IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 June 2023, accepted 5 July 2023, date of publication 12 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294572

RESEARCH ARTICLE

An Improved Point-Level Supervision Method for Temporal Action Localization

SANGJIN LEE[1], JAEBIN LIM[1], JINYOUNG MOON[2], AND CHANHO JUNG[1]
[1]Department of Electrical Engineering, Hanbat National University, Daejeon 34158, South Korea
[2]Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea

Corresponding author: Chanho Jung (peterjung@hanbat.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT), Development of Previsional Intelligence Based on Long-Term Visual Memory Network, under Grant 2020-0-00004.

ABSTRACT Recently, with the expansion of the video platform market, research has been actively conducted on temporal action localization (TAL) for detecting actions in atypical videos. Most learning methods for TAL include full and weak supervision (weak supervision with only action classes) approaches. Full supervision requires considerable time for labeling and weak supervision exhibits low localization performance owing to the lack of informative annotations. To solve this problem, point-level weak supervision using single-point timestamps within the temporal interval of action instances has been proposed, which demonstrates superior performance to weakly-supervised methods using only action classes of action instances. In this study, we proposed an improved point-level supervision mechanism that provides point-level annotations for each action and background instance. In addition, a widely used multiple instance learning (MIL)-based framework was used to verify the proposed method, and pseudo-labels were used for action instance boundary learning. Also, the background point loss was designed to leverage the added point-level annotations. The datasets used in the experiment were THUMOS14, GTEA, BEOID, and ActivityNet1.2, and improved results were obtained compared to existing point-level supervision. The code is available from https://github.com/sang9390/An-Improved-Point-Level-Supervision-Method-for-TAL.

INDEX TERMS Temporal action localization, multiple instance learning, fully-supervised learning, weakly-supervised learning.

I. INTRODUCTION

Recently, research on the weakly-supervised methods in temporal action localization (TAL) has been actively conducted. Weakly-supervised methods aim to detect and classify action instances in an untrimmed video, given video-level annotations with action classes of action instances within the video. Fully-supervised methods require complete annotations, including action instance boundaries and action classes. Therefore, a full supervision approach necessitates repeatedly watching videos to accurately label action boundaries, indicating a time-consuming labeling operation [1]. However, for weakly-supervised methods, very little time is required for labeling owing to the difference in necessary informative annotations [2].

Weakly-supervised methods assume that action segments in a video have a significant influence on video-level classification. However, these methods lack temporal contextual information, indicating that background segments are likely to significantly affect video-level classification [3]. It is evident that certain components do not comply with the aforementioned assumption. Therefore, weakly-supervised methods show lower localization performance than fully-supervised methods [4], [5], [6], [7]. To tackle this problem, SF-Net [2] and BackTAL [3] introduced the action instance-based point-level supervision and background instance-based point-level supervision methods, respectively. SF-Net [2] confirmed that it takes 45, 50, and 300 seconds to generate video-level, point-level, and

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

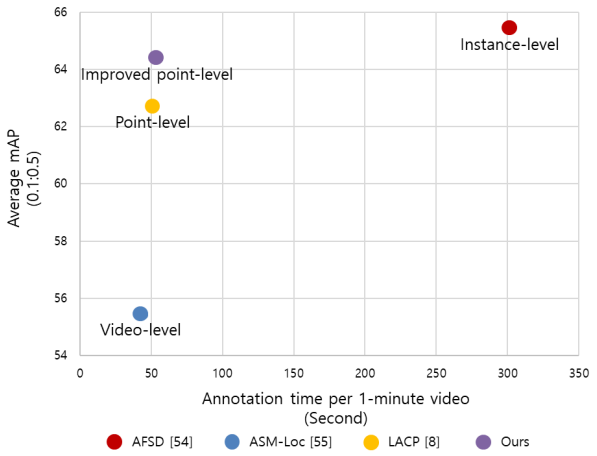

FIGURE 1. Performance and annotation time per 1-minute video comparison of various supervision methods on THUMOS14.

instance-level annotations, respectively, based on a 1-minute video. BackTAL [3] confirmed that most errors that occur in action localization results based on weakly-supervised methods occur in the background. To annotate video-level supervision, one only needs to watch the entire video once. Similarly, to add an annotation to point-level supervision, it is necessary to watch the entire video once and record a point when an action or background is recognized. The annotation process can be completed with an annotation tool, thus requiring a similar amount of time or point-level supervision as for video-level supervision. However, frame-level supervision annotation must accurately record the start and end frames of an action instance; therefore, the video must be watched while rolling back, significantly increasing the annotation cost. Each point-level supervision method demonstrated high localization performance compared to video-level supervision methods and demonstrated significant potential for development.

However, point-level supervision methods still exhibit a large performance gap compared to fully-supervised methods at high intersection-over-union (IoU) thresholds. In this paper, we argue that this is because temporal context information remains insufficient for point-level weakly-supervised TAL. Based on this claim, we propose an improved point-level supervision mechanism.

In our method, we propose leveraging both action and background points. Improved point-level annotations are extracted from the annotations of the fully-supervised methods. The action instance is extracted using a Gaussian distribution and the center point is extracted for the background instance. To assess the proposed method, the MIL-based framework was used as a baseline. Among the MIL-based frameworks, LACP [8], which demonstrated high performance in point-level supervision, was used. The framework learns point-level annotations to extract the action scores, which are then used to generate pseudo-labels (i.e., sequences) and learn the action instance boundaries. Also, background instance loss is proposed to utilize the added background instance-based point-level annotations. We can more clearly distinguish the temporal context by learning improved point-level annotations with very little annotation time added compared to existing methods and show significant performance improvement (see Figure 1).

In this paper, the experiments were conducted using four datasets and demonstrated that the proposed method generally outperformed existing point-level supervision methods. The datasets used were THUMOS14 [9], BEOID [10], GTEA [11], and ActivityNet 1.2 [12]. THUMOS14 comprises 20 action classes, BEOID 30 action classes, GTEA 7 action classes, ActivityNet 1.2 100 action classes.

The contributions of this study can be summarized as follows:

1) We propose a point-level supervision based on action and background instances for the temporal action localization task.
2) A background point loss is proposed to effectively utilize background point annotations.
3) We have performed extensive experiments on the three benchmarks and demonstrated that the proposed method outperforms LACP [8], the state-of-the-art (SOTA) method in point-level supervision.

II. RELATED WORK

Section II introduces the latest research trends in TAL [5], [13], [14]. First, we examine the fully-supervised method and the widely used one-stage [15], [16], [17] and two-stage [18], [19], [20], [21], [22] frameworks. Second, weakly-supervised methods are investigated. Finally, we explore the point-level supervision method, which is the basis of the proposed method.

A. FULLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

learns based on accurate annotation for each frame. Existing methods are largely divided into one-stage and two-stage frameworks and primarily use a two-stage framework. The one-stage framework simultaneously infers action instance boundaries and classes, while the two-stage framework creates action instance proposals and classifies them separately. There are several methods of creating proposals. One method uses the sliding-window technique [19], [23], [24], [25], [26], [27], [28]. Another method extracts the probability that each frame is the start or end point of an action instance and then uses a combination of probable action instances as a proposal or proposes an anchor mechanism [22], [29], [30], [31], [32]. The latest fully-supervised methods, such as TALLFormer [33] and RCL [34], exhibit outstanding localization performance.

B. WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

employs video-level labels as annotation to reduce the laborious labeling process of fully-supervised methods. Based on the MLP framework, a method is devised to select and learn segments that influence video-level classification, similar to approaches such as UntrimmedNet [35] and STPN [36]. This

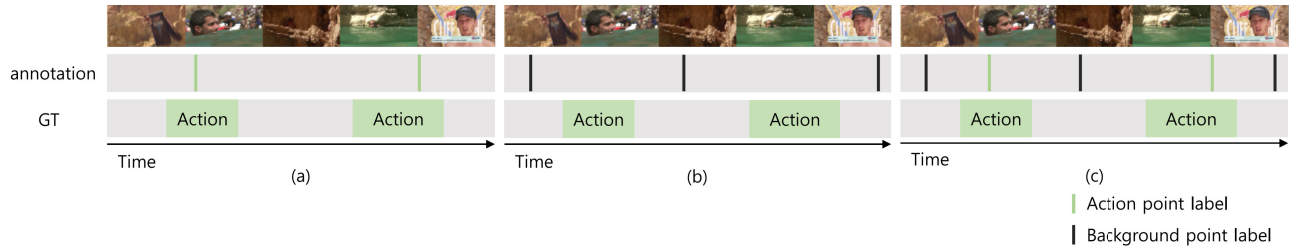

FIGURE 2. Point-level supervision types. (a) Action instance based [2]. (b) Background instance based [3]. (c) Action and background instance based (ours).

method also aims to localize action instances by utilizing specific action score thresholds. Subsequently, various methods for leveraging weak supervision annotations have been studied. One method focuses on background instance modeling [3], and the EM-MIL approach [37] leverages pseudo-label extraction to learn the boundaries of an action instance. Despite the various approaches described above, weakly-supervised TAL demonstrates extremely poor performance compared to the fully-supervised method, owing to the lack of temporal context information. To solve this fundamental problem, several methods have used external information such as action count [7], [38] or audio [39].

C. SINGLE-FRAME SUPERVISED TEMPORAL ACTION LOCALIZATION

is a type of weakly-supervised method where the time required for the labeling operation is similar to that of the weakly-supervised method, and the localization performance is similar to that of the full instruction method. A representative example is SF-Net [2]; using point-level annotations and extracted pseudo-labels, a higher localization performance was obtained compared to that of the weakly-supervised method. BackTAL [3] argued that the localization performance of SF-Net [2] was inferior to that of the fully-supervised method because of background errors; a background instance point annotation was proposed to improve this. In addition, it exhibits better localization performance than SF-Net [2]. LACP [8] argued that existing point-level instruction methods do not consider the completeness of the action instance. To provide completeness, we contrast the action instance with the background instance. This approach exhibits a significant performance improvement at a low IoU threshold. However, it still exhibits lower localization performance than the fully-supervised method [33], [34].

III. METHOD

Section III describes our improved point-level supervision method in detail. After that, the TAL framework used in this study and the proposed background point loss are examined.

Problem Setting: In this study, the task of instance-click supervision for TAL was set based on the research concerning SF-Net [2] and BackTAL [3]. Given an input video, point and class labels were provided for each action and background instance. The corresponding label is expressed as follows: $B^{ins} = \left(t_i, y_{t_i}\right)_{i=1}^{M^{ins}}$, where the i-th instance is labeled with the Y_{t_i}-class in the t_i-th frame. Y_{t_i} represents the category of $C+1$ including the background in the total number of classes and the one-hot labeling method is used. M^{ins} is the total number of instances including the number of action and background instances in the video, aiming to distinguish each action instance boundary and class in the training and test videos using the above annotation. The differences with existing point-level annotations are shown in Figure 2.

A. BASELINE FRAMEWORK

The framework used in this study is based on LACP [8], as shown in Figure 3. Subsequently, we briefly describe the pipeline.

1) EXTRACTED FEATURES

We use RGB and optical flow images as the inputs. The feature extractor uses I3D [38], which is widely used in TAL tasks and has achieved satisfactory performance. The following steps briefly describe the extraction of these features. First, the input images are divided into 16 frames to create snippets. The snippet is then fed into an I3D feature extractor [38] to generate the extracted features X. A feature has a size of $D \times T$, where D is the number of channels and T is the total number of snippets in the video.

2) EMBEDDED FEATURES

The TAL task requires the localization of actions along the temporal domain. Therefore, the model consists of 1D convolutions with strengths in temporal feature modeling. Embedded features F are generated by inputting the extracted features X into a 1D convolutional layer. The extracted features X are then embedded to optimize the TAL task. The size, $D \times T$, is equal to the number of extracted features X.

3) ACTION SCORES

The embedded features F are input into the 1D convolutional layer. Subsequently, the generated feature is input into the sigmoid function to obtain the action scores P, which have a magnitude of $C \times T$. P represents the probability that the snippet belongs to the class c.

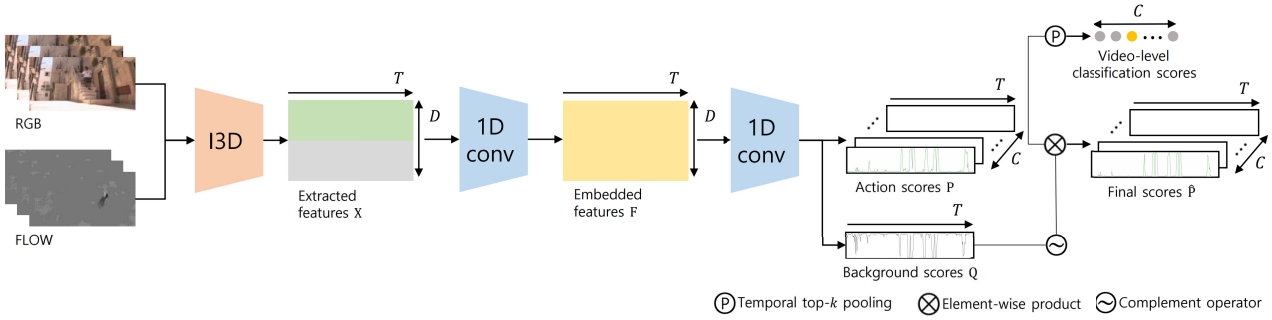

FIGURE 3. Baseline framework.

4) BACKGROUND SCORES

The embedded features F are input into the 1D convolutional layer. Afterward, the generated feature is input into the sigmoid function to obtain the background scores Q. The size is $1 \times T$ and Q_t is the probability that the snippet is in the background.

5) FINAL SCORES

In order to have completeness, the final scores $\hat{P}$ are extracted using the element-wise product of P and $1 - Q$. This is the same size as P.

6) VIDEO-LEVEL CLASSIFICATION SCORES

This is a video-level classification score obtained by performing temporal top-k pooling on the action scores P.

7) PSEUDO GT

As the pseudo GT outputs through the optimal sequence search of LACP [8], it is possible to obtain an optimal sequence (i.e., the pseudo-ground truth) that is most likely to be similar to the action instance.

B. BACKGROUND POINT LOSS

In this section, we briefly review the loss in the existing method [8] and describe the proposed background point loss.

The losses used in the existing method are as follows:

$$L_{total} = \lambda_1 L_{video} + \lambda_2 L_{point} + \lambda_3 L_{score} + \lambda_4 L_{feat}, \quad (1)$$

where λ_* denotes the hyperparameter. Assessing each loss, L_{video} is the loss value measured by comparing the video-level annotation and classification scores. L_{score} is a loss value measured by comparing the pseudo-labels. L_{feat} is a loss value measured by comparing the pseudo-labels. L_{point} can be divided into the following two types:

$$L_{point} = L^{act}_{point} + L^{bkg}_{point} \quad (2)$$

The background point loss is calculated using binary cross-entropy, and the focal loss [22] is employed. The expression is defined as follows:

$$L^{bkg}_{point} = -\frac{1}{M^{bkg}} \sum_{\forall t \in B^{bkg}} \Big(\sum_{c=1}^{C} \hat{p}_t[c]^{\beta} \log\big(1 - \hat{p}_t[c]\big) + (1 - q_t)^{\beta} \log q_t \Big), \quad (3)$$

where $B^{bkg} = \{t_j\}_{j=1}^{M^{bkg}}$ denotes the set of pseudo-background points extracted based on the score of q_t. M^{bkg} denotes the total number of background points extracted. β uses 2, which is the same number as the focal loss [22], as the focusing parameter.

However, when background point loss is employed, point annotation of each added background instance cannot be used. In addition, there is a possibility of adversely affecting localization performance due to learning through a set of pseudo-background points. Therefore, in this paper, we propose a GT background point loss method based on background point annotation. The GT background point loss is defined as follows:

$$L^{bkg\_gt}_{point} = -\frac{1}{M^{bkg\_gt}} \sum_{\forall t \in B^{bkg\_gt}} \Big(\sum_{c=1}^{C} \hat{p}_t[c]^{\beta} \log\big(1 - \hat{p}_t[c]\big) + (1 - q_t)^{\beta} \log q_t \Big), \quad (4)$$

where $B^{bkg\_gt} = \{t_j\}_{j=1}^{M^{bkg\_gt}}$ denotes the GT background point set. M^{bkg_gt} represents the total number of GT background points. β is the same as that in equation (3). The total point-level loss is defined as the sum of equations (2) and (3), as follows:

$$\tilde{L}_{point} = L^{act}_{point} + L^{bkg}_{point} + L^{bkg\_gt}_{point} \quad (5)$$

In summary, the learning in this study is based on the following equation:

$$L_{total} = \lambda_1 L_{video} + \lambda_2 \tilde{L}_{point} + \lambda_3 L_{score} + \lambda_4 L_{feat} \quad (6)$$

The reasoning process is configured in the same way as in LACP [8]. First, the class to be detected is selected based on the video-level classification score threshold Θ^{vid}. Subsequently, the proposal is extracted based on the segment-level threshold Θ^{seg} at $\hat{P}$. We also use non-maximum suppression (NMS) to remove overlapping proposals.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

In this paper, THUMOS14 [9], BEOID [10], GTEA [11], and ActivityNet 1.2 [12] datasets were used to verify the performance of the proposed method. THUMOS14 [9] consists of validation and test sets of 200 and 213 unstructured

TABLE 1. State-of-the-art comparison on THUMOS14.

Supervision	Method	mAP@IoU(%) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG (0.1:0.5)	AVG (0.3:0.7)
Frame-level	BMN [22]	-	-	56.0	47.4	38.8	29.7	20.5	-	38.5
	G-TAD [43]	-	-	54.5	47.6	40.2	30.8	23.4	-	39.3
	BC-GNN [44]	-	-	57.1	49.1	40.4	31.2	23.1	-	40.2
	Zhao et al. [31]	-	-	53.9	50.7	45.4	**38.0**	**28.5**	-	**43.3**
	P-GCN [20]	**69.5**	**67.8**	**63.6**	**57.8**	**49.1**	-	-	**61.6**	-
Video-level	Lee et al. [6]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	32.9
	CoLA [45]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1
	AUMN [46]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4
	TS-PCA [47]	67.6	61.1	53.4	43.4	34.3	**24.7**	**13.7**	52.0	33.9
	UGCT [48]	**69.2**	**62.9**	**55.5**	**46.5**	**35.9**	23.8	11.4	**54.0**	**34.6**
Point-level (background)	BackTAL [3]	-	-	54.4	45.5	36.3	26.2	14.8	-	35.4
Point-level (action)	Moltisanti et al. [49]	24.3	19.9	15.9	12.5	9.0	-	-	16.3	-
	SF-Net [2]	68.3	62.3	52.8	42.2	30.5	20.6	12.0	51.2	31.6
	Ju et al. [41]	72.3	64.7	58.2	47.1	35.9	23.0	12.8	55.6	35.4
	PCL [50]	74.6	70.2	63.3	56.5	45.3	-	-	61.7	-
	LACP [8]	75.7	71.4	64.6	56.5	45.3	34.5	**21.8**	62.7	44.5
Point-level (action & background)	Ours	**77.0**	**73.3**	**66.8**	**57.8**	**47.1**	**34.8**	21.1	**64.4**	**45.5**

TABLE 2. State-of-the-art comparison on GTEA and BEOID.

Dataset	Method	mAP@IoU(%) 0.1	0.3	0.5	0.7	AVG (0.1:0.7)
GTEA	SF-Net [2]	58.0	37.9	19.3	11.9	31.0
	Ju et al. [41]	59.7	38.3	21.9	18.1	33.7
	Li et al. [42]	60.2	44.7	28.8	12.2	36.4
	LACP [8]	63.9	55.7	33.9	20.8	43.5
	PCL [50]	65.2	**56.8**	34.3	**21.2**	44.9
	Ours	**68.6**	55.4	**37.4**	**21.2**	**45.6**
BEOID	SF-Net [2]	62.9	40.6	16.7	3.5	30.9
	Ju et al. [41]	63.2	46.8	20.9	5.8	34.9
	Li et al. [42]	71.5	40.3	20.3	5.5	34.4
	LACP [8]	76.9	61.4	42.7	25.1	51.8
	PCL [50]	**78.7**	63.3	44.1	26.9	53.3
	Ours	73.4	**64.4**	**48.5**	**27.2**	**53.7**

TABLE 3. State-of-the-art comparison on ActivityNet 1.2.

Supervision	Method	mAP@IoU(%) 0.5	0.75	0.95	AVG (0.5:0.95)
Video-level	ASL [51]	40.2	-	-	25.8
	Lee et al. [4]	41.2	25.6	6.0	25.9
	ACSNet [52]	40.1	**26.1**	**6.8**	26.0
	D2-Net [53]	42.3	25.5	5.8	26.0
	CoLA [45]	**42.7**	25.7	5.8	**26.1**
Point-level (background)	BackTAL [3]	41.5	**27.3**	4.7	27.0
Point-level (action)	SF-Net [2]	37.8	-	-	22.8
	LACP [8]	44.0	26.9	5.9	26.8
Point-level (action & background)	Ours	**44.6**	26.7	**6.1**	**27.2**

TABLE 4. Comparison of point-level labels from different point distributions on THUMOS14.

Method	Distribution	mAP@IoU(%) 0.3	0.5	0.7	AVG (0.1:0.7)
SF-Net [2]	Uniform	52.0	**30.2**	**11.8**	40.5
	Gaussian	47.4	26.2	9.1	36.7
	Center (reproduced)	**52.8**	**30.2**	11.3	**40.9**
LACP [8]	Uniform	60.4	42.6	20.2	49.3
	Gaussian	**64.6**	**45.3**	**21.8**	**52.8**
	Center (reproduced)	64.2	45.2	20.6	52.6
Ours	Uniform	66.9	46.0	19.4	53.6
	Gaussian	66.8	**47.1**	**21.1**	54.0
	Center	**67.6**	46.7	20.4	**54.2**

TABLE 5. Comparison of point-level labels from different point distributions on GTEA.

Method	Distribution Action/Background	mAP@IoU(%) 0.3	0.5	0.7	AVG (0.1:0.7)
LACP [8] (reproduced)	Uniform/-	54.0	29.1	16.2	40.4
	Gaussian/-	**55.7**	**37.7**	**18.7**	**44.1**
	Center/-	48.9	28.2	15.1	40.1
Ours	Uniform/Uniform	**55.8**	34.0	18.1	43.5
	Uniform/Gaussian	54.0	36.2	20.4	43.9
	Uniform/Center	54.3	**38.4**	**20.7**	**44.8**
Ours	Gaussian/Uniform	53.8	27.3	15.2	40.3
	Gaussian/Gaussian	48.2	30.2	17.8	40.1
	Gaussian/Center	**55.4**	**37.4**	**21.2**	**45.6**
Ours	Center/Uniform	54.1	**37.6**	**19.7**	44.1
	Center/Gaussian	53.3	36.5	19.6	**44.3**
	Center/Center	**55.1**	34.7	18.5	44.2

videos and has a total of 20 classes. In addition, the validation set is typically used for training and the test set is used for testing [4], [35], [36]. In this study, 210 videos were used for the test, excluding three videos from the test set based on the LACP guidelines [8]. BEOID [10] comprises 30 action classes and 58 videos. GTEA [11] comprises 28 videos and seven action classes; we also used 21 and seven videos for training and testing, respectively. GTEA [11] and BEOID [10] follow the data segmentation method of [2]. ActivityNet 1.2 [12] has 100 action classes and consists of 4,819 training, 2,383 validation, and 2,480 test videos.

2) EVALUATION METRICS

We employed evaluation indicators that are mainly used for TAL. After obtaining the average precision (AP) score according to the IoU threshold, the mean average precision (mAP) was calculated.

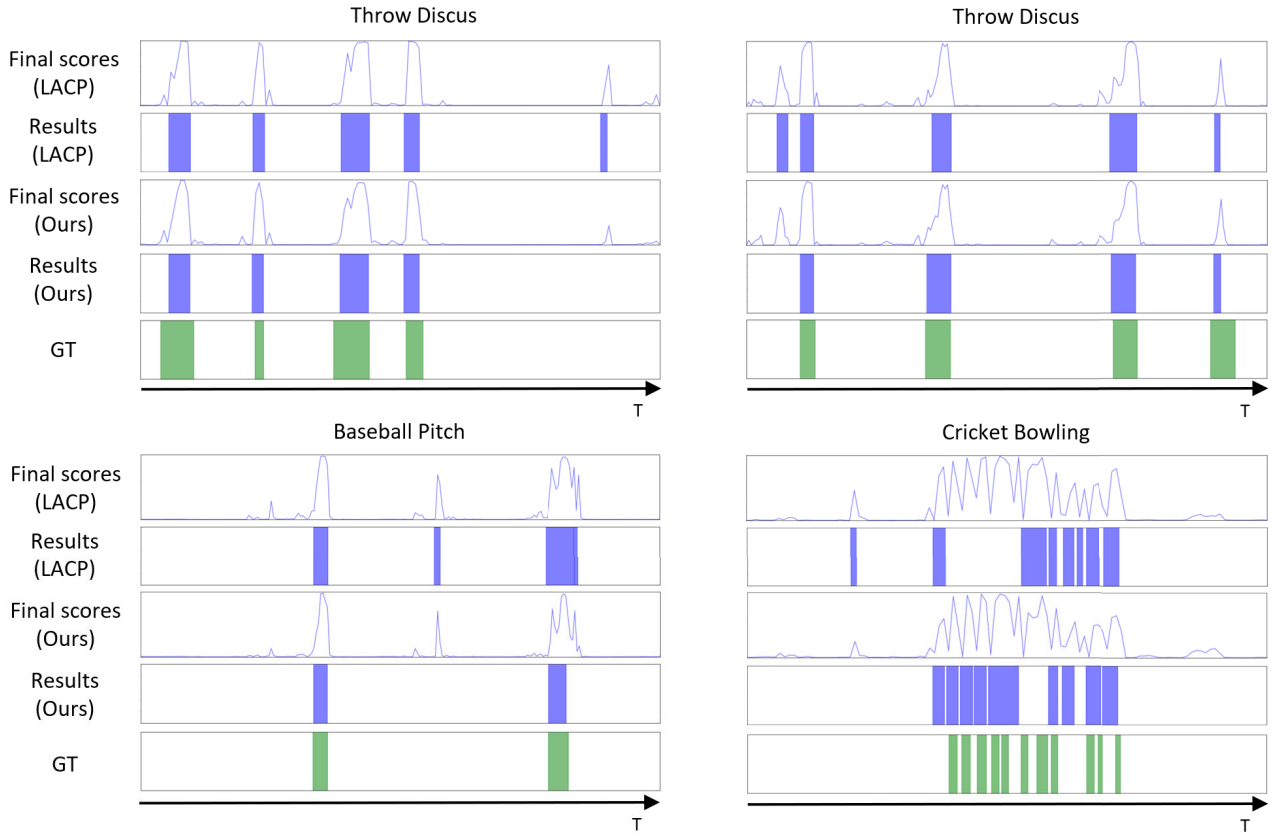

FIGURE 4. Qualitative result (THUMOS14).

TABLE 6. Comparison of point-level labels from different point distributions on BEOID.

Method	Distribution Action/Background	mAP@IoU(%) 0.3	0.5	0.7	AVG (0.1:0.7)
LACP [8] (reproduced)	Uniform/-	60.1	40.0	13.3	48.3
	Gaussian/-	55.4	38.1	14.7	46.4
	Center/-	**64.8**	**47.7**	**20.8**	**53.8**
Ours	Uniform/Uniform	**64.3**	45.5	**21.3**	**52.4**
	Uniform/Gaussian	64.2	**45.8**	19.5	51.9
	Uniform/Center	62.9	44.4	17.4	51.6
Ours	Gaussian/Uniform	**64.6**	44.7	16.9	51.3
	Gaussian/Gaussian	63.7	41.9	20.9	52.2
	Gaussian/Center	64.4	**48.5**	**27.2**	**53.7**
Ours	Center/Uniform	**66.6**	**51.0**	24.8	**56.4**
	Center/Gaussian	59.7	40.2	20.3	50.2
	Center/Center	66.2	48.7	**28.9**	56.2

TABLE 7. Ablation study on THUMOS14.

L^{bkg}_{point}	$L^{bkg}_{point\_gt}$	mAP@IoU (%) 0.1	0.3	0.5	0.7	AVG
✓		76.1	65.1	45.2	20.5	52.7
	✓	76.0	65.3	45.1	20.4	52.7
✓	✓	**77.0**	**66.8**	**47.1**	**21.1**	**54.0**

In this section, the existing SOTA method and the proposed method are compared.

Table 1 shows the performances of the proposed method based on different supervised methods on the THUMOS14 [9] dataset. The performance is shown to be higher when the IoU threshold is 0.6 or less than LACP [8], which has been used as a baseline framework. In addition, there is a slight performance gap compared to the fully-supervised method. Table 2 compares the performance of the proposed point-level supervision method and the state-of-the-art (SOTA) method for the GTEA [11] and BEOID [10] datasets. Our method demonstrates superior performance compared to existing methods overall, except for specific IoU thresholds, for both datasets. Table 3 presents the performance comparison of our proposed method with previous SOTA approaches on the ActivityNet 1.2 dataset [12]. The comparison results demonstrate that our proposed method achieves higher performance compared to the existing SOTA methods, it can be seen that the proposed method is generally helpful in improving performance.

TABLE 8. Comparison of performance by convolutional layer on THUMOS14.

Conv. Type	Method	mAP@IoU(%) 0.3	0.5	0.7	AVG (0.1:0.7)
1D	LACP [8]	64.6	45.3	21.8	52.8
	Ours	**66.8**	**47.1**	**21.1**	**54.0**
2D	LACP [8]	42.9	29.0	14.8	36.1
	Ours	**47.8**	**33.1**	**15.9**	**39.3**

B. COMPARISON WITH STATE-OF-THE-ART METHODS

C. QUANTITATIVE COMPARISON

Tables 4, 5, and 6 compare the performance changes in various datasets according to the label distribution (THUMOS14

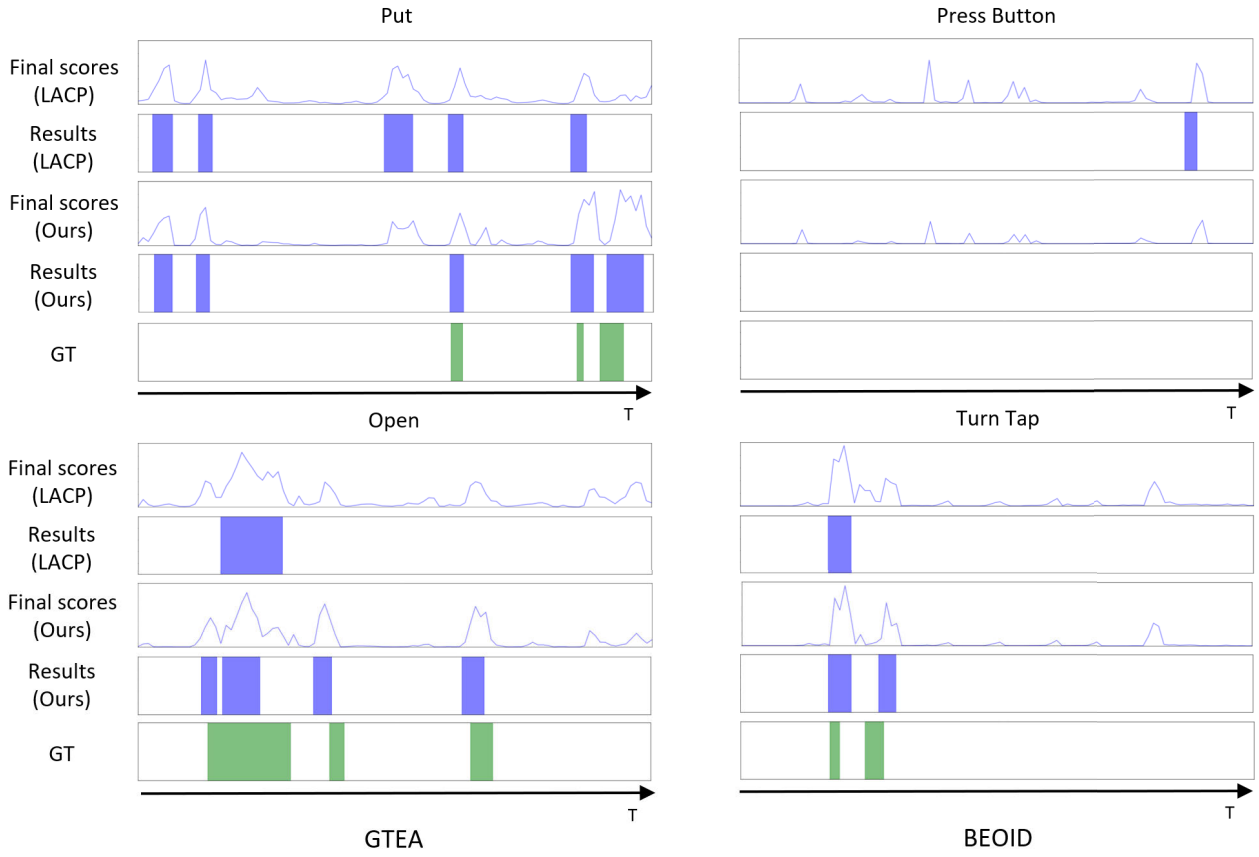

FIGURE 5. Qualitative result (GTEA and BEOID).

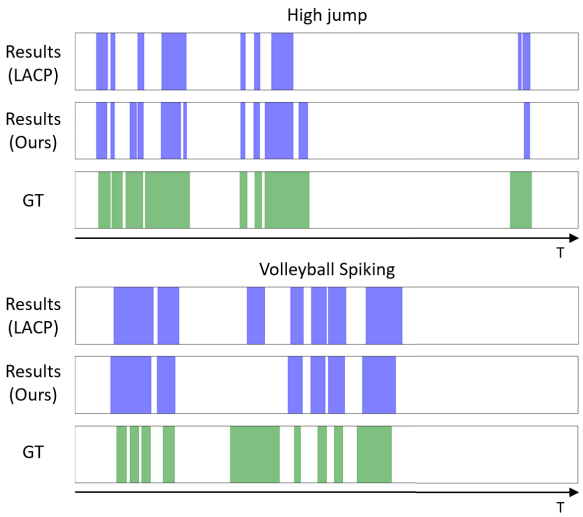

FIGURE 6. The failure cases of our proposed method (THUMOS14).

[9], BEOID [10], and GTEA [11]). "Center" represents the label extracted from the midpoint of the action or background instance; all other labels represent simulated labels from that distribution. Tables 4, 5, and 6 demonstrate that the proposed method outperforms LACP [8] for most distributions. In particular, a significant performance improvement is shown at an IoU threshold of 0.3, indicating that the proposed method improves the ability to distinguish the temporal context. Table 7 compares the performance changes in THUMOS14 according to the changes in the background point loss. When L_{point}^{bkg} or $L_{point\_gt}^{bkg}$ is used, the performance is similar to each other; when the two are used concurrently, a performance improvement of 1.3 (average mAP) is achieved. Table 8 presents the experimental results conducted to demonstrate the importance of the existing 1D convolution layer. The experiment was carried out by replacing the 1D convolution layer with a 2D convolution layer in the baseline framework. As a result of the experiment, it can be confirmed that the performance decreases when using a 2D convolutional layer instead of a 1D convolutional layer.

D. QUALITATIVE COMPARISON

This section presents qualitative comparisons of the proposed and LACP [8] methods performed in this study. Through Figures 4 and 5, it was confirmed that the proposed method had fewer false positives and false negatives. In the case of LACP [8], it is often found that the final score is high in the background segment or the final score is low in the action segment. On the other hand, the proposed method produces more accurate detection results by contrasting the difference

between the action frame and the background frame. Through this, it was confirmed that our point-level supervision helps to learn the temporal context well.

E. FAILURE CASES, LIMITATIONS, AND FUTURE WORK

Figure 6 shows failure cases of our proposed approach. From the figure, we can see that our scheme is still vulnerable to false positives and over-completeness errors. Although these limitations may be evident for certain videos. On the other hand, as shown in Figures 4 and 5, the proposed approach demonstrates superior qualitative results on various datasets (i.e., TUMOS14 [9], BEOID [10], and GTEA [11] compared to existing state-of-the-art methods. The reason for the aforementioned inaccurate detection results is that the proposed approach does not accurately learn the boundaries of the working instances. This suggests a lack of precise boundary learning, which may be attributed to the insufficient completeness guidance provided by the GT background points. To address these limitations in the future, our goal is to conduct further research to utilize GT background points for more effective completeness guidance.

V. CONCLUSION

In this paper, we proposed an instance point supervision method that provides action and background instance point annotations. In addition, a new background point loss was designed to leverage the added annotation. The proposed method demonstrated the best performance among point-level supervisory methods on four benchmark datasets. In addition, it exhibited a localization performance similar to that of the fully-supervised method. Therefore, the improved point-level supervision mechanism demonstrated its potential for development.

ACKNOWLEDGMENT

(Sangjin Lee and Jaebin Lim are co-first authors.)

REFERENCES

[1] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8667–8677.

[2] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou, "SF-Net: Single-frame supervision for temporal action localization," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2020, pp. 420–437.

[3] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-click supervision for temporal action localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9814–9829, Dec. 2022.

[4] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11320–11327.

[5] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1854–1862.

[6] P. X. Nguyen, D. Ramanan, and C. Fowlkes, "Weakly-supervised action localization with background modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5501–5510.

[7] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9070–9078.

[8] P. Lee and H. Byun, "Learning action completeness from points for weakly-supervised temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13628–13637.

[9] Y. G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. (2014). *THUMOS Challenge: Action Recognition With a Large Number of Classes*. [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[10] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas, "Discovering task relevant objects and their modes of interaction from multi-user egocentric video," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 3.

[11] P. Lei and S. Todorovic, "Temporal deformable residual networks for action segmentation in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6742–6751.

[12] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. 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.

[13] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2782–2795, Nov. 2013.

[14] D. Zhang, J. Han, G. Cheng, and M. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[15] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 988–996.

[16] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 344–353.

[17] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-TAD: Sub-graph localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10156–10165.

[18] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5794–5803.

[19] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1130–1139.

[20] R. Zeng, W. Huang, C. Gan, M. Tan, Y. Rong, P. Zhao, and J. Huang, "Graph convolutional networks for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7093–7102.

[21] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3648–3656.

[22] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3888–3897.

[23] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S. Chang, "CDC: Convolutional-De-Convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1417–1426.

[24] Z. Shou, D. Wang, and S. 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.

[25] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," 2017, *arXiv:1703.02716*.

[26] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7477–7484.

[27] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3093–3102.

[28] 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.

[29] L. Chuming, L. Jian, W. Yabiao, T. Ying, L. Donghao, C. Zhipeng, W. Chengjie, L. Jilin, H. Feiyue, and J. Rongrong, "Fast learning of temporal action proposal via dense boundary generator," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11499–11506.

[30] S. Wang, Z. Miao, W. Xu, C. Ma, and M. Li, "Boundary sensitive and category sensitive network for temporal action proposal generation," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 3–19.

[31] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian, "Bottom-up temporal action localization with mutual regularization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 539–555.
[32] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3599–3608.
[33] F. Cheng and G. Bertasius, "TALLFormer: Temporal action localization with a long-memory transformer," 2022, *arXiv:2204.01680*.
[34] Q. Wang, Y. Zhang, Y. Zheng, and P. Pan, "RCL: Recurrent continuous localization for temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13556–13565.
[35] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6402–6411.
[36] 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.
[37] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 5502–5511.
[38] 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.
[39] J. T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–17.
[40] K. E. Ko and K. B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 226–234, Jan. 2018.
[41] C. Ju, P. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Point-level temporal action localization: Bridging fully-supervised proposals to weakly-supervised losses," 2020, *arXiv:2012.08236*.
[42] Z. Li, Y. A. Farha, and J. Gall, "Temporal action segmentation from timestamp supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8361–8370.
[43] Y. Bai, Y. Wang, Y. Tong, Y. Yang, Q. Liu, and J. Liu, "Boundary content graph neural network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.
[44] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "CoLA: Weakly-supervised temporal action localization with snippet contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16005–16014.
[45] W. Luo, T. Zhang, W. Yang, J. Liu, T. Mei, F. Wu, and Y. Zhang, "Action unit memory network for weakly supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9964–9974.
[46] Y. Liu, J. Chen, Z. Chen, B. Deng, J. Huang, and H. Zhang, "The blessings of unlabeled background in untrimmed videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6172–6181.
[47] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 53–63.
[48] D. Moltisanti, S. Fidler, and D. Damen, "Action recognition from single timestamp supervision in untrimmed videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9907–9916.
[49] P. Li, J. Cao, and X. Ye, "Prototype contrastive learning for point-supervised temporal action detection," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118965.
[50] 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.
[51] Z. Liu, L. Wang, Q. Zhang, W. Tang, J. Yuan, N. Zheng, and G. Hua, "ACSNet: Action-context separation network for weakly supervised temporal action localization," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2233–2241.
[52] S. Narayan, H. Cholakkal, M. Hayat, F. S. Khan, M. Yang, and L. Shao, "D2-Net: Weakly-supervised action localization via discriminative embeddings and denoised activations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13588–13597.
[53] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3319–3328.
[54] B. He, X. Yang, L. Kang, Z. Cheng, X. Zhou, and A. Shrivastava, "ASM-Loc: Action-aware segment modeling for weakly-supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13915–13925.

SANGJIN LEE is currently pursuing the master's degree with the Department of Electrical Engineering, Hanbat National University, Daejeon, Republic of Korea. His research interests include computer vision, machine learning, pattern recognition, and image processing.

JAEBIN LIM is currently pursuing the master's degree with the Department of Electrical Engineering, Hanbat National University, Daejeon, Republic of Korea. His research interests include computer vision, machine learning, pattern recognition, and image processing.

JINYOUNG MOON received the B.S. degree in computer engineering from Kyungpook National University, Daegu, Republic of Korea, in 2000, and the M.S. degree in computer science and the Ph.D. degree in industrial and systems engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2002 and 2018, respectively. Since 2002, she has been with the Visual Intelligence Research Section, Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. Since 2019, she has also been with the ICT Department, University of Science and Technology (UST), where she is currently an Assistant Professor. Her research interests include action recognition, action detection, temporal moment localization, and video quality assurance.

CHANHO JUNG received the B.S. and M.S. degrees in electronic engineering from Sogang University, Seoul, Republic of Korea, in 2004 and 2006, respectively, and the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2013. From 2006 to 2008, he was a Research Engineer with the Digital Television Research Laboratory, LG Electronics, Seoul. From 2013 to 2016, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon. Since 2016, he has been with the Department of Electrical Engineering, Hanbat National University, Daejeon, where he is currently an Associate Professor. His research interests include computer vision, machine learning, embedded systems, pattern recognition, and image processing.