w_i) \cdot (w_r^*)^{\beta}}, \quad A_i \in \mathcal{A}_{r,ass}, \tag{6}$$

where β is hyper-parameter. Then add the corrected instance A_r^{aug} of A_r^* to the corrected pseudo-label set $\mathcal{A}^{cor}$, and add w_r^{cor} to corrected weight set $\mathcal{W}^{cor}$.

When all instance in $\mathcal{A}^*$ is corrected, the online set $\mathcal{A}^*$ is updated as $\mathcal{A}^{cor}$. The online weight $\mathcal{W}^*$ is updated as $\mathcal{W}^{cor}$. Through this updating, the $\mathcal{A}^*$ and $\mathcal{W}^*$ is corrected.

Missing Instance Compensation MIC aims to further update the pseudo-label set online, leveraging the potential missing information in neighboring pseudo-label instances, which are generated based on teacher predictions. Although AIC effectively utilizes contextual information within the pseudo-labels, it only focuses on correcting existing action instances in noisy pseudo-labels and does not address the task of compensating missing action instances. We refer to **Supplementary** for the main workflow of the MIC.

MIC takes teacher predictions $\mathcal{P}_t$ and the online pseudo-label set $\mathcal{A}^*$ as inputs. First, MIC filters out action instances with low confidence from $\mathcal{P}_t$, saving as $\mathcal{P}_{filt}^{com}$. Then it applies NMS to select low IoU instance on $\mathcal{P}_{filt}^{com}$ to obtain $\mathcal{P}_{keep}^{com}$. Next, for each instance P_n in $\mathcal{P}_{keep}^{com}$, MIC computes the IoU with every action instance in $\mathcal{A}^*$. If the maximum IoU is less than η_2, P_n is included in the compensated pseudo-label set $\mathcal{A}^{com}$. The adaptive weight w_n^{com} of $A_n \in \mathcal{A}^{com}$ is obtained through:

$$w_n^{com} = e^{\min(\max(A_n(\theta_n),0),1)}, \tag{7}$$

where $A_n(\theta_n)$ is the confidence score of A_n. w_n^{com} is added into compensated weight $\mathcal{W}^{com}$.

Through the MIC module, missing action instances are filled. Similar before, we update A^* as $\mathcal{A}^{com}$, $\mathcal{W}^*$ as $\mathcal{W}^{com}$ again to achieve online label compensation.

High-Quality Pseudo-Label Mining To fully utilize the supervision information from accurate pseudo-labels, we introduce adaptive weights for each action instance by assigning higher optimization weights to high-quality instances and vice versa. Specially, we propose the High-Quality Pseudo-Label Mining (HPM) Loss, which is formulated as:

$$\begin{aligned} L &= \frac{\lambda}{N_{pos}} \sum_{l,t} \mathbb{1}_{\{c_t^l > 0\}} \cdot w_t \cdot (\sigma_{IoU} L_{cls} + L_{reg}) \\ &+ \frac{1}{N_{reg}} \sum_{l,t} \mathbb{1}_{\{c_t^l = 0\}} \cdot L_{cls}, \end{aligned} \tag{8}$$

where λ represents the weight of positive sample loss, N_{pos} and N_{neg} denote the number of positive and negative samples, and l stands for the number of classification heads and Trident-head heads. For each classification head and Trident-head, we compute foreground classification loss L_{cls}, regression loss L_{reg}, and background classification loss sequentially. In this paper, L_{cls} is focal loss and L_{reg} is IoU loss. It's worth noting that using adaptive weights w_t for each foreground action instance gives more importance to high-quality instances during the optimization process.

Experiment

Settings and Datasets

Datasets. We evaluate our method on THUMOS14 and ActivityNet v1.2, following (Yang et al. 2021b). **THUMOS14** contains 200 validation and 213 test samples. The validation set is used for training, and the test set for evaluation. **ActivityNet v1.2** includes 4,819 training, 2,383 validation, and 2,489 test videos across 100 action classes, averaging 1.5 instances per video. We use the training set for training and the validation set for evaluation.

Implementation. We use the I3D model for feature extraction and ASM-Loc (He et al. 2022) as the default noise label generator unless stated otherwise. Additional dataset-specific details are in the supplementary material. We follow standard protocols with IoU thresholds to measure localization accuracy and report mIoU for pseudo-label quality across foreground and background categories.

Comparison with State-Of-Art Methods

We compare NoCo with the existing methods on ActivityNet v1.2 and THUMOS14.The result is shown in Table 1.

Results on THUMOS14. NoCo outperforms all WTAL baseline methods. Its average mAP is 6.1% higher than CO2-Net and 5.8% higher than ASM-Loc . Additionally, NoCo surpasses the latest methods CASE , AHIM , DDG-Net , and the current leading method PivoTAL, pushing the average mAP to 50.9%, setting a new state-of-the-art performance. Even when compared with FTAL methods, NoCo outperforms TAL-Net at all IoUs. Furthermore, we also applied NoCo to point-supervised temporal action localization methods LACP and HR-Pro, achieving 3.4% and 0.9% average mAP improvements over the baselines, respectively, resulting in current state-of-the-art performance.

Results on ActivityNet v1.2. Our method outperforms existing methods across all metrics. Compared to the baseline ASM-Loc, our method improves mAP by 4.2%, and compared to the baseline CO2-Net, it improves by 3.7%. NoCo also surpasses Ju et al.(Ju et al. 2023), achieving a new state-of-the-art localization accuracy of 30.7%. Our method

Sup.	Method	mAP@IoU(%)							AVG mAP(%)		
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.5	0.3:0.7	0.1:0.7
Fully	SSN (Zhao et al. 2017)	60.3	56.2	50.6	40.8	29.1	-	-	47.4	-	-
	TAL-Net(Chao et al. 2018)	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
Point-level	SF-Net (Ma et al. 2020)	68.3	62.3	52.8	42.2	30.5	20.6	12.0	51.2	31.6	41.2
	DCM (Ju et al. 2021)	70.2	63.5	55.6	44.7	32.3	22.0	12.3	53.3	33.4	42.9
	BackTAL(Yang et al. 2021a)	-	-	54.4	45.5	36.3	26.2	14.8	-	35.4	-
	PCL (Li, Cao, and Ye 2023)	74.6	70.2	63.3	55.9	44.4	-	-	61.7	-	-
	LACP (Lee and Byun 2021)	75.7	71.4	64.6	56.5	45.3	34.5	21.8	62.7	44.5	52.8
	LACP+NoCo	78.8	74.6	68.9	60.1	49.6	37.2	24.2	66.4	48.0	56.2
	HR-Pro (Zhang et al. 2024a)	85.6	81.6	74.3	64.3	52.2	39.8	24.8	71.6	51.1	60.4
	HR-Pro+NoCo	**86.0**	**82.1**	**75.6**	**65.7**	**53.6**	**40.6**	**25.3**	**72.6**	**52.2**	**61.3**
Video-level	CoLA (Zhang et al. 2021)	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	DELU (Chen et al. 2022)	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.5	37.4	46.4
	Zhou et al. (Zhou et al. 2023)	74.0	69.4	60.7	51.8	42.7	26.2	13.1	59.7	38.9	48.3
	Ju et al. (Ju et al. 2023)	73.5	68.8	61.5	53.8	42.0	29.4	16.8	59.9	40.7	49.4
	PivoTAL (Rizve et al. 2023)	74.1	69.6	61.7	52.1	42.8	30.6	16.7	60.1	40.8	49.7
	CASE (Liu et al. 2023)	72.3	67.1	59.2	49.4	37.7	24.2	13.7	57.1	36.8	46.2
	ISSF (Yun et al. 2024)	72.4	66.9	58.4	49.7	41.8	25.5	12.8	57.8	37.6	46.8
	CO_2-Net (Hong et al. 2021)	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
	CO_2-Net+NoCo	74.6	70.3	63.0	53.9	44.0	31.7	17.4	61.2	42.0	50.7
	ASM-Loc (He et al. 2022)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	ASM-Loc+NoCo	**75.2**	**70.7**	**63.0**	**54.1**	**44.0**	**31.7**	**17.7**	**61.4**	**42.1**	**50.9**

Table 1: Comparisons with state-of-the-art methods. AVG is the average mAP under the IoU thresholds 0.1:0.7:0.1 for THUMOS14 dataset. '-' means that the corresponding results are nor reported in the original papers.

Sup.	Method	mAP@IoU(%)			
		0.5	0.75	0.95	AVG
Fully	SSN (Zhao et al. 2017)	41.3	27.0	6.1	26.6
Point-level	SF-Net (Ma et al. 2020)	37.8	-	-	22.8
	BackTAL (Yang et al. 2021a)	41.5	27.3	4.7	27.0
	BackTAL+NoCo	49.8	31.6	6.1	30.8
	LACP (Lee and Byun 2021)	44.0	26.0	5.9	26.8
	LACP+NoCo	**50.2**	**30.1**	**6.8**	**30.7**
Video-level	CoLA (Zhang et al. 2021)	42.7	25.7	5.8	26.1
	DELU (Chen et al. 2022)	44.2	26.7	5.4	26.9
	Ren et al. (Ren et al. 2023)	44.2	26.1	5.3	25.5
	Ju et al. (Ju et al. 2023)	48.3	29.3	6.1	29.6
	CASE (Liu et al. 2023)	43.8	27.2	6.7	27.9
	CO_2-Net (Hong et al. 2021)	43.3	26.3	5.2	26.4
	CO_2-Net+NoCo	48.9	30.0	6.6	30.1
	ASM-Loc (He et al. 2022)	43.4	26.6	5.3	26.5
	ASM-Loc+NoCo	**49.1**	**30.5**	**6.7**	**30.7**

Table 2: Detection results on ActivityNet1.2

even outperforms some previous fully supervised methods, such as SSN, demonstrating the feasibility of noise correction modeling. Additionally, we applied NoCo to point-supervised temporal action localization methods BackTAL and LACP, achieving average mAP improvements of 3.8% and 3.9%, respectively, compared to the baseline.

The consistent results on ActivityNet v1.2 and THUMOS14 demonstrate the effectiveness of our NoCo.

Ablation Studies

We conducted a series of ablation experiments to evaluate the contributions of each component in our method. The results, presented in Table 3, are labeled from *a* to *g* for clarity.

#	CALA	T-S	AIC	MIC	HPM	AVG mAP(%)
a						35.8
b	✓					37.1 (+1.3)
c	✓	✓				37.8 (+2.0)
d	✓	✓	✓			40.0 (+4.2)
e	✓	✓		✓		39.6 (+3.8)
f	✓	✓	✓	✓		41.9 (+6.1)
g	✓	✓	✓	✓	✓	42.1 (+6.3)

Table 3: Ablation study of core components,'T-S' denotes the Teacher-Student module.

Baseline. *Experiment a* combines the weakly supervised model ASM-Loc with the fully supervised TriDet model for pseudo-label retraining on the THUMOS14 dataset, without any correction modules. Pseudo-labels were generated by filtering action instances with confidence scores above 0.5. However, this threshold-based method introduced substantial noise, resulting in minimal performance improvement.

Influence of CALA. *Experiment b* applies CALA, achieving a 1.3% mAP increase, demonstrating its ability to enhance pseudo-labels with contextual information.

Influence of Customized Teacher-Student Training. *Experiment c* introduces a customized teacher-student framework without pseudo-label correction modules. The teacher's labels guide the student, improving performance by 0.7% due to the student model's robustness from prior teacher-student iterations.

Influence of AIC. *Experiment d* adds AIC, boosting perfor-

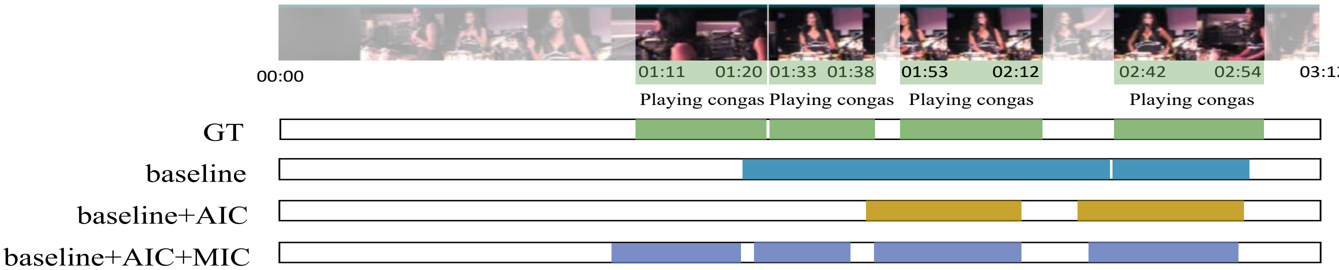

Figure 3: Visualization of ground-truth and predictions.

WTAL	CALA	T-S	MIC	AIC	HPM	AVG mAP(%)
UM						30.3
	✓					31.6 (+1.3)
	✓	✓				32.3 (+2.0)
	✓	✓	✓			33.9 (+3.6)
	✓	✓	✓	✓		36.1 (+5.8)
	✓	✓	✓	✓	✓	**36.4 (+6.1)**
CO_2-Net						36.0
	✓					37.0 (+1.0)
	✓	✓				37.7 (+1.7)
	✓	✓	✓			40.2 (+4.2)
	✓	✓	✓	✓		41.8 (+5.8)
	✓	✓	✓	✓	✓	**42.0 (+6.0)**

Table 4: Performances under two typical WTAL algorithms on THUMOS14 dataset.

mance by 2.2%. AIC corrects inaccurately localized pseudo-labels, refining the student model's learning.

Influence of MIC. *Experiment e* incorporates MIC, resulting in a 1.8% improvement. MIC compensates for missing action instances in pseudo-labels, enhancing learning.

Influence of Combined MIC and AIC. *Experiment f* applies both MIC and AIC, yielding a 4.1% performance increase over *experiment c*. AIC corrects overly complete instances, while MIC fills in missing actions, addressing the one-to-many issue.

Influence of High-Quality Pseudo-Label Mining. *Experiment g* adds high-quality pseudo-label mining loss, providing a 0.2% performance boost and reaching the highest mAP of 42.1%.

Generalizability Analysis of WTAL Models

Our noise correction framework is decoupled from weakly supervised models, making it applicable to any existing temporal action localization method. To evaluate its generalizability, we apply our framework to the UM (Lee et al. 2021) and CO_2-Net (Hong et al. 2021) models, with results shown in Table 4.

For the UM model, the average mAP on THUMOS14 is 30.3%. The CALA module improves this by 1.3%, and the Teacher-Student module contributes an additional 2.0%. The MIC, AIC, and HPM modules further increase the performance by 3.6%, 5.8%, and 6.1%, respectively.

Similarly, for CO_2-Net, the mAP starts at 36.0%. The sequential introduction of the CALA, Teacher-Student, MIC, AIC, and HPM modules results in improvements of 1.0%, 1.7%, 4.2%, 5.8%, and 6.0%, respectively.

These results highlight the strong generalizability of our noise correction framework across various weakly-supervised models.

	mAP(0.3:0.7)↑	speed(video/s)↑	FLOPS↓
Baseline	35.8	1.64	124.79G
NoCo	42.1	6.67	20.12G

Table 5: NoCo outperforms existing methods in terms of both localization average mAP and inference speed.

Analysis of Inference Speed

NoCo gains a speed advantage by running only the student model during inference. We compare the inference performance and speed in Table 5. It can be observed that NoCo is not only faster than the baseline WTAL method, showing a threefold improvement, but also achieves more accurate localization results.

Qualitative Results

Figure 3 illustrates the visual comparison between the base model and NoCo for the action *Playing congas*. We observe that when introducing AIC, the action boundary is more precise compared to the baseline. Furthermore, with the addition of MIC, missing action instances are compensated for, and many-to-one problem is solved. These results provide evidence for the effectiveness of proposed AIC and MIC.

Conslusion

In this paper, we propose a progressive framework for correcting noisy pseudo-labels. We introduce a CALA module to obtain more accurate action boundaries. Subsequently, we establish an online noise correction framework based on a teacher-student training strategy. This framework incorporates the AIC and MIC module to tackle the instances-missing and many-to-one problem. Furthermore, we introduce an HPM loss to discover high-quality pseudo-labels. Our approach achieves state-of-the-art performance on THUMOS14 and ActivityNet v1.2, with a threefold improvement in inference speed.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012,KJZD20230923114916032, GJHZ20240218113604008), Beijing Key Lab of Networked Multimedia and National Natural Science Foundation of China under Grant 62202302.

References

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1130–1139.

Chen, M.; Gao, J.; Yang, S.; and Xu, C. 2022. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, 192–208. Springer.

Chen, Y.; Huang, X.; Zhang, Q.; Li, W.; Zhu, M.; Yan, Q.; Li, S.; Chen, H.; Hu, H.; Yang, J.; et al. 2024. GIM: A Million-scale Benchmark for Generative Image Manipulation Detection and Localization. *arXiv preprint arXiv:2406.16531*.

Das, D.; and Lee, C. G. 2018. Graph matching and pseudo-label guided deep unsupervised domain adaptation. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, 342–352. Springer.

He, B.; Yang, X.; Kang, L.; Cheng, Z.; Zhou, X.; and Shrivastava, A. 2022. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13925–13935.

Hong, F.-T.; Feng, J.-C.; Xu, D.; Shan, Y.; and Zheng, W.-S. 2021. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM international conference on multimedia*, 1591–1599.

Ju, C.; Zhao, P.; Chen, S.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13455–13464.

Ju, C.; Zheng, K.; Liu, J.; Zhao, P.; Zhang, Y.; Chang, J.; Tian, Q.; and Wang, Y. 2023. Distilling Vision-Language Pre-training to Collaborate with Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14751–14762.

Lee, P.; and Byun, H. 2021. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13648–13657.

Lee, P.; Wang, J.; Lu, Y.; and Byun, H. 2021. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1854–1862.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, P.; Cao, J.; and Ye, X. 2023. Prototype contrastive learning for point-supervised temporal action detection. *Expert Systems with Applications*, 213: 118965.

Liu, D.; Jiang, T.; and Wang, Y. 2019. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1298–1307.

Liu, Q.; Wang, Z.; Rong, S.; Li, J.; and Zhang, Y. 2023. Revisiting Foreground and Background Separation in Weakly-supervised Temporal Action Localization: A Clustering-based Approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10433–10443.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, Z.; Wang, L.; Tang, W.; Yuan, J.; Zheng, N.; and Hua, G. 2021a. Weakly supervised temporal action localization through learning explicit subspaces for action and context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2242–2250.

Liu, Z.; Wang, L.; Zhang, Q.; Tang, W.; Yuan, J.; Zheng, N.; and Hua, G. 2021b. Acsnet: Action-context separation network for weakly supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2233–2241.

Lu, F.; Dong, S.; Zhang, L.; Liu, B.; Lan, X.; Jiang, D.; and Yuan, C. 2024a. Deep Homography Estimation for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10341–10349.

Lu, F.; Lan, X.; Zhang, L.; Jiang, D.; Wang, Y.; and Yuan, C. 2024b. CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16772–16782.

Lu, F.; Zhang, L.; Lan, X.; Dong, S.; Wang, Y.; and Yuan, C. 2024c. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*.

Luo, Z.; Guillory, D.; Shi, B.; Ke, W.; Wan, F.; Darrell, T.; and Xu, H. 2020. Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning. In *European Conference on Computer Vision*, 729–745.

Ma, F.; Zhu, L.; Yang, Y.; Zha, S.; Kundu, G.; Feiszli, M.; and Shou, Z. 2020. Sf-net: Single-frame supervision for temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 420–437. Springer.

Narayan, S.; Cholakkal, H.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13608–13617.

Neubeck, A.; and Van Gool, L. 2006. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, 850–855. IEEE.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6752–6761.

Pardo, A.; Alwassel, H.; Caba, F.; Thabet, A.; and Ghanem, B. 2021. Refineloc: Iterative refinement for weakly-supervised action localization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3319–3328.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 563–579.

Pleiss, G.; Zhang, T.; Elenberg, E.; and Weinberger, K. Q. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33: 17044–17056.

Ren, H.; Yang, W.; Zhang, T.; and Zhang, Y. 2023. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2394–2404.

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.-Y.; Lee, Y. J.; Schwing, A. G.; and Kautz, J. 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10598–10607.

Rizve, M. N.; Mittal, G.; Yu, Y.; Hall, M.; Sajeev, S.; Shah, M.; and Chen, M. 2023. PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22992–23002.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 4325–4334.

Weng, Z.; Yang, X.; Li, A.; Wu, Z.; and Jiang, Y.-G. 2022. Semi-supervised vision transformers. In *European Conference on Computer Vision*, 605–620. Springer.

Yang, L.; Han, J.; Zhao, T.; Lin, T.; Zhang, D.; and Chen, J. 2021a. Background-click supervision for temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9814–9829.

Yang, W.; Zhang, T.; Mao, Z.; Zhang, Y.; Tian, Q.; and Wu, F. 2021b. Multi-scale structure-aware network for weakly supervised temporal action detection. *IEEE Transactions on Image Processing*, 30: 5848–5861.

Yang, W.; Zhang, T.; Yu, X.; Qi, T.; Zhang, Y.; and Wu, F. 2021c. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 53–63.

Yun, W.; Qi, M.; Wang, C.; and Ma, H. 2024. Weakly-Supervised Temporal Action Localization by Inferring Salient Snippet-Feature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6908–6916.

Zhai, Y.; Wang, L.; Tang, W.; Zhang, Q.; Yuan, J.; and Hua, G. 2020. Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization. In *European Conference on Computer Vision*, 37–54.

Zhang, C.; Cao, M.; Yang, D.; Chen, J.; and Zou, Y. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16010–16019.

Zhang, H.; Wang, X.; Xu, X.; Qing, Z.; Gao, C.; and Sang, N. 2024a. Hr-pro: Point-supervised temporal action localization via hierarchical reliability propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7115–7123.

Zhang, Q.; Liu, X.; Li, W.; Chen, H.; Liu, J.; Hu, J.; Xiong, Z.; Yuan, C.; and Wang, Y. 2024b. Distilling Semantic Priors from SAM to Efficient Image Restoration Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25409–25419.

Zhang, Q.; and Qi, Y. 2024. Can MLLMs Guide Weakly-Supervised Temporal Action Localization Tasks? *arXiv preprint arXiv:2411.08466*.

Zhao, T.; Han, J.; Yang, L.; and Zhang, D. 2022. Equivalent classification mapping for weakly supervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3019–3031.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*, 2914–2923.

Zhou, J.; Huang, L.; Wang, L.; Liu, S.; and Li, H. 2023. Improving Weakly Supervised Temporal Action Localization by Bridging Train-Test Gap in Pseudo Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23003–23012.