Towards Adaptive Pseudo-label Learning for Semi-Supervised Temporal Action Localization

Feixiang Zhou, Bryan Williams, and Hossein Rahmani*

Lancaster University, UK
{f.zhou3,b.williams6, h.rahmani}@lancaster.ac.uk

Abstract. Alleviating noisy pseudo labels remains a key challenge in Semi-Supervised Temporal Action Localization (SS-TAL). Existing methods often filter pseudo labels based on strict conditions, but they typically assess classification and localization quality separately, leading to suboptimal pseudo-label ranking and selection. In particular, there might be inaccurate pseudo labels within selected positives, alongside reliable counterparts erroneously assigned to negatives. To tackle these problems, we propose a novel Adaptive Pseudo-label Learning (APL) framework to facilitate better pseudo-label selection. Specifically, to improve the ranking quality, Adaptive Label Quality Assessment (ALQA) is proposed to jointly learn classification confidence and localization reliability, followed by dynamically selecting pseudo labels based on the joint score. Additionally, we propose an Instance-level Consistency Discriminator (ICD) for eliminating ambiguous positives and mining potential positives simultaneously based on inter-instance intrinsic consistency, thereby leading to a more precise selection. We further introduce a general unsupervised Action-aware Contrastive Pre-training (ACP) to enhance the discrimination both within actions and between actions and backgrounds, which benefits SS-TAL. Extensive experiments on THUMOS14 and ActivityNet v1.3 demonstrate that our method achieves state-of-the-art performance under various semi-supervised settings.

Keywords: Temporal action localization · Video understanding · Semi-supervised learning · Action recognition

1 Introduction

Temporal action localization (TAL) aims to localize temporal boundaries of action instances and identify corresponding categories from an untrimmed video. Existing fully-supervised TAL methods [19, 34, 45, 52] have achieved promising performance by utilizing a large amount of labeled data. However, manually annotating temporal boundaries and class labels for large-scale datasets is very time-consuming and expensive. To address this issue, semi-supervised TAL (SS-TAL) methods [11, 26] have been proposed, where a large number of unlabeled videos and only a few labeled videos are leveraged for model training.

* Corresponding author

One of SS-TAL frameworks [11, 33, 43] normally combines existing proposal-based TAL models with semi-supervised learning (SSL) approaches. However, this strategy suffers from the error propagation problem caused by sequential localization and classification design, resulting in accumulated errors. As an alternative, a proposal-free model [26] is proposed to address this problem, which designs a parallel localization and classification architecture. Despite the performance improvement, it fails to effectively suppress noisy pseudo labels. A recent work [46] elaborates on the importance of location biases and category errors, which focuses on alleviating the label noise problem. It combines adaptively learned class scores with subsequent manually calculated location scores (*i.e.*, boundary variance) to measure the overall quality of pseudo labels but ignores the potential synergies or correlations between them due to its divergence from an end-to-end paradigm. Thus, how to effectively represent localization reliability remains an open question. More importantly, the above methods lack the ability to identify potential false positives and false negatives from pseudo labels selected based on fixed or dynamical score thresholds, which hinders the full exploitation of positive instances generated on unlabeled data.

To address the above issues, in this paper, we propose a novel Adaptive Pseudo-label Learning (APL) framework for SS-TAL. Specifically, instead of manually computing the location score based on the predicted boundaries, the proposed Adaptive Label Quality Assessment (ALQA) jointly learns classification confidence and localization reliability of action instances by adopting an end-to-end learning architecture. For localization sub-task in most one-stage TAL methods [32, 52], a Distance-IoU (DIoU) loss [56] is normally performed to regress offsets from current frames to boundaries. Intuitively, the DIoU can evaluate the temporal intersection over union (overlap rate) between two segments (*i.e.*, tIoU) and temporal normalized distance (tND) between predicted and ground-truth (GT) boundaries, which can be a natural indicator to assess localization quality. Motivated by this, we design two parallel branches to predict tIoU and tND, enabling the learning of localization reliability from different perspectives. The predictions are then dynamically divided into positives and candidates based on the joint score of classification and localization.

Apart from the ALQA, we also propose an Instance-level Consistency Discriminator (ICD) to refine the selected pseudo labels on unlabeld videos by removing ambiguous positives (pseudo labels with high joint scores but wrong categories) and mining potential positives (pseudo labels with low joint scores but correct categories) from candidates. More concretely, the ICD is first trained using all labeled action instances to encourage feature consistency between different instances. Afterwards, during inference, each instance selected by the ALQA and the corresponding labeled instances having the same category are fed into the trained ICD to yield similarity probabilities, which serve as an auxiliary similarity score to further select high-quality pseudo labels.

Self-supervised pre-training [26] has been investigated to improve the SS-TAL performance. However, this method relies on a customized and complex pretext task, which makes it difficult to be compatible with other backbones.

More importantly, it aims to distinguish between actions and backgrounds but ignores discrimination between different actions. To this end, we design a general unsupervised ACP to enhance representation learning, which consists of coarse- and fine-grained contrasts. The former is implemented by contrasting binary-class frames (0 and 1 indicate foreground (*i.e.*, action) and background in an untrimmed video) sampled from each video of a mini-batch, while the latter is performed by elaborately contrasting multi-class frames sampled from all videos of a mini-batch. In summary, the main contributions are as follows:

- We propose a SS-TAL framework APL, in which high-quality pseudo labels are adaptively selected to boost semi-supervised learning. Extensive experiments demonstrate that APL surpasses all the previous methods and achieves state-of-the-art performance.
- We propose an ALQA module to facilitate more direct interaction between classification and localization via a joint learning paradigm, ensuring a robust quality assessment of pseudo labels.
- We design an ICD for pseudo-label refinement, which aims to eliminate ambiguous positives and mine potential positives by leveraging the inherent consistency between distinct action instances.
- We introduce a unsupervised ACP to enhance frame-level representation and improve the discrimination both within actions and between actions and backgrounds.

2 Related Works

Temporal Action Localization. TAL involves simultaneously localizing and identifying action instances from untrimmed videos. Similar to the development of object detection [30, 31, 41], existing fully-supervised approaches can be divided into two categories, namely, two-stage methods [1, 20, 22, 29, 38, 49] and one-stage methods [16, 32, 34, 35, 52]. Two-stage methods in TAL typically follow a proposal-generation and action-classification paradigm. Previous two-stage methods [9, 18, 20, 22, 55] usually focused on action proposal generation, where anchor-based works [9,10,18] classify actions from specific anchor windows while boundary-based methods [20,22,55] predict the boundary probability and apply boundary matching mechanism to produce candidate proposals. Recent efforts have aimed at refining proposals by exploring temporal correlations between them using graph networks [51, 54] or self-attention mechanisms [29, 58]. Two-stage pipeline has shown effectiveness in handling complex temporal structures but may exhibit limitations in terms of computational efficiency. The one-stage methods integrate action localization and classification into a single network without using action proposals. Most previous works [19, 21] utilized convolutional neural networks (CNNs) for feature encoding. Inspired by the recent success of DETR [5], transformer-based models [34, 35, 52] have been designed for localization and classification, achieving new state-of-the-art performance.

Semi-Supervised Learning. SSL that aims to improve model generalization and performance using a small amount of labeled data and a large amount of

unlabeled data has been widely applied in various computer vision tasks, such as image classification [37, 40], action recognition or segmentation [36, 47, 57], object detection [8, 24] and semantic segmentation [28, 50]. Consistency regularization [13, 25, 39] and pseudo-labeling [14, 37] are two main paradigms in SSL. Consistency regularization methods, such as Mixmatch [2], aim to enforce consistency between predictions made on different perturbations of the same input. Pseudo-labeling methods exploit unlabeled data by training on self-generated predictions, *i.e.*, pseudo labels. However, they are susceptible to low-quality pseudo labels due to inaccuracies in the model's predictions, leading to incorrect labels being assigned to the unlabeled data. As a result, another line of work [7,8,15,23,44,48] has attempted to tackle label noise. Our method addresses this issue in SS-TAL with a novel framework, where ALQA jointly learns classification confidence and localization reliability to better evaluate pseudo-label quality, while ICD eliminates ambiguous positives and mines potential positives to refine pseudo labels.

Semi-Supervised Temporal Action Localization. Despite promising results in TAL, SS-TAL still remains insufficiently explored. A common pipeline of existing methods [11, 33, 43] is to incorporate SSL methods into fully-supervised TAL models. However, the accumulated errors caused by localization error propagation are inevitable. SPOT [26] addresses this problem with a proposal-free framework where localization and classification heads are constructed in a parallel manner. A very recent study, *i.e.*, NPL [46], aims to handle noisy pseudo labels by reranking the predictions according to the learned classification score and manually computed localization score.

3 Proposed Method

Problem Definition. In SS-TAL, we aim to train a model using a small amount of labeled videos $\{V_i\}_{i=1}^{N_l}$ and a large amount of unlabeled videos $\{U_i\}_{i=1}^{N_u}$, where N_l and N_u indicate the numbers of labeled and unlabeled videos, respectively. Each labeled video V_i is composed of a set of action instances $\mathcal{I}_i = \{I_j = (X_j, (t_{s,j}, t_{e,j}, y_j))\}_{j=1}^{M_i}$, where M_i is the number of action instances, and $t_{s,j}$, $t_{e,j}$ are the starting and ending time of the j-th action instance I_j. $y_j \in Y = \{0, 1, ..., K-1\}$ is the class label, where K denotes the number of classes. X_j represents the features of I_j. Following the common practice in previous work [46, 52], we adopt video features extracted by a pre-trained video encoder (*e.g.*, I3D network [6]) as the input of our model, and sample a fixed number of frames for training. Therefore, each video V_i can be represented as $V_i \in \mathbb{R}^{D \times T}$, where D is the feature dimension and T is the total number of frames.

Approach Overview. Following the recent work [46] for SS-TAL, we leverage a one-sage detector, *i.e.*, ActionFormer [52] as our baseline, where each video frame is directly supervised by the corresponding labels, including distances to action boundaries and the action category.

In this paper, we propose a novel architecture, named APL, to adaptively explore high-quality pseudo labels, as shown in Fig. 1. Specifically, the proposed

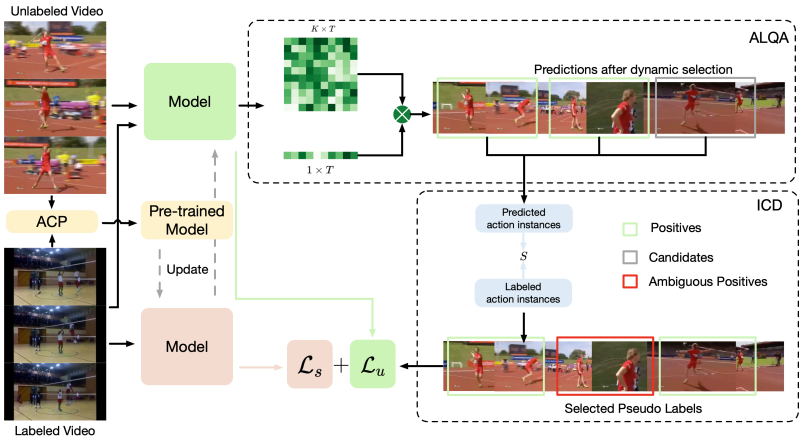

Fig. 1: The overview of the proposed APL framework. We first leverage both labeled and unlabeled videos for ACP without using any GT labels, which enhances the frame-level representation. We then update the pre-trained model by using a small amount of labeled videos and generate pseudo labels for unlabeled videos, where ALQA jointly learns classification confidence $\hat{P}_{cls} \in \mathbb{R}^{K \times T}$ and localization reliability $\hat{P}_{diou} \in \mathbb{R}^{1 \times T}$ before dynamically selecting pseudo labels according to their joint score. Finally, we propose an ICD to refine the pseudo-label selection by removing ambiguous positives and mining potential positives.

ALQA jointly learns classification confidence and localization reliability in a fully adaptive fashion, thus achieving better ranking and selection of pseudo labels (Sec. 3.1). Besides, the ICD aims to further refine the pseudo-label selection by discovering false positives and false negatives (Sec. 3.2). Finally, the ACP is also introduced to generate more discriminative frame-level representation (Sec. 3.3).

3.1 Adaptive Label Quality Assessment

In SS-TAL, the accurate assessment of classification and localization quality of pseudo labels is paramount. Recently, the determination of localization scores has relied on manual computations based on predicted boundaries [46]. In contrast, our proposed method, ALQA, introduces a novel paradigm by jointly learning the classification confidence and localization reliability of action instances, as shown in Fig. 2(a). By jointly learning classification confidence and localization reliability, ALQA provides a unified framework for assessing the quality of pseudo labels. This allows for more accurate and reliable evaluation, as it considers the interaction between classification and localization.

As aforementioned, the regression head of most TAL methods, *e.g.*, ActionFormer [52], involves the use of the DIoU loss for learning action boundaries. We leverage the intuitive qualities of DIoU, which evaluates both tIoU and tND between predicted and GT boundaries. Based on the original DIoU loss [56] for

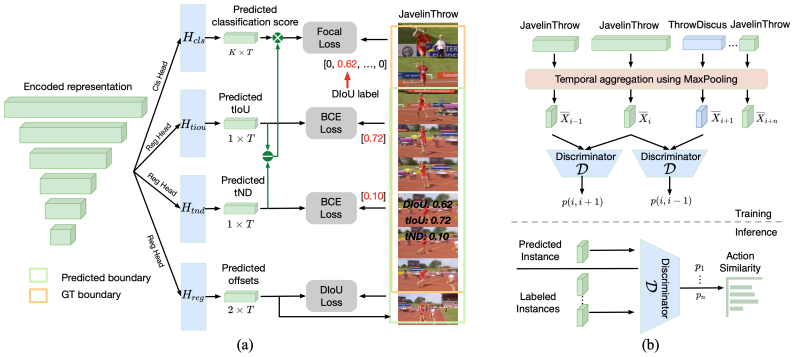

Fig. 2: Illustration of (a) Adaptive Label Quality Assessment and (b) Instance-level Consistency Discrimination. (a) We evaluate localization reliability by designing two parallel branches (heads) to predict tIoU and tND, respectively, leading to a joint score of classification and localization. (b) We aggregate temporal features of labeled action instances using Maxpooling, then use a discriminator $\mathcal{D}$ to learn the similarity probability between two instance pairs. During inference, $\mathcal{D}$ provides similarity scores between predicted instances and labeled instances of the same action category.

bounding box regression, the DIoU loss used for TAL [52] is formulated as:

$$DIoU = tIoU - tND = tIoU - \frac{\rho^2(c_{pre}, c_{gt})}{d^2} \tag{1}$$

where d is the length of the smallest temporal box covering the predicted and GT boundaries and $\rho(c_{pre}, c_{gt})$ is the distance between central points of the predicted and GT boundaries.

Based on the above observation, we design two parallel branches that can be seamlessly integrated into the original regression head of the baseline. They are dedicated to predicting tIoU and tND, respectively, offering distinct perspectives to learn localization reliability. Formally, given the encoded multi-scale representation F (here we take the scale with T frames for example), the predicted tIoU ($\hat{P}_{tiou}$) and tND ($\hat{P}_{tnd}$) can be formulated as:

$$\hat{P}_{tiou} = \text{sigmod}(H_{tiou}(F)) \in \mathbb{R}^{1\times T}, \hat{P}_{tnd} = \text{sigmoid}(H_{tnd}(F)) \in \mathbb{R}^{1\times T} \tag{2}$$

where H_{tiou} and H_{tnd} have a similar structure as the regression head H_{reg} of the baseline, consisting of 1D convolutions and layer normalization. H_{tiou}, H_{tnd} and H_{reg} share the same parameters, except for the final layer.

In the baseline, the original regression head aims to predict offsets from current frames to predicted boundaries, and the DIoU loss defined in Eq. (1) is used to minimize the tIoU and tND between predictions and ground truth. In this way, we take the tIoU (P_{tiou}) and tND (P_{tnd}) calculated in this branch as the ground truth of our proposed tIoU and tND branches, respectively, and the overall loss of localization quality can be defined as follows:

$$\mathcal{L}_{locq} = \text{BCE}\left(\hat{P}_{tiou}, P_{tiou}\right) + \text{BCE}\left(\hat{P}_{tnd}, P_{tnd}\right) \tag{3}$$

where BCE denotes the binary cross entropy. The outputs, *i.e.*, $\hat{P}_{tiou}$ and $\hat{P}_{tnd}$, are probability values between 0 and 1. Thus, we use BCE loss to measure the difference between the model's output and the ground truth, guiding the model optimization. For the classification branch, the joint score $\hat{P}$ (between 0 and 1) can be obtained by combining the classification score $\hat{P}_{cls}$ predicted by the original classification head and the localization reliability consisting of the $\hat{P}_{tiou}$ and $\hat{P}_{tnd}$, defined as:

$$\hat{P} = \hat{P}_{diou} \odot \hat{P}_{cls} = \max[\hat{P}_{tiou} - \hat{P}_{tnd}, \epsilon] \odot \hat{P}_{cls} \tag{4}$$

where $\odot$ denotes element-wise product. The difference $\hat{P}_{diou}$ between $\hat{P}_{tiou}$ and $\hat{P}_{tnd}$ represents the overall localization reliability. Particularly, to avoid negative reliability probabilities during training, we set the difference to a small positive value ϵ when it is negative. This is because a negative difference indicates a relatively large temporal distance between the predictions and the ground truth, representing a low reliability probability.

Similar to the baseline, the objective function in the classification branch is also based on focal loss (FL) [35, 52]. However, the classification target of these methods normally uses one-hot label encoding, which is less consistent with the predicted joint score. Inspired by the soft label mechanism [17, 23], we replace the one-hot label with a combination of predicted tIoU, tND and the one-hot label to facilitate unified optimization, promoting direct interaction between classification and localization. Thus, we form a DIoU-based soft label $P = \{0, \ldots, P_{diou}, \ldots, 0\}$, where $P_{diou} = P_{tiou} - P_{tnd}$ (we also set P_{diou} to a small positive value during training when it is negative). The overall focal loss combining classification and localization qualities is formulated as:

$$\mathcal{L}_{cls} = \mathrm{FL}\left(\hat{P}, P\right) \tag{5}$$

After generating predictions on unlabeled data, we apply Soft-NMS [3] to remove redundant and low-quality instances. Unlike previous methods that rely on fixed confidence thresholds or frequency [46] of different actions in pseudo labels to select positive instances, our approach introduces a selection mechanism with adaptive criteria. Specifically, given the sets of predicted instances $\hat{\mathcal{I}}_u = \{\hat{I}_i\}_{i=1}^{L_{nms}}$ and corresponding joint scores $\hat{\mathcal{P}}_u = \{\hat{P}_i\}_{i=1}^{L_{nms}}$ where L_{nms} is the number of instances after NMS, we construct positives $\hat{\mathcal{I}}_{pos}$ and candidates $\hat{\mathcal{I}}_{can}$ based on the thresholds of joint scores:

$$\hat{\mathcal{I}}_{pos} = \{\hat{I}_i \mid \hat{P}_i >= \tau_{pos}\}_{i=1}^{L_{nms}}, \hat{\mathcal{I}}_{can} = \{\hat{I}_i \mid \hat{\tau}_{neg} < \hat{P}_i < \tau_{pos}\}_{i=1}^{L_{nms}} \tag{6}$$

where τ_{neg} is fixed to 0.15 to directly remove low-quality instances. τ_{pos} is dynamically computed based on the mean and standard deviation of joint scores $\{\hat{P}_i \mid \hat{P}_i > \tau_{neg}\}_{i=1}^{L_{nms}}$ of the remaining instances.

3.2 Instance-level Consistency Discrimination

Although ALQA can achieve better ranking and selection of pseudo labels, there still exist some ambiguous positives and potential positives in $\hat{\mathcal{I}}_{pos}$ (defined in Eq.

(6)) and $\hat{\mathcal{I}}_{can}$, respectively. To tackle this problem, we propose the ICD to learn and leverage inter-instance intrinsic consistency, ensuring a more comprehensive and accurate identification of positive instances within the predictions.

As shown in Fig. 2(b), during training on labeled videos, we sample the i-th instance of one mini-batch and the corresponding input feature and action category are defined as $X_i \in \mathcal{X}_i^b = \{X_i\}_{i=1}^{M_b}$ and y_i, respectively, where M_b is the total number of action instances within a mini-batch. Two feature sets are then constructed as:

$$\mathcal{G}_i = \{X_r | y_r = y_i, 1 \leq r \leq M_b\}, \overline{\mathcal{G}}_i = \{X_q | y_q \neq y_i, 1 \leq q \leq M_b\} \tag{7}$$

where $\mathcal{G}_i$ comprises features of instances sharing the same action category as the i-th instance, whereas $\overline{\mathcal{G}}_i$ encompasses features of instances with distinct action categories. Notably, in our implementation, we sample a fixed number of instances for the two sets to ensure a balanced training in Eq. (8).

Subsequently, the features (*e.g.*, X_i *vs* X_r) of instances with the same action type should be similar, while those (*e.g.*, X_i *vs* X_q) with different action types should be dissimilar. Therefore, the overall objective function of the ICD is formulated as follows:

$$\mathcal{L}_{icd} = -\mathbb{E}_{X_i \sim \mathcal{X}_i}\left[\mathbb{E}_{X_r \sim \mathcal{G}_i}[\log \mathcal{D}(\overline{X}_i, \overline{X}_r)] + \mathbb{E}_{X_q \sim \overline{\mathcal{G}}_i}[\log(1 - \mathcal{D}(\overline{X}_i, \overline{X}_q))]\right] \tag{8}$$

where $\mathbb{E}$ denotes the expectation. $\overline{X}_i = \text{MAP}(X_i)$ and $\text{MAP}(\cdot)$ is the MaxPooling operation that aggregates temporal features of action instances. $\mathcal{D}$ is the discriminator that aims to predict the probability of pair-wise instances being similar, which works as follows:

$$\mathcal{D}(\overline{X}_i, \overline{X}_r) = \text{MLP}([\overline{X}_i; \overline{X}_r]), \mathcal{D}(\overline{X}_i, \overline{X}_q) = \text{MLP}([\overline{X}_i; \overline{X}_q]) \tag{9}$$

where $[;]$ represents the concatenation operation along the feature dimension. MLP maps the input features from $2D$ to 1. The objective is to ensure that the output probability of $\mathcal{D}(\overline{X}_i, \overline{X}_r)$ approaches 1, while that of $\mathcal{D}(\overline{X}_i, \overline{X}_q)$ tends toward 0. Consequently, we leverage the BCE loss for the optimization of our ICD, which is independent of the baseline's training procedure.

In the inference phase, the ICD is employed to produce similarity scores, reflecting the overall similarity between a predicted instance and the labeled instances belonging to the same category. Mathematically, when considering the action category $\hat{y}_i$ and the corresponding input feature $\hat{X}_i$ of a predicted instance, we begin by selecting all GT instances with the same category from labeled videos. The feature set of these instances, denoted as $\mathcal{X}_i^l$, can be expressed as $\mathcal{X}_i^l = \{X_j | y_j = \hat{y}_i\}_{j=1}^{M_l}$, where M_l denotes the total number of instances, and y_j represents the action class of the j-th instance. We then pass the predicted instance with feature $\hat{X}_i$ and each instance from $\mathcal{X}_i^l$ to the ICD, resulting in the average similarity score:

$$S_i = \frac{1}{|\mathcal{X}_i^l|} \sum_{X_j \in \mathcal{X}_i^l} \mathcal{D}(\text{MAP}(\hat{X}_i), \text{MAP}(X_j)) \tag{10}$$

The computed similarity score helps identify ambiguous positives from $\hat{\mathcal{I}}_{pos}$ and reveal potential positives from $\hat{\mathcal{I}}_{can}$. Since the overall quality of pseudo labels from $\hat{\mathcal{I}}_{can}$ is lower than those from $\hat{\mathcal{I}}_{pos}$, we set a relatively high threshold ς_{icd} (> 0.5) to select positive instances from $\hat{\mathcal{I}}_{can}$. To retain more positives from $\hat{\mathcal{I}}_{pos}$, instances with a similarity score below τ_{icd} are excluded (see Fig. 3).

3.3 Action-aware Contrastive Pre-training

While the self-supervised pre-training [26] has demonstrated efficacy for SS-TAL, the reliance on intricate pretext tasks and neglect of distinctions between actions limit its application and performance. The proposed ACP seeks to provide a general and unsupervised alternative to enhance frame-level representation, which can be combined with various backbone models.

The ACP involves two types of contrasts, *i.e.*, coarse- and fine-grained contrasts (Fig. S1 in **Supp. C**). The former aims at contrasting pair-wise frames within an untrimmed video, with a specific emphasis on distinguishing actions and backgrounds. Our ACP is performed on the multi-scale representation encoded by the FPN neck of ActionFormer [52]. To implement the coarse-grained contrast, we first upsample each representation to match the temporal length of the input and then concatenate these representations along the feature dimension to generate a new representation. We then partition the representation of each video in a mini-bath into N equal segments. Subsequently, a single frame is randomly selected from each partition, forming the representation set of each video as the input for our ACP. Since the ACP follows a unsupervised setting, no GT action labels are provided to guide contrastive learning. Hence we perform K-means clustering on the corresponding input features to generate initial action classes and the number of clusters is set to 2. Formally, let $f_i \in \mathcal{F} = \{f_i\}_{i=1}^{N}$ and $l_i \in \{0, 1\}$ denote the i-th feature of the representation set $\mathcal{F}$ and the corresponding clustering labels, the positive sets $\mathcal{P}_i$ and negative sets $\mathcal{N}_i$ are constructed as $\mathcal{P}_i = \{f_j | l_j = l_i\}_{i=1}^{N}$ and $\mathcal{N}_i = \{f_j | l_j \neq l_i\}_{i=1}^{N}$, respectively. We then use infoNCE loss [27] for coarse-grained contrast, which is defined as follows:

$$\mathcal{L}_{conc} = -\frac{1}{N_c} \sum_{f_i} \sum_{f_j \in \mathcal{P}_i} \log \frac{\exp(\text{sim}(f_i, f_j)/\varsigma)}{\exp(\text{sim}(f_i, f_j)/\varsigma) + \sum_{f_* \in \mathcal{N}_i} \exp(\text{sim}(f_i, f_*)/\varsigma)} \tag{11}$$

where $N_c = \sum_i |\mathcal{P}_i|$, $\text{sim}(\cdot)$ is the inner product between two normalized vectors and $\varsigma > 0$ is a temperature parameter.

Complementing the coarse-grained contrast, the fine-grained contrast aims to improve the discrimination between actions by elaborately contrasting frames from all videos in a mini-batch. To achieve this, we combine the representation sets of all videos and obtain the fine-grained clustering labels. Different from binary-class labels used in coarse-grained contrast, we assign multi-class labels to the frames of the combined set, where $l_i \in \{0, 1, ..., B-1\}$. Here, B is the number of clusters and is determined based on the batch size and datasets. Similarly, the fine-grained contrast loss $\mathcal{L}_{conf}$ can be computed using Eq. (11), and the overall pre-trained loss is represented as $\mathcal{L}_{acp} = \mathcal{L}_{conc} + \mathcal{L}_{conf}$. Note that $\mathcal{L}_{acp}$ is also employed for fine-tuning on labeled videos with GT action classes and unlabeled videos with pseudo labels after pre-training (see Tab. S4).

Finally, the overall objective loss for our SS-TAL framework is designed as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}^s + \beta\mathcal{L}^u \\ &= \frac{1}{N^s_{pos}}\sum_t(\mathcal{L}^s_{cls} + \lambda_{reg}\mathbb{1}_{in_t}\mathcal{L}^s_{reg} + \lambda_{locq}\mathbb{1}_{in_t}\mathcal{L}^s_{locq}) + \lambda_{acp}\mathcal{L}^s_{acp} + \mathcal{L}_{icd} \\ &+ \beta(\frac{1}{N^u_{pos}}\sum_t(\mathcal{L}^u_{cls} + \lambda_{reg}\mathbb{1}_{in_t}\mathcal{L}^u_{reg} + \lambda_{locq}\mathbb{1}_{in_t}\mathcal{L}^u_{locq}) + \lambda_{acp}\mathcal{L}^u_{acp})\end{aligned} \tag{12}$$

where β is the weight of unsupervised loss, which is set to 2. $\mathbb{1}_{in_t}$ is an indicator denoting whether the t-th frame is within a GT action or background. N^*_{pos} is the number of frames within action segments. $\mathcal{L}^*_{reg}$ indicates the DIoU loss and λ_{reg} is set to the default value in [52]. Both λ_{locq} and λ_{acp} are set to 0.1.

4 Experiments

Datasets and Evaluation. We evaluate our method on two challenging TAL benchmarks, *i.e.*, THUMOS14 [12] and ActivityNet v1.3 [4]. We report the mean average precision (mAP) at different tIoU thresholds. The thresholds are [0.3:0.1:0.7] for THUMOS14 and [0.5:0.05:0.95] for ActivityNet v1.3. Following [46], we randomly select 10%, 20%, 40% and 60% of the training videos as labeled data and the remaining as unlabeled data.
Implementation Details. Similar to [46], our SS-TAL framework is also based on the detector ActionFormer [52]. In addition, we combine the proposed components with BMN [20], which is a two-stage proposal-based detector. For fair comparisons, we use two popular backbones, *i.e.*, TSN [42] and I3D [6] pretrained on Kinetics to extract the video features. More implementation details of our semi-supervised learning are provided in the supplementary material.

4.1 Comparison with State-of-the-art Methods

THUMOS14. Following [46], we combine our APL with ActionFormer [52] and BMN [20]. For ActionFormer (using I3D features), we add two regression heads to predict tIoU and tND, which measure the localization reliability of pseudo labels. As shown in Tab. 1, our APL achieves superior performance and suppresses the state-of-the-art methods in mAP at different thresholds, which demonstrates its effectiveness. In particular, APL achieves 42.8% in the average mAP on THUMOS14 with 20% labeled data, which outperforms NPL by a large margin, namely 5% absolute improvement. For the two-stage detector, we incorporate the proposed components into BMN (using TSN features) to facilitate the selection of action proposals. The results in Tab. 1 show that the two-stage detector can also benefit from the proposed APL. Notably, APL obtains almost 2% average mAP improvement compared with the recent anchor-free SPOT [26] when using 10%, 20% and 40% labeled data.
ActivityNet v1.3. For the ActivityNet v1.3 dataset, we also adopt the I3D and TSN as our backbone features. With I3D features, our method reaches an average

Table 1: Comparison with the state-of-the-art methods on THUMOS14 and ActivityNet v1.3. We report mAP (%) at different tIoU thresholds. ActF refers to ActionFormer [52]. * means using only labeled training videos.

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* [52]	I3D	28.5	22.9	14.1	8.2	4.1	15.6	47.8	24.2	1.7	25.6
	ActF + MixUp [53]	I3D	29.7	24.2	14.5	9.6	5.4	16.7	49.4	27.9	3.1	28.8
	NPL (ActF) [46]	I3D	32.8	29.6	20.1	11.7	7.2	20.3	51.9	33.4	3.6	32.5
	APL (ActF)	I3D	**35.1**	**31.7**	**25.6**	**19.1**	**11.0**	**24.5**	**52.2**	**33.9**	**6.7**	**33.5**
	SSP [11]	TSN	44.2	34.1	24.6	16.9	9.3	25.8	38.9	28.7	8.4	27.6
	SSTAP [43]	TSN	45.6	35.2	26.3	17.5	10.7	27.0	40.7	29.6	**9.0**	28.2
	SPOT [26]	TSN	49.4	40.4	31.5	22.9	12.4	31.3	49.9	31.1	8.3	32.1
	NPL (BMN) [46]	TSN	50.0	41.7	33.5	23.6	13.4	32.4	50.9	32.0	7.9	32.6
	APL (BMN)	TSN	**51.5**	**42.5**	**34.6**	**24.4**	**13.5**	**33.3**	**51.5**	**32.4**	8.2	**33.0**
20%	ActF* [52]	I3D	49.1	41.6	32.6	21.5	12.1	31.4	51.2	34.3	3.8	32.9
	ActF + MixUp [53]	I3D	51.2	43.2	34.0	23.9	14.1	33.3	52.9	34.7	3.9	33.3
	NPL (ActF) [46]	I3D	54.5	47.1	39.3	29.7	18.5	37.8	53.1	35.8	3.9	33.8
	APL (ActF)	I3D	**59.2**	**54.2**	**44.5**	**34.2**	**22.1**	**42.8**	**53.5**	**36.1**	**7.1**	**34.5**
	SPOT [26]	TSN	52.6	43.9	34.1	25.2	16.2	34.4	51.7	32.0	6.9	32.3
	NPL (BMN) [46]	TSN	53.9	45.6	36.2	26.9	16.5	35.8	52.1	32.9	7.9	32.9
	APL (BMN)	TSN	**54.8**	**45.9**	**37.1**	**28.5**	**16.9**	**36.6**	**52.4**	**33.3**	**8.3**	**33.4**
40%	ActF* [52]	I3D	69.0	60.4	49.3	31.5	19.3	45.9	53.2	35.7	3.8	34.2
	ActF + MixUp [53]	I3D	69.7	61.9	52.4	34.4	20.1	47.7	53.1	36.0	4.3	34.5
	NPL (ActF) [46]	I3D	71.9	65.4	55.7	40.9	23.4	51.5	53.6	36.5	4.6	35.3
	APL (ActF)	I3D	**73.2**	**68.2**	**59.1**	**44.9**	**28.7**	**54.8**	**53.8**	**36.7**	**8.2**	**35.5**
	SPOT [26]	TSN	54.4	45.8	37.2	29.7	19.4	37.3	53.3	33.0	6.6	33.2
	NPL (BMN) [46]	TSN	56.2	46.7	38.8	30.3	19.5	38.3	53.4	**33.9**	8.1	33.8
	APL (BMN)	TSN	**57.0**	**47.1**	**39.5**	**32.7**	**20.1**	**39.3**	**53.5**	33.8	**8.5**	**34.1**
60%	ActF* [52]	I3D	71.5	65.6	59.9	47.3	32.7	55.4	53.9	36.1	5.7	35.0
	ActF + MixUp [53]	I3D	72.2	67.5	61.2	48.7	34.0	56.7	54.1	36.4	5.7	35.2
	NPL (ActF) [46]	I3D	74.5	69.9	62.8	51.1	36.6	59.0	54.3	**36.7**	6.5	35.8
	APL (ActF)	I3D	**77.3**	**73.1**	**65.0**	**52.4**	**37.6**	**61.1**	**54.4**	**36.7**	**8.4**	**36.0**
	SSP [11]	TSN	53.2	46.8	39.3	29.7	19.8	37.8	49.8	34.5	7.0	33.5
	SSTAP [43]	TSN	56.4	49.5	41.0	30.9	21.6	39.9	50.1	34.9	7.4	34.0
	SPOT [26]	TSN	58.9	50.1	42.3	33.5	22.9	41.5	52.8	35.0	8.1	35.2
	NPL (BMN) [46]	TSN	59.0	51.4	42.9	34.3	23.3	42.2	53.9	35.8	8.5	35.7
	APL (BMN)	TSN	**59.7**	**51.6**	**43.2**	**34.9**	**23.6**	**42.6**	**54.2**	**36.2**	**8.6**	**35.9**

mAP of 33.5% with 10% labeled data, outperforming the closest competitor NPL by 1%. It is noteworthy that our APL significantly improves the mAP when the tIoU threshold is set to 0.95 (mAP@0.95), resulting in improvements of 3.1%, 3.3%, 3.6% and 1.9% under different semi-supervised settings. The results are still comparable when using TSN features. Our method receives the best mAP@0.5 and average mAP compared to other SS-TAL frameworks, which can be attributed to the improvement in the quality of pseudo labels.

4.2 Ablation Study

To further verify the efficacy of our contributions, we conduct comprehensive ablation studies on the THUMOS14 dataset, including each component of our method and the choice of hyper-parameters. ActionFormer [52] based on I3D [6] is used as the localization framework.

Table 2: Effectiveness of three main components on THUMOS14, using 10% and 40% labeled videos. '+' means training by the proposed method.

Label	Method	mAP(%)			
		0.3	0.5	0.7	Avg.
10%	baseline	26.6	17.8	6.1	17.0
	+ALQA	31.9	22.9	8.1	21.3
	+ICD	33.6	24.6	10.2	23.3
	+ACP	**35.1**	**25.6**	**11.0**	**24.5**
40%	baseline	65.9	51.7	22.8	48.8
	+ALQA	71.0	56.7	26.8	52.6
	+ICD	72.0	58.1	27.7	53.9
	+ACP	**73.2**	**59.1**	**28.7**	**54.8**

Table 3: The effect of different localization reliability learning strategies on THUMOS14 with 10% and 40% labeled videos.

Label	Method	mAP(%)			
		0.3	0.5	0.7	Avg.
10%	tIoU (1b)	28.9	20.7	7.6	19.4
	DIoU (1b)	30.7	20.8	8.0	20.0
	tIoU+tND (2b)	**31.9**	**22.9**	**8.1**	**21.3**
40%	tIoU (1b)	68.4	54.8	24.4	50.7
	DIoU (1b)	70.3	55.4	26.0	51.8
	tIoU+tND (2b)	**71.0**	**56.7**	**26.8**	**52.6**

Effectiveness of each component. We demonstrate the effectiveness of three proposed components in APL, including ALQA, ICD, and ACP. In Tab. 2, the baseline denotes the case where the pseudo labels are filtered according to a fixed classification confidence threshold (*i.e.*, 0.3). Under the 10% labeling ratio, we can see that compared with the baseline, our ALQA brings about a 4.3% absolute improvement in the average mAP, proving the effectiveness of the module by jointly learning classification and localization quality. After being equipped with ICD, it boosts the performance by 2% of average mAP, which demonstrates the effectiveness of ICD in refining the selection of pseudo labels. When further applying our pre-training strategy, the performance is increased to 24.5% mAP. This substantiates that our ACP effectively facilitates SS-TAL. We can observe a similar trend when utilizing 40% labeled videos.

Ablation on ALQA. In this section, we present the ablation results for different localization reliability learning strategies in Tab. 3. tIoU (1b) denotes designing a single branch (1b) to learn the tIoU. With 10% labeled data, the performance is increased from 19.4% mAP to 20% mAP by replacing the tIoU prediction with DIoU (Eq. (1)) prediction (1b). This shows that temporal distance can assist in evaluating the localization quality. We also find that the average mAP is further improved by 1.3% when we design two different branches (2b) for predicting tIoU and tND, respectively. The main reason can be that having two separate branches enables the model to adaptively adjust its focus on different aspects of action localization, resulting in improved performance.

Ablation on ICD. To study how ICD affect performance, we separately apply EAP, MPP and their combination (*i.e.*, EAP + MPP) to optimize the pseudo labels. In Tab. 4, we see that MPP plays a more important role in filtering pseudo labels than EAP and merging them can gain further promotion.

We also investigate the hyperparameters in ICD. Fig. 3(a) shows the performance curve of average mAP corresponding to threshold τ_{icd} on THUMOS14 with 10% labeled data. The average mAP gradually improves as τ_{icd} increases, but it slightly drops when τ_{icd} reaches 0.3. This is because setting τ_{icd} too large may lead to the removal of more true positives. Fig. 3(b) presents the average

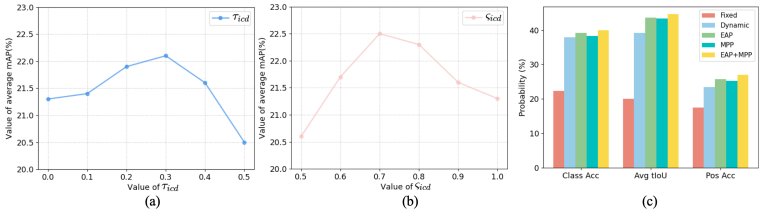

Fig. 3: (a) and (b) The effect of different hyperparameters (*i.e.*, τ_{icd} and ς_{icd}) settings. (c) Ablation studies on the quality of pseudo labels when using 10% labeled videos. Class Acc: action classification accuracy. Avg tIoU: average tIoU. Pos Acc: accuracy of positive predictions.

Table 4: The effect of ICD on THUMOS14 with 10% and 40% labeled videos. EAP and MPP refer to eliminating ambiguous positives and mining potential positives.

Label	Method	mAP(%)			
		0.3	0.5	0.7	Avg.
10%	No ICD	31.9	22.9	8.1	21.3
	EAP	31.7	23.6	9.9	22.1
	MPP	32.4	23.5	10.0	22.5
	EAP+MPP	**33.6**	**24.6**	**10.2**	**23.3**
40%	No ICD	71.0	56.7	26.8	52.6
	EAP	71.5	57.3	27.2	53.2
	MPP	71.7	57.7	27.5	53.3
	EAP+MPP	**72.0**	**58.1**	**27.7**	**53.9**

Table 5: The effect of coarse- and fine-grained contrasts on THUMOS14 with 10% and 40% labeled videos.

Label	Loss	mAP(%)			
		0.3	0.5	0.7	Avg.
10%	No ACP	33.6	24.6	10.2	23.3
	$\mathcal{L}_{conc}$	34.6	25.5	10.6	24.1
	$\mathcal{L}_{conf}$	34.4	25.1	10.5	23.8
	$\mathcal{L}_{conc}$+$\mathcal{L}_{conf}$	**35.1**	**25.6**	**11.0**	**24.5**
40%	No ACP	72.0	58.1	27.7	53.9
	$\mathcal{L}_{conc}$	72.5	58.6	28.4	54.4
	$\mathcal{L}_{conf}$	72.3	59.0	27.9	54.4
	$\mathcal{L}_{conc}$+$\mathcal{L}_{conf}$	**73.2**	**59.1**	**28.7**	**54.8**

mAP for different values of ς_{icd}. Our method achieves the highest performance when ς_{icd} is set to 0.7. Thus, we set τ_{icd} to 0.3 and ς_{icd} to 0.7.

Ablation on ACP. To go deeper into our ACP, we conducted three experiments: coarse-grained contrast loss only, fine-grained contras loss only, and the complete pre-training loss. From Tab. 5, either $\mathcal{L}_{conc}$ or $\mathcal{L}_{conf}$ can improve the performance compared to the baseline (No ACP), and when we combine them together, the average mAP experiences a notable increase of 1.2% and 0.9% in the case of the 10% and 40% settings, respectively. This improvement can be attributed to the enhanced capability of the model to discriminate between actions and backgrounds, as well as between different actions.

Ablation on the quality of pseudo labels. We further study the quality of pseudo labels in terms of classification accuracy (Class Acc), average temporal IoU (Avg tIoU) w.r.t ground truth and accuracy of positive predictions (Pos Acc). Here positive predictions mean that the estimated instance has the same action class as the ground truth and tIoU is above 0.5. Specifically, we consider three cases: first, where τ_{pos} in Eq. (6) is fixed at 0.3 to filter pseudo labels; second, where τ_{pos} is dynamically computed; third, where the pseudo labels are enhanced by EAP, MPP and EAP+MPP, respectively. Note that the results reported in Fig. 3(c) are calculated based on 90% unlabeled videos. From Fig.

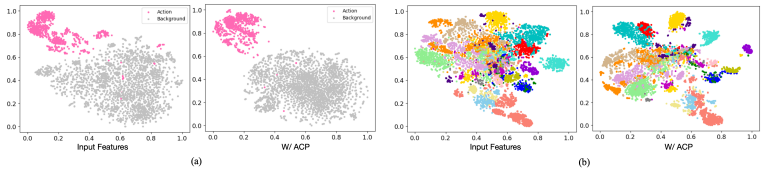

Fig. 4: The effect of our ACP on THUMOS14. (a) t-SNE visualization of action and background features. (b) t-SNE visualization of features for different actions. The legend for different actions is provided in the supplementary material.

3(c), we can observe that the quality of pseudo labels is very poor across all metrics when using a fixed threshold, while it is improved by a large margin after employing a dynamical threshold. Based on the positives and negatives initially divided by the dynamical τ_{pos}, both EAP and MPP contribute to improving the pseudo-label quality and their combination leads to more reliable pseudo labels.

Qualitative results. To better illustrate the effectiveness of ACP, we visualize some qualitative results on THUMOS14 in Fig. 4. Specifically, we choose a subset of the validation set and visualize the learned representation (Sec. 3.3). In Fig. 4(a), the visualization demonstrates better separation between action and background features, indicating that ACP effectively enhances the discrimination between action-related and background information. Additionally, Fig. 4(b) shows the result where distinct clusters can be observed for each action category. This suggests that ACP contributes to improved feature representation, allowing for better discrimination between different actions. Please refer to **Supp. E** for more ablation studies and visualization results.

5 Conclusion

In this paper, we explore the label noise problem in SS-TAL with a novel framework. We introduce ALQA to jointly learn classification confidence and localization reliability, providing more reliable joint scores for pseudo-label ranking. Additionally, ICD is designed to improve pseudo-label quality by eliminating ambiguous positives and mining potential positives. Finally, we propose ACP pretraining to enhance discrimination within and between actions and backgrounds, serving as a versatile component for SS-TAL methods. Extensive experiments on two benchmarks demonstrate new state-of-the-art performance.

Limitation and future work. In the ICD, different similarity score thresholds may affect the quality of pseudo-label refinement. Exploring more adaptive strategies to identify positive instances is a potential avenue for future work. Additionally, the generalization ability of our approach across different TAL frameworks could be further explored.

Acknowledgements

The research is supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. Bai, Y., Wang, Y., Tong, Y., Yang, Y., Liu, Q., Liu, J.: Boundary content graph neural network for temporal action proposal generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 121–137. Springer (2020)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
3. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M., Zhuang, Y.: Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14381–14390 (2022)
8. Chen, B., Li, P., Chen, X., Wang, B., Zhang, L., Hua, X.S.: Dense learning based semi-supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4815–4824 (2022)
9. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: Daps: Deep action proposals for action understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 768–784. Springer (2016)
10. Gao, J., Chen, K., Nevatia, R.: Ctap: Complementary temporal action proposal generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 68–83 (2018)
11. Ji, J., Cao, K., Niebles, J.C.: Learning temporal action proposals with fewer labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7073–7082 (2019)
12. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/` (2014)
13. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)

14. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
15. Li, P., Purkait, P., Ajanthan, T., Abdolshah, M., Garg, R., Husain, H., Xu, C., Gould, S., Ouyang, W., van den Hengel, A.: Semi-supervised semantic segmentation under label noise via diverse learning groups. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1229–1238 (2023)
16. Li, W., Wang, W., Chen, X., Wang, J., Li, G.: A joint model for action localization and classification in untrimmed video with visual attention. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 619–624. IEEE (2017)
17. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in Neural Information Processing Systems **33**, 21002–21012 (2020)
18. Lin, C., Li, J., Wang, Y., Tai, Y., Luo, D., Cui, Z., Wang, C., Li, J., Huang, F., Ji, R.: Fast learning of temporal action proposal via dense boundary generator. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11499–11506 (2020)
19. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2021)
20. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3889–3898 (2019)
21. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 988–996 (2017)
22. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
23. Liu, C., Zhang, W., Lin, X., Zhang, W., Tan, X., Han, J., Li, X., Ding, E., Wang, J.: Ambiguity-resistant semi-supervised learning for dense object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15579–15588 (2023)
24. Liu, L., Zhang, B., Zhang, J., Zhang, W., Gan, Z., Tian, G., Zhu, W., Wang, Y., Wang, C.: Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7370–7379 (2023)
25. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1979–1993 (2018)
26. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Semi-supervised temporal action detection with proposal-free masking. In: European Conference on Computer Vision. pp. 663–680. Springer (2022)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684 (2020)

29. Qing, Z., Su, H., Gan, W., Wang, D., Wu, W., Wang, X., Qiao, Y., Yan, J., Gao, C., Sang, N.: Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 485–494 (2021)
30. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
32. Shao, J., Wang, X., Quan, R., Zheng, J., Yang, J., Yang, Y.: Action sensitivity learning for temporal action localization. arXiv preprint arXiv:2305.15701 (2023)
33. Shi, B., Dai, Q., Hoffman, J., Saenko, K., Darrell, T., Xu, H.: Temporal action detection with multi-level supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8022–8032 (2021)
34. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866 (2023)
35. Shi, D., Zhong, Y., Cao, Q., Zhang, J., Ma, L., Li, J., Tao, D.: React: Temporal action detection with relational queries. In: European conference on computer vision. pp. 105–121. Springer (2022)
36. Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K., Das, A.: Semi-supervised action recognition with temporal contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10399 (2021)
37. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
38. Su, T., Wang, H., Wang, L.: Multi-level content-aware boundary detection for temporal action proposal generation. IEEE Transactions on Image Processing (2023)
39. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3132–3141 (2021)
40. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
41. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
42. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
43. Wang, X., Zhang, S., Qing, Z., Shao, Y., Gao, C., Sang, N.: Self-supervised learning for semi-supervised temporal action proposal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1905–1914 (2021)
44. Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., Xia, S.T.: Iterative learning with open-set noisy labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8688–8696 (2018)

45. Weng, Y., Pan, Z., Han, M., Chang, X., Zhuang, B.: An efficient spatio-temporal pyramid transformer for action detection. In: European Conference on Computer Vision. pp. 358–375. Springer (2022)
46. Xia, K., Wang, L., Zhou, S., Hua, G., Tang, W.: Learning from noisy pseudo labels for semi-supervised temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10160–10169 (2023)
47. Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., Jiang, Y.G.: Svformer: Semi-supervised video transformer for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18816–18826 (2023)
48. Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., Zhang, W., Wang, C., Zeng, L.: Class-aware contrastive semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14421–14430 (2022)
49. Yang, H., Wu, W., Wang, L., Jin, S., Xia, B., Yao, H., Huang, H.: Temporal action proposal generation with background constraint. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3054–3062 (2022)
50. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7236–7246 (2023)
51. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7094–7103 (2019)
52. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)
53. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
54. Zhao, C., Thabet, A.K., Ghanem, B.: Video self-stitching graph network for temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13658–13667 (2021)
55. Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., Tian, Q.: Bottom-up temporal action localization with mutual regularization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 539–555. Springer (2020)
56. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12993–13000 (2020)
57. Zhou, F., Jiang, Z., Zhou, H., Li, X.: Smc-nca: Semantic-guided multi-level contrast for semi-supervised action segmentation. arXiv preprint arXiv:2312.12347 (2023)
58. Zhu, Z., Tang, W., Wang, L., Zheng, N., Hua, G.: Enriching local and global contexts for temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13516–13525 (2021)