Boosting Semi-Supervised Temporal Action Localization by Learning from Non-Target Classes

Kun Xia, *Member, IEEE,* Le Wang, *Senior Member, IEEE,* Sanping Zhou, *Member, IEEE,* Gang Hua, *Fellow, IEEE,* Wei Tang, *Member, IEEE.*

***Abstract*—The crux of semi-supervised temporal action localization (SS-TAL) lies in excavating valuable information from abundant unlabeled videos. However, current approaches predominantly focus on building models that are robust to the error-prone target class (*i.e.* the predicted class with the highest confidence) while ignoring informative semantics within non-target classes. This paper approaches SS-TAL from a novel perspective by advocating for learning from non-target classes, transcending the conventional focus solely on the target class. The proposed approach involves partitioning the label space of the predicted class distribution into distinct subspaces: target class, positive classes, negative classes, and ambiguous classes, aiming to mine both positive and negative semantics that are absent in the target class, while excluding ambiguous classes. To this end, we first devise innovative strategies to adaptively select high-quality positive and negative classes from the label space, by modeling both the confidence and rank of a class in relation to those of the target class. Then, we introduce novel positive and negative losses designed to guide the learning process, pushing predictions closer to positive classes and away from negative classes. Finally, the positive and negative processes are integrated into a hybrid positive-negative learning framework, facilitating the utilization of non-target classes in both labeled and unlabeled videos. Experimental results on THUMOS14 and ActivityNet v1.3 demonstrate the superiority of the proposed method over prior state-of-the-art approaches.**

***Index Terms*—Temporal Action Localization Semi-Supervised Learning.**

I. Introduction

TEMPORAL Action Localization (TAL) attempts to temporally locate and recognize action instances of interest in untrimmed videos. It is a fundamental yet challenging task in computer vision, with a wide range of applications, such as security surveillance [1], [2] and human behavior analysis [3], [4]. Traditional TAL approaches [5]–[8] rely heavily on large-scale, well-annotated datasets, a process that is both tedious and time-consuming in practice. Hence, Weakly-Supervised Temporal Action Localization (WS-TAL) [9]–[12] where only video-level labels are available receives attention. However, WS-TAL [13], [14] imposes an intractable problem of distinguishing between actions and backgrounds due to missing instance-level annotations. In response to these challenges, recent efforts have been directed toward Semi-Supervised Temporal Action Localization (SS-TAL), aiming to train models using only a limited number of labeled samples with instance-level annotations and a substantial amount of unlabeled data.

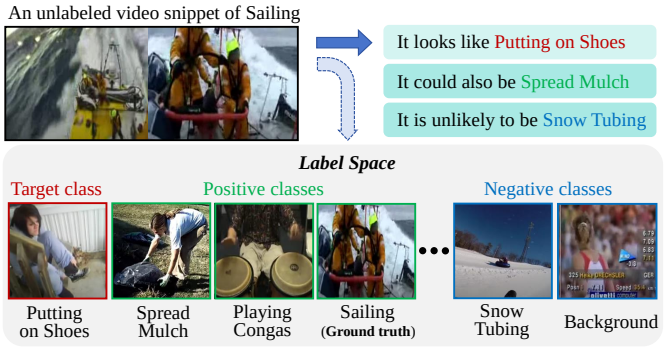


Fig. 1. Illustration of unreliable predictions on an unlabeled video snippet. A common practice is to treat the action class with the highest confidence , *i.e.*, "Putting on Shoes" as its target class for model optimization, while the ground truth label , *i.e.*, "Sailing" is buried in the non-target classes.

Recent advancements of SS-TAL [15]–[18] have demonstrated notable success, leveraging two well-known semi-supervised learning paradigms: consistency regularization and self-training. Consistency regularization approaches [15], [16] aim to generate reliable predictions through a teacher model to guide the learning process of the student model. However, learning a decent teacher model with limited labeled data is as challenging as the SS-TAL task itself. More recently, self-training approaches [17], [18] tailored for SS-TAL have dominated this area, attaining state-of-the-art performance. These approaches iteratively use the current model to assign pseudo labels to unlabeled videos and train a new model on both the labeled videos and the pseudo-labeled videos.

Despite achieving promising results, existing approaches simply utilize the *target class* (*i.e.* the predicted class with the highest confidence) as the pseudo label, which has two significant drawbacks. First, the target class tends to be highly noisy, given that the model is trained on a limited amount of labeled data. Second, the *non-target classes* are entirely disregarded, even though they often contain valuable cues about the action. An illustrative example is depicted in Figure 1. A

Manuscript received xxx; revised xxx; accepted xxx. Date of publication xxx; date of current version xxx. This work was supported in part by National Science and Technology Major Project under Grant 2024YFB4708100, National Natural Science Foundation of China under Grants 62088102 and U24A20325, Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80, and Fundamental Research Funds for the Central Universities under Grant XTR072022001. *(Corresponding author: Le Wang.)*

Kun Xia is with the School of Computer Science and Technology, Xi'an Jiaotong University, Le Wang and Sanping Zhou are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China. (e-mail: {kunxia, lewang, spzhou}@xjtu.edu.cn)

Gang Hua is with Dolby Laboratories, Bellevue, WA 98004, USA. (e-mail: ganghua@gmail.com)

Wei Tang is with the Department of Computer Science, University of Illinois, Chicago, IL 60607, USA. (e-mail: tangw@uic.edu)

video snippet of "Sailing" is mistakenly assigned the target class "Putting on Shoes" for self-training, leading to noisy pseudo labels, while the semantics of the ground truth label are buried among the ignored non-target classes.

In this paper, we approach Semi-Supervised Temporal Action Localization from a novel perspective by learning informative semantics from non-target classes, moving beyond the traditional focus on the target class. Given a predicted class probability distribution on unlabeled data, we often observe two phenomena. First, when the ground truth label does not align with the target class, it frequently falls within other top-ranked classes in the prediction. Second, it is highly unlikely that the low-confidence or bottom-ranked classes contain the ground truth label.

Building upon this observation, we partition the *label space* of the predicted class probability distribution into four subspaces: *target class*, *positive classes*, *negative classes*, and *ambiguous classes*. As mentioned earlier, the target class is defined as the highest-confidence class. Positive classes encompass non-target classes with high confidences, often covering the ground truth class. Negative classes comprise non-target classes with low confidences, making them unlikely to contain the ground truth class. The remaining non-target classes form the ambiguous classes.

While the idea of learning from non-target classes is intriguing, two key challenges need to be addressed: *How should the non-target classes, especially the positive and negative classes, be identified from the predicted class distribution? How can the model effectively learn from these non-target classes?* In response to the first challenge, we devise innovative strategies to *adaptively* select high-quality positive and negative classes from the label space. This involves modeling both the confidence and rank of a class in relation to those of the target class. To tackle the second challenge, we introduce novel positive and negative losses designed to push the prediction closer to the positive classes and push it away from the negative classes. Consequently, *positive learning* empowers the model to extract richer semantics relevant to the true class but absent in the target class, while *negative learning* reinforces the model's belief of which classes are incorrect. Given the high uncertainty and noise associated with ambiguous classes, we exclude them from the training process. Finally, we integrate the positive and negative learning processes into a hybrid positive-negative learning framework to leverage the non-target classes across both labeled and unlabeled videos.

The main contributions of this paper are summarized as follows:

- This paper introduces a novel paradigm for SS-TAL by emphasizing learning from non-target classes, transcending the conventional focus solely on the target class. The approach involves partitioning the label space of the predicted class distribution into different subspaces, aiming to mine both positive and negative semantics that are absent in the target class, while excluding ambiguous classes.
- Key aspects of this novel paradigm include identifying the positive and negative classes and learning from these non-target classes. The paper introduces innovative strategies for adaptively selecting high-quality positive and negative classes from the label space. Additionally, new positive and negative losses are proposed to guide the non-target learning effectively. These processes are integrated into a hybrid positive-negative learning framework, facilitating the utilization of non-target classes in both labeled and unlabeled videos.
- We evaluate the proposed approach on THUMOS14 and ActivityNet v1.3 under a wide range of training settings. Extensive experiments demonstrate that our approach surpasses the previous state-of-the-art methods.

The rest of the paper is organized as follows. Section II discusses related work. We present the technical details of the proposed method in Section III. Experimental results and discussions are presented in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

In this section, we review previous works related to ours, which we categorize into three parts: (1) Fully-supervised temporal action localization, (2) Semi-supervised temporal action localization and (3) Learning on pseudo labels.

Fully- and Weakly-Supervised Temporal Action Localization has witnessed significant advancements in recent years through using plentiful well-annotated videos. Concretely, early *anchor-based* methods [19]–[21] typically employ the multi-scale anchors and attach a classification head and a boundary regression head to refine these pre-defined anchors. G-TAD [20] updates features of all snippets in a video via a graph network, and classifies each pre-defined anchor and then regresses their boundaries. However, anchor-based methods are sensitive to hyper-parameters such as the number and scales of anchors. Moreover, the temporal durations of the proposals are inflexible, and their temporal boundaries are imprecise. *Anchor-free* methods [22]–[26] directly regress the boundary locations or perform frame-level action classification to reduce the complexity. BSN [22] and BMN [27] intensively predict frame-level probability sequences to generate action proposals. However, boundary-based methods often suffer from noisy boundary locations and thus generate unreliable results. BU-MR [23] introduces two regularization terms (Intra-phase Consistency and Inter-phase Consistency) to enforce consistent predictions in and between starting, continuing, and ending phases. BSN++ [28] exploits complementary boundary regression and relation modeling for temporal proposal generation and achieves state-of-the-art performance. Current prevailing *Transformer-based* methods [29]–[32] tackle temporal action localization in a Transformer encoder-decoder framework, which models action instances as a set of learnable action queries. To reduce the cost of manual annotation, many weakly-supervised TAL approaches had already achieved relatively advanced performance using only video-level labels. [33] proposes a learnable dictionary of class centroids to enforce semantic consistency among snippets with similar representations, improving feature discriminability under video-level supervision. [34] develops a hierarchical attention framework with metric learning to model snippet-level

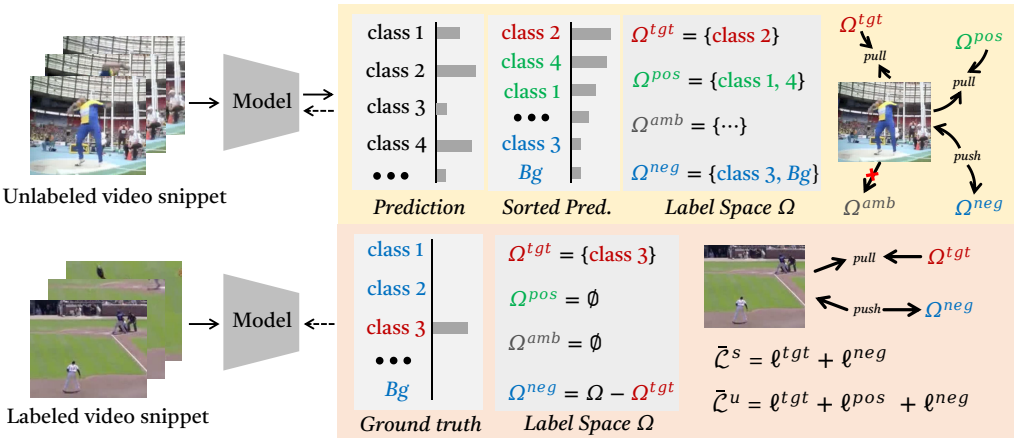


Fig. 2. **An overview of our proposed Non-target Classes Learning framework**. It follows the self-training paradigm, which iteratively uses the current model to assign pseudo labels to unlabeled videos and trains a new model on both the labeled videos and the pseudo-labeled videos. Given an unlabeled video snippet, the current model predicts a probability distribution of all classes. Our method adaptively partitions the label space Ω into a target class Ω^{tgt}, positive classes Ω^{pos}, negative classes Ω^{neg}, and ambiguous classes Ω^{amb}, by modeling both the confidence and rank of a class in relation to those of the target class. Based on the label space partition, we design the new positive learning loss ℓ_{pos} and negative learning loss ℓ_{neg} to mine positive and negative semantics that are absent in the target class, while excluding ambiguous classes.

structural information and suppress action-context confusion. [35] introduces a semantic query mechanism and pseudo-labels to refine snippet-level features and enhance action completeness in a point-supervised setting. [36] introduces an Equivalent Classification Mapping (ECM) mechanism, which unifies pre-classification and post-classification pipelines via a shared classifier to learn a more robust and consistent representation. [12] proposes a global-local attention mechanism that integrates inter-segment similarity and local receptive fields to enhance the discovery of foreground action segments. [37] presents the SODA framework, which addresses core challenges in WTAL through an astute background response strategy for suppression and a self-distillation learning strategy for discovering more complete action frames.

Semi-Supervised Temporal Action Localization leverages valuable information from the unlabeled data with lower annotation cost. Existing arts [15]–[18], [38] benefit from the development of general *semi-supervised learning* [39], [40] and follow two frameworks, *i.e.*, consistency regularization and self-training. Ji *et al.* [15] design two essential types of sequential perturbations to make consistent action proposal predictions for both teacher and student models. Nag *et al.* [17] develop a proposal-free temporal masking model to solve the localization error propagation problem. Xia *et al.* [18] tackle the label noise problem and present a noise-tolerant framework to update the model with reliable pseudo labels that are strictly screened. Different from existing methods, this paper provides a new perspective for SS-TAL, by learning informative semantics from non-target classes that are ignored by existing approaches.

Learning on Pseudo Labels is an important yet key technology in semi-supervised learning. However, most approaches [41]–[46] are limited to learning directly from the target class, so it is inevitable that the model will be misled by noisy pseudo labels. Chen *et al.* [41] present a proposal self-assignment for pseudo label assignment, which injects the proposals from student into teacher and generates accurate pseudo labels to match each proposal in the student model accordingly. Apart from above methods, the complementary label has been used to specify a class that a sample does not belong to [47]. Yu *et al.* [48] theoretically analyze the problem of biased complementary labels and propose to estimate transition probabilities with no bias. Chen *et al.* [45] introduce an additional entropy meaning loss which enforces a uniform distribution of non-target classes to avoid them competition with target class and a negative learning method which selects k categories as complementary labels based on the prediction consistency for low confidence samples. Kim *et al.* [49] aim at learning clean data with ground truth labels while training noise data with a randomly selected label as a complementary label. [50] tackles timestamp-supervised and unsupervised temporal action segmentation by designing frame-level positive/negative sample sets for contrastive learning in the feature space. Differing from existing methods, we introduce a novel negative learning approach that adaptively selects richer negative classes based on the confidence of the target class. These negative classes are more informative, reducing the risk of selecting the true label. Additionally, our new positive learning method extracts additional semantics relevant to the true class that may be absent in the target class.

III. Method

In this section, we first describe the problem setting of SS-TAL and our motivation in Sec. III-A and Sec. III-B, respectively. An overview of our method is illustrated in Figure 2. We introduce the target class-based learning in Sec. III-C. Our adaptive negative learning and positive learning strategies are introduced in Sec. III-D and Sec. III-E, respectively.

A. Preliminaries

Problem Setting. Given a smaller set of N^l labeled videos $\{X_i^l, Y_i^l\}_{i=1}^{N^l}$ and a larger set of N^u unlabeled videos $\{X_i^u\}_{i=1}^{N^u}$,

semi-supervised temporal action localization (SS-TAL) aims to improve action detection by effectively learning from both labeled and unlabeled data. The annotation Y_i^l of each labeled video contains the start time, end time, and action category of each action instance.

Feature Embedding. For a video X, following conventions [26], [51], we extract its snippet-level features $\{\boldsymbol{x}_i\}_{i=1}^{N^v}$ from consecutive frames by a fine-tuned two-stream network, where N^v is the number of video snippets.

Baseline Model. Recent works [17], [18] formulate SS-TAL as a snippet-level classification task. Our method also adopts the proposal-free framework with self-training for SS-TAL, which locates action instances by a classification head and a mask head optimized by a mask learning loss $\mathcal{L}^m$, a refinement loss $\mathcal{L}^{ref}$, and a feature reconstruction loss $\mathcal{L}^{rec}$, as in prior arts [17], [18]. The learning objective is to minimize the loss function below:

$$\mathcal{L} = \mathcal{L}^s + \alpha\mathcal{L}^u + \mathcal{L}^m + \mathcal{L}^{ref} + \mathcal{L}^{rec}, \tag{1}$$

where $\mathcal{L}^s$ and $\mathcal{L}^u$ denote the supervised loss and the unsupervised loss applied on labeled videos and unlabeled videos, respectively, and α is a hyper-parameter. The main purpose of the action detection model is to learn the parameters θ of a model $\mathbb{F}(\cdot;\theta)$ by optimizing a cross-entropy (CE) loss function on both labeled and unlabeled data:

$$\mathcal{L}^s = \frac{1}{N^v}\sum_{i=1}^{N^v} \ell^{ce}\left(\mathbb{F}\left(\boldsymbol{x}_i^l;\theta\right), \boldsymbol{y}_i^l\right), \tag{2}$$

$$\mathcal{L}^u = \frac{1}{N^v}\sum_{i=1}^{N^v} \ell^{ce}\left(\mathbb{F}\left(\boldsymbol{x}_i^u;\theta\right), \boldsymbol{y}_i^u\right), \tag{3}$$

where $\boldsymbol{x}_i^l$ and $\boldsymbol{x}_i^u$ are respectively the i-th snippet feature vectors of a labeled video and an unlabeled video. $\boldsymbol{y}_i^l \in \mathbb{R}^{C+1}$ and $\boldsymbol{y}_i^u \in \mathbb{R}^{C+1}$ are the one-hot vectors of their ground truth label and pseudo label, respectively, including C action classes and a background class.

B. Motivation

Existing approaches simply utilize the *target class* (*i.e.* the predicted class with the highest confidence) as the pseudo label. The target class tends to be highly noisy, given that the model is trained on a limited amount of labeled data, thereby significantly degrading the self-training. This paper moves beyond the traditional focus on the target class and addresses SS-TAL from a novel perspective, by learning informative semantics from non-target classes. The motivation for our approach stems from two key observations regarding a predicted class probability distribution on unlabeled data. First, when the ground truth label does not align with the target class, it frequently falls within other top-ranked classes in the prediction. Second, it is highly unlikely that the low-confidence or bottom-ranked classes contain the ground truth label.

Building upon these observations, we divide the label space of the predicted class probability distribution on an unlabeled video snippet into four subspaces:

$$\Omega = \{1, \ldots, C+1\} = \Omega^{tgt} \cup \Omega^{pos} \cup \Omega^{neg} \cup \Omega^{amb}, \tag{4}$$

where Ω^{tgt} only holds the target class while Ω^{pos}, Ω^{neg} and Ω^{amb} are the positive classes, negative classes, and ambiguous classes, respectively. Positive classes encompass non-target classes with high confidences, often covering the ground truth class. Negative classes comprise non-target classes with low confidences, making them unlikely to contain the ground truth class. The remaining non-target classes form the ambiguous classes. Complementary to traditional target class-based learning (Sec. III-C), *negative learning* (Sec. III-D) reinforces the model's belief of which classes are incorrect, while *positive learning* (Sec. III-E) empowers the model to extract richer semantics relevant to the true class but absent in the target class. Given the high uncertainty and noise associated with ambiguous classes, we exclude them from self-training.

C. Learning from Target Class

Existing approaches first obtain the probability distribution $\boldsymbol{p} = \mathbb{F}(\boldsymbol{x}^u;\theta)$ from an unlabeled snippet $\boldsymbol{x}^u$, and then use $argmax_c(p_c)$ as its target class to construct the one-hot pseudo label vector $\boldsymbol{y}^u$. The learning objective is formulated as the cross-entropy loss between the model prediction and the target class:

$$\ell^{tgt} = -\sum_{c=1}^{C+1} y_c^u \log p_c, \tag{5}$$

where C is the number of action classes and $y_c^u \in \{0,1\}$ represents whether the target class is present. The model is trained by maximizing the log-likelihood of the target class.

D. Learning from Negative Classes

As previously discussed in Sec. III-B, the model may exhibit uncertainty regarding whether a video snippet belongs to the noisy target class but can be fairly certain that it does not belong to negative classes. To effectively learn negative information, a negative class is chosen from non-target classes, and the model is then trained using a negative learning loss given by:

$$\tilde{\ell}^{neg} = -\log\left(1 - p_{c^{neg}}\right), \tag{6}$$

which aims to minimize the log-likelihood on the negative class. However, selecting suitable negative classes is challenging. On the one hand, only selecting one negative class is insufficient to learn valuable negative information. On the other hand, regarding all non-target classes as negative classes would carry the risk of negatively learning the ground truth semantics buried in non-target classes. Therefore, we design an adaptive negative learning strategy to tackle this challenge.

Specifically, let $\boldsymbol{p} = [p_1, \ldots, p_{C+1}]$ denote the class probability distribution predicted on an unlabeled video snippet. Then, we sort it in ascending order of the confidence:

$$\hat{\boldsymbol{p}} = sorted(\boldsymbol{p}) = [min(\boldsymbol{p}), \ldots, max(\boldsymbol{p})], \tag{7}$$

where $max(\boldsymbol{p})$ corresponds to the confidence of the target class. The higher $max(\boldsymbol{p})$ is, the more certain the model is that the target class aligns with the ground truth class. It also means that we can treat more non-target classes as negative classes for learning negative information. This line of reasoning motivates us to design an adaptive negative learning strategy by taking the

confidence of the target class as reference. Concretely, we first compute the cumulative probability of its bottom-k classes. If the cumulative probability is less than $max(\boldsymbol{p})$, these k classes will be treated as negative classes that contribute equivalently to negative learning, which could be formulated as:

$$\Omega^{neg} = \left\{ k : \sum_{c=1}^{k} \hat{p}_c \leqslant max(\boldsymbol{p}) \right\}, \tag{8}$$

where Ω^{neg} holds k negative classes that meet the above criteria. When $max(\boldsymbol{p})$ is very high, it suggests that low-confidence classes carry a lower risk of containing ground truth semantics; therefore, the model will involve more low-confidence classes into Ω^{neg} for negative learning. When $max(\boldsymbol{p})$ is very low, the ground truth semantics will be more likely to be buried in low-confidence classes; therefore, the model will only select a few negative classes since the cumulative probability of bottom-k classes is small. Based on the negative classes, we reformulate the negative learning loss as:

$$\ell^{neg} = - \sum_{c \in \Omega^{neg}} \log\left(1 - p_c\right). \tag{9}$$

Our proposed adaptive negative learning will enable the model to effectively learn underlying negative information from as many negative classes as possible.

E. Learning from Positive Classes

Learning from the remaining non-target classes (excluding negative classes) is intriguing as the ground truth semantics are buried among them. However, learning positive information from all remaining non-target classes is suboptimal since ambiguous classes would confuse the model. Therefore, we leverage the confidence of the target class as an informative indicator to select positive classes:

$$\Omega^{pos} = \left\{ k : \hat{p}_k \geqslant \lambda \cdot max(\boldsymbol{p}) \right\}, \tag{10}$$

where Ω^{pos} holds k positive classes that meet the above criteria and λ is a hyper-parameter. In this way, the model will only select the classes whose confidences are close to the target class since they are likely to share similar information related to the ground truth class. Based on the positive classes, we formulate the positive learning loss as:

$$\ell^{pos} = - \sum_{c \in \Omega^{pos}} \log p_c. \tag{11}$$

The positive learning empowers the model to extract richer semantics relevant to the true class but absent in the target class.

F. Hybrid Positive-Negative Learning

Finally, we integrate the proposed negative learning and positive learning into our semi-supervised TAL framework. In training, for all labeled data, the ground truth labels are treated as the target classes with no doubt. The remaining classes, *i.e.*, all non-target classes, will act as negative classes for negative learning, since they are completely unrelated to the ground truth label. Thus, we apply the cross-entropy loss and negative loss for all labeled data. For unlabeled data, we apply the cross-entropy loss for target classes, the positive and negative losses for positive and negative classes as mentioned above, respectively. The overall loss function is shown below:

$$\mathcal{L} = \bar{\mathcal{L}}^s + \bar{\mathcal{L}}^u + \mathcal{L}^m + \mathcal{L}^{ref} + \mathcal{L}^{rec}, \tag{12}$$

where the supervised loss $\bar{\mathcal{L}}^s$ contains the cross-entropy loss ℓ^{tgt} and the negative learning loss ℓ^{neg}, *i.e.*, $\bar{\mathcal{L}}^s = \ell^{tgt} + \ell^{neg}$. The unsupervised loss $\bar{\mathcal{L}}^u$ contains the positive learning loss ℓ^{pos} and the negative learning loss ℓ^{neg} as well as ℓ^{tgt}, *i.e.*, $\bar{\mathcal{L}}^u = \ell^{tgt} + \ell^{neg} + \ell^{pos}$. In addition, the SS-TAL model is mainly composed of a classification head and a mask head, which is further optimized by the mask learning loss $\mathcal{L}^m$, the refinement loss $\mathcal{L}^{ref}$, and the feature reconstruction loss $\mathcal{L}^{rec}$, as in [17], [18]. The key distinction between our proposed full method (Eq. 12) and the baseline (Eq. 1) is that the latter does not include our novel negative loss ℓ^{neg} and positive loss ℓ^{pos} in $\mathcal{L}^s$ and $\mathcal{L}^u$.

In inference, the model generates action instance predictions for each testing video by the classification and mask predictions, as in SPOT [17]. More specifically, we can obtain candidate snippets by using a classification threshold and a localization threshold on the classification and mask heads, respectively. Therefore, only the video snippets with high class probabilities and mask scores are selected as top scoring snippets. We use a set of thresholds to produce sufficient candidates. For each candidate, we compute its confidence score by multiplying the classification probability and mask score. In post-processing, Soft-NMS [52] is finally applied to obtain top scoring results.

G. Discussion

Our proposed hybrid positive-negative learning framework introduces key innovations that *move beyond the standard paradigm of generating independent pseudo-labels for each class*. While conventional semi-supervised methods assign a single pseudo-label (the target class) per snippet, our approach innovates in two primary ways.

First, instead of focusing on a single, often noisy, target class, we structure the entire predicted label distribution into distinct subspaces, including target, positive, negative, and ambiguous classes. This allows for a systematic mining of semantic information from and across non-target classes, which are typically discarded.

Second, the novelty lies not merely in using multiple classes, but in how we select and learn from them. We adaptively identify high-quality positive and negative classes based on their confidence and rank relative to the target class, modeling inter-class relationships. Subsequently, we introduce novel loss functions that explicitly pull predictions towards the selected positive classes and push them away from the negative classes. This relational learning provides richer supervisory signals and enhances robustness to label noise.

Therefore, our method's novelty is not in generating per-class pseudo-labels with fixed thresholds, but in formulating a holistic learning strategy that leverages the structural information within the entire label distribution to guide the model more effectively.

TABLE I
MAIN RESULTS ON THUMOS14 AND ACTIVITYNET V1.3 WITH DIFFERENT PERCENTAGES OF LABELED VIDEOS, WHERE BASELINE REFER TO THE BASELINE MODEL WITHOUT POSITIVE AND NEGATIVE LEARNING LOSSES. NOTABLY, SSP AND SSTAP EMPLOY UNTRIMMEDNET [51] TRAINED WITH 100% CLASS LABELS FOR PROPOSAL CLASSIFICATION.

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%	SSP [15]	TSN	44.2	34.1	24.6	16.9	9.3	25.8	38.9	28.7	8.4	27.6
	SSTAP [16]	TSN	45.6	35.2	26.3	17.5	10.7	27.0	40.7	29.6	9.0	28.2
	SPOT [17]	TSN	49.4	40.4	31.5	22.9	12.4	31.3	49.9	31.1	8.3	32.1
	NPL [18]	TSN	50.0	41.7	33.5	23.6	13.4	32.4	50.9	32.0	7.9	32.6
	APL [53]	TSN	51.5	42.5	34.6	24.4	13.5	33.3	51.5	32.4	8.2	33.0
	Baseline	TSN	49.6	40.8	33.2	23.0	12.9	31.9	51.2	31.8	7.3	32.1
	Ours	TSN	**52.4**	**43.5**	**35.6**	**24.9**	**14.7**	**34.2**	**53.0**	**34.4**	**9.2**	**34.5**
20%	SPOT [17]	TSN	52.6	43.9	34.1	25.2	16.2	34.4	51.7	32.0	6.9	32.3
	NPL [18]	TSN	53.9	45.6	36.2	26.9	16.5	35.8	52.1	32.9	7.9	32.9
	APL [53]	TSN	54.8	45.9	37.1	28.5	16.9	36.6	52.4	33.3	8.3	33.4
	Baseline	TSN	53.5	45.2	36.2	27.1	16.2	35.6	51.8	32.2	7.0	32.4
	Ours	TSN	**55.2**	**46.9**	**38.0**	**28.7**	**17.2**	**37.2**	**53.5**	**34.7**	**9.4**	**34.8**
40%	SPOT [17]	TSN	54.4	45.8	37.2	29.7	19.4	37.3	53.3	33.0	6.6	33.2
	NPL [18]	TSN	56.2	46.7	38.8	30.3	19.5	38.3	53.4	33.9	8.1	33.8
	APL [53]	TSN	57.0	47.1	39.5	**32.7**	20.1	39.3	53.5	33.8	8.5	34.1
	Baseline	TSN	54.8	45.9	37.3	29.9	19.1	37.4	53.5	33.2	6.9	33.4
	Ours	TSN	**57.5**	**48.0**	**39.6**	31.5	**21.4**	**39.6**	**54.1**	**35.6**	**9.4**	**35.4**
60%	SSP [15]	TSN	53.2	46.8	39.3	29.7	19.8	37.8	49.8	34.5	7.0	33.5
	SSTAP [16]	TSN	56.4	49.5	41.0	30.9	21.6	39.9	50.1	34.9	7.4	34.0
	SPOT [17]	TSN	58.9	50.1	42.3	33.5	22.9	41.5	52.8	35.0	8.1	35.2
	NPL [18]	TSN	59.0	51.4	42.9	34.3	23.3	42.2	53.9	35.8	8.5	35.7
	APL [53]	TSN	59.7	51.6	43.2	34.9	23.6	42.6	54.2	36.2	8.6	35.9
	Baseline	TSN	58.7	50.0	42.6	33.7	23.0	41.6	52.9	34.9	7.9	35.0
	Ours	TSN	**59.9**	**52.6**	**43.9**	**35.7**	**24.0**	**43.2**	**54.4**	**35.8**	**9.5**	**35.9**

IV. EXPERIMENTS

In this section, we first introduce the evaluation datasets and metrics. Then, we describe the implementation details of our framework and compare our method with previous state-of-the-art methods under conventional evaluation protocols. Lastly, we conduct comprehensive ablation studies on each component to validate the effects.

A. Datasets and Metrics

Evaluation Datasets. Following conventions [26], [29], we evaluate our proposed method on two challenging TAL benchmarks, *i.e.*, THUMOS14 [54] and ActivityNet v1.3 [55]. THUMOS14 [54] contains 200 validation videos and 213 testing videos, including 20 action categories. It is very challenging since each video has more than 15 action instances. Following the common setting [56], we use the validation set for training and evaluate on the testing set. ActivityNet v1.3 [55] is a large-scale benchmark for video-based action localization. It contains 10k training videos and 5k validation videos corresponding to 200 different actions. Following the standard practice [57], we train our method on the training set and test it on the validation set.

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

B. Implementation Details

Following the conventional setting [16], [17], [32], we extract each video snippet feature over every fixed consecutive frames by TSN [58] pre-trained on Kinetics [59]. The temporal dimension is fixed at 100 and 256 for ActivityNet v1.3 and THUMOS14, respectively. Our action localization framework adopts the popular proposal-free approach SPOT [17], which is mainly composed of a classification head and a mask head. Our main contributions focus on the classification head, which originally adopts the cross-entropy loss for target classes. Also, we employ another anchor-free approach Actionformer [29] with the I3D backbone [60] for fair comparisons.

For semi-supervised setting, we first pre-train our model on the training set for 12 epochs and then we fine-tune the pre-trained model for 15 epochs with a learning rate of 10^{-4} for ActivityNet v1.3 and 10^{-5}, and a cosine learning rate decay is used. Following SPOT [17], we adopt the same label sharpening operator and the threshold set for mask. The Soft-NMS [52] is performed on ActivityNet v1.3 and THUMOS14 with a threshold of 0.6 and 0.4, respectively. $\alpha = 1$. For the labeling ratios, we introduce four SS-TAL settings with different label sizes. Following NPL [18], we randomly select 10%, 20%, 40%, and 60% training videos as the labeled set and

the remaining as the unlabeled set. Both labeled and unlabeled sets are accessible for SS-TAL model training.

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

The main results are reported in Table I, where we report mAP at different tIoU thresholds and average mAP. We can observe that our method achieves stable performance improvements over previous works across all data splits on both datasets. We also present the performance of our baseline model in Tabel I. We can see that our main contributions achieve significant performance gains, benefiting from the superiority of the proposed framework.

Specifically, for the THUMOS14 dataset, it is a challenging TAL benchmark due to dense action instances and ambiguous semantics. Our method still outperforms all other comparable methods in all labeled ratios, indicating that the performance gains from our positive and negative learning strategies. Especially, our method obtains remarkable performance when the number of labeled data is very limited (with only 10% or 20% labeled videos). It demonstrates that our method could learn underlying valuable information from non-target classes.

The superiority of the proposed method is more emphasized for ActivityNet v1.3, which is a more large-scale video dataset so as to provide a larger label space for effective hybrid positive-negative learning. As depicted in Table I, our method shows a distinct improvement compared to all other methods. Additionally, the improvements suggests that indirectly learning from positive and negative classes further benefits SS-TAL.

In addition, our experiments on THUMOS14 (dense, short actions) and ActivityNet v1.3 (sparse, longer videos) provided the following insights.

For THUMOS14, the model tends to produce predictions with a sharper confidence distribution (higher confidence for the top class). Here, our scheme more aggressively prunes ambiguous classes, focusing on a clearer separation between high-confidence positives and low-confidence negatives. For ActivityNet v1.3, predictions often have a flatter confidence distribution. Our adaptive scheme responds by being more conservative, often resulting in a broader set of classes being considered as positive or negative, which helps capture the more varied and co-occurring semantics in complex, long-term videos.

D. Ablation Study

Effectiveness of loss terms. To prove our core insight, *i.e.*, learning underlying informative semantics from non-target classes, we conduct experiments in Table II to ablate each loss step by step. Above all, we use ℓ^{tgt} trained model as our baseline, achieving average mAP of 31.4% and 32.1% on THUMOS14 and Activitynet v1.3, respectively. Applying the proposed ℓ^{neg} consistently improves the baseline by a large margin on both benchmarks, arguably since the potential negative information improves the snippet-level semantic discrimination. In addition, the proposed ℓ^{pos} also significantly improve the performance of the model by excavating the ground truth label related semantics from the positive classes.

TABLE II
ABLATION STUDY OF DIFFERENT LOSSES ON THUMOS14 AND ACTIVITYNET V1.3 WITH 10% LABELS, WHERE ℓ^{tgt} IS THE VANILLA CROSS-ENTROPY LOSS FOR SNIPPET-LEVEL ACTION CLASSIFICATION, AND ℓ^{neg} AND ℓ^{pos} ARE THE PROPOSED NEGATIVE AND POSITIVE LEARNING LOSSES FOR EXCAVATING COMPLEMENTARY INFORMATION.

ℓ^{tgt}	ℓ^{neg}	ℓ^{pos}	THUMOS14 (%)						ActivityNet v1.3 (%)			
			0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
✓			49.6	41.1	33.2	22.7	12.9	31.9	51.2	31.8	7.3	32.1
✓	✓		50.6	42.7	34.4	24.0	14.3	33.2	52.3	32.7	8.3	33.4
✓	✓	✓	**52.4**	**43.5**	**35.6**	**24.9**	**14.7**	**34.2**	**53.0**	**34.4**	**9.2**	**34.5**

TABLE III
ABLATION STUDY OF SS-TAL RESULTS ON THUMOS14 USING I3D FEATURES AND ACTIONFORMER [29], WHERE THE LABEL RATIO IS 10% AND ⋆ REPRESENTS ONLY USING LABELED VIDEOS.

Method	Backbone	THUMOS14 (%)			
		0.3	0.5	0.7	Avg.
ActF⋆ [29]	I3D	28.5	14.1	4.1	15.6
NPL (ActF) [18]	I3D	32.8	20.1	7.2	20.3
APL (ActF) [53]	I3D	35.1	25.6	11.0	24.5
Ours (ActF)	I3D	**36.2**	**26.9**	**11.9**	**25.6**

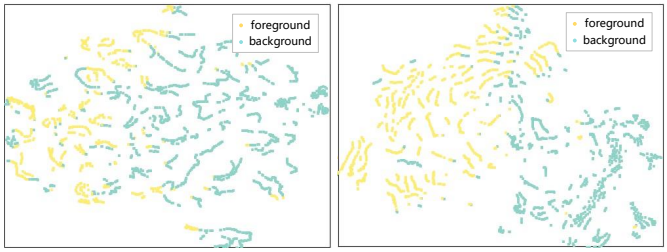
Generality on different features and architectures. Our main experiments are based on features extracted by a TSN backbone. To further demonstrate the generality of our proposed hybrid positive-negative learning loss, we conduct additional experiments on two distinct setups: (1) using features from a fundamentally different I3D backbone [60], and (2) integrating our loss into the powerful Actionformer detector [29], which has a Transformer-based architecture. As shown in Table III, our method provides consistent performance gains over the strong baselines across all setups. Since our method operates solely on the predicted class probability distribution without modifying the feature extraction or architectures, these results confirm that its effectiveness is both feature-agnostic and model-agnostic.

Empirical study of hyper-parameter λ. Ground truth classes are often hidden in positive classes, which is ignored by target-class-based learning methods. In contrast, we introduce a hyper-parameter λ that adaptively selects the number of positive classes based on the confidence of the sample. Then, we conduct an ablation study to vary the value of λ and delve into its impact on the performance. From Table IV-D, it can be observed that higher λ choosing fewer positive classes may make it difficult to fully learn the informative semantics via positive learning while lower λ choosing more classes as positive classes may carry the risk of involving unreliable ambiguous classes.

Qualitative analysis of the positive and negative learning. Learning complementary information from non-target classes contributes to improving the class-level representation. To verify this point, we present the visualizations of foreground-background features and foreground-instance features in Figure 3 and Figure 4, respectively. On the one hand, from Figure 3, the proposed hybrid positive-negative learning separates foreground and background features more clearly by excavating the

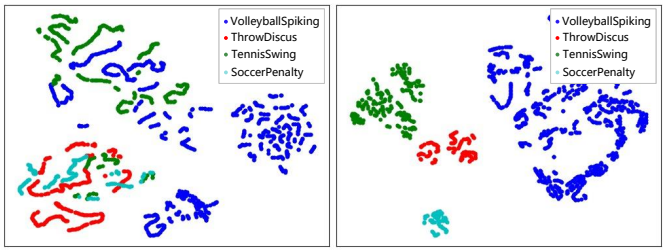
TABLE IV
EMPIRICAL STUDY OF HYPER-PARAMETER λ ON THUMOS14 WITH 10% LABELS, WHERE λ AFFECTS THE NUMBER OF THE POSITIVE CLASSES FOR POSITIVE LEARNING.

λ value	THUMOS14 (%)			
	0.3	0.5	0.7	Avg.
0.90	51.6	35.1	14.6	33.8
0.85	**52.4**	**35.6**	**14.7**	**34.2**
0.75	50.3	34.6	14.2	33.2
0.60	48.4	32.9	12.8	30.8
0.50	46.5	29.7	9.6	27.9



without positive-negative learning with positive-negative learning

Fig. 3. Effect of our method on foreground-background subtask. We present the visualization of foreground feature and background feature on an unlabeled THUMOS14 video.



without positive-negative learning with positive-negative learning

Fig. 4. Effect of our method on foreground-instance subtask. We present the visualization of features of four challenging classes on THUMOS14.

TABLE V
COMPARISON WITH THE SOFT PSEUDO-LABEL METHOD [61] AND COMPLEMENTARY LABEL [47]. THE COMPARISON RESULTS VERIFY THE SUPERIORITY OF OUR METHOD OVER PREVIOUS SEMI-SUPERVISED TECHNOLOGIES.

Method	THUMOS14 (%)			
	0.3	0.5	0.7	Avg.
soft pseudo label	49.5	33.2	12.7	31.7
complementary label	50.0	33.5	13.1	32.1
Ours	**52.4**	**35.6**	**14.7**	**34.2**

ground truth semantics hidden in positive classes. On the other hand, from Figure 4, we can observe that the model produces a much clearer boundary of each class with the hybrid positive-negative learning. It shows that our method could improve the generalization ability of the model.

Comparison with other semi-supervised approaches. To validate the superiority of our method over previous semi-supervised technologies, we explore the soft pseudo-label method [61] and the complementary label method [47] (learning from a random non-target class). We incorporate their main ideas into our work. The comparison results in Table V indicate that the soft pseudo-label produced by the model is quite noisy and includes limited extra knowledge beyond the target label. In contrast to the complementary label, our hybrid positive-negative learning can adaptively extract richer, more informative action semantics from unlabeled videos while reducing the risk of choosing the noisy label.

TABLE VI
SS-TAL RESULTS USING I3D FEATURES ON THUMOS14, WHERE WE USE 10% AND 60% LABELED DATA.

Labels	Method	THUMOS14 (%)		
		0.3	0.5	0.7
10%	SSP [15]	43.1	25.5	9.6
	SSTAP [16]	45.3	27.5	11.0
	SPOT [17]	49.1	31.7	12.6
	Ours	**51.3**	**34.8**	**14.9**
60%	SSP [15]	53.5	39.7	20.4
	SSTAP [16]	55.9	41.6	22.0
	SPOT [17]	58.7	42.4	23.1
	Ours	**59.8**	**44.1**	**24.3**

TABLE VII
MEAN AND VARIANCE OF OUR METHOD'S PERFORMANCE OVER 5 DATA FOLDS.

label ratio	Method	THUMOS14 (%)				
		0.3	0.4	0.5	0.6	0.7
10%	Ours	52.1±0.5	43.5±0.3	35.7±0.3	24.8±0.2	14.6±0.2
20%	Ours	55.0±0.4	46.8±0.3	38.0±0.2	28.6±0.2	17.0±0.2
40%	Ours	57.4±0.2	47.9±0.2	39.7±0.2	31.5±0.1	21.3±0.1
60%	Ours	59.9±0.2	52.5±0.1	43.9±0.2	35.7±0.1	24.1±0.1

Besides, we further compare the SS-TAL results using another popular backbone I3D [60] features in TAL, which is more powerful than TSN. The experimental results are shown in Table VI. By introducing the proposed positive and negative losses, the performance gains support our point, confirming that the superiority of our method is feature-agnostic.

Mean and Variance of the Performance. Table VII shows the mean and variance of our method's performance over 5 data folds. With the increase of labeled data, the performance of the model becomes more stable.

Visualization results. As shown in Figure 5, we provide some qualitative results by previous work SPOT [17] and our approach, where the model is trained with 10% and 40% labeled data on both THUMOS14 and ActivityNet v1.3. Benefiting from using non-target classes, our method can locate and recognize the target actions more accurately, demonstrating the superiority of our method.

More Analysis on Label Space. This paper approaches SS-TAL from a novel perspective by advocating for learning from non-target classes, transcending the conventional focus solely on the target class. From Figure 6, it can be observed

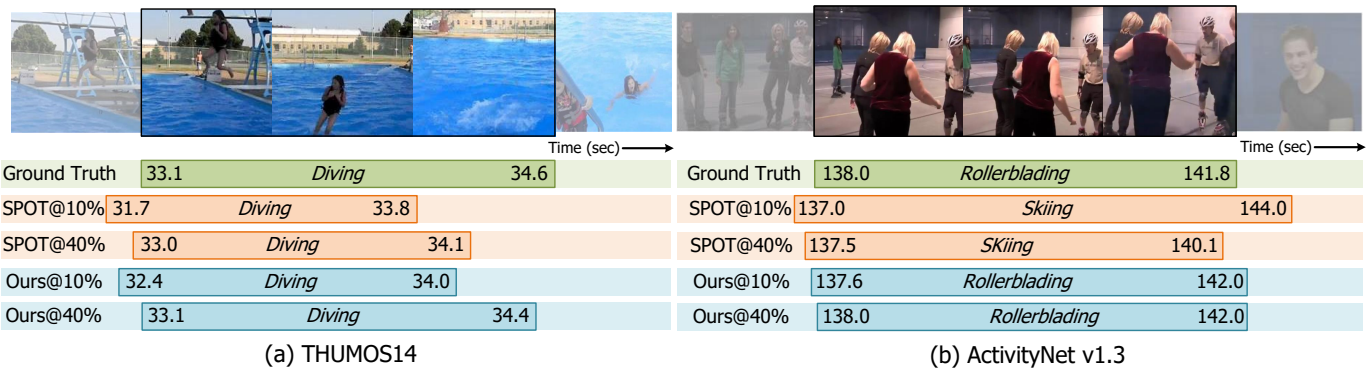


Fig. 5. Qualitative SS-TAL result comparison of our proposed method with SPOT [17] on two untrimmed videos from (a) THUMOS14 and (b) ActivityNet v1.3, respectively.

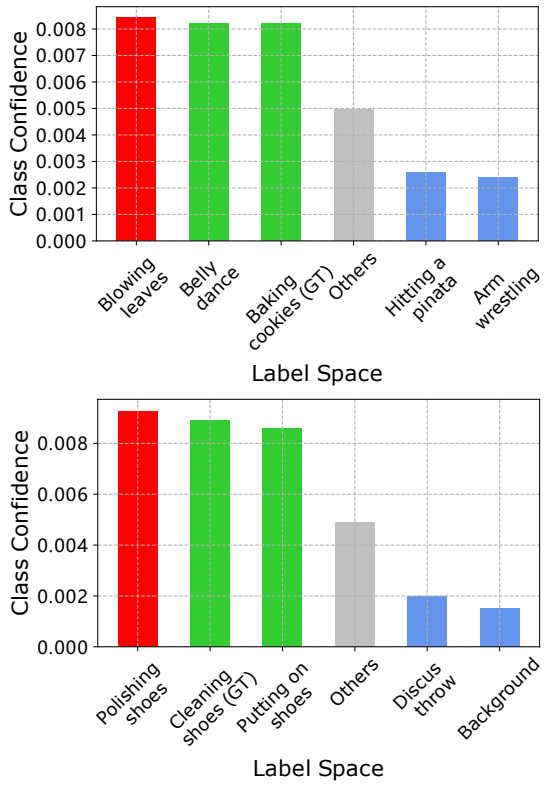


Fig. 6. Visualization examples of label space, where we selectively visualize several positive classes (green bar) and negative classes (blue bar).

that the target class tends to be highly noisy, resulting in the model degeneration. Thus, it is important that learning informative semantics from non-target classes. The proposed positive learning could empower the model to extract richer semantics relevant to the true class but absent in the target class, while the negative learning could reinforce the model's belief of which classes are incorrect. So, exploring the valuable information from label space is a promising topic worthy of further study.

V. Conclusion

In this paper, we introduce a novel paradigm for SS-TAL by emphasizing learning from non-target classes, transcending the conventional focus solely on the target class. The approach fist partitions the entire label space of the predicted class distribution into different subspaces, aiming to mine both positive and negative semantics that are absent in the target class, while excluding ambiguous classes. Then, we develop innovative strategies for adaptively selecting high-quality positive and negative classes from the label space. Additionally, new positive and negative losses are proposed to guide the non-target learning effectively. The extensive experiments on two popular benchmarks with consistent performance gains demonstrate the effectiveness of our method.

References

[1] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Higcin: Hierarchical graph-based cross inference network for group activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 45, no. 6, pp. 6955–6968, 2020.

[2] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 591–600.

[3] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 284–299.

[4] Z. Qing, S. Zhang, Z. Huang, X. Wang, Y. Wang, Y. Lv, C. Gao, and N. Sang, "Mar: Masked autoencoders for efficient action recognition," *IEEE Transactions on Multimedia (T-MM)*, vol. 26, pp. 218–233, 2023.

[5] J. Yang, P. Wei, Z. Ren, and N. Zheng, "Gated multi-scale transformer for temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, vol. 26, pp. 5705–5717, 2023.

[6] Q. Li, G. Zu, H. Xu, J. Kong, Y. Zhang, and J. Wang, "An adaptive dual selective transformer for temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, 2024.

[7] M.-G. Gan and Y. Zhang, "Temporal attention-pyramid pooling for temporal action detection," *IEEE Transactions on Multimedia (T-MM)*, vol. 25, pp. 3799–3810, 2022.

[8] K. Xia, L. Wang, Y. Shen, S. Zhou, G. Hua, and W. Tang, "Exploring action centers for temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, vol. 25, pp. 9425–9436, 2023.

[9] Y. Shao, F. Zhang, and C. Xu, "Snippet-to-prototype contrastive consensus network for weakly supervised temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, vol. 26, pp. 6717–6729, 2024.

[10] Y. Zhao, H. Zhang, Z. Gao, W. Gao, M. Wang, and S. Chen, "A novel action saliency and context-aware network for weakly-supervised temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, vol. 25, pp. 8253–8266, 2023.

[11] Y. Wang, S. Zhao, and S. Chen, "Action-semantic consistent knowledge for weakly-supervised action localization," *IEEE Transactions on Multimedia (T-MM)*, 2024.

[12] T. Zhang, R. Li, P. Feng, and R. Zhang, "Integration of global and local knowledge for foreground enhancing in weakly supervised temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, vol. 26, pp. 8476–8487, 2024.

[13] B. Li, R. Liu, T. Chen, and Y. Zhu, "Weakly supervised temporal action detection with temporal dependency learning," *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, vol. 32, no. 7, pp. 4473–4485, 2021.

[14] B. Wang, X. Zhang, and Y. Zhao, "Exploring sub-action granularity for weakly supervised temporal action localization," *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, vol. 32, no. 4, pp. 2186–2198, 2021.

[15] J. Ji, K. Cao, and J. C. Niebles, "Learning temporal action proposals with fewer labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7073–7082.

[16] X. Wang, S. Zhang, Z. Qing, Y. Shao, C. Gao, and N. Sang, "Self-supervised learning for semi-supervised temporal action proposal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1905–1914.

[17] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Semi-supervised temporal action detection with proposal-free masking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[18] K. Xia, L. Wang, S. Zhou, G. Hua, and W. Tang, "Learning from noisy pseudo labels for semi-supervised temporal action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 10 160–10 169.

[19] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1130–1139.

[20] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 156–10 165.

[21] Q. Wang, Y. Zhang, Y. Zheng, and P. Pan, "RCL: Recurrent continuous localization for temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 566–13 575.

[22] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[23] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian, "Bottom-up temporal action localization with mutual regularization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 539–555.

[24] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Learning salient boundary feature for anchor-free temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3320–3329.

[25] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, "Learning to refactor action and co-occurrence features for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 874–13 883.

[26] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Temporal action detection with global segmentation mask learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[27] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "Bmn: Boundary-matching network for temporal action proposal generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3889–3898.

[28] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, "Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 3, 2021, pp. 2602–2610.

[29] C. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing moments of actions with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 492–510.

[30] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing (T-IP)*, vol. 31, pp. 5427–5441, 2022.

[31] D. Shi, Y. Zhong, Q. Cao, J. Zhang, L. Ma, J. Li, and D. Tao, "React: Temporal action detection with relational queries," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 105–121.

[32] J. Kim, M. Lee, and J.-P. Heo, "Self-feedback detr for temporal action detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[33] Y. Wang, Y. Li, and H. Wang, "Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 878–18 887.

[34] Y. Wang and S. Zhao, "Weakly-supervised action localization by hierarchical attention mechanism with multi-scale fusion strategies," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[35] Y. Wang, S. Zhao, and S. Chen, "Sql-net: Semantic query learning for point-supervised temporal action localization," *IEEE Transactions on Multimedia (T-MM)*, 2024.

[36] T. Zhao, J. Han, L. Yang, and D. Zhang, "Equivalent classification mapping for weakly supervised temporal action localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 45, no. 3, pp. 3019–3031, 2022.

[37] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "Soda: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *International Journal of Computer Vision (IJCV)*, vol. 129, no. 8, pp. 2474–2498, 2021.

[38] X. Ding, N. Wang, X. Gao, J. Li, X. Wang, and T. Liu, "KFC: An efficient framework for semi-supervised temporal action localization," *IEEE T-IP*, vol. 30, pp. 6869–6878, 2021.

[39] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020, pp. 596–608.

[40] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1196–1205.

[41] B. Chen, W. Chen, S. Yang, Y. Xuan, J. Song, D. Xie, S. Pu, M. Song, and Y. Zhuang, "Label matching semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 381–14 390.

[42] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, "Instant-teaching: An end-to-end semi-supervised object detection framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4081–4090.

[43] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5051–5059.

[44] Y. Jin, J. Wang, and D. Lin, "Semi-supervised semantic segmentation via gentle teaching assistant," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2022, pp. 2803–2816.

[45] Y. Chen, X. Tan, B. Zhao, Z. Chen, R. Song, J. Liang, and X. Lu, "Boosting semi-supervised learning by exploiting all unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7548–7557.

[46] P. Qiao, Z. Wei, Y. Wang, Z. Wang, G. Song, F. Xu, X. Ji, C. Liu, and J. Chen, "Fuzzy positive learning for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 465–15 474.

[47] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017.

[48] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 68–83.

[49] Y. Kim, J. Yim, J. Yun, and J. Kim, "Nlnl: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 101–110.

[50] Y.-C. Chen and W.-T. Chu, "Positive and negative set designs in contrastive feature learning for temporal action segmentation," *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, vol. 34, no. 11, pp. 11 156–11 168, 2024.

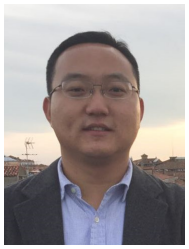
[51] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4325–4334.

[52] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 5561–5569.

[53] F. Zhou, B. Williams, and H. Rahmani, "Towards adaptive pseudo-label learning for semi-supervised temporal action localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2025, pp. 320–338.

[54] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," 2014.
[55] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970.
[56] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional module for temporal action localization in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 10, pp. 6209–6223, 2022.
[57] X. Liu, Y. Hu, S. Bai, F. Ding, X. Bai, and P. H. Torr, "Multi-shot temporal event localization: a benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 596–12 606.
[58] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.
[59] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang, "Cuhk & ethz & siat submission to activitynet challenge 2016," *arXiv preprint arXiv:1608.00797*, 2016.
[60] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.
[61] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*, 2020, pp. 1–8.

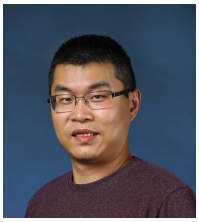
Kun Xia received the Ph.D. degree in Control Science and Engineering from Xi'an Jiaotong University, Xi'an, China, in 2024. He received the B.E. degree in automation from the Shenyang University of Technology, Shenyang, China, in 2017, and the M.E. degree in control science and engineering from Northeastern University, Shenyang, China, in 2020. From 2022 to 2023, he was a visiting Ph.D. student with University of Illinois Chicago, IL, USA. He is currently an Assistant Professor with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, image/video processing, and analysis and understanding.



Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in Control Science and Engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a visiting Ph.D. student with Stevens Institute of Technology, Hoboken, New Jersey, USA. From 2016 to 2017, he was a visiting scholar with Northwestern University, Evanston, Illinois, USA. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, and machine learning. He is the author of more than 80 peer reviewed publications in prestigious international journals and conferences. He is an area chair of ICCV'2025, ICLR'2025, WACV'2024&2025, ICPR'2022&2024, and CVPR'2022. He is an associate editor of PR, MVA, and PRL.

Sanping Zhou (Member, IEEE) received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2020. From 2018 to 2019, he was a visiting Ph.D. student at the Robotics Institute, Carnegie Mellon University. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include machine learning and computer vision, with a focus on object detection, image segmentation, visual tracking, multitask learning, and metalearning.

Gang Hua (Fellow, IEEE) received the B.S. and M.S. degrees in Automatic Control Engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 1999 and 2002, respectively. He received the Ph.D. degree in Electrical Engineering and Computer Science at Northwestern University, Evanston, Illinois, USA, in 2006. He is currently the Vice President of the Multimodal Experiences Research Lab at Dolby Laboratories. His research focuses on computer vision, pattern recognition, machine learning, robotics, towards general Artificial Intelligence, with primary applications in cloud and edge intelligence. Before that, he was the CTO of Convenience Bee, and the Managing Director and Chief Scientist of its research branch in US, Wormpex AI Research (2018-2024). He also served in various roles at Microsoft (2015-18) as the Science/Technical Adviser to the CVP of the Computer Vision Group, Director of Computer Vision Science Team in Redmond and Taipei ATL, and Senior Principal Researcher/Research Manager at Microsoft Research . He was an Associate Professor at Stevens Institute of Technology (2011-15). During 2014-15, he took an on leave and worked on the Amazon-Go project. He was a Visiting Researcher (2011-14) and a Research Staff Member (2010-11) at IBM Research T. J. Watson Center, a Senior Researcher (2009-10) at Nokia Research Center Hollywood, and a Senior Scientist (2006-09) at Microsoft Live labs Research. He is an associate editor of TPAMI and MVA. He is a general chair of ICCV'2027 and a program chair of CVPR'2019&2022. He is the author of more than 200 peer reviewed publications in prestigious international journals and conferences. He holds 35 US patents and has 15 more US patents pending. He is the recipient of the 2015 IAPR Young Biometrics Investigator Award. He is an IEEE Fellow, an IAPR Fellow, and an ACM Distinguished Scientist.



Wei Tang (Member, IEEE) received the Ph.D. degree in electrical engineering from Northwestern University, Evanston, IL, USA, in 2019. He received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2012 and 2015 respectively. He is currently an Assistant Professor with the Department of Computer Science, University of Illinois at Chicago. His research interests include computer vision, pattern recognition, and machine learning.