Weakly-Supervised Temporal Action Localization by Progressive Complementary Learning

Jia-Run Du, Jia-Chang Feng, Kun-Yu Lin, Fa-Ting Hong, Zhongang Qi, Ying Shan, Jian-Fang Hu, and Wei-Shi Zheng

Abstract—Weakly-Supervised Temporal Action Localization (WSTAL) aims to localize and classify action instances in long untrimmed videos with only video-level category labels as supervision. A critical challenge of WSTAL is the large gap between video-level supervision and unavailable snippet-level supervision. Prevailing methods typically assign pseudo labels to snippets, but these methods suffer from significant noise caused by the pseudo snippet-level labels. In this work, we address the WSTAL from a novel category exclusion perspective, which gradually enhances the snippet-level supervision to bridge the gap. Our proposed Progressive Complementary Learning (ProCL) is inspired by the fact that, video-level labels precisely indicate the categories that all snippets surely do not belong to, which is ignored by previous works. Accordingly, we first exclude these surely non-existent categories by the deterministic complementary learning. And then, we introduce the entropy-based pseudo complementary learning that is able to exclude more categories for snippets of less ambiguity. Furthermore, for the remaining ambiguous snippets, we attempt to reduce the ambiguity by distinguishing foreground actions from the background. Extensive experimental results show that our method achieves new state-of-the-art performance on THUMOS14, ActivityNet1.3, and MultiTHUMOS benchmarks.

Index Terms—Weakly-supervised learning, temporal action localization, progressive complementary learning.

I. INTRODUCTION

TEMPORAL action localization aims to localize and classify action instances in an untrimmed video. It is important for applications of video retrieval [1], [2], anomaly detection [3] and highlight detection [4]. Fully-supervised temporal action localization [5], [6], [7], [8] require the start and end timestamps of each action instance for training. It is time and labor-consuming to annotate all action instances in long untrimmed videos.

In this work, we focus on Weakly-Supervised Temporal Action Localisation (WSTAL) [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], which does not require precise boundary annotations but only the action categories in the video. Typically, existing methods [11], [12], [13], [17], [19], [23], [24], [25], [26] exploit the Multi-Instance Learning (MIL) paradigm [27], where each untrimmed video is considered as a labeled bag of snippet samples and video-level prediction is aggregated from snippets. Also, some methods [9], [16], [18], [23] provide effective supervision beyond video-level labels by introducing additional information, e.g., frequency, pose. Despite the progress of above methods, the absence of snippet-level supervision still blocks the weakly-supervised temporal action localization.

To obtain snippet-level supervision, one of the most effective types is based on pseudo labels [11], [17], which assigns pseudo labels for unlabeled snippets. For example, Huang et al. [24], [28] assign pseudo labels by thresholding temporal attention scores with a pre-set threshold, and He et al. [29], [30] assign pseudo labels based on predicted action proposals. However, some action instances of different categories look visually similar, e.g., snippets of “ThrowDiscus” and “Shotput”. Therefore, it is challenging to exactly label action categories for a snippet, and incorrect pseudo labels would impair the localization performance.

To bridge the large gap between video-level supervision and unavailable snippet-level supervision, we note a fact that, the video-level category labels precisely indicate categories that all snippets surely do not belong to. As shown in Figure 1, given a video with video-level labels “CricketShot” and “CricketBowling”, we know that all snippets in the video surely do not belong to “Diving”, “HighJump”, “GolfSwing”, etc. Thus, we can label all snippets as “Not Diving”, “Not HighJump”, “Not GolfSwing”, etc. Moreover, for each snippet, it is intuitively easier to exclude some non-existent categories than exactly predicting the ground-truth categories (i.e., excluding all non-existent categories). As the snippet marked by the right red box in Figure 1, it is difficult to accurately predict its action category, since the visually small...
athletes lead to confusion between the action and background. From another perspective, it is easier to confirm that this snippet does not belong to “CricketBowling”, because the athlete in this snippet holds the cricket bat.

Accordingly, in this work, we propose a novel method that gradually enhances snippet-level supervision, named **Progressive Complementary Learning** (ProCL). Our ProCL gradually excludes the categories that snippets should not belong to, which we term **complementary categories**. According to the video-level labels, we first exclude the categories that all snippets surely do not belong to, which we term **deterministic complementary categories**. Then, to further enhance the snippet-level supervision, we attempt to exclude more categories than those deterministic complementary categories. After identifying snippets of less ambiguity, we assign pseudo complementary labels to these snippets, aiming to exclude the categories that snippets likely do not belong to, which we term **pseudo complementary categories**.

For the remaining ambiguous snippets, we propose to reduce the ambiguity by a Foreground-Background Discrimination loss, which provides additional snippet-level supervision by distinguishing foreground actions from the background. Furthermore, considering the information collaboration between multi-scale snippet sequences, we propose the Multi-scale Pseudo Complementary Learning loss to improve the quality of pseudo complementary labels.

Overall, the main contributions are as follows:

- We propose a Progressive Complementary Learning (ProCL) method that progressively enhances the snippet-level supervision from the perspective of category exclusion.
- We propose deterministic complementary learning that provides noise-free snippet-level supervision for all snippets in the video.
- We propose an entropy-based pseudo complementary labeling method and propose foreground background discrimination loss to provide further snippet-level supervision.

- Extensive experiments on THUMOS14, ActivityNet1.3, and more challenging MultiTHUMOS demonstrate the superiority of our ProCL over the state-of-the-art methods. Notably, our method significantly outperforms existing pseudo-label-based methods in providing snippet-level supervision.

**II. RELATED WORKS**

**Fully-Supervised Temporal Action Localization** aims to find the temporal intervals of action instances from long untrimmed videos and classify them, which is an important task in the video understanding domain [31], [32], [33], [34], [35]. To address it, accurate timestamp annotation for each action instance is required in the given video during training. Many methods [5], [6], [7], [8], [36], [37], [38] are based on the proposal-then-classification paradigm, which generate proposals via bottom-up [6], [7], [8], [37], [38], [39] or top-down [5], [36] manner followed by multi-class classification. Some methods explore a one-stage paradigm, which directly generates class-aware proposals [40], [41], [42], or group the action and boundary predictions [43].

Although achieving significant results, the requirement of accurate timestamp annotation in fully-supervised temporal action localization limits their scalability and practicability in real-world application scenarios.

**Weakly-Supervised Temporal Action Localization** aims to localize and classify action instances by learning with only video-level category labels. Most works focus on utilizing pre-extracted features for both RGB and optical flow modalities and learning with the Multiple-Instance Learning (MIL) strategy. Li et al. [44] propose to model long-range temporal dependencies and collaborates with the MIL paradigm.

Sun et al. [45] propose a slow-motion enhanced network that focuses on localizing slow-motion. Some works [17], [19], [45], [46], [47], [48] learn attention weights for each snippet and then threshold the weights for generating action proposals. Hu et al. [49] propose re-ranking the proposals by injecting proposal-aware contextual information into each snippet. And several works [25], [50], [51], [52], [53], [54] propose co-action similarity learning to enclose the video-level features between the video pairs with common classes. Some works [24], [55], [56], [57] pursue the modality-wise consensus for consistent predictions, which is extended by Hong et al. [12] for pre-extracted features re-calibration via a cross-modal consensus mechanism. Besides, some works [11], [13], [26], [58], [59], [60] seek to explicitly model the background activity for better foreground-background separation. There are some methods that introduce additional auxiliary models [61] or information [9], [16], [18], [23] to provide rich semantics or supervision.

Though the works mentioned above make great progress, their performance is still inferior to fully-supervised methods. One of the key factors for this result is the lack of snippet-level supervision. To address this issue, some methods [11], [13], [24], [28], [47], [62] assign pseudo labels on attention weights by a preset threshold. And some methods [17], [29], [30] assign pseudo labels based on predicted action proposals. All of them provide snippet-level supervision and
achieve performance gains. Nevertheless, these methods have a common problem that assigning category labels directly to snippets is prone to introduce noise due to misclassification, which impairs the performance of localization. In contrast, we adopt an ideology of exclusion induction to progressively exclude categories that the snippets should not belong to based on different confidence levels. This provides effective snippet-level supervision for the model. And, our method does not require the introduction of additional auxiliary information, e.g., frequency, pose.

**Complementary Learning** trains a model using complementary category labels that the pattern does not belong to. It was first proposed by Ishida et al. [63] to reduce the cost of data collection for the multi-classification task, which uses complementary labels to provide valid information for limited data. Then, subsequent methods explore how to effectively use complementary labels [64], [65], [66], [67], [68], [69]. In addition, some other methods adopt the complementary labels to alleviate the noisy label problem [70], [71], [72], [73]. Different from the above methods, our proposed Progressive Complementary Learning is specifically designed for weakly-supervised temporal action localization. In our work, we consider two task-oriented properties as prior knowledge for progressive category exclusion, i.e., actions and the background cannot co-exist in one snippet and action instances of the same category have varied durations.

### III. PROGRESSIVE COMPLEMENTARY LEARNING

In this section, we detail our Progressive Complementary Learning (ProCL), which gradually enhances the snippet-level supervision from a category exclusion perspective.

#### A. Problem Formulation and Preliminaries of MIL

1) Problem Formulation: Weakly-Supervised Temporal Action Localization (WSTAL) assumes that there are \(N\) untrimmed videos denoted by \(\{V^{(i)}\}_{i=1}^{N}\), where \(V^{(i)} = \{y^{(i)}_1, \ldots, y^{(i)}_{C}\}\) is the video-level category labels of the video \(V^{(i)}\). Specifically, \(C\) denotes the number of action categories, and \(y^{(i)}_1 = 1\) means the video \(V^{(i)}\) contains at least one action instance of the \(j\)-th category, while \(y^{(i)}_j = 0\) if the video \(V^{(i)}\) does not contain any action instance of the \(j\)-th category. In what follows, the superscript of all symbols will be omitted for convenience, e.g., \(y^{(i)}\) will be replaced by \(y\). WSTAL aims to train a model that can predict a set of action proposals during the inference phase and denotes them in the form of \(\{(t_e, t_c, \psi, c)\}\), where \(t_e\) and \(t_c\) represent the start and end time, \(c\) is the predicted action category, and \(\psi\) represents the confidence score.

2) Preliminaries of Multi-Instance Learning for WSTAL: Following the previous WSTAL works [12], [13], [26], [62], the background category and a temporal attention branch are introduced for background suppression, and the multi-instance learning loss is adopted for video-level learning. Specifically, we encode features \(X\) using several layers of temporal convolution to capture the temporal relationships between snippets and generate the class activation sequence denoted as \(S = \{s_{t,c}\}_{t=C+1}^{T} \in \mathbb{R}^{T \times (C+1)}\) through the classification head, where \(s_{t,c}\) represents the activation score of the \(c\)-th category for the \(t\)-th snippet. The background category is denoted as the \(C+1\)-th category for each snippet. Correspondingly, we extend \(Y\) to \(\hat{Y} = \{y_1, \ldots, y_C, 0\} \in \mathbb{R}^{C+1}\), which represents the existence of background context.

Simultaneously, we input \(X\) into the attention head to get class-agnostic attention scores \(A = \{a_t\}_{t} \in \mathbb{R}^{T}\), and \(a_t\) denotes the probability that the \(t\)-th snippet is an action. Next, we can obtain the background-free class activate sequence denoted as \(\tilde{S} = \{\tilde{s}_{t, c}\}_{t=C+1}^{T} \in \mathbb{R}^{T \times (C+1)}\) by suppressing the background context via \(A\), where \(\tilde{s}_{t, c} = s_{t, c} \cdot a_t\). Similarly, we can obtain background-free video-level labels \(\hat{Y} = \{y_1, \ldots, y_C, 0\} \in \mathbb{R}^{C+1}\), which represents the inexistence of background context.

Overall, the video-level classification loss is a Multi-Instance Learning (MIL) loss as follows,

\[
L_{MIL} = L_{CE}(\hat{Y}, \Phi(\Gamma(S))) + L_{CE}(\hat{Y}, \Phi(\Gamma(\tilde{S}))),
\]

where \(L_{CE}\) is the categorical cross-entropy error function, \(\Phi(\cdot)\) represents the category-wise softmax function, and \(\Gamma(\cdot)\) is the temporal Top-K pooling following the previous methods [12], [26], [74] to obtain the video-level prediction scores \(\Gamma(S) \in \mathbb{R}^{C+1}\). The multiple-instance learning utilizes only video-level supervision, whereas our ProCL progressively enhances the snippet-level supervision for WSTAL.

#### B. Model Overview

Figure 2 illustrates the overall framework of our method, our ProCL gradually excludes the categories that snippets should not belong to, based on different confidence levels.

Firstly, according to the video-level labels, we adopt Deterministic Complementary Learning to exclude categories that all snippets in the video surely do not belong to, providing noisy-free snippet-level supervision (Section III-C). Then, to further enhance the snippet-level supervision, we propose entropy-based Pseudo Complementary Labeling for snippets of less ambiguity to exclude more categories. Furthermore, the Foreground Background Discrimination loss is adopted to disambiguate the ambiguous snippets for coarsely enhancing the snippet-level supervision (Section III-D). Moreover, we utilize the collaborative information of multi-scale snippet sequences to further improve the quality of the snippet-level supervision (Section III-E).

#### C. Deterministic Complementary Learning

By analyzing the WSTAL problem, we note a fact that the video-level labels precisely indicate the categories that all snippets surely do not belong to, and we term these categories deterministic complementary categories. Therefore, we propose the Deterministic Complementary Learning loss to exclude these deterministic complementary categories for all snippets in a video.

Specifically, the snippet-level classification scores \(P = \{p_{t,c}\}_{t=C+1}^{T} \in \mathbb{R}^{T \times (C+1)}\) are first obtained by performing the category-wise softmax function on class activation sequence
S, and the Deterministic Complementary Learning (DCL) loss is formulated as follows:

$$L_{DCL} = \frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{C+1} (1 - y_{t,c}) \log (1 - p_{t,c}),$$  

(2)

where $y_{t,c}$ is the expansion of the video-level label $\bar{Y}_c$ in the temporal dimension, i.e., $y_{t,c} = \bar{Y}_c$.

Through the Deterministic Complementary Learning loss, we simultaneously suppress the activation of all deterministic complementary categories for all snippets. Although the pseudo-label-based methods also provide snippet-level supervision, the noise problem of directly assigning pseudo labels deteriorates the performance of localization. In contrast, our deterministic complementary categories obtained from the video-level labels are free of noise and provide precise snippet-level supervision.

D. Pseudo Complementary Learning

Although above we have excluded the categories that all snippets surely do not belong to, there are still multiple categories remaining for the multi-label temporal action localization task. To further enhance the snippet-level supervision, we propose to exclude some non-existent categories from those video-level categories for the snippets of less ambiguity. Intuitively, it is easier to exclude some categories than exactly predict the ground-truth categories (i.e., excluding all non-existent categories).

Therefore, we propose Pseudo Complementary Learning to further exclude some categories for snippets in the video. It consists of three parts, namely identifying the snippets of less ambiguity by category information entropy, excluding categories (i.e., snippets likely not belong to) for snippets of less ambiguity, and disambiguation between foreground and background for ambiguous snippets.

1) Ambiguity Identification by Information Entropy: The ambiguous snippets are identified by the category information entropy of the class activation scores $S$. If the category information entropy is higher, the snippet is more likely to be ambiguous. Specifically, we denote the indicator $F = \{f_t\} \in \mathbb{R}^T$ of ambiguous snippets as follows:

$$f_t = \begin{cases} 1, & h_t > \theta, \\ 0, & \text{otherwise}, \end{cases}$$

(3)

where $\theta$ is a preset threshold to estimate the ambiguity of the $t$-th snippet. The $h_t$ represents the category information entropy of $t$-th snippet computed as follows:

$$h_t = -s_{t,fg} \cdot \log(s_{t,fg}) - s_{t,bg} \cdot \log(s_{t,bg}),$$

(4)

$$s_{t,fg} = \sum_{c \in G, c \neq bg} s_{t,c},$$

(5)

where $G \in \mathbb{R}^G$ is the category set except deterministic complementary categories, “fg” denotes the set of foreground action categories and “bg” denotes background category.

2) Background-Aware Pseudo Complementary Labeling: For the snippets of less ambiguity (i.e., $f_t = 0$), we assign the snippet-level pseudo complementary labels $R = \{r_{t,c}\}^{T \times (C+1)} \in \mathbb{R}^{T \times (C+1)}$ to them. Our purpose is to exclude the categories.

Fig. 2. Overview of our proposed Progressive Complementary Learning (ProCL). Given a video, the class activation sequence $S$ and the class-agnostic attention scores $A$ are generated by the classification and the attention heads. ProCL progressively excludes categories that should not belong to, for gradually enhancing snippet-level supervision. First, according to the video-level labels, we exclude the categories that all snippets surely do not belong to (i.e., deterministic complementary categories) by the Deterministic Complementary Learning loss $L_{DCL}$. Then, for snippets of less ambiguity, we exclude some categories that the snippets likely do not belong to (i.e., pseudo complementary categories) by the Multi-scale Pseudo Complementary Learning loss $L_{MPCL}$. Furthermore, for the remaining ambiguous snippets, we disambiguate them by the Foreground Background Discrimination loss $L_{FBD}$, which aims to provide further snippet-level supervision. Besides, the Multiple-Instance Learning loss $L_{MIL}$ is adopted for video-level supervision. Best viewed in color.
that snippets likely not belong to, which we term **pseudo complementary categories**. The process is formulated as follows:

\[
    r_{t,c} = \begin{cases} 
    0, & \text{if } \forall c \in \mathcal{C}, f_t = 0 \text{ and } s_{t,c} < \mu_t, \\
    1, & \text{otherwise},
    \end{cases}
\]  

(6)

where \(\mu_t\) represents the mean value of activation scores in \(t\)-th snippet for the categories contained in \(\mathcal{C}\). Considering the fact that the foreground actions and background cannot co-exist in the same snippet, we further exclude categories with conflict. Specifically, after executing Eq. (6) to exclude the pseudo complementary categories, if a snippet still has the background and at least one foreground action remaining, we exclude the one with a lower prediction score. Moreover, it is worth noting that the foreground could be a set of action categories, and when the background is excluded, multiple action categories may be left.

Based on the pseudo complementary labels \(\mathcal{R} \in \mathbb{R}^{T \times (C+1)}\) obtained above and the snippet-level classification scores \(\mathcal{P} \in \mathbb{R}^{T \times (C+1)}\), the Pseudo Complementary Learning (PCL) loss is given as follows:

\[
    L_{PCL} = \frac{1}{N} \sum_{t=1}^{T} \sum_{c \in \mathcal{G}} \left( f_t = 0 \right) \sum_{c \in \mathcal{G}} -(1 - r_{t,c}) \log (1 - p_{t,c}),
\]  

(7)

where \(N = T - \sum_{t=1}^{T} f_t\) is the number of snippets of less ambiguity, and \(\left( f_t = 0 \right)\) means the ambiguous snippets are not involved in the \(L_{PCL}\).

3) **Disambiguation Between Foreground and Background:**

For the remaining ambiguous snippets, it is inferior to use Pseudo Complementary Labeling because the model is not confident about the prediction of these snippets. Therefore, we propose a Foreground Background Discrimination loss to reduce the ambiguity of the ambiguous snippets, which provides further snippet-level supervision for distinguishing foreground from background.

Specifically, for the ambiguous snippets (i.e., \(f_t = 1\)), we use the class-agnostic attention scores \(\mathcal{A}\) generated by the attention branch and the background category activation scores \(\mathcal{S}_{bg} = \mathcal{S}_{C+1}\) in class activation sequence for foreground-background disambiguation learning. As shown in Figure 2, background attention scores can be obtained via \(\mathcal{B} = (1 - \mathcal{A}) \in \mathbb{R}^{T}\), i.e., \(b_t = 1 - a_t\) for \(t\)-th snippet. And, considering that the background prediction of the two modules (i.e., \(\mathcal{S}_{bg}\) and \(\mathcal{B}\)) should be consistent, we propose the Foreground-Background Discrimination (FBD) loss to disambiguate the ambiguous snippets, which is formulated as follows:

\[
    L_{FBD} = \frac{1}{N} \sum_{t=1}^{T} \sum_{c \in \mathcal{G}} \left( f_t = 1 \right) E(p_{t,bg}, b_t)L_{BCE}(b_t, P_{t,bg})
    + \frac{1}{N} \sum_{t=1}^{T} \sum_{c \in \mathcal{G}} \left( f_t = 1 \right) E(p_{t,bg}, b_t)L_{BCE}(p_{t,bg}, b_t),
\]  

(8)

where \(L_{BCE}\) is a binary cross-entropy loss, and \(p_{t,bg}\) is the classification score of the background category of \(t\)-th snippet in the snippet-level classification scores \(P\). Besides, \(\left( f_t = 1 \right)\) indicates that only ambiguous snippets are involved, and

\[
    N = \sum_{t=1}^{T} f_t. E(\cdot, \cdot) = \exp(-KL(\cdot, \cdot)) \text{ is a weight term based on Kullback-Leibler divergence.}
\]

**E. Multi-Scale Complementary Labeling**

All of the above are based on single-scale snippet sequences, i.e., sampling a sequence of \(T\) snippets uniformly from each video. However, in real-world videos, action instances of the same category could be performed within different temporal durations. Taking “Diving” as an example, it is obvious that jumping from platforms of different heights require various durations, but they are all “Diving” actions. Based on the above consideration, we sample two extra snippet sequences to further improve the quality of the snippet-level supervision.

In detail, we first densely sample snippets \(\mathcal{X}^D \in \mathbb{R}^{2T \times D}\) and sparsely sample snippets \(\mathcal{X}^S \in \mathbb{R}^{N \times D}\). And, we feed the multi-scale snippet sequences (i.e., dense sampling, normal sampling, and sparse sampling) into the classification head with shared parameters to obtain the corresponding class activation sequence \(\mathcal{S}_D = \{s_{t,c}^d\} \in \mathbb{R}^{2T \times (C+1)}\), \(\mathcal{S}_N = \{s_{t,c}^n\} \in \mathbb{R}^{T \times (C+1)}\), and \(\mathcal{S}_S = \{s_{t,c}^s\} \in \mathbb{R}^{N \times T \times (C+1)}\) respectively. Then, the nearest neighbor interpolation is used to re-scale them into \(T\)-snippet sequences. And then, the mean class activate sequence \(\mathcal{S}_{bg} = \{s_{t,c}^B\} \in \mathbb{R}^{T \times C+1}\) and the variance class activation sequence \(\sigma_B = \{s_{t,c}^S\} \in \mathbb{R}^{T \times C+1}\) are obtained as follows:

\[
    s_{t,c}^B = \frac{\sum_{t=1}^{T} s_{t,c}^d}{T},
\]

\[
    s_{t,c}^S = \frac{\sum_{t=1}^{T} (s_{t,c}^d - s_{t,c}^B)^2}{T},
\]

(9)

where \(s_{t,c}^d\) and \(s_{t,c}^S\) denote the mean and variance of the activation score corresponding to the \(c\)-th category of the \(t\)-th snippet, respectively.

Notably, \(\mathcal{S}_D\) indicates the average class activation sequence that incorporates multi-scale information, and \(\sigma_B\) indicates the inconsistency between sequences of different scales. Subsequently, we obtain pseudo complementary labels \(\hat{\mathcal{R}}\) and indicator \(\hat{\mathcal{R}}\) of ambiguous snippets for the multi-scale case using the same approach as Section III-D. Overall, the Multi-scale Pseudo Complementary Learning (MPCL) loss is given as follows:

\[
    L_{MPCL} = \frac{1}{N} \sum_{t=1}^{T} \sum_{c \in \mathcal{G}} \left( f_t = 0 \right) D(\hat{\mathcal{R}}_t, \mathcal{S}_D^c, \mathcal{S}_N^c),
\]

(10)

\[
    D(L, U, J) = \frac{1}{|\mathcal{G}|} \sum_{c \in \mathcal{G}} \exp(-u_c)(1 - l_c)\log(1 - j_c),
\]

(11)

where \(\left( f_t = 0 \right)\) means the ambiguous snippets are not involved in \(L_{MPCL}\), \(u_c, j_c\) denotes the elements of \(c\)-th category in \([L, U, J]\), and \(\exp()\) represents a weight term based on \(\sigma_B\), i.e., a larger variance \(\sigma_B\) leads to more possible misjudgment, and thus the weight should be reduced.

**F. Overall Training Objective and Inference**

Overall, the training objective of our ProCL is given as follows:

\[
    L = L_{MIL} + L_{DCL} + L_{MPCL} + L_{FBD},
\]

(12)

where the coefficient for each loss is set to 1.
In the inference stage, following the standard process [13], we first generate $A$ and $\hat{S}$ for the given video. Then the video-level classification scores are obtained based on $\hat{S}$ to determine which action categories exist in the video by thresholding $\rho$. Next, class-agnostic action proposals are obtained from $A$ following the [13]. For the set of candidate action proposals $\{(t_s, t_e, \psi, c)\}$, we use non-maximal suppression to remove redundant proposals. Notably, our $L_{MPCL}$ introduces no extra cost in the inference phase and takes more computation only in training phase.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate our method on three public benchmarks, i.e., THUMOS14 [84], ActivityNet1.3 [85], and MultiTHUMOS [86] for temporal action localization.

THUMOS14 dataset contains 200 validation videos and 213 test videos of 20 action classes. It is a challenging benchmark with around 15.5 action instances per video and whose videos have diverse durations. Following standard split, we use the validation videos as training set and the test videos as test set.

ActivityNet1.3 is a large dataset that covers 200 action categories, with a training set of 10,024 videos and a validation set of 4,926 videos. It contains around 1.5 action instances per video. Following standard split, we use the training and validation sets for training and testing, respectively.

MultiTHUMOS is an extended version of the THUMOS14 dataset with dense, multi-category action instance annotations. It is more realistic and challenging, which contains 38,690 annotations of 65 action categories with an average of 1.5 labels per frame and 10.5 action classes per video. We use the standard split with 200 validation videos for training and 213 test videos for testing.

Evaluation metric. We evaluate our method with mean average precision (mAP) under multiple temporal Intersection over Union (IoU) thresholds, which are the standard evaluation metrics for temporal action localization. Our results are measured by the officially released evaluation method [85]. Moreover, for the MultiTHUMOS dataset, we additionally adopt the frame-mAP as an evaluation metric, following previous works [87], [88].

B. Implementation Details

We apply I3D [32] pre-trained on Kinetics-400 [95] to extract both RGB and optical flow features, which are concatenated into 2048-dimensional snippet-level features. Each snippet contains continuous non-overlapping 16 frames from the original untrimmed video. Following Hong et al. work [12], the classification head consists of two consecutive blocks (i.e., a block contains Conv1D, LeakyReLU, and Dropout) followed by a Conv1D. The attention head consists of a block followed by a Conv1D and Sigmoid. In the training stage, we set $T = 560$ and $T = 90$ for THUMOS14 and ActivityNet1.3 respectively, while all snippets are taken during inference. Adam optimizer with 0.001 weight decay rate and 0.00003 learning rate is used. The number of training iterations for THUMOS14 and ActivityNet1.3 is set to 10000 and 20000, $\gamma$ of top-k pooling that $k = T/\gamma$ is set to 7 and 10, $\rho$ is set to 0.2 and 0.15. We use identical hyper-parameters on THUMOS14 and MultiTHUMOS. Also, we perform the experiments multiple times with different seeds to report the mean values. The method is implemented in PyTorch [96] and all experiments are performed on an NVIDIA GTX 1080Ti GPU.

C. Comparison With State-of-the-Arts

In Table I, we compare our ProCL with the state-of-the-art WSTAL methods on THUMOS14. We find that our method outperforms all previous methods in terms of AVG 0.1:0.5 and AVG 0.1:0.7. In particular, our approach outperforms all methods that adopt pseudo labeling to obtain snippet-level supervision, including both ways of assigning pseudo labels with preset thresholds [11], [24], [79] and based on predicted action proposals [29], [30]. Also, our approach surpasses some methods [16], [23] that use additional information to provide more supervision. Moreover, the results of our ProCL are even comparable with fully-supervised methods in terms of mAP@0.1 and mAP@0.2, and surpass SSN [6]. The results demonstrate the superior performance of our ProCL, which obtains relatively accurate snippet-level supervision by progressive category exclusion. Furthermore, we conduct experiments on UntrimmedNet [46] (UNT) features of the THUMOS14 dataset. As shown in Table II, our proposed ProCL also achieves state-of-the-art performance, which is consistent with the results on the I3D features. This consistent result not only verifies the effectiveness of our approach but also shows that our method is robust to different types of features.

We also conduct experiments on the ActivityNet1.3 dataset, and the results are reported in Table III. Consistent with the results on THUMOS14, our method also achieves state-of-the-art performance, surpassing the latest methods. This further demonstrates the effectiveness of our ProCL.

In addition, we conduct experiments on the MultiTHUMOS dataset (more realistic and challenging), and the results are reported in Table IV. From the results, we find that our ProCL outperforms the latest method (i.e., DDG-Net [83]) and pseudo-label-based methods (i.e., UM [11] and ASM-Loc [30]), which is attributed to our ProCL providing effective snippet-level supervision based on the category exclusion perspective.

D. Ablation Studies

1) Quantitative Analysis of Each Component: In Table V, we show the quantitative analysis of the different components of our ProCL. We use “AVG 0.1:0.7” for the performance metric, which is the average of mAP values for different IoU thresholds [0.1:0.1:0.7]. Exp.1 is the baseline only trained with $L_{MIL}$, and Exp.7 is our full method. By using the Deterministic Complementary Learning loss $L_{DCL}$, Exp.3 improves by 3.6% compared with Exp.1, which attributes to accurately exploiting the snippet-level supervision indicated by video-level labels. Surprisingly, our $L_{DCL}$ outperforms the
previous pseudo-label-based approach \( L_{PL} \) in Exp.2 by 2.1%, which benefits from the fact that our \( L_{DCL} \) is noise-free and provides the appropriate snippet-level supervision for model training. Exp.4 shows 0.8% performance improvement over Exp.3, which demonstrates that our Pseudo Complementary Learning further provides effective snippet-level supervision for the model. By introducing the Foreground Background Discrimination loss, Exp.6 achieve further performance gains by 0.9%. Moreover, Exp.7 shows that the information fusion of multi-scale snippet sequences further improves the action localization performance.

2) Label Precision of Different Labeling Methods: In Figure 3 we compare the precision of the labels that are assigned by the previous pseudo-label-based method [30] and our pseudo complementary labeling. With just the single-scale snippet sequence, our method has a 20% precision improvement. By comparing PCL and MPCL, it is found that the collaborative information of multi-scale snippet sequences indeed contributes to improving the quality of pseudo complementary labels. Not only that, but we also found that the precision of our pseudo complementary labels has improved with iterative refinement. This indicates that the pseudo complementary labels contribute to the predictions of the model, and vice versa. These further demonstrate that our pseudo complementary labeling is superior to previous pseudo labeling methods, which alleviates the noise problem introduced by assigning category labels directly.

E. Qualitative Analysis

1) Qualitative Analysis of Each Component: In Figure 4, we visualize the qualitative results for the different components, where “GT” represents the ground truth. Compared with the second row, the results of the third row show that the activation of some background snippets is obviously suppressed, which demonstrates that our deterministic complementary categories are able to improve the performance
of action localization. Meanwhile, from the fourth row of the figure, we find that our pseudo complementary labels also show a positive effect on the localization results. Furthermore, consistent with the quantitative results, with the introduction of $L_{FBD}$ (the fifth row in the figure), the background snippets are further suppressed, demonstrating that our Foreground Background Discrimination loss has the ability to disambiguate ambiguous snippets.

2) Qualitative Analysis by Category Prediction Scores: In Figure 5, we visualize the prediction scores of two similar actions (i.e., “ThrowDiscus” and “Shotput”) in the category activation sequence $S$. We find that the pseudo-label-based method [30] incorrectly predicts an instance of “ThrowDiscus” as “Shotput” (i.e., denoted by red dash boxes), even though this video does not contain any instance of the “Shotput”. This is because the “ThrowDiscus” and “Shotput” actions are visually very similar. In contrast, our method correctly distinguishes “ThrowDiscus” from “Shotput”, as the activation of “Shotput” are suppressed. This result demonstrates that our ProCL can provide the correct snippet-level supervision information, which effectively suppresses the prediction of categories that snippets should not belong to.

3) Qualitative Analysis of Foreground Background Discrimination: In Figure 6, we visualize the localization

| Supervision | Method | Publication | mAP@ IOU (%) | AVG | mAP@ IOU (%) |
|-------------|--------|------------|--------------|-----|--------------|
| Pull        | TAL-Net [94] | CVPR 2018 | 38.2 ± 0.3 | 30.2 | 30.2 ± 0.3 |
|             | BSN [7] | ECCV 2018 | 46.5 ± 0.0 | 30.0 | 30.0 ± 0.0 |
|             | P-GCN [37] | ICCV 2019 | 42.9 ± 0.1 | 27.0 | 27.0 ± 0.1 |
| Weak        | TSRRNet [10] | AAAI 2019 | 33.1 ± 0.7 | 21.8 | 21.8 ± 0.7 |
|             | STAR [9] | AAAI 2019 | 31.1 ± 0.8 | - | - |
|             | PretShowNet [23] | AAAI 2020 | 34.8 ± 0.9 | 22.5 | 22.5 ± 0.9 |
| Weak²       | BasNet [36] | AAAI 2020 | 34.5 ± 0.4 | 22.2 | 22.2 ± 0.4 |
|             | A2CL-PF [57] | ECCV 2017 | 36.8 ± 0.2 | 22.5 | 22.5 ± 0.2 |
|             | $f$ TSCN [24] | ECCV 2017 | 33.3 ± 1.4 | 21.7 | 21.7 ± 1.4 |
|             | ACSNet [14] | AAAI 2021 | 36.3 ± 0.8 | 23.9 | 23.9 ± 0.8 |
|             | $f$ UM [11] | AAAI 2021 | 37.0 ± 0.3 | 23.7 | 23.7 ± 0.3 |
|             | AUMN [22] | CVPR 2019 | 38.3 ± 0.5 | 23.5 | 23.5 ± 0.5 |
|             | $f$ UGC [55] | CVPR 2019 | 39.1 ± 0.4 | 23.8 | 23.8 ± 0.4 |
|             | PACN [48] | ICCV 2021 | 37.6 ± 0.6 | 26.0 | 26.0 ± 0.6 |
|             | Wang et al. [34] | TVT 2021 | 37.1 ± 0.4 | 26.1 | 26.1 ± 0.4 |
|             | Li et al. [44] | TVT 2021 | 40.9 ± 0.7 | 25.6 | 25.6 ± 0.7 |
| Weak        | FTCL [77] | CVPR 2022 | 40.0 ± 0.3 | 24.8 | 24.8 ± 0.3 |
|             | DCC [78] | CVPR 2022 | 38.8 ± 0.3 | 24.3 | 24.3 ± 0.3 |
|             | $f$ Huang et al. [79] | CVPR 2022 | 40.6 ± 0.3 | 25.0 | 25.0 ± 0.3 |
|             | $f$ ASCL-LoC [30] | CVPR 2022 | 41.0 ± 0.4 | 25.1 | 25.1 ± 0.4 |
|             | $f$ SMM [45] | TVT 2022 | 41.7 ± 0.4 | 26.0 | 26.0 ± 0.4 |
|             | DGCGN [80] | MM 2022 | 37.2 ± 0.8 | 23.9 | 23.9 ± 0.8 |
|             | Boost-WTAL [61] | CVPR 2023 | 41.8 ± 0.0 | 26.0 | 26.0 ± 0.0 |
|             | P-MIL [82] | CVPR 2023 | 41.8 ± 0.4 | 25.5 | 25.5 ± 0.4 |
|             | PBNet [60] | TVT 2023 | 39.9 ± 0.4 | 25.3 | 25.3 ± 0.4 |
|             | LPR [69] | TVT 2023 | 41.4 ± 0.3 | 25.4 | 25.4 ± 0.3 |
| ProCL       | Ours | TVT 2023 | 38.3 ± 0.6 | 24.5 | 24.5 ± 0.6 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE V

| Exp | $\mathcal{L}_{\text{MIL}}$ | $\mathcal{L}_{\text{DCL}}$ | $\mathcal{L}_{\text{PCL}}$ | $\mathcal{L}_{\text{FBD}}$ | $\mathcal{L}_{\text{MPCL}}$ | $\mathcal{L}_{\text{PL}}$ | mAP@IoU (%) | AVG |
|-----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-------------|-----|
| 1   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 66.1 60.8 52.2 44.1 37.1 24.7 12.9 | 42.7 |
| 2   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 67.8 63.5 53.8 45.8 38.1 26.3 15.8 | 44.2 |
| 3   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 72.2 66.3 56.7 48.0 39.2 27.1 14.5 | 46.3 |
| 4   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 72.5 66.5 57.5 48.8 40.5 28.3 15.7 | 47.1 |
| 5   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 72.5 66.7 58.9 47.9 40.1 27.5 14.2 | 48.8 |
| 6   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 73.7 68.6 59.6 49.4 41.1 27.9 15.5 | 48.0 |
| 7   | ✓               | ✓               | ✓               | ✓               | ✓               | ✓               | 74.4 69.0 60.1 50.5 41.9 28.6 15.9 | 48.6 |

Fig. 3. The label precision of different labeling methods, where PL denotes the pseudo labels obtained by following previous pseudo-label-based method [30], PCL and MPCL denotes our pseudo complementary labels obtained from single-scale snippet sequence and multi-scale snippet sequences, respectively. Best viewed in color.

Fig. 4. Visualization of ablation studies, where “+” indicates a new component is added upon the previous experiment. Red dash boxes note areas of significant improvement. Best viewed in color.

4) Visualization of the Localization Results: In Figure 7, we visualize the localization results between the Baseline and our ProCL. We observe that our method is able to accurately localize actions compared to the Baseline, which is attributed to our method providing more effective snippet-level supervision. Specifically, by disambiguating the foreground from background and identifying ambiguous snippets, our approach effectively mitigates the problem of incorrectly recognizing ambiguous background snippets as actions (Figure 7 (a)). Moreover, by providing more effective snippet-level supervision, it alleviates the problem that only the most discriminative snippets can be recognized under the traditional Multiple-Instance Learning paradigm (Figure 7 (c)).

F. More Analysis

1) Analysis of Hyper-Parameter $\theta$: In Figure 8, we analyze the value of threshold $\theta$ that used in pseudo complementary
Fig. 7. Visualization of localization results. This figure shows the predicted proposals with the highest confidence score corresponding to the number of ground truths. The "GT" denotes the ground truth, Baseline denotes the model trained with \( L_{MIL} \), and ProCL is our full method. Transparent frames represent the background frames.

Fig. 8. Analysis of the threshold \( \theta \). The figure shows the performance with different thresholds \( \theta \) that are used to identify ambiguous snippets. AVG 0.1:0.7 denotes the average mAP for IoU thresholds 0.1:0.1:0.7. Base means that this experiment adopts both \( L_{MIL} \) and \( L_{DCL} \). Base+PCL means the addition of pseudo complementary learning loss \( L_{PC} \) to Base. ProCL means the full method, and LPR [49] is the state-of-the-art method.

Fig. 9. Performance of different discrepancy estimate functions in \( L_{FBD} \). AVG 0.1:0.7 denotes the average mAP for IoU thresholds 0.1:0.1:0.7. MAE, BCE, MSE, JS, and KL denote Mean Absolute Error, Binary Cross-Entropy, Mean Square Error, Jensen–Shannon divergence, and Kullback-Leibler divergence are adopted in \( L_{FBD} \) of our full method, respectively. Also, Base denotes that this experiment adopts multi-instance learning loss \( L_{MIL} \) and deterministic complementary learning \( L_{DC} \).

2) Different Implementations of \( L_{FBD} \): In Figure 9, we adopt different discrepancy estimate functions to align the background attention score \( B \) (i.e., \( B = 1 - A \)) and the prediction scores \( S_{bg} \) of the background category in the class activation sequence \( S \). We find that by using different discrepancy estimation functions in \( L_{FBD} \), the performance surpasses most of the previous methods, where MAE, KL, and BCE outperform the latest state-of-the-art method LPR [49] (i.e., 47.1 of AVG 0.1:0.7). Moreover, the best performance is achieved by using BCE (Ours) as the discrepancy estimate function. We conclude that the information entropy term in the KL and JS slightly blocks the optimization of the model.

labeling (i.e., in Eq. (3)). Based on the design of foreground-background information entropy, \( \theta \) is used to balance \( L_{FBD} \) and \( L_{PC} \). Specifically, if \( \theta \) is too high or too low, no snippet will be identified as ambiguous snippet or all snippets will be identified as ambiguous snippets, which will lead to failure of \( L_{FBD} \) or \( L_{PC} \), and the experimental results in Figure 8 are consistent with this discussion. In addition, we find that \( \theta = 0.69 \) achieves the best performance (i.e., ProCL). Moreover, from the ProCL and LPR [49], it can be seen that our ProCL outperforms the latest state-of-the-art method in different values of \( \theta \).
Fig. 10. The cumulative cover rate of training videos with iteration. The “cover rate” denotes the proportion of activations that are suppressed over the total non-target categories activation map. “real-time” indicates the cover rate in each iteration.

Fig. 11. The concurrency rate of training videos with iteration. The “concurrency rate” denotes the fraction of snippets that satisfy the concurrency condition (i.e., all non-target categories are suppressed simultaneously) over all snippets.

Fig. 12. The Average mAP with iteration. The AVG is the average mAP for IoU threshold from 0.1 to 0.7 with 0.1 increment. The horizontal coordinate of this figure is consistent with Figures 10, and 11.

3) Difference and Complementarity Analysis of $L_{MIL}$ and $L_{DCL}$: Here we qualitatively and quantitatively illustrate the differences and complementarities of $L_{MIL}$ and $L_{DCL}$. For convenience, we term the categories existing in the video (i.e., video-level labels $\tilde{Y} = 1$) as target categories and categories not existing (i.e., video-level labels $\tilde{Y} = 0$) as non-target categories (i.e., deterministic complementary categories).

(1) Implementation differences. As shown in Figure 13, the $L_{MIL}$ utilizes the target category labels (i.e., $\tilde{Y} = 1$) and $L_{DCL}$ utilizes the non-target category labels (i.e., $\tilde{Y} = 0$) to compute the loss. They make use of video-level labels $\tilde{Y}$ from different perspectives. Besides, $L_{MIL}$ provides only video-level supervision while $L_{DCL}$ provides snippet-level supervision.

(2) Effect differences. As shown in Figure 13, the $L_{MIL}$ mainly improves the video-level activation (i.e., generated by Top-K pooling) of target categories, while $L_{DCL}$ suppresses the snippet-level activation of all non-target categories for all snippets simultaneously. Although $L_{MIL}$ suppresses some video-level activation (i.e., generated by Top-K pooling) of non-target categories, it is incomplete in both temporal (i.e., Top-K snippets, typically 12.5%) and categorical (i.e., Top-K cannot contain all non-target categories for each snippet) dimensions. As shown in Figure 10, compared with $L_{DCL}$, the $L_{MIL}$ (i.e., Top-K pooling) neither covers all snippets in real-time nor cumulatively. Besides, as shown in Figure 11, the $L_{MIL}$ cannot guarantee that all non-target categories are suppressed simultaneously for each snippet. By contrast, $L_{DCL}$ completely suppresses the activation of all non-target categories in both temporal and categorical dimensions.

(3) Performance complement. As shown in Figure 12, the combination of $L_{MIL}$ and $L_{DCL}$ not only accelerates the convergence, but also significantly improves the performance (i.e., improve by 3.6%) of localization. Incorporating $L_{MIL}$ with $L_{DCL}$ makes better use of the given annotation.

TABLE VI

| Exp | $\lambda_{DCL}$ | $\lambda_{MPCL}$ | $\lambda_{FB}$ | AVG |
|-----|----------------|-----------------|----------------|-----|
| 1   | 1.0            | 0.1             | 0.5            | 48.2|
| 2   | 1.0            | 0.5             | 0.1            | 46.9|
| 3   | 0.5            | 0.1             | 1.0            | 48.0|
| 4   | 0.5            | 1.0             | 0.1            | 46.2|
| 5   | 0.1            | 0.5             | 1.0            | 46.6|
| 6   | 0.1            | 1.0             | 0.5            | 45.8|
| 7   | 1.5            | 1.0             | 0.5            | 47.9|
| 8   | 0.5            | 0.1             | 1.5            | 48.0|
| 9   | 0.1            | 1.5             | 1.0            | 46.4|
| 10  | 1.5            | 0.1             | 0.5            | 48.1|
| 11  | 0.1            | 0.5             | 1.5            | 47.1|
| 12  | 0.5            | 1.5             | 0.1            | 46.4|
| 13  | 1.0            | 1.0             | 1.0            | 48.6|

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
information and $L_{DCL}$ providing more effective snippet-level supervision to the model.

4) Analysis of Different Loss Weights: In Figure 14 and Table VI, we tune different weights of our proposed $L_{DCL}$, $L_{MPCL}$, and $L_{FBD}$ losses. Specifically, Figure 14 shows the results of tuning a single loss weight while the other loss weights are set to 1. Table VI shows some results for different combinations of the loss weights. From the results in Figure 14 and Table VI, we find that the best result is achieved when the weights of all losses are set to 1. This demonstrates that all three losses (i.e., $L_{DCL}$, $L_{MPCL}$, and $L_{FBD}$) proposed by us have a positive effect on the model.

V. Conclusion

In this work, we propose a Progressive Complementary Learning (ProCL) method that progressively enhances the snippet-level supervision from the perspective of category exclusion. Specifically, our ProCL gradually excludes the categories that snippets should not belong to, based on different confidence levels. Moreover, we propose three snippet-level losses for weakly-supervised temporal action localization, bridging the gap between video-level supervision and unavailable snippet-level supervision, without using additional auxiliary models (e.g., GloVe [98]) or information (e.g., frequency or pose). Extensive experiments demonstrate that ProCL can provide effective snippet-level supervision to the model, significantly improving the weakly-supervised temporal action localization.

Despite the progress of our ProCL, it still faces challenges in addressing scenarios with overlapping subtle actions (i.e., multi-labels) and long-duration actions (i.e., duration longer than 2 minutes). These scenarios are difficult to address for weakly-supervised methods due to the absence of precise timestamp annotations. Nevertheless, our ProCL outperforms the existing works in these scenarios involved in the MultiTHUMOS and ActivityNet1.3 datasets. In our future work, we will investigate more effective ways to enhance snippet-level supervision and cooperate multi-modal information to further address the scenarios with overlapping subtle actions and long duration actions.

REFERENCES

[1] A. Ciptadi, M. S. Goodwin, and J. M. Rehg, “Movement pattern histogram for action recognition and retrieval,” in Proc. ECCV, 2014, pp. 695–710.

[2] M. Ramezani and F. Yaghmaee, “A review on human action analysis in videos for retrieval applications,” Artif. Intell. Rev., vol. 46, no. 4, pp. 485–514, Dec. 2016.

[3] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, “MIST: Multiple instance self-training framework for video anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14004–14013.

[4] F.-T. Hong, X. Huang, W.-H. Li, and W.-S. Zheng, “MINI-Net: Multiple instance ranking network for video highlight detection,” in Proc. ECV, 2020, pp. 345–360.

[5] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage CNNs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1049–1058.

[6] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, “Temporal action detection with structured segment networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2933–2942.

[7] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, “BSN: Boundary sensitive network for temporal action proposal generation,” in Proc. ECV, 2018, pp. 3–19.

[8] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, “BMN: Boundary-matching network for temporal action proposal generation,” in Proc. ICCV, 2019, pp. 3889–3898.

[9] Y. Xu et al., “Segregated temporal assembly recurrent networks for weakly supervised multiple action detection,” in Proc. AAAI, 2019, pp. 9070–9078.

[10] X.-Y. Zhang, H. Shi, C. Li, K. Zheng, X. Zhu, and L. Duan, “Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision,” in Proc. AAAI, 2019, vol. 33, no. 1, pp. 9227–9234.

[11] P. Lee, J. Wang, Y. Lu, and H. Byun, “Weakly-supervised temporal action localization by uncertainty modeling,” in Proc. AAAI, 2021, pp. 1854–1862.

[12] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng, “Cross-modal consensus network for weakly supervised temporal action localization,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 1591–1599.

[13] A. Islam, C. Long, and R. Radke, “A hybrid attention mechanism for weakly-supervised temporal action localization,” in Proc. AAAI, 2021, pp. 1637–1645.

[14] Z. Liu et al., “ACSNet: Action-context separation network for weakly supervised temporal action localization,” in Proc. AAAI, 2021, pp. 2233–2241.

[15] X. Liu, J.-Y. Lee, and H. Jin, “Learning video representations from correspondence proposals,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4268–4276.

[16] F. Ma et al., “SF-Net: Single-frame supervision for temporal action localization,” in Proc. ECV, 2020, pp. 420–437.

[17] Z. Luo et al., “Weakly-supervised action localization with expectation-maximization multi-instance learning,” in Proc. ECV, 2020, pp. 729–745.

[18] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, “3C-Net: Category count and center loss for weakly-supervised action localization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8678–8686.

[19] P. Nguyen, B. Han, T. Liu, and G. Prasad, “Weakly supervised action localization by sparse temporal pooling network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6752–6761.

[20] R. Zeng, C. Gan, P. Chen, W. Huang, Q. Wu, and M. Tan, “Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5797–5808, Dec. 2019.
X. Yu, T. Liu, M. Gong, and D. Tao, “Learning with biased complementary labels,” in Proc. ECCV, 2018, pp. 68–83.

L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, “Learning with multiple complementary labels,” in Proc. ICML, 2020, pp. 3072–3081.

Y. Gao and M.-L. Zhang, “Discriminative complementary-label learning with weighted loss,” in Proc. ICML, 2021, pp. 3587–3597.

D.-B. Wang, L. Feng, and M.-L. Zhang, “Learning from complementary labels via partial-output consistency regularization,” in Proc. 30th Int. Joint Conf. Artif. Intell., Aug. 2021, pp. 3075–3081.

B. Han et al., “SIGUA: Forgetting may make learning with noisy labels more robust,” in Proc. ICML, 2020, pp. 4006–4016.

Y. Kim, J. Yim, J. Yun, and J. Kim, “NLNL: Negative learning for noisy labels,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 101–110.

K. Kim, J. Yoon, H. Shon, and J. Kim, “Joint negative and positive learning for noisy labels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9437–9446.

X. Xia et al., “Part-dependent label noise: Towards instance-dependent label noise,” in Proc. NeurIPS, 2020, pp. 7597–7610.

M. Chen, J. Gao, S. Yang, and C. Xu, “Dual-evidential learning for weakly-supervised temporal action localization,” in Proc. ECCV, 2022, pp. 216–232.

J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, “Weakly supervised action selection learning in video,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 7583–7592.

Z. Yang, J. Qin, and D. Huang, “ACGNet: Action complement graph network for weakly-supervised temporal action localization,” in Proc. AAAI, 2022, pp. 3090–3098.

J. Gao, M. Chen, and C. Xu, “Fine-grained temporal contrastive learning for weakly-supervised temporal action localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 19967–19977.

J. Li, T. Yang, W. Ji, J. Wang, and L. Cheng, “Exploring denoised cross-video contrast for weakly-supervised temporal action localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 19882–19892.

L. Huang, L. Wang, and H. Li, “Weakly supervised temporal action localization via representative snippet knowledge propagation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 3262–3271.

H. Shi, X.-Y. Zhang, C. Li, L. Gong, Y. Li, and Y. Bao, “Dynamic graph modeling for weakly-supervised temporal action localization,” in Proc. 30th ACM Int. Conf. Multimedia, Oct. 2022, pp. 3820–3828.

Y. Wang, Y. Li, and H. Wang, “Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 155, Jun. 2023, pp. 18878–18887.

H. Ren, W. Yang, T. Zhang, and Y. Zhang, “Proposal-based multiple instance learning for weakly-supervised temporal action localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 2394–2404.

X. Tang, J. Fan, C. Luo, Z. Zhang, M. Zhang, and Z. Yang, “DDG-Net: Discriminability-driven graph network for weakly-supervised temporal action localization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2023, pp. 6622–6632.

Y.-G. Jiang et al. (2014). THUMOS Challenge: Action Recognition With A Large Number of Classes. [Online]. Available: http://msrc.uic.edu/THUMOS14/

F. C. Heilbron, V. Escorcia, B. Ghanem, and I. C.Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 961–970.

S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, “Every moment counts: Dense detailed labeling of actions in complex videos,” Int. J. Comput. Vis., vol. 126, nos. 2–4, pp. 375–389, Apr. 2020.

A. Piergiovanni and M. S. Ryoo, “Learning latent super-events to detect multiple activities in videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5304–5313.

A. Piergiovanni and M. Ryoo, “Temporal Gaussian mixture layer for videos,” in Proc. ICML, 2019, pp. 5152–5161.

Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, “AutoLoc: Weakly-supervised temporal action localization in untrimmed videos,” in Proc. ECCV, 2018, pp. 154–171.
Zhongang Qi received the B.S. and Ph.D. degrees in information science and electronic engineering from Zhejiang University, Hangzhou, China. He was a Research Associate with the School of Electrical Engineering and Computer Science, Oregon State University. He is currently a Principal Researcher with the ARC Laboratory, Tencent PCG. His research interests include multimedia and relational data mining, machine learning, and computer vision.

Ying Shan is currently a Distinguished Scientist with Tencent PCG, where he is the Director of the ARC Laboratory. Before joining Tencent PCG, he was with Microsoft Research as a Post-Doctoral Researcher, SRI International (Sarnoff Subsidiary) as a Senior MTS, and Microsoft Bing Ads as a Principal Scientist Manager. He has published over 70 papers in top conferences and journals in the areas of computer vision, machine learning, and data mining, served as ACMs of CVPR and a Senior PC of KDD, and holds a number of U.S./International patents. He is currently leading research and development efforts in web search and content AI for a suite of social media and content distribution products.

Jian-Fang Hu received the B.S. and Ph.D. degrees from the School of Mathematics, Sun Yat-sen University, Guangzhou, China, in 2010 and 2016, respectively. He is currently an Associate Professor with Sun Yat-sen University. He has published several scientific papers in international conferences and journals, including ICCV, CVPR, NeurIPS, ECCV, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. His research interests include video-text analysis, human mesh modeling, and human action recognition.

Wei-Shi Zheng is now a Full Professor with Sun Yat-sen University. He has now published more than 200 papers, including more than 150 publications in main journals (TPAMI, IJCV, SIGGRAPH, TIP) and top conferences (ICCV, CVPR, ECCV, NeurIPS). His research interests include person association and activity understanding, AI robotics learning, and the related weakly supervised/unsupervised and continuous learning machine learning algorithms. He has ever served as area chairs of ICCV, CVPR, ECCV, BMVC, and NeurIPS. He is an Associate Editors/on the editorial board of IEEE-TPAMI, Artificial Intelligence Journal, Pattern Recognition. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He is a Cheung Kong Scholar Distinguished Professor, a recipient of the NSFC Excellent Young Scientists Fund, a Fellow of IAPR, and a recipient of the Royal Society-Newton Advanced Fellowship of U.K.