UNLEASHING THE POTENTIAL OF ADJACENT SNIPPETS FOR WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION

Qinying Liu, Zilei Wang, Ruoxi Chen, and Zhilin Li

{lydyc,ruoxi16,lizhilin}@mail.ustc.edu.cn; zlwang@ustc.edu.cn

ABSTRACT

Weakly-supervised temporal action localization (WTAL) intends to detect action instances with only weak supervision, *e.g.*, video-level labels. The current *de facto* pipeline locates action instances by thresholding and grouping continuous high-score regions on temporal class activation sequences. In this route, the capacity of the model to recognize the relationships between adjacent snippets is of vital importance which determines the quality of the action boundaries. However, it is error-prone since the variations between adjacent snippets are typically subtle, and unfortunately this is overlooked in the literature. To tackle the issue, we propose a novel WTAL approach named Convex Combination Consistency between Neighbors (C³BN). C³BN consists of two key ingredients: a micro data augmentation strategy that increases the diversity in-between adjacent snippets by convex combination of adjacent snippets, and a macro-micro consistency regularization that enforces the model to be invariant to the transformations *w.r.t.* video semantics, snippet predictions, and snippet representations. Consequently, fine-grained patterns in-between adjacent snippets are enforced to be explored, thereby resulting in a more robust action boundary localization. Experimental results demonstrate the effectiveness of C³BN on top of various baselines for WTAL with video-level and point-level supervisions. Code is at C3BN.

Index Terms— Weakly-supervised temporal action localization, adjacent snippets

1. INTRODUCTION

Temporal action localization [1, 2, 3] intends to localize action instances and recognize their categories in videos. In recent years, numerous works delve into the fully supervised TAL and gain significant improvement. However, these methods require tremendous manual frame-level annotations, which is expensive and time-consuming. Recently, weakly-supervised TAL (WTAL)[4, 5] has received increasing attention, as it allows us to detect the action instances with only weak supervision, *e.g.*, video-level labels [4] and point-level labels [6, 7]. In particular, video-level labels are the most commonly used.
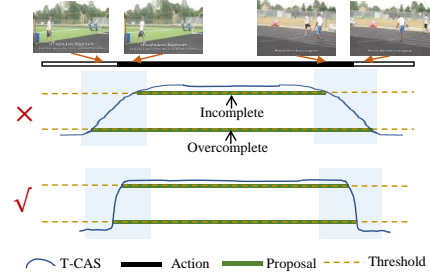
Mainstream WTAL methods [4, 8], regardless of the types

Fig. 1. **Motivation of our method.** Due to the vague distinctions between adjacent snippets, an under-performing model may produce similar activations for these snippets, resulting in incomplete/overcomplete proposals (see the upper T-CA). In this paper, we expect the model to correctly classify adjacent snippets, thereby localizing accurate boundaries (see the lower T-CAS).

of weak supervisions, employ a video action classification model to learn the Temporal Class Activation Sequence (T-CAS). After training, they utilize the T-CAS to localize action in a bottom-up fashion [1] derived from the watershed algorithm. Specifically, it consists of two main steps. ❶ *boundary localization*: generating action proposals by thresholding and merging the continuous action regions of T-CAS with multiple thresholds; ❷ *proposal evaluation*: calculating proposal-level scores by aggregating snippet-level scores within the regions. Recent methods pay many efforts to learn accurate snippet-level scores by various techniques, *e.g.*, pseudo-labeling [9, 10] and contrastive learning [11, 12]. In other words, these methods focus on the semantic relationships between each snippet and global class centers/other snippets. Despite the progress, we argue that they may be sub-optimal since what really matters in ❶ is the relationship between adjacent snippets [13]. As depicted in Fig. 1, adjacent snippets are usually similar in content and thus have close activation, which may cause incomplete or over-complete proposals. Hence, it is necessary to enable the model to be sensitive enough to the fine-grained distinctions between adjacent snippets.

To counteract this issue, we introduce a plug-and-play training strategy dubbed Convex Combination Consistency Between Neighbors (C³BN) for WTAL. The idea of our work stems from MixUp [14], where the classification model trained on the mixture of image pairs achieves promising per-

formance. In light of this, to enhance the ability of the WTAL model to distinguish adjacent snippets, we propose a micro [1] data augmentation strategy, where the pairs of adjacent snippets (termed as *parent snippets*) are mixed by convex combination to generate a set of new snippets (termed as *child snippets*). However, there are still two problems that need to be handled before using the child snippets. The first problem is how to feed the child snippets into WTAL models. Unlike conventional MixUp which treats images as independent instances, most WTAL models take the snippet sequences as input, followed by a few temporal convolution layers to enlarge the temporal receptive field. Therefore, we have to define the temporal orders of the child snippets before they can be processed by the models. To address the challenge, we propose to take advantage of the *temporal continuity* prior in videos [15]: the scenes usually change smoothly and continuously along temporal dimension. This property implies that the temporal location of each child snippet lies in-between that of its parent snippets. With the virtual locations, we arrange the child snippets of a video into a new sequence, which can be viewed as a locally deformed version of the original sequence.

The second problem is how to utilize the child snippets to promote model training. In MixUp, the mixed sample is assigned with the mixture of the ground-truth labels of the original samples, encouraging the model to behave linearly in-between samples. In our case, however, only weak labels of snippets are available. To this end, we develop a macro-micro consistency regularization, which makes use of both weak supervision and linear behaviour to regularize the model training . Specifically, we introduce three consistency regularization terms to exploit different relationships between child and parent snippets *w.r.t.* video semantics, snippet predictions and snippet representations, thereby facilitating model training from *macro view* to *micro view* and from *low-level representations* to *high-level semantics*. In this way, more fine-grained cues in-between adjacent snippets are preserved, eventually improving the robustness of boundary localization.

The idea behind C^3BN is generic and conceptually complementary to other methods, which is justified by the performance promotion on a variety of base approaches and datasets. More importantly, extensive quantitative and qualitative results verify the efficacy of C^3BN in ❶ boundary localization. Hence, our contributions are: 1) We propose to consider the potential of adjacent snippets in WTAL and then design a micro data augmentation strategy by convex combination of adjacent snippets. 2) We propose three regularization terms to enhance the consistency properties *w.r.t.* video semantics, snippet predictions and snippet features. 3) Our method can be easily plugged into existing WTAL methods with either video-level supervision or point-level supervision.

---

[1]By 'micro', we mean the proposed data augmentation strategy is on snippets rather than videos.

2. RELATED WORK

Data Augmentation aims to enlarge the train set using transformations. Conventional image transformations include cropping, flips, rotation, *etc*. Recent studies consider employing multiple images for augmentation. MixUp [14] proposes to combine the pixel values and labels of two images by linear interpolation. It has been proven effective for the classification task, which is followed by [16]. Our method employ the idea of instance mixtures with task-specific designs. Concretely, we achieve the mixture operation on two snippets within a video rather than on two different videos likewise MixUp, making the perturbations to snippets more controllable for incorporating the proposed method into the existing WTAL frameworks.

MixUp trains a model by linearly interpolating two training examples and their labels [14]. It is proven effective for the classification task, followed by different variants. For instance, [16] extends the linear interpolation from input-level to feature-level. Recently, extensive methods are proposed to incorporate MixUp with semantic segmentation [17], self-supervised learning [18], *etc*. Our method also employs the idea of instance mixtures, but it is a not trivial extension of previous methods. For example, the original MixUp mixes two randomly selected images. Extending it directly from image to snippet will cause the locations of the generated snippets undefined. In addition, according to [17], another alternative is to mix two random videos snippet-by-snippet. It is also not feasible for WTAL as the video lengths of two videos fed into the models may be different [19] in practice. *Different from above methods, we achieve the mixture operation on two adjacent snippets within a video, yielding more controllable perturbation to snippets for incorporating the method into the existing WTAL models.*

Consistency regularization is a crucial technique in semi-supervised learning. It is assumed that a classifier should output the same class probability for an unlabeled sample even after it is augmented. Prior works [20, 21] apply the consistency regularization on different augmentations of an unlabeled sample. After that, several variants [22, 23] are further proposed to extend its applications. Among them, Mix-Match [22] also uses MixUp by mixing unlabeled samples and their pseudo-labels.The differences between our method and them are *1) MixMatch randomly mixes two examples, while we only mix the adjacent snippets; 2) MixMatch guesses the hard pseudo-labels of unlabeled samples and relies on a complicated ensemble of multiple predictions to improve the quality of pseudo-labels, whereas we do not guess the hard pseudo-labels of unlabeled samples, thereby reducing undesirable label noise [24, 25].*

Self-supervised contrastive learning has attracted much attention in representation learning. The widely adopted contrastive learning optimizes the model by instance discrimination [26, 27]. Specifically, it learns to embed the features of
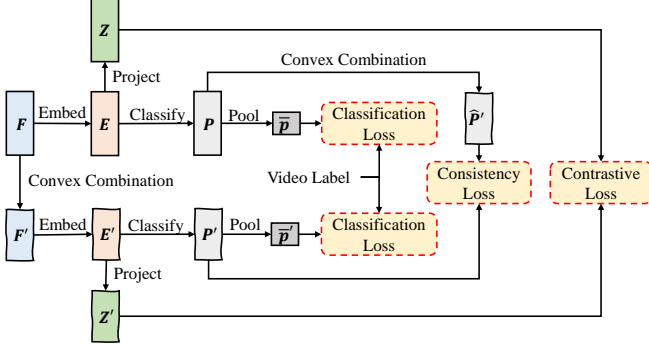
Fig. 2. **Overview of our method**. We first perform convex combination on input snippet sequence $\boldsymbol{F}$ to produce the augmented one $\boldsymbol{F}'$. The procedure is denoted as micro data augmentation. Then we simultaneously feed the $\boldsymbol{F}$ and $\boldsymbol{F}'$ into the model and compute four regularization loss terms. We call it macro-micro consistency regularization.

differently augmented versions of the same image to be similar, while being dissimilar if they came from different images. Some recent works [18, 28] have incorporated the idea of MixUp with contrastive learning. Our method is different from these methods in: *1) They regard the mixed samples as queries and the original samples as keys, while we additionally consider a reverse operation to exchange their roles. 2) In our method, the negative samples come from the same video as the positive samples, they can be viewed as hard negative samples, which is important in contarstive learning [27].*

Weakly-supervised temporal action localization aims to tackle TAL in the weakly-supervised setting. Untrimmed-Net [29] is the pioneering work for it. In addition, there are some attempts [8] to explore WTAL with only point-level action supervision, where each action instance is annotated with only a frame. Recently, [7] proposes a new WTAL setting with point-level background supervision, which annotates a frame in each background segment. In this work, we consider the former two types of supervision with more followers.

Despite of different supervisions, most methods follow a localization-by-classification procedure, which formulates WTAL as a video classification task. Under this pipeline, an important component is to select snippets with high probabilities of actions. In general, there are two groups of strategies: multiple instance learning (MIL)-based methods [12, 30] and attention-based methods [31, 32]. The former obtains the video-level scores from T-CAS by applying a pooling on the top-k values for each class. The latter introduces the attention weights to eliminate background snippets. Recently, some WTAL methods have also noticed contrastive learning [12, 11]. The difference between the above methods and ours is indeed obvious. *They rely on pseudo-labels for defining positive and negative pairs. In contrast, we formulate it as an instance discrimination task, which is simpler and more generic.*

3. OUR METHOD

In this section, we first review the basic pipeline of the mainstream WTAL methods which we adopt as our baselines. Then we elaborate on our proposed C³BN, which contains two core components: micro data augmentation and macro-micro consistency regularization, as depicted in Fig. 2.

3.1. Preliminaries

For an untrimmed video $\boldsymbol{V}$, we can only access its video-level label $\boldsymbol{y} = \{y_c\}_{c=1}^C$, where C represents the number of action classes. A common practice is to employ a video classification model to predict the video-level label $\boldsymbol{y}$. Specifically, we first divide $\boldsymbol{V}$ into T non-overlapping snippets and then extract snippet-wise features $\boldsymbol{F} \in \mathrm{R}^{T \times D_f}$ using the pre-trained feature extractor. Since the extractor is not trained from scratch for the WTAL task, we further use several temporal convolution layers for mapping $\boldsymbol{F}$ to task-specific feature embedding $\boldsymbol{E} = [\boldsymbol{e}_1, ..., \boldsymbol{e}_T] \in \mathrm{R}^{T \times D_e}$. Afterward, $\boldsymbol{E}$ is fed into a snippet classifier to output a snippet prediction sequence $\boldsymbol{P} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_T] \in \mathrm{R}^{T \times C}$, where each snippet has its own scores $\boldsymbol{p}_t \in \mathrm{R}^C$. After this, we aggregate the snippet scores to obtain video-level scores $\bar{\boldsymbol{p}} \in \mathrm{R}^C$. There are two main mechanisms in the literature for this purpose: MIL-based [12] and attention-based [19]. The former applies a temporal top-k pooling to select high-score snippets while the latter uses snippet-wise attention weights to aggregate snippets via attentive pooling. We refer to **Supplementary** for details. Then, we formulate a video classification loss as follows:

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{c=1}^{C} y_c \log \bar{p}_c. \tag{1}$$

3.2. Micro Data Augmentation

Upon the T snippets $\{\boldsymbol{f}_t\}_{t=1}^T$, we perform convex combination on adjacent snippets to generate $T-1$ augmented snippets. Formally,

$$\boldsymbol{f}_t' = \alpha_t \boldsymbol{f}_t + (1 - \alpha_t)\boldsymbol{f}_{t+1} \quad \forall t \in \{1, .., T-1\}, \tag{2}$$

where $\boldsymbol{f}_t'$ is the child snippet and $\{\boldsymbol{f}_t, \boldsymbol{f}_{t+1}\}$ are its parent snippets. The weight $\alpha_t \in (0, 1)$ is randomly sampled from a beta distribution $\mathrm{Beta}(\gamma, \gamma)$. Here γ is a preset scalar. According to the temporal continuity of videos, the location of $\boldsymbol{f}_t'$ lies in between t and $t + 1$. Consequently, $\boldsymbol{f}_t'$ is always in front of $\boldsymbol{f}_{t+1}'$. Based on this principle, we stack the child snippets $\{\boldsymbol{f}_t'\}_{t=1}^{T-1}$ along temporal dimension to form a 1D feature map dubbed $\boldsymbol{F}' \in \mathrm{R}^{(T-1) \times D_f}$.

Remark. Notably, our proposed micro data augmentation is fundamentally different from the native MixUp. Specifically, MixUp randomly selects two images for mixing, which is not applicable to the snippets in WTAL as it will render the temporal locations of the mixed snippets undefined. In contrast, we intend to improve the boundary localization in

WTAL and thus propose to mix the adjacent snippets, meanwhile taking advantage of the natural continuity within videos to define the locations of the mixed snippets. Besides, we propose various consistency regularization terms to encode more task-specific knowledge, which is also distinct from MixUp.

3.3. Macro-Micro Consistency Regularization

To effectively exploit the child snippets during training, we derive three combinatorial rules to collaboratively regularize the learning procedure.

Video semantic consistency. The semantic label of each child snippet is expected to be the combination of its parents' labels. Then we can deduce that the video-level label of child sequence F' is consistent with that of parent sequence F. Therefore, it is feasible to utilize the known video-level labels to regularize the child sequences. Specifically, we feed F' into network and get a video classification loss named L'_{cls} in the same form as L_{cls}. Since F' is a locally deformed version of F, the usage of L'_{cls} is supposed to help to improve the robustness of the model. However, such macro regularization ignores the relationship between individual child snippet and parent snippet. To this end, we propose to further regularize the network from micro perspectives.

Snippet prediction consistency. Inspired by MixUp, we encourage the model to behave linearly in-between adjacent snippets, thereby enhancing the ability of the model to classify adjacent snippets. Without ground-truth labels, we propose to take the predictions of the snippets as their "soft-labels". Thereafter, we introduce a consistency regularization term to enforce the soft-labels/prediction of a child snippet to be consistent with the same convex combination of soft-labels/predictions of its parent snippets.

By feeding F and F' into the model, we can obtain two snippet prediction sequences, namely P and P'. To bridge P and P', we apply convex combination on P to obtain the shifted version of P dubbed $\widehat{P}'$. Formally,

$$\hat{p}'_t = \alpha_t p_t + (1 - \alpha_t) p_{t+1} \quad \forall t \in \{1, .., T-1\}. \tag{3}$$

Then we apply the MSE loss to enforce the consistency between P' and $\widehat{P}'$. Formally,

$$\mathcal{L}_{cons} = \frac{1}{T-1} \sum_{t=1}^{T-1} ||p'_t - \hat{p}'_t||_2^2. \tag{4}$$

Snippet feature contrastive-consistency. Recent methods [33, 34] demonstrate that feature contrastive learning is complementary to classifier learning. In light of these works, we propose to further regularize the intermediate features of the model via contrastive learning. In particular, we develop a contrastive-consistency regularization that integrates the consistency regularization into the contrastive learning scheme. By "contrastive", we mean that the model is forced to distinguish the parent/child snippets of each child/parent snippet from other parent/child snippets. By "consistency", we mean

that we enforce the model to learn the degree of proximity between child and parent snippets. Here we jointly achieve both of goals, and let the snippet features be aware of the relative similarity in-between adjacent snippets in comparing with other snippets. Consequently, the features would gradually capture necessary fine-grained discriminability to distinguish subtle differences between adjacent snippets.

First, to avoid the conflict between the instance-based contrastive learning and the underlying semantics within the feature embedding E, we append a projection head [26], comprised by a FC layer and a L2 normalization, to map E into a low-dimensional unit hypersphere in which the contrastive learning is performed. As a result, E serves as a medium for information transfer between the classifier and the projection head, allowing the classifier to leverage discriminative fine-grained patterns captured by the projection head to extract accurate class-level patterns. Let us denote the output by $Z = [z_1, .., z_t] \in \mathrm{R}^{T \times D_z}$. Here D_z is the number of channels with $D_z < D_f$. Similarly, we can obtain the counterpart of E' denoted by $Z' = [z'_1, .., z'_t]$.

Then, taking each child snippet z'_t as a query, we define that: 1) its parent snippets z_t and z_{t+1} are its semi-positive keys with the probability of α_t and $1 - \alpha_t$, respectively. 2) other snippets in Z are negative keys. Thus we can construct a soft contrastive loss as follows:

$$
\begin{aligned}
\mathcal{L}_{cont} = -\frac{1}{T-1} \sum_{t=1}^{T-1} \alpha_t & \log \frac{\exp(z'^{\top}_t z_t / \rho)}{\sum_{\tau=1}^{T} \exp(z'^{\top}_t z_\tau / \rho)} + \\
& (1 - \alpha_t) \log \frac{\exp(z'^{\top}_t z_{t+1} / \rho)}{\sum_{\tau=1}^{T} \exp(z'^{\top}_t z_\tau / \rho)},
\end{aligned} \tag{5}
$$

where ρ is the temperature coefficient.

Eq.(5) only considers the unilateral reference from Z' to Z. To explore more fine-grained patterns and enhance the consistency regularization between Z and Z', we propose a bilateral reference mechanism to further take the reference from Z to Z' into consideration. That is, we treat the elements of Z as queries and the elements of Z' as keys. Meanwhile, the snippet-to-snippet relations remain unchanged. As a result, we can compute another contrastive loss dubbed $\mathcal{L}'_{cont}$ in a similar way to Eq.(5).

Remark. The three regularization terms are introduced to work collaboratively, comprehensively promoting the model training from *macro view* to *micro view* and from *low-level representations* to *high-level semantics*. Intuitively, these constraints together will encourage the model to exploit the various relationships between parent and child snippets, eventually facilitating the exploration of fine-grained distinctions between adjacent snippets. We will show later in Sec. B the efficacy and compatibility of these terms. It is noteworthy that the concrete differences between our proposed regularization strategies and related methods are highlighted in Sec. 2.

Table 1. Comparisons of performance on THUMOS14. The AVG represents average mAP under IoU thresholds of 0.1:0.7. We re-implement all the adopted baselines for C^3BN.

Method	mAP @ IoU (%)							AVG
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
WUM [39]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	41.9
AUMN [40]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	41.5
CoLA [12]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	40.9
UGCT [41]	69.2	62.9	55.5	46.5	35.9	23.8	54.0	43.6
DCC [11]	69.0	63.8	55.9	45.9	35.7	24.3	13.7	44.0
RSKP [9]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	45.1
ASM-Loc [42]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	45.1
Li *et al.* [43]	69.7	64.5	58.1	**49.9**	39.6	27.3	14.2	46.1
MIL	56.0	46.4	37.3	30.3	22.0	15.0	8.2	30.7
+ C^3BN	63.0	56.7	48.0	39.8	29.9	19.2	10.2	38.1$_{+7.4}$
BaSNet [37]	62.0	54.6	44.6	35.7	25.9	17.0	8.9	35.5
+ C^3BN	64.3	58.4	49.7	40.6	30.8	19.9	12.1	39.4$_{+3.9}$
FACNet [19]	71.8	64.0	53.7	42.5	30.7	20.9	12.2	42.3
+ C^3BN	**72.6**	**66.5**	56.4	43.8	32.6	21.0	12.7	43.7$_{+1.5}$
DELU [38]	70.1	64.5	56.0	47.6	40.2	27.8	15.0	45.9
+ C^3BN	71.6	66.0	**58.2**	49.3	**41.0**	**27.9**	**15.3**	**47.0**$_{+1.1}$

4. EXPERIMENTS

In this section, we empirically validate the effectiveness of C^3BN. Due to space limitation, we refer to **Supplementary** for more details for all experiments. Moreover, we only report the results on WTAL with video-level supervision and refer to **Supplementary** for WTAL with point-level supervision.

4.1. Dataset and Metrics

THUMOS14 [35] contains untrimmed videos with 20 classes. By convention, we use the 200 videos in validation set for training and 213 videos in test set for evaluation. *ActivityNet v1.3* [36] is a large-scale dataset with 200 categories. By convention, we train on the training set with 10, 024 videos and test on validation set with 4, 926 videos. The mean Average Precision (mAP) values under different temporal intersection over union (tIoU) thresholds are used as metrics.

4.2. Ablation study

Effectiveness on different baselines To validate the generic effectiveness of C^3BN, we incorporate C^3BN into different WTAL methods. Specifically, we plug C^3BN into four baselines, including the aforementioned MIL-based baseline (named **MIL**) and three off-the-shelf well-performing approaches, *i.e.*, **BaSNet** [37], **FACNet** [19], and recent **DELU** [38](ECCV 2022). - Table 1 shows the performance comparison. We can observe that C^3BN consistently improves the performance of all baselines by 7.4%, 3.9%, 1.5%, and 1.1% on AVG mAP for MIL, BaSNet, FACNet, and DELU, respectively. These results clearly confirm the generalization ability of C^3BN.

Contribution of each regularization term Our C^3BN introduces several regularization/loss terms during training. To verify the contribution of each regularization term, we conduct a detailed analysis in Table 3. Here, we regard BaSNet and FACNet as the baselines to conduct the ablation study

Table 2. Results on ActivityNet v1.3. AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95.

Method	mAP @ IoU			
	0.5	0.75	0.95	AVG
WUM [39]	37.0	23.9	5.7	23.7
AUMN [40]	38.3	23.5	5.2	23.5
UGCT [41]	39.1	22.4	5.8	23.8
DCC [11]	38.8	24.2	5.7	24.3
RSKP [9]	40.6	24.6	5.9	25.0
ASM-Loc [42]	41.0	24.9	6.2	25.1
BaSNet	35.6	21.0	5.3	21.7
+ C^3BN	37.3	22.4	5.4	23.0$_{+1.3}$
FACNet [19]	40.1	24.2	5.8	24.7
+ C^3BN	**45.2**	**26.9**	5.9	**27.3**$_{+2.6}$

Table 3. Ablation studies of the proposed regularization terms on THUMOS14.

#	Loss terms				Baselines	
	$\mathcal{L}'_{cls}$	$\mathcal{L}_{cons}$	$\mathcal{L}_{cont}$	$\mathcal{L}'_{cont}$	BaSNet	FACNet
1					35.5	42.3
2	✓				35.9	42.6
3		✓			37.7	43.0
4			✓	✓	36.5	42.8
5	✓	✓			38.0	43.0
6	✓		✓	✓	36.8	42.9
7		✓	✓	✓	38.8	43.5
8	✓	✓		✓	39.1	43.4
9	✓	✓	✓		38.9	43.5
10	✓	✓	✓	✓	39.4	43.7

Table 4. Contribution of C^3BN to ❶ boundary localization and ❷ proposal evaluation. "❶Base+❷Base" represents the baseline, *i.e.*, BaSNet and FACNet. "❶C^3BN+❷Base" indicates that we combine our ❶ and the ❷ of baseline. Likewise for "❶Base+❷C^3BN" and "❶C^3BN+❷C^3BN".

	❶Base+❷Base	❶C^3BN+❷Base	❶Base+❷C^3BN	❶C^3BN+❷C^3BN
BaSNet	35.5	38.0	37.1	39.4
FACNet	42.3	43.2	42.8	43.7

due to their favorable efficiency and flexibility. Comparing the rows #1-4, we can see that each regularization term contributes to the performance. Furthermore, it can be seen that the micro consistency regularization terms (*i.e.*, $\mathcal{L}_{cons}$, $\mathcal{L}_{cont}$ and $\mathcal{L}'_{cont}$) bring larger gains than the macro term (*i.e.*, $\mathcal{L}'_{cls}$). This indicates that fine-grained information is pretty important in WTAL.

Complementarity of regularization terms In rows #5-6 of Table 3, we evaluate the performance of combining any two of the regularization terms. We can see that combining two terms consistently outperforms each of them. Moreover, after combining all the terms, the model obtains the best performance, as shown in row #10. These results evidently demonstrate the complementary relations of the regularization terms.

Effectiveness of bilateral reference mechanism We propose a bilateral reference mechanism in the snippet feature contrastive-consistency regularization, resulting in two loss terms (*i.e.*, $\mathcal{L}_{cont}$ and $\mathcal{L}'_{cont}$). In rows #8-9 of Table 3, we provide the results where only one of them is used. It can be seen that the combination of $\mathcal{L}_{cont}$ and $\mathcal{L}'_{cont}$ outperforms using only of them, validating the superiority of our proposed bilateral reference mechanism.

Necessity of projection head   We adopt a projection head to transform the embeddings into a new latent space so that the instance-based contrastive learning would not directly hurt the inherent semantics of the embeddings. To show the necessity of our design, we conduct an experiment where the projection head is removed. The experimental results show that it leads to a performance degradation of 1.1% on BaSNet (from 39.4% to 38.3%) and 0.9% on FACNet (from 43.7% to 42.8%). This evidently justifies that the projection head is essential in our method.

4.3. Comparisons with state-of-the-arts (SOTAs)

Table 1 and Table 2 show the comparison between our method and previous approaches on THUMOS14 and ActivityNet v1.3, respectively. It can be seen that after integrating recent strong WTAL baselines, our method achieves the SOTA performances on both datasets. In **Supplementary**, we also report the results on *ActivityNet v1.2* [36] (a subset of ActivityNet v1.3), which is used in some previous methods [38, 43].

4.4. Evaluation for Motivation

In this section, we provide experimental results to deliver more insights for our motivation (depicted in Fig. 1).

To begin with, we investigate the contribution of C^3BN to "❶ boundary localization" and "❷ proposal evaluation" respectively in Table 4. To be specific, in the test phase, we alternately replace the results of ❶ and ❷ of baseline with that of baseline+C^3BN (ours). As we can see, the performances significantly increase (from 35.5% to 38.0% on BaSNet and from 42.3% to 43.2% on FACNet) once replacing the ❶ of baselines with our ❶. As a comparison, the replacement on ❷ only brings the performance to 37.1% on BaSNet and 42.8% on FACNet. These results indicate that C^3BN is particularly beneficial for boosting boundary localization.

To understand how C^3BN improves the boundary localization, we first compute the absolute score difference between each pair of adjacent snippets, *i.e.*, $d_t = |p_{t+1} - p_t|$. Next, we calculate the average entropy of d_t of all pairs, *i.e.*, $H(d_t) = \text{mean}(-d_t \log d_t)$. Intuitively, the $H(d_t)$ reflects the distribution of the differences, *i.e.*, the smaller the $H(d_t)$, the more polarized the differences, and the more discriminative and confident the model is about the relations of adjacent snippets. Experimental results show that the corresponding $H(d_t)$ of BaSNet and FACNet is 0.0876 and 0.0673 respectively while that of "BaSNet + C^3BN" and "FACNet + C^3BN" is 0.0684 and 0.0592 respectively. Hence, we conjecture that the boundary localization is improved because C^3BN renders the model more confident (discriminative) in distinguishing the relations of adjacent snippets.

In **Supplementary**, we provide extensive qualitative results to demonstrate it.

5. CONCLUSION

In this paper, we propose a universal training strategy dubbed C^3BN for weakly-supervised action localization. Concretely, C^3BN first produces new snippets by convex combination between adjacent snippets, and then uses them to regularize the model with three regularization terms, *i.e.*, video semantic consistency, snippet prediction consistency and snippet feature contrastive-consistency. The empirical results validate that C^3BN is applicable to various WTAL methods with video-level supervision and point-level supervision, and helps establish the new SOTA results on all the evaluated datasets.

6. REFERENCES

[1] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin, "Temporal action detection with structured segment networks," in *ICCV*, 2017. 1

[2] Qinying Liu, Zilei Wang, and Shenghai Rong, "Improve temporal action proposals using hierarchical context," *Pattern Recognition*, vol. 140, pp. 109560, 2023. 1

[3] Qinying Liu and Zilei Wang, "Progressive boundary refinement network for temporal action detection," in *AAAI*, 2020. 1

[4] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *CVPR*, 2017. 1, 9

[5] Qinying Liu, Zilei Wang, Shenghai Rong, Junjie Li, and Yixin Zhang, "Revisiting foreground and background separation in weakly-supervised temporal action localization: A clustering-based approach," in *ICCV*, 2023. 1

[6] Davide Moltisanti, Sanja Fidler, and Dima Damen, "Action recognition from single timestamp supervision in untrimmed videos," in *CVPR*, 2019. 1

[7] Le Yang, Junwei Han, Tao Zhao, Tianwei Lin, Dingwen Zhang, and Jianxin Chen, "Background-click supervision for temporal action localization," *TPAMI*, 2021. 1, 3

[8] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou, "Sf-net: Single-frame supervision for temporal action localization," in *ECCV*, 2020. 1, 3, 9, 10, 11

[9] Linjiang Huang, Liang Wang, and Hongsheng Li, "Weakly supervised temporal action localization via representative snippet knowledge propagation," in *CVPR*, 2022. 1, 5

[10] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *CVPR*, 2021. 1

[11] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng, "Exploring denoised cross-video contrast for weakly-supervised temporal action localization," in *CVPR*, 2022. 1, 3, 5

[12] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou, "Cola: Weakly-supervised temporal action localization with snippet contrastive learning," in *CVPR*, 2021. 1, 3, 5, 8, 9, 11

[13] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim, "Uboco: Unsupervised boundary contrastive learning for generic event boundary detection," in *CVPR*, 2022. 1

[14] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018. 1, 2

[15] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid, "Action tubelet detector for spatio-temporal action localization," in *ICCV*, 2017. 2

[16] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *ICML*, 2019. 2

[17] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang, "Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization," in *BMVC*, 2020. 2

[18] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun, "Mixco: Mix-up contrastive learning for visual representation," *arXiv preprint arXiv:2010.06300*, 2020. 2, 3

[19] Linjiang Huang, Liang Wang, and Hongsheng Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *ICCV*, 2021. 2, 3, 5, 9, 11

[20] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2017. 2

[21] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *NeurIPS*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, Eds., 2016. 2

[22] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS*, 2019. 2

[23] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020. 2

[24] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang, "Humble teachers teach better students for semi-supervised object detection," in *CVPR*, 2021. 2

[25] Qinying Liu and Zilei Wang, "Collaborating domain-shared and target-specific feature clustering for cross-domain 3d action recognition," in *ECCV*, 2022. 2

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020. 2, 4

[27] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka, "Contrastive learning with hard negative samples," in *ICLR*, 2020. 2, 3

[28] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee, "i-mix: A domain-agnostic strategy for contrastive representation learning," in *ICLR*, 2020. 3

[29] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *CVPR*, 2017. 3, 8

[30] Md Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C Leu, "Action completeness modeling with background aware networks for weakly-supervised temporal action localization," in *ACMMM*, 2020. 3

[31] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han, "Weakly supervised action localization by sparse temporal pooling network," in *CVPR*, 2018. 3

[32] Ashraful Islam, Chengjiang Long, and Richard Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *AAAI*, 2021. 3

[33] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He, "Multi-level feature learning for contrastive multi-view clustering," in *CVPR*, 2022. 4

[34] Kien Do, Truyen Tran, and Svetha Venkatesh, "Clustering by maximizing mutual information across views," in *ICCV*, 2021. 4

[35] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," 2014. 5, 9

[36] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015. 5, 6, 9

[37] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun, "Background suppression network for weakly-supervised temporal action localization," in *AAAI*, 2020. 5, 9

[38] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu, "Dual-evidential learning for weakly-supervised temporal action localization," in *ECCV*, 2022. 5, 6, 9, 11

[39] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *AAAI*, 2021. 5, 11

[40] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang, "Action unit memory network for weakly supervised temporal action localization," in *CVPR*, 2021. 5

[41] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *CVPR*, 2021. 5

[42] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava, "Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization," in *CVPR*, 2022. 5

[43] Ziqiang Li, Yongxin Ge, Jiaruo Yu, and Zhongming Chen, "Forcing the whole video as background: An adversarial learning strategy for weakly temporal action localization," in *ACMMM*, 2022. 5, 6, 11

[44] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes, "Weakly-supervised action localization with background modeling," in *ICCV*, 2019. 8

[45] Daochang Liu, Tingting Jiang, and Yizhou Wang, "Completeness modeling and context separation for weakly supervised action localization," in *CVPR*, 2019. 8

[46] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu, "Weakly supervised action selection learning in video," in *CVPR*, 2021. 9

[47] Pilhyeon Lee and Hyeran Byun, "Learning action teness from points for weakly-supervised temporal action localization," in *ICCV*, 2021. 9, 10, 11

[48] C Zach, T Pock, and H Bischof, "A duality based approach for realtime tv-l 1 optical flow," *PR*, 2007. 9

[49] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017. 9

[50] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 9

[51] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *ECCV*, 2018. 9

[52] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis, "Soft-nms–improving object detection with one line of code," in *ICCV*, 2017. 9

[53] A Calway, W Mayol-Cuevas, D Damen, O Haines, and T Leelasawassuk, "Discovering task relevant objects and their modes of interaction from multi-user egocentric video," in *BMVC*, 2015. 9

[54] Peng Lei and Sinisa Todorovic, "Temporal deformable residual networks for action segmentation in videos," in *CVPR*, 2018. 9

[55] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian, "Divide and conquer for single-frame temporal action localization," in *ICCV*, 2021. 10, 11

[56] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua, "Two-stream consensus network for weakly-supervised temporal action localization," in *ECCV*, 2020. 11

[57] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, "D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations," in *ICCV*, 2021. 11

[58] Zichen Yang, Jie Qin, and Di Huang, "Acgnet: Action complement graph network for weakly-supervised temporal action localization," in *AAAI*, 2022. 11

Contents

A. ADDITIONAL DETAILS OF OUR METHOD

Additional details of baseline.   Given the video labels, we first aggregate the snippet scores to obtain video class scores for computing a video classification loss. There are two main strategies in the literature for this purpose: MIL-based methods [29, 12] and attention-based methods [44, 45].

The MIL-based methods average the top-k snippet *logit* scores (dubbed $\boldsymbol{S} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_T] \in \mathrm{R}^{T \times C}$) along temporal dimension for each class to build the video class score:

$$\bar{s}_c = \frac{1}{k} \max_{\substack{\boldsymbol{l} \subset \{1,..,T\} \\ |\boldsymbol{l}| = k}} \sum_{\tau \in \boldsymbol{l}} s_{\tau,c} \quad \forall c \in \{1,..,C\}, \qquad (6)$$

where k is a hyper-parameter proportional to the video length T, *i.e.*, $k = \max(1, T//r)$, and r is a pre-defined parameter. Thereafter, we obtain the probability for each class by applying the $\mathrm{Softmax}$ function to the aggregated scores:

$$\bar{p}_c = \frac{\exp(\bar{s}_c)}{\sum_{i=1}^{C} \exp(\bar{s}_i)} \quad \forall c \in \{1,..,C\}. \qquad (7)$$

The attention-based methods first learn a set of snippet-wise attention weights (dubbed $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_T] \in \mathrm{R}^T$). Then the attention weights are used to aggregate snippet-level scores into video-level scores as follows,
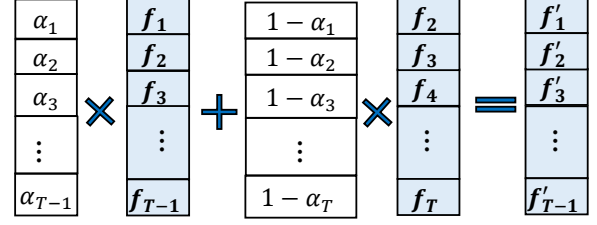

Fig. 3. **Illustration of micro data augmentation.** The $\{\boldsymbol{f}_t\}_{t=1}^{T-1}$ indicates the parent sequence and the $\{\boldsymbol{f}'_t\}_{t=1}^{T-1}$ represents the child sequence.
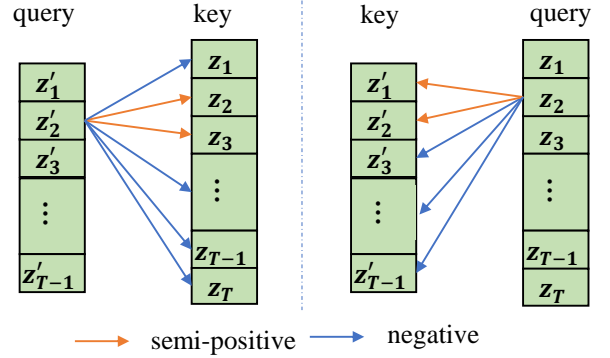

Fig. 4.   **Illustration of snippet feature contrastive-consistency.** It contains two parts: the child snippets are query while the parent ones are key (left); the parent snippets are query while the child ones are key (right).

$$\bar{\boldsymbol{p}} = \frac{1}{\sum_{t=1}^{T} \lambda_t} \sum_{t=1}^{T} \lambda_t \boldsymbol{p_t}. \qquad (8)$$

Illustration of micro data augmentation and snippet feature contrastive-consistency .   In Fig. 3 and Fig. 4, we shows more details for micro data augmentation and snippet feature contrastive-consistency respectively.

Training objective.   To train the entire model in an end-to-end fashion, we optimize the following loss

$$\mathcal{L} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}'_{cls} + \lambda_2 \mathcal{L}_{cons} + \lambda_3 (\mathcal{L}_{cont} + \mathcal{L}'_{cont}). \qquad (9)$$

Here λ_* indicates the weight term. $\mathcal{L}_{base} = \mathcal{L}_{cls} + \mathcal{L}_{other}$ represents the objective function of the baseline, where $\mathcal{L}_{other}$ represents the sum of other losses apart from $\mathcal{L}_{cls}$.

B. EXPERIMENTS

B.1. Implementation Details

Implementation details of baselines   **MIL** is a simple MIL-based baseline, which has been introduced Sec. A. Briefly, it first performs snippet classification over the input

sequences to obtain T-CAS, then utilizes a top-k pooling operation following [4] to build video-level scores. At last, it minimizes a cross-entropy loss with the video-level labels. We implement MIL based on the code provided by [12], except that we disable the contrastive loss used in [12]. **BaSNet** [37] introduces an additional class to the snippet-level classifier for modeling background. Besides, it utilizes a class-agnostic attention layer to highlight foreground snippets and suppress the background snippets. The performance for BaSNet implemented by [37] is not stable on THUMOS14. Thus, we re-implement it on the basis of the code of [46, 19]. We refer to our released code for details. **FACNet** [19] proposes to force the foreground score output by the snippet-level classifier and that output by the attention layer to be consistent. We implement the FACNet in a similar way to the BaSNet. **DELU** [38] extends the traditional paradigm of evidential deep learning to adapt to the weakly-supervised multi-label classification goal. **SF-Net** [8] mines pseudo action and background frames by adaptively expanding each annotated single frame to its nearby frames. **LACP** [47] takes the points as seeds and searchs for the optimal sequence that is likely to contain complete action instances while agreeing with the seeds. For DELU, SF-Net and LACP, we use their official code to implement them. All the baselines are implemented on the Pytorch library.

Training details For the feature extraction, we first sample RGB frames at 25 fps for each video and apply the TV-L1 algorithm [48] to generate optical flow frames. Then, we divide each video into non-overlapping snippets with consecutive 16 frames. Thereafter, we perform the I3D network [49] pre-trained on the Kinetics dataset [50] to obtain the snippet-level feature. The proposed C^3BN and the baseline models are jointly trained in an end-to-end manner. Here, we only provide the details about the specific hyperparameters of C^3BN. The temperature ρ is set as 0.1 and the output dimension of projection head D_z is set as 128 and the γ is set as 2. Since the amplitudes of basic loss in different baselines are different, the loss weights λ_2 and λ_3 are set differently on different baselines, except that the λ_1 is always set by 1. On BaSNet, the λ_2 and λ_3 are set as 10 and 0.1, respectively. On FACNet, the λ_2 and λ_3 are set as 1 and 0.2, respectively. On DELU, the λ_2 and λ_3 are set as 0.1 and 0.3, respectively. On SF-Net, the λ_2 and λ_3 are set as 10 and 0.2, respectively. On LACP, the λ_2 and λ_3 are set as 2 and 0.1, respectively. All experiments are conducted on one GTX 3090 GPU (24 GB).

Inference details The proposed C^3BN is a training strategy, which introduces no overhead in the test phase. As addressed in main paper, for existing WTAL methods, there exist two core procedures in the test phase, *i.e.*, boundary localization and proposal evaluation. Despite of this, the actual test paradigms in different baselines are slightly different. We hereby take our implementation on BaSNet as an example for illustration. In the inference stage, we first threshold on the video-level scores to determine the action categories.

And then for the selected action class, we apply a threshold strategy on the T-CAS to obtain action proposals. After obtaining the action proposals, we calculate the class-specific score for each proposal using the outer-inner-contrastive technique [51]. To enrich the proposal pool, multiple thresholds are applied. The Non-Maximum Suppression (NMS) is used to remove duplicated proposals, where SoftNMS [52] is particularly adopted.

Additional details of datasets. For W-TAL with video-level supervision, we conduct experiments on two popular benchmark datasets: *THUMOS14* [35] and *ActivityNet v1.3* [36]. THUMOS14 contains untrimmed videos with 20 classes. The video length varies from a few seconds to several minutes and multiple action instances may exist in a single video, which makes it very challenging. By convention, we use the 200 videos in validation set for training and the 213 videos in test set for evaluation. ActivityNet v1.3 is a large-scale dataset with 200 categories. Since the annotations for the test set are not released, following the common practice, we train on the trainining set with 10, 024 videos and test on validation set with 4, 926 videos. *ActivityNet v1.2* [36] is a subset of ActivityNet v1.3, and covers 100 action categories with 4, 819 and 2, 383 videos in the training and validation sets, respectively

For W-TAL with point-level supervision, three public datasets are commonly used, including *THUMOS14*, *BEOID* [53], and *GTEA* [54]. GTEA contains 28 videos of 7 fine-grained types of activities in the kitchen. There are 58 videos from 30 action classes in BEOID. We follow [8] to split the training and test sets.

B.2. Qualitative Results

To gain further insights, we visualize a couple of samples for comparing the snippet-level predictions of the baseline model and that of our method in Fig. 5. From the solid boxes of Fig. 5, it can be seen that our method is able to generate accurate action boundaries, while the baseline suffers from poor discrimination around boundaries. These visualized examples evidently and intuitively verify our motivation.

In Fig. 6, we show some visualized results of the proposal generation. It can be seen that compared with the baseline model, the boundary of the proposals generated by our method is closer to the ground-truth action boundaries. This further demonstrates the superiority of our method in boundary localization.

Additionally, Fig. 7 illustrates some failure cases of our method. The failure cases are caused by 1) low quality of images (see the top row); 2) ambiguous action boundary annotation (see the middle row) ; 3) indistinguishable body motions (see the bottom row). These challenging cases are our future work.
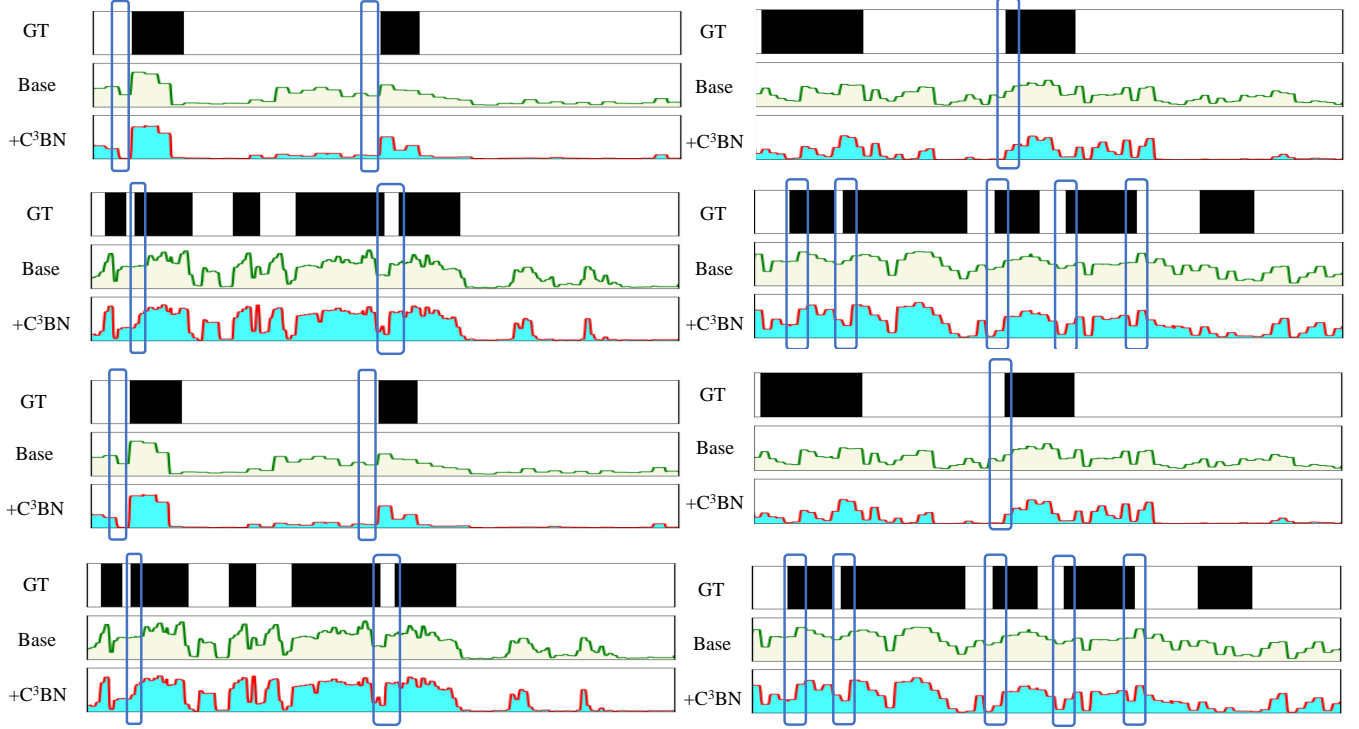
Fig. 5. **Qualitative results of T-CAS.** "GT" denotes ground-truth annotation. "Base" denotes the T-CAS predicted by BaSNet while "+C^{3}BN" denotes that predicted by "BaSNet + C^{3}BN". The solid boxes indicate some noteworthy regions.
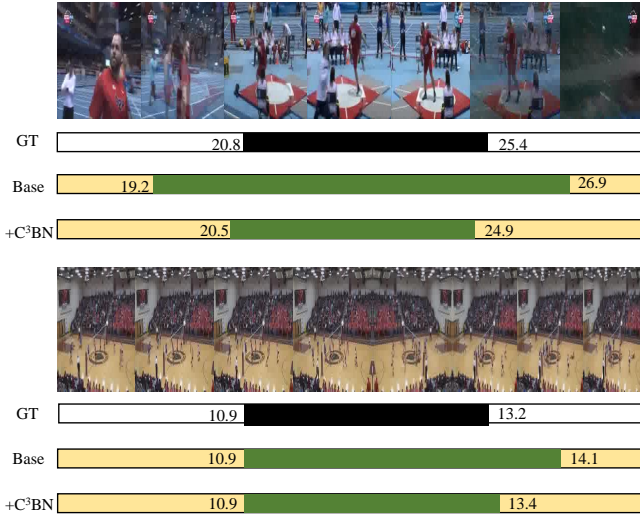

Fig. 6. **Qualitative results of proposals.** "GT" denotes ground-truth annotation. "Base" denotes the proposals output by BaS-Net while "+C^{3}BN" denotes the proposals generated by "BaS-Net + C^{3}BN".

B.3. Results on WTAL with Point-Level Labels

There are also a few works proposed for WTAL with point-level supervision, *e.g.*, SFNet [8], LACP [47], and DC [55].
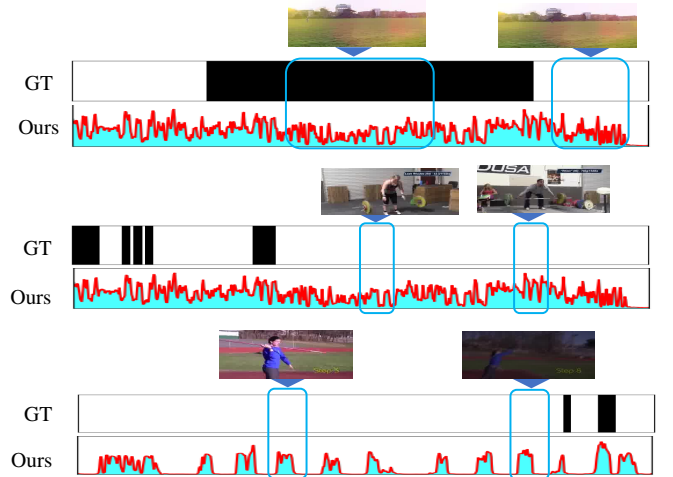

Fig. 7. **Samples of failure cases.** We highlight the regions with wrong predictions by dashed boxes.

Our C^{3}N is generic and is expected to work well for this task. We hereby take **SFNet** [8] and **LACP** [47] as the baselines, and conduct experiments on three benchmark datasets: THUMOS14, BEOID, and GTEA.

We compare the proposed approach with recent methods for WTAL with point-level supervision in Table 5. It can be

Table 5. Comparison on THUMOS14, GTEA and BEOID. AVG indicates the average mAP at IoU thresholds 0.1:0.7.

Dataset	Method	0.1	0.3	0.5	0.7	AVG
THUMOS	DC [55]	72.8	58.1	34.5	11.9	44.3
	SF-Net [8]	69.9	53.6	29.9	10.0	40.9
	Our C^3BN	73.8	57.3	30.7	10.3	$43.3_{+2.4}$
	LACP [47]	75.5	64.0	44.5	20.6	52.3
	Our C^3BN	**76.0**	**65.7**	**47.6**	**22.6**	$\mathbf{54.1}_{+1.8}$
GTEA	DC [55]	**59.7**	38.3	21.9	18.1	33.7
	SF-Net [8]	53.8	38.0	21.9	18.2	32.3
	Our C^3BN	55.1	**40.7**	**22.9**	**18.2**	$\mathbf{34.2}_{+1.9}$
BEOID	DC [55]	63.2	46.8	20.9	5.8	34.9
	SF-Net [8]	60.3	43.2	21.7	11.0	33.9
	Our C^3BN	65.5	44.0	26.3	9.7	$37.0_{+3.1}$
	LACP [47]	81.4	73.1	45.8	21.7	56.6
	Our C^3BN	**82.1**	**73.3**	**47.4**	**23.3**	$\mathbf{57.6}_{+1.0}$

Table 6. Results on ActivityNet v1.2. AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95.

	0.5	0.75	0.95	AVG
TSCN [56]	37.6	23.7	5.7	23.6
WUM [39]	41.2	25.6	6.0	25.9
CoLA [12]	42.7	25.7	5.8	26.1
D2-Net [57]	42.3	25.5	5.8	26.0
ACGNet [58]	41.8	26.0	5.9	26.1
Li et al.[43]	41.6	24.8	5.4	25.2
DELU [38]	**44.2**	26.7	5.4	26.9
FACNet [19]	41.2	26.2	5.9	26.3
+ Our C^3BN	43.9	**27.1**	**6.3**	$\mathbf{27.4}_{+1.1}$

seen that our C^3BN improves the performances of SF-Net [8] and LACP [47] by a large margin. Besides, our method also outperforms the recently proposed DC [55] and achieves SOTA performance on all three datasets. These results validate that our C^3BN is compatible with WTAL with different weak supervisions.

B.4. Results on ActivityNet v1.2

In some previous WTAL methods (especially early methods), ActivityNet v1.2 is a preferred dataset rather than ActivityNet v1.3. Hence, we also evaluate our method on this benchmark. The results are presented in Table 6. After combining with FACNet [19], our method significantly outperforms previous methods, including the very recent method DELU (ECCV 2022). The favorable performances on all the benchmarks demonstrate the overall superiority of our proposed method.