Exploring Denoised Cross-video Contrast for Weakly-supervised Temporal Action Localization

Jingjing Li[1,†], Tianyu Yang[2,†], Wei Ji[1,*], Jue Wang[2], Li Cheng[1]
[1]University of Alberta, Canada [2]Tencent AI Lab, Shenzhen, China
{jingjin1, wji3, lcheng5}@ualberta.ca, tianyu-yang@outlook.com, arphid@gmail.com

Abstract

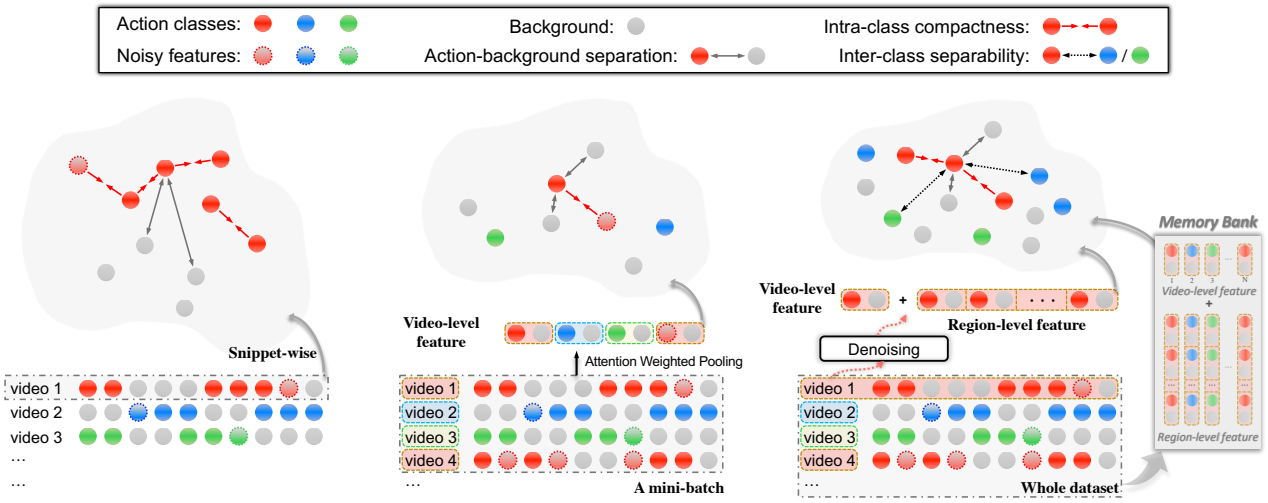
Weakly-supervised temporal action localization aims to localize actions in untrimmed videos with only video-level labels. Most existing methods address this problem with a "localization-by-classification" pipeline that localizes action regions based on snippet-wise classification sequences. Snippet-wise classifications are unfortunately error prone due to the sparsity of video-level labels. Inspired by recent success in unsupervised contrastive representation learning, we propose a novel denoised cross-video contrastive algorithm, aiming to enhance the feature discrimination ability of video snippets for accurate temporal action localization in the weakly-supervised setting. This is enabled by three key designs: 1) an effective pseudo-label denoising module to alleviate the side effects caused by noisy contrastive features, 2) an efficient region-level feature contrast strategy with a region-level memory bank to capture "global" contrast across the entire dataset, and 3) a diverse contrastive learning strategy to enable action-background separation as well as intra-class compactness & inter-class separability. Extensive experiments on THUMOS14 and ActivityNet v1.3 demonstrate the superior performance of our approach.

1. Introduction

As a fundamental yet challenging computer vision task, temporal action localization aims to localize the occurrences of prescribed action categories in untrimmed videos. It has received extensive research attention due to its wide applications in surveillance [49], video summarization [32], and highlight detection [55], *etc*. Many existing methods [4, 7, 28, 43, 56, 66, 68] are based on fully-supervised training, which rely heavily on densely annotated frame labels that are typically laborious and time-consuming to acquire. On the other hand, it is much easier for users to provide video-level tags describing scene context and content. This naturally gives rise to the weakly-supervised temporal action localization, or WS-TAL, where cheap video-level tags are utilized as an alternative supervision signal [38,41,50]. Most existing WS-TAL methods [18,25,38, 39, 41, 50, 60, 64] follow a *"localization-by-classification"* pipeline: a snippet-wise classification is carried out over time to generate the Temporal Class Activation Sequence, also called T-CAS or T-CAM [38, 41]; this is followed by selecting snippets with high responses to localize the plausible action regions. Given the sparsity nature of video-level labels, however, snippet-wise classifications are often error-prone, which may severely damage the final localization performance.

To learn a good T-CAS for action localization, it becomes crucial to enhance the feature discrimination ability of various video snippets in snippet-wise classification. Generally, the snippet feature embedding space is expected to satisfy two properties: 1) action snippets should be separable from the background snippets that do not belong to any action classes, *i.e., action-background separation*; 2) action snippets from a same class should be closer than those from different classes, *i.e., intra-class compactness & inter-class separability*. This has led to several prior studies [36, 41, 64] exploring deep metric learning [15, 26] or contrastive learning [5] to foster learning discriminative features. As illustrated in Fig. 1 (a) & (b), their focus is mostly on action-background separation, by pushing action features of a specific class to be close and pulling action features away from the background ones, either within individual videos [64], or within a carefully-designed mini-batch [36, 41]. They unfortunately fail to capture the inter-class separability, and ignore the useful *"global"* contrast across training videos in the entire dataset. Given the lack of frame-level annotations, snippet-wise pseudo-labels [64] or attention-based mechanisms [36, 41] are often used internally as a substitute. As illustrated in Fig. 1 (a), action-background separation is performed based on pseudo-labels over the snippets of each video. In Fig. 1 (b), attention-pooled video-level features from a mini-batch are engaged in the feature contrastive training process. Due to the noisy pseudo-labels or false activations in the learned attention se-

Jingjing Li does this work during an internship at Tencent AI Lab.
† Equal contribution. * Corresponding author.



(a) Within Video Contrast (b) Within Mini-batch Contrast (c) Our Denoised Cross-video Contrast

Figure 1. **Different contrastive learning schemes. (a)** Exploiting snippet-wise contrastive learning within single video to separate snippet-wise actions from backgrounds with pseudo-labels (*e.g.*, [64]). **(b)** Exploiting deep metric learning within mini-batch to separate video-level actions from backgrounds with attention-weighted pooling (*e.g.*, [36, 41]). **(c)** Our denoised cross-video contrastive algorithm with 1) pseudo-label denoising module, 2) region-level feature contrastive learning across entire dataset, and 3) action-background separation, as well as intra-class compactness & inter-class separability.

quence, these strategies would inevitably give rise to noisy contrastive features. Incorporating these noisy contrastive features may unnecessarily complicate the snippet feature training, and result in suboptimal performance of action localization.

The above observations motivate us to propose a novel Denoised Cross-video Contrastive (DCC) algorithm tailored for weakly-supervised temporal action localization. As illustrated in Fig. 1 (c), it contains three key ideas. First, to account for the pseudo-label noises that are ubiquitous in weakly-supervised TAL, a pseudo-label denoising (PLD) module is devised to reduce the negative impacts of noisy contrastive features. By down-weighting the confidence scores of incorrect pseudo-labels, more accurate contrastive features can be generated. Second, to capture "global" contrast across the entire dataset, we propose a *region-level* feature contrast strategy which, together with a *region-level* memory bank, allow our learned model to preserve "global" informative features across the entire dataset. Third, a diverse contrastive training strategy is proposed to enforce contrasts *between actions and backgrounds*, and *between different action classes*. It is capable of promoting action-background separation, inter-class separation and intra-class compactness. Note that our DCC algorithm is performed only during training, so it does not incur additional computational cost in testing.

Here we summarize our main contributions. (1) A novel denoised cross-video contrastive algorithm is proposed for weakly-supervised TAL. It reduces the influence of noisy contrastive features; it also captures "global" contrast across the entire dataset, and simultaneously promotes action-background separation, inter-class separability as well as intra-class compactness. As a result, the discrimination ability of snippet features is significantly enhanced. (2) Extensive experiments on THUMOS14 and ActivityNet v1.3 datasets demonstrate the superior performance of our approach over the state-of-the-art methods. Specifically, we observe a 16.7% improvement over the baseline in terms of average mAP of IoU thresholds from 0.1 to 0.7 on THUMOS14, a significant amount without incurring extra computation cost in inference.

2. Related Work

Temporal action localization (TAL). Fully-supervised TAL has been extensively studied over the years. They can be roughly classified into two categories, namely two-stage methods and one-stage methods. Two-stage models [4, 7, 10, 21, 22, 24, 43, 45, 56, 62, 68] first generate action proposals, then classify them by temporal boundary regression. One-stage methods [1, 23, 28, 65, 66], on the contrary, directly predict frame-level action labels. The fully-supervised paradigm unfortunately relies on densely annotated labels at the frame-level, which may be prohibitively expensive to acquire.

Weakly-supervised TAL is drawing increasing attention, as the video-level labels are comparably at low cost. UntrimmedNet [50] performs per-clip classification, then selects important clips for video label generation via a soft or hard attention. STPN [38] introduces sparsity loss to assist sparse selection of video snippets. To facilitate the detection of complete actions, [33, 46, 69] propose to remove the discriminative action parts or randomly hide

video snippets to press the models in exploring complementary action regions. Liu *et al.* [25] design a multi-branch network and a diversity loss to discover distinct temporal snippets. To improve feature discriminability, deep metric learning algorithms are explored in [33, 37, 41] to encourage action features of the same class to stay similar and to distinguish the activity-related snippets from the backgrounds. CoLA [64] proposes a snippet contrast loss to refine the hard snippet representation in feature space and make them more distinguishable. Meanwhile, explicit background modeling is introduced in [18, 39] with an auxiliary background class. Nguyen *et al.* [39] generate background attention from the foreground attention in order to pool background frames for training the background class; BaSNet [18] designs an asymmetrical training strategy to suppress background snippet activations. In [19], background frames are modeled as out-of-distribution samples. The action-context separation problem has been considered in DGAM [42] and CMCS [25]. More recently, attempts are made in [30, 40, 58, 63] to generate frame-level pseudo-labels for iterative network training. The pioneer work of [40] proposes an iterative refinement approach by estimating and training with pseudo frame-level ground-truth at each iteration. Zhai *et al.* [63] generate frame-level pseudo-labels by considering two-stream consensus and designing an attention normalization loss to promote polarizing the attention predictions. Expectation-Maximization [34] is employed in [30] to alternatively train key instance assignment module and foreground classification module. Yang *et al.* [58] train RGB and optical flow streams using pseudo-labels generated from each other, with an uncertainty-aware learning module to alleviate noises in pseudo-labels. Our approach also tackles pseudo-label noise issue, while it is based on clustering-based confidence voting and is used to generate more accurate contrastive features. Action-foreground consistency is explored in [13] with a hybrid attention to improve boundary accuracy. Lou *et al.* [29] propose an action unit memory bank to learn action units specific classifiers. The differences of our approach with existing methods are discussed in Sec. 3.4.

Contrastive Learning. As an important branch of deep metric learning [15], contrastive learning [5, 9, 11, 12, 53] have recently made impressive progress in unsupervised representation learning. These approaches learn representations in a discriminative manner by contrasting positive pairs against negative ones: two augmentations of the same image may be viewed as a positive pair, while two different images are considered as a negative pair. However, *false* negative samples are inevitably brought in [5] due to the lack of label information [6]. Prannay *et al.* [16] introduce supervised contrastive loss for image classification, showcasing the benefits of engaging label information in constructing positive and negative pairs. Moreover, several latest studies extend contrastive loss to a variety of downstream tasks, *e.g.,* semantic segmentation [51, 67] and object detection [47, 52, 54], and lead to new state-of-the-art performance.

3. Methodology

In this section, we first describe our baseline method in Sec. 3.1, then detail the proposed Denoised Cross-video Contrastive (DCC) algorithm in Sec. 3.2. This is followed by introducing the overall training objective and our inference process in Sec. 3.3. Finally, we discuss the differences with existing works in Sec. 3.4.

3.1. Baseline Setup

Fig. 2 (upper) presents the pipeline of our baseline algorithm. Given a training video sample $\{\boldsymbol{v}, \boldsymbol{y}\}$, where $\boldsymbol{y} \in \mathbb{R}^C$ stands for the action label of video $\boldsymbol{v}$ and C is the number of action categories, we sample a fixed number of T non-overlapping snippets, each with 16 frames, for each video and then extract snippet-wise features using the pre-trained feature extractor (*e.g.*, I3D [3])[1]. Next, we apply several layers of temporal convolution layers on the pre-trained features to introduce some temporal involvement between snippets and output the base Temporal Class Activation Sequence (T-CAS) $\mathcal{A}^b \in \mathbb{R}^{T\times(C+1)}$ using a classification head. Here we additionally predict a background class for each snippet to better model background. Following BaSNet [18], a parallel branch termed as *foreground selection* module is introduced to learn the class-agnostic foreground probability $\mathcal{Q} \in \mathbb{R}^{T\times 1}$, which can be regarded as temporal attention for actions. By multiplying $\mathcal{Q}$ with $\mathcal{A}^b$ temporally, we obtain the T-CAS $\mathcal{A}^f \in \mathbb{R}^{T\times(C+1)}$, which filters out non-action predictions. Following multiple instance learning [8], we apply a temporal top-k pooling followed by a softmax on both $\mathcal{A}^b$ and $\mathcal{A}^f$ to generate the video level prediction $\boldsymbol{p}^b, \boldsymbol{p}^f \in \mathbb{R}^{C+1}$, respectively.

By using snippet-wise binary cross entropy loss, we calculate the MIL loss as,

$$\begin{aligned}\mathcal{L}_{\text{MIL}} = -\sum_{c=1}^{C+1} (&\boldsymbol{y}_c^b \log \boldsymbol{p}_c^b + (1-\boldsymbol{y}_c^b)\log(1-\boldsymbol{p}_c^b) \\ &+ \boldsymbol{y}_c^f \log \boldsymbol{p}_c^f + (1-\boldsymbol{y}_c^f)\log(1-\boldsymbol{p}_c^f)),\end{aligned} \tag{1}$$

where $\boldsymbol{y}^b, \boldsymbol{y}^f$ are the corresponding labels of $\boldsymbol{p}^b, \boldsymbol{p}^f$ by introducing the background labels. Concretely, $\boldsymbol{y}_c^b = \boldsymbol{y}_c^f = \boldsymbol{y}_c$ for $1 \le c \le C$. $\boldsymbol{y}_{C+1}^b$ is set to 1 because that all training videos contain background snippets and $\boldsymbol{y}_{C+1}^f$ is set to 0 since background snippets are filtered in $\mathcal{A}^f$. To enforce the foreground scores to be more polarized, we also apply

[1] We use two modalities, *i.e.*, RGB and optical flow, as the input of feature extractor.

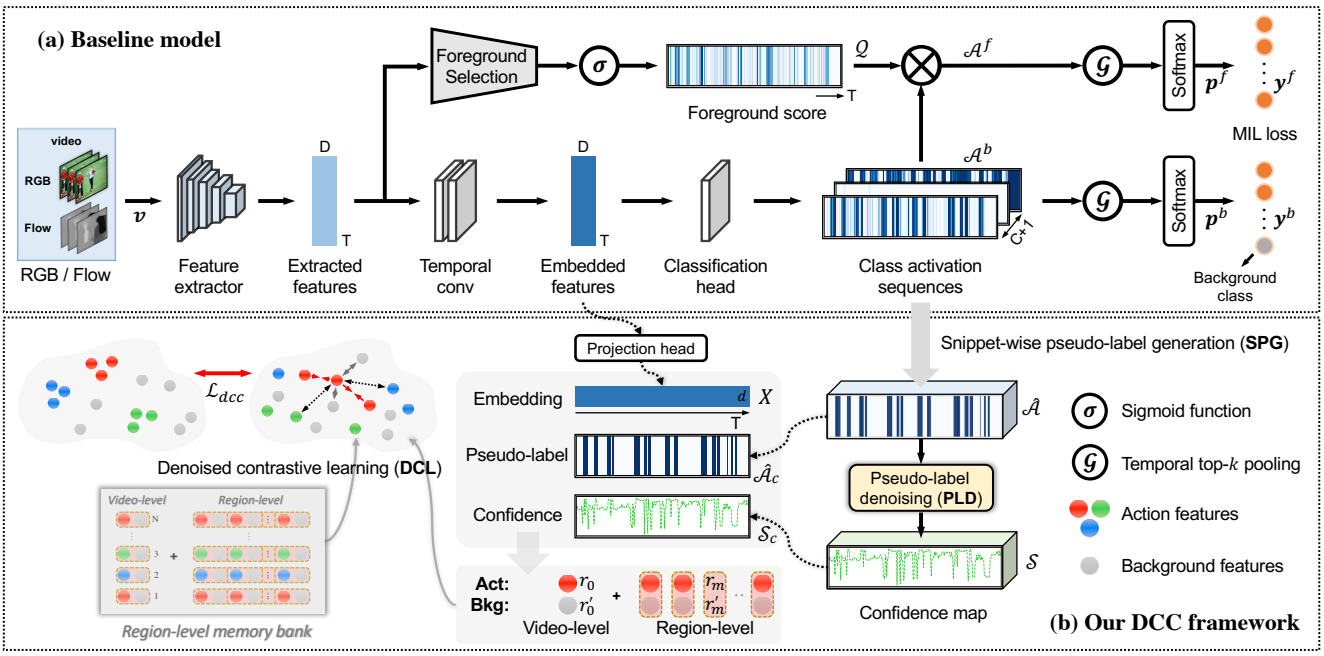


Figure 2. **The overall architecture** of our approach. The upper stream (a) presents the baseline model trained with conventional multiple instance learning loss with background modeling. We propose (b) denoised cross-video contrastive (DCC) algorithm in the bottom stream, aiming to shape the snippet feature embedding space and generate better Temporal Class Activation Sequence (T-CAS) for temporal action localization.

a L_1 normalization loss [38] on $\mathcal{Q}$, $\mathcal{L}_{\text{norm}} = \frac{1}{T}\sum_{t=1}^{T}|\mathcal{Q}_t|$. The final loss of this baseline method can be formulated as,

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{MIL}} + \gamma\mathcal{L}_{\text{norm}}, \tag{2}$$

where γ is a balance factor and is set to 1e-5 following [18].

3.2. Denoised Cross-video Contrastive Algorithm

An overview of our DCC is illustrated in Fig. 2 (bottom). Our pipeline includes three components which are Snippet-wise Pseudo-label Generation (SPG), Pseudo-label Denoising (PLD) and Denoised Contrastive Learning (DCL). SPG aims to estimate the snippet-wise label for action and background region extraction in videos and PLD is designed to emphasize confident video regions while suppress unreliable ones to alleviate noisy issue of snippet-wise label. DCL is in charge of constructing denoised contrastive features and generating positive and negative feature pairs for contrastive learning.

Snippet-wise Pseudo-label Generation. To determine the required action or background portions under the weakly-supervised setting, we opt to generate pseudo-label $\widehat{\mathcal{A}}$ by thresholding $\mathcal{A}^b$ as in [30]. A softmax function along the category dimension $\varepsilon(\cdot)$ is first applied on $\mathcal{A}^b$ to map the logits to probability scores. This process is formulated as

$$\widehat{\mathcal{A}}_{t,c} = \Phi(\varepsilon(\mathcal{A}^b)_{t,c}; \theta_c), \tag{3}$$

where θ_c is the threshold value for class c and is set to the mean value of $\varepsilon(\mathcal{A}^b)_c$ along temporal dimension; Φ is the thresholding operation where $\widehat{\mathcal{A}}_{t,c}$ is 1 if $\varepsilon(\mathcal{A}^b)_{t,c} \geq \theta_c$, and 0 otherwise.

Pseudo-label Denoising. To address the issue of noisy estimated snippet-wise pseudo-label $\widehat{\mathcal{A}}$, we design a Pseudo-label Denoising (PLD) module aiming to assign each video snippet a confidence score that estimates the probability of its pseudo-label being a trustworthy true label. Intuitively, video snippets within the same cluster are more likely to maintain the same category label; so the outliers, *i.e.*, video snippets whose pseudo-labels are inconsistent with the majority in each cluster, have a high probability of being misclassified and should be assigned with lower confidence scores.

Concretely, we cluster the embedding features using the basic K-*means* algorithm [20] with the number of cluster center set to K. After feature clustering, each snippet will be assigned to a cluster center, which is denoted by $\{E_t\}_{t=1}^{T}$ where $E_t \in [1, K]$. The confidence score of pseudo-label $\widehat{\mathcal{A}}_{t,c}$ can be calculated by a confidence voting strategy,

$$\mathcal{S}_{t,c} = \frac{\sum_{k=1}^{T}\mathbb{1}(E_t = E_k \ \wedge \ \widehat{\mathcal{A}}_{t,c} = \widehat{\mathcal{A}}_{k,c})}{\sum_{k=1}^{T}\mathbb{1}(E_t = E_k)}, \tag{4}$$

where $\mathbb{1}(condition)$ is the indicator function, *i.e.*, a function that returns 1 if the condition is satisfied, and 0 otherwise. $\wedge$ means the *and* operation. This strategy takes as confidence score the percentage of snippets in cluster center E_t that have the same pseudo-label as the t-th snippet.

Denoised Contrastive Learning. The estimated pseudo-labels and the confidence scores computed in PLD module are then engaged to generate contrastive features. To capture "global" contrast across the entire dataset, we propose a *region-level* feature contrast strategy which, together with a *region-level* memory bank, allow our learned model to preserve "global" informative features across the entire dataset. As illustrated in Fig. 2, following [5], we first append a projection head after the embedded features to get more compact representation, termed as $X \in \mathbb{R}^{T\times d}$, where d is the dimension of projected feature, for constrative learning. Then we compute the denoised action video feature F by multiplying the projected feature X by pseudo-label $\widehat{\mathcal{A}_c}$ and its corresponding confidence score in an element-wise manner:

$$F_{t,i} = \widehat{\mathcal{A}}_{t,c} \times \mathcal{S}_{t,c} \times X_{t,i}, \tag{5}$$

where c is the video label of encoded X. For background feature F', we alter the pseudo-label $\widehat{\mathcal{A}}_{t,c}$ with $1 - \widehat{\mathcal{A}}_{t,c}$ accordingly,

$$F'_{t,i} = (1 - \widehat{\mathcal{A}}_{t,c}) \times \mathcal{S}_{t,c} \times X_{t,i}. \tag{6}$$

Next, we evenly divide the denoised action video feature F into $\mathcal{M}$ *action region features* along temporal dimension, denoted as $F \Rightarrow \{R_m\}_{m=0}^{\mathcal{M}}$, where we set $R_0 = F$ by treating video feature as a relatively large region feature. Finally, we temporally average pooling these region features to obtain their corresponding vectors $\{r_m\}_{m=0}^{M}$ for contrastive learning. Similarly, the *background region features* $\{r'_m\}_{m=0}^{M}$ are also generated. At the same time, a *region-level* memory bank is introduced to store the *region features* of all training videos, which enables our model to learn "global" contrast from the entire dataset.

Given these denoised region-level features, we then apply a diverse contrastive training strategy to both enforce contrasts *between actions and backgrounds*, and *between different action classes*. The positive/negative sample pairs are constructed from two sources, *i.e.*, within video and cross video. In detail, given a denoised action region feature r_m, its positive sample set $\mathcal{P}_m$ includes: 1) action region features from the same video with the same class label; 2) action region features from other videos with the same class label. Its negative sample set $\mathcal{N}_m$ consists of: 1) background region features from the same video; 2) background region features from other videos; 3) action region features from other videos but with different class label. Equipped with the InfoNCE [12] loss, we can formulate the contrastive learning as,

$$\mathcal{L}_{\text{dcc}} = -\frac{1}{M}\sum_{m=0}^{M} \log \frac{\sum_{r_m^+ \in \mathcal{P}_m} \exp(r_m \cdot r_m^+/\tau)}{\sum_{r_m^\pm \in \mathcal{P}_m \cup \mathcal{N}_m} \exp(r_m \cdot r_m^\pm/\tau)}, \tag{7}$$

where τ is the temperature parameter. Note that all the embeddings in the loss function are l_2-normalized. With $\mathcal{L}_{\text{dcc}}$, the model is able to capture action-background separation, intra-class compactness, and inter-class separability.

3.3. Overall Training Objective and Inference

The overall training objective of our model is

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{base}} + \beta\mathcal{L}_{\text{dcc}}, \tag{8}$$

where β is a balancing factor. Since the contrastive features are less informative in the early training stage, we gradually increase β from 0.1 to 10000 during network training to focus more on the MIL loss at the early training stage and regularize the feature space learning at later stage. We note that the DCC algorithm is only applied during training and will be removed at inference time. Thus it does not introduce any extra computation at deployment stage.

In the inference stage, we first threshold on the video-level prediction $\boldsymbol{p}^f$ with threshold θ_v to determine the action categories to be localized. For each selected category, we threshold the T-CAS $\mathcal{A}_b$ with θ_l to obtain candidate action proposals. To enrich the proposal pool, multiple thresholds are applied and Non-Maximum Suppression (NMS) is used to remove duplicated proposals.

3.4. Discussion

Deep metric learning [15,26] and contrastive learning [5] are also explored in [33, 36, 37, 41, 64] for temporal action localization, and the differences are discussed as follows: (1) [64] designs a snippet-wise contrastive loss to refine the hard action or background snippet features. They only consider the action-background separation within singe videos. While in our DCC, a diverse contrastive learning strategy is proposed to simultaneously contrast action-background, and different classes. Moreover, our *region-level* feature contrast enables the model to learn "global" contrast across entire dataset. (2) [36, 41] exploit deep metric learning techniques to enforce action-background separation across video-level features in a mini-batch while our method captures the region-level contrast in the entire dataset and also learns inter-class separability. (3) [41, 64] fail to address the noisy contrastive feature issue, whereas in our method, a novel pseudo-label denoising module is designed to generate better contrastive features. (4) In [33, 37], contrasts are made between noisy attention-pooled video features and the class-specific central features, while in our DCC, contrasts are made between the denoised region-level features and abundant "global" features in a novel region-level memory bank.

4. Experiments

4.1. Datasets and Evaluation Metrics

Empirical analysis are carried out on two popular benchmark datasets including THUMOS14 [14] and ActivityNet

v1.3 [2]. ***THUMOS14*** includes untrimmed videos with 20 categories. The videos are densely annotated with frame-wise labels, in which their temporal lengths vary greatly. Note that we only use the video-level labels in WS-TAL. By convention [18, 41], we use 200 videos in the validation set for training and 213 videos in the test set for evaluation. ***ActivityNet v1.3*** [2] is a superset of ActivityNet v1.2, which consists of 10024 training videos, 4926 validation videos and 5044 testing videos belonging to 200 action categories. Since the annotations for the test set are not released, following the common practice [33, 63], we train our model on the training set and evaluate it on the validation set.

Following the standard protocol, the mAP (mean Average Precision) under different IoU (Intersection-over-Union) is used together with the benchmark code provided by ActivityNet[2] for evaluation.

4.2. Implementation Details

The network is implemented in PyTorch toolbox on a PC with a single Tesla P40 GPU. Optical flow frames are generated using TV-L1 algorithm [61]. Following [18, 64], the number of sampled snippets T is set to 750 and 50 for THUMOS14 and ActivityNet v1.3, respectively. For fair comparisons, the I3D [3] feature extractor is not fine-tuned. The foreground selection module contains two fully-connected layers with ReLU [35] activation. The projection head [5] is implemented in a similar way with the output dimension d set to 512. We use Adam optimizer [17] with learning rate 0.0001. τ is set to 0.1 following [5]. The cluster center K and region number $\mathcal{M}$ are both set to 5 for THUMOS14, and 2 for ActivityNet v1.3. The training takes 4 hours for THUMOS14 and 15 hours for ActivityNet v1.3. The GPU memory consumption for THUMOS14 is about 3.5GB. Empirically, to avoid model collapse where all snippets are classified to be background, we adopt a two-stage training mode, *i.e.*, the baseline network is first trained to generate pseudo-labels, with which we then optimize the whole network from scratch. θ_v is set to 0.2. θ_l spans from 0 to 0.9 with a step size of 0.025.

4.3. Ablation Studies

In this section, we provide detailed analysis on the effectiveness of our core model designs, using THUMOS14.

Effect of each component. Table 1 presents the comparison results of eliminating different modules of DCC. The DCC model without denoising improves the baseline performance greatly by 12.2% (from 37.7% to 42.3%) in terms of average mAP of IoU thresholds from 0.1 to 0.7, verifying the effectiveness of our method to improve feature discriminability. More detailed analyses and visualizations of cross-video contrastive algorithm are in the following subsections. When equipped with PLD module, our DCC further improves the action localization performance by 4%. The average confidence scores for both correct and incorrect labels are shown in Table 2. It is observed that, the correct pseudo-label achieves higher average confidence score than these incorrect ones, verifying the effectiveness of our PLD module to distinguish correct pseudo-labels from incorrect ones. In Table 3, we experiment with different number of cluster K using K-means. It is observed that under a wide range of K, the results are all better than the model without considering the noisy issue, which further demonstrates the usefulness and robustness of our proposed pseudo-label denoising module.

Action-background separation. To study the effectiveness of our model in capturing action-background separation, we conduct a comparison experiment with results presented in Table 4. It is observed that the action localization performance is significantly improved by an absolute value of 1.1% for average mAP@0.1:0.7, verifying the effectiveness of our model to learn action-background separation. Moreover, in Fig. 3, we visualize the embedded features of a video example from the THUMOS14 test set for baseline and our DCC, respectively. The embeddings are projected to 2-dimensional space using t-SNE tool [48] for visualization. As we can see, our method can better separate actions from backgrounds than the baseline model.

Intra-class compactness & inter-class separability. To investigate the importance of modeling intra-class compactness & inter-class separability, we further introduce the contrast between different action classes to enforce inter-class separability (3^{rd} row in Table 4). These results show that

Table 1. Ablation studies on THUMOS14 test set. "DCC w/o denoising" means only cross-video contrastive learning are adopted. "DCC" is our final model with denoised cross-video contrastive algorithm.

Ablation Models	mAP@IoU(%) 0.1	0.3	0.5	0.7	Avg (0.1:0.7)
Baseline	61.7	48.2	29.3	10.9	37.7
DCC w/o denoising	67.3	53.9	33.8	12.5	42.3
DCC (Ours)	**69.0**	**55.9**	**35.7**	**13.7**	**44.0**

Table 2. The average confidence scores of both correct and incorrect labels computed on THUMOS14 training set.

Average conf	Action	Background	All snippets
Correct label	0.662	0.849	0.799
Incorrect label	0.577	0.707	0.638
Diff (Δ)	**+0.085**	**+0.142**	**+0.161**

Table 3. Comparison results with different K on THUMOS14. We report the average mAP under IoU thresholds from 0.1 to 0.7. "w/o denoise" means without pseudo-label denoising module.

Cluster K	w/o denoise	3	5	10	15	50	100
mAP@Avg	42.3	43.3	**44.0**	43.8	43.9	43.5	43.2

[2]https://github.com/activitynet/ActivityNet/

Table 4. Ablation studies of different contrastive learning designs on THUMOS14. "Intra&Inter": intra-class compactness and inter-class separability. "Act-bkg": action-background separation.

Contrastive features		Contrastive strategies		mAP@IoU(%)	
Video-level	Region-level	Act-bkg	Intra & Inter	**0.5**	**Avg**
✗	✗	✗	✗	29.3	37.7
✓	✗	✓	✗	30.7	38.8
✓	✗	✓	✓	31.5	39.6
✓	✓	✓	✓	**33.8**	**42.3**

Table 5. Analysis on region number $\mathcal{M}$ on THUMOS14. We report the average mAP under IoU thresholds from 0.1 to 0.7. We also show the number of features in the memory bank, where N_v is the total number of training videos.

Region number $\mathcal{M}$	**1**	**3**	**5**	**10**
Number of features	$2N_v$	$8N_v$	$12N_v$	$22N_v$
mAP@Avg	39.6	41.2	**42.3**	41.7

Table 6. Analysis on contrastive features from different videos on THUMOS14.

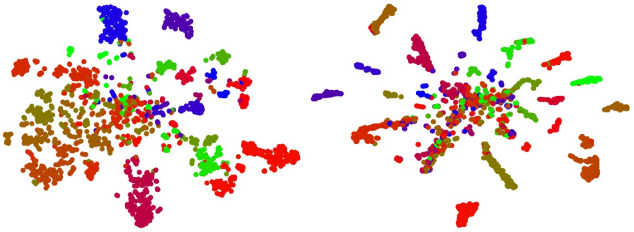
Ablation Models	**Avg (0.1:0.7)**
Baseline (w/o contrast)	37.7
Intra-video Contrast	38.5
Inter-video Contrast w/o memory (Mini-batch)	39.7
Inter-video Contrast w/ memory (Entire dataset)	42.3

the performance of the average mAP@0.1:0.7 is further improved by an absolute value of 0.8% thanks to modeling intra-class compactness & inter-class separability. We then visualize the learned feature distribution of various classes in Fig. 4, where the left part shows the feature space of the model trained with baseline MIL loss and the right part shows the feature space of our DCC model. It is observed that the snippet embeddings of our model are more compact and well separated, which can produce more discriminative features and improve action localization performance.

Different level of contrastive features. When using only video-level features (3^{rd} row in Table 4) for contrastive learning, the model gets 39.6% average mAP@0.1:0.7. With our *region-level* features (4^{th} row), we achieve significant performance gain (39.6% $\rightarrow$ 42.3% for average mAP), which strongly verifies the effectiveness of our region-level contrastive feature design. In addition, in Table 5, we evaluate the effect of different region number $\mathcal{M}$, which represents the granularity for segmenting videos. The larger the $\mathcal{M}$ value is, the finer the feature granularity is. Experimental results in Table 5 suggest that: (1) within a relatively coarse granularity, larger $\mathcal{M}$ usually leads to higher mAP score since more features are retained in the contrastive training process; (2) too fine-grained granularity ($M > 5$) does not further improve the performance. We conjecture that this is because too fine-grained features are prone to introduce noisy contrastive features and lead to suboptimal contrastive training.

(a) Baseline (b) Ours

Figure 3. T-SNE visualizations of action-background separation.



(a) Baseline (b) Ours

Figure 4. T-SNE visualization of intra-class compactness and inter-class separability.

Table 7. Generalization analysis on different backbones.

Ablation Models	**mAP@IoU(%) 0.1**	**0.3**	**0.5**	**0.7**	**Avg (0.1:0.7)**
STPN	57.0	42.8	24.7	10.0	33.5
STPN+DCC	**64.5**	**51.6**	**32.5**	**11.3**	**40.5**
BaSNet	61.7	48.2	29.3	10.9	37.7
BaSNet+DCC	**69.0**	**55.9**	**35.7**	**13.7**	**44.0**

Contrastive features from different videos. Table 6 presents the ablation experiments for verifying the contributions of various contrastive features within the same video, within mini-batch, and across entire dataset. It is observed that our "Inter-video Contrast" on *entire dataset* is shown to significantly boost the performance over "Intra-video Contrast" and "Inter-video Contrast" on *Mini-batch*. This demonstrates the superiority of our DCC in learning better snippet embeddings by exploiting "global" contrast across the entire dataset. Meanwhile, the mAP scores are gradually increased as more video features are engaged in contrastive training. This observation is consistent with many recent unsupervised contrastive learning works [5, 12, 53].

Generalization analysis. We verify the generalization ability of our DCC algorithm by applying it to two recent baseline models, STPN [38] and BaSNet [18]. Experimental results are presented in Table 7. After integrating with DCC, the performances of these two methods are significantly improved by 20.9% and 16.7% on the average mAP@0.1:0.7 scores. This verifies the good generalization ability of our approach on different backbones.

Table 8. Performance comparison on THUMOS14 testing set. The 'Avg' columns show average mAP under IoU thresholds of 0.1:0.5 and 0.1:0.7. † indicates access to newly-collected data or additional annotations. ∗ means using I3D features.

Supervision	Methods	Publication	mAP(%)@IoU 0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg (0.1:0.5)	Avg (0.1:0.7)
Full	S-CNN [45]	CVPR'16	47.7	43.5	36.3	28.7	19.0	10.3	5.3	35.0	27.3
	SSN [68]	ICCV'17	66.0	59.4	51.9	41.0	29.8	-	-	49.6	-
	TAL-Net [4]	CVPR'18	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	45.1
	BSN [24]	ECCV'18	-	-	53.5	45.0	36.9	28.4	20.0	-	-
	GTAN [28]	CVPR'19	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-
Weak†	BM* [39]	ICCV'19	64.2	59.5	49.1	38.4	27.5	17.3	8.6	29.8	37.8
	3C-Net* [37]	ICCV'19	59.1	53.5	44.2	34.1	26.6	-	8.1	43.5	-
	STAR* [57]	AAAI'19	68.8	60.0	48.7	34.7	23.0	-	-	47.0	-
	SF-Net* [31]	ECCV'20	71.0	63.4	53.2	40.7	29.3	18.4	9.6	51.5	40.8
Weak	UntrimNet [50]	CVPR'17	44.4	37.7	28.2	21.1	13.7	-	-	29.0	-
	STPN* [38]	CVPR'18	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	27.0
	W-TALC* [41]	ECCV'18	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-
	AutoLoc [44]	ECCV'18	-	-	35.8	29.0	21.2	13.4	5.8	-	-
	CleanNet [27]	ICCV'19	-	-	37.0	30.9	23.9	13.9	7.1	-	-
	Liu *et al.** [25]	CVPR'19	57.4	50.8	41.2	32.1	23.1	15.0	7.0	40.9	32.4
	BaSNet* [18]	AAAI'20	58.2	52.3	44.6	36.0	27.0	18.6	10.4	43.6	35.3
	DGAM* [42]	CVPR'20	60.0	56.0	46.6	37.5	26.8	17.6	9.0	45.6	37.0
	EMMIL* [30]	ECCV'20	59.1	52.7	45.5	36.8	30.5	22.7	**16.4**	45.0	37.7
	TSCN* [63]	ECCV'20	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	37.8
	A2CL-PT* [33]	ECCV'20	61.2	56.1	48.1	39.0	30.1	19.2	10.6	46.9	37.8
	UM* [19]	AAAI'21	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	41.9
	CoLA* [64]	CVPR'21	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	40.9
	AUMN* [29]	CVPR'21	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	41.5
	FAC-Net* [13]	ICCV'21	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	42.2
	D2Net* [36]	ICCV'21	65.7	60.2	52.3	43.4	**36.0**	-	-	51.5	-
	DCC (**Ours**)*	-	**69.0**	**63.8**	**55.9**	**45.9**	35.7	**24.3**	13.7	**54.1**	**44.0**

Table 9. Performance comparison on ActivityNet v1.3 dataset. The average mAP is computed on thresholds 0.5:0.05:0.95.

Supervision	Methods	mAP(%)@IoU 0.5	0.75	0.95	Avg
Weak	STPN [38]	29.3	16.9	2.6	16.3
	CMCS [25]	34.0	20.9	**5.7**	21.2
	BM [39]	36.4	19.2	2.9	19.5
	TSM [59]	30.3	19.0	4.5	-
	BaSNet [18]	34.5	22.5	4.9	22.2
	TSCN [63]	35.3	21.4	5.3	21.7
	A2CL-PT [33]	36.8	22.0	5.2	22.5
	AUMN [29]	38.3	23.5	5.2	23.5
	DCC (**Ours**)	**38.8**	**24.2**	**5.7**	**24.3**

4.4. Comparison with State-of-the-Arts

We compare our method with state-of-the-art approaches under different level of supervisions on THUMOS14 test set in Table 8. Note that "Full" means training using frame-wise annotations; "Weak†" indicates using newly collected data [39] or additional annotations [31, 37, 57]. Our method outperforms recently proposed weakly-supervised methods, *e.g.* UM [19] and FAC- Net [13], with a large margin. The average mAP of IoU thresholds from 0.1 to 0.7 even reaches 44.0%, bringing the state-of-the-art to a new level. Our method also outperforms weak† approaches at almost all IoU thresholds, and obtains competitive results even compared with fully-supervised methods, which substantially closes the gap between weakly-supervised TAL and fully-supervised one. Evaluation on ActivityNet v1.3 benchmark is displayed in Table 9. We report the mAP score at various IoU thresholds and report the average mAP for IoU thresholds from 0.5 to 0.95 with a step size of 0.05. As can be seen, our method performs favourably compared with state-of-the-art approaches.

5. Conclusion

In this paper, we propose a novel denoised cross-video contrastive algorithm tailored for weakly-supervised temporal action localization. Our key insight is to enhance the feature discrimination ability by three critical ingredients, namely pseudo-label denoising module in addressing noisy contrastive features, region-level feature contrast strategy and region-level memory bank to capture "global" cross-video contrast, and a diverse contrastive learning strategy to regularize the snippet embedding representation. Extensive experiments on two benchmarks demonstrate the superior performance of our approach.

Acknowledgement. This research was supported by CCF-Tencent Open Fund, the University of Alberta Start-up Grant, UAHJIC Grants, and NSERC Discovery Grants (No. RGPIN-2019-04575).

References

[1] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Procedings of the British Machine Vision Conference (BMVC)*, 2019. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 3, 6

[4] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1130–1139, 2018. 1, 2, 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 1, 3, 5, 6, 7

[6] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 3

[7] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5793–5802, 2017. 1, 2

[8] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997. 3

[9] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in Neural Information Processing Systems*, 27:766–774, 2014. 3

[10] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017. 2

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006. 3

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 3, 5, 7

[13] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8002–8011, 2021. 3, 8

[14] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. 5

[15] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019. 1, 3, 5

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 34(7):11320–11327, 2020. 1, 3, 4, 6, 7, 8

[19] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. *arXiv preprint arXiv:2006.07006*, 2020. 3, 8

[20] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. 4

[21] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11499–11506, 2020. 2

[22] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3889–3898, 2019. 2

[23] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM International Conference on Multimedia (ACM MM)*, pages 988–996, 2017. 2

[24] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2, 8

[25] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1298–1307, 2019. 1, 3, 8

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 507–516, 2016. 1, 5

[27] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908, 2019. 8

[28] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 344–353, 2019. 1, 2, 8

[29] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9969–9979, 2021. 3, 8

[30] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 4, 8

[31] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8

[32] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005. 1

[33] Kyle Min and Jason J. Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5, 6, 8

[34] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996. 3

[35] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010. 6

[36] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 13608–13617, 2021. 1, 2, 5, 8

[37] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8679–8687, 2019. 3, 5, 8

[38] Phuc Nguyen, Bohyung Han, Ting Liu, and Gautam Prasad. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6752–6761, 2018. 1, 2, 4, 7, 8

[39] Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5502–5511, 2019. 1, 3, 8

[40] Alejandro Pardo, Humam Alwassel, Fabian Caba, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3319–3328, 2021. 3

[41] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 588–607, 2018. 1, 2, 3, 5, 6, 8

[42] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1009–1019, 2020. 3, 8

[43] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5734–5743, 2017. 1, 2

[44] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 162–179, 2018. 8

[45] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1058, 2016. 2, 8

[46] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553, 2017. 2

[47] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2291–2301, 2021. 3

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[49] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013. 1

[50] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6402–6411, 2017. 1, 2, 8

[51] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 3

[52] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *arXiv preprint arXiv:2106.02637*, 2021. 3

[53] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018. 3, 7

[54] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8392–8401, 2021. 3

[55] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1267, 2019. 1

[56] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5783–5792, 2017. 1, 2

[57] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 9070–9078, 2019. 8

[58] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 53–63, 2021. 3

[59] Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. Temporal structure mining for weakly supervised action detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5522–5531, 2019. 8

[60] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. *arXiv preprint arXiv:1905.08586*, 2019. 1

[61] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223, 2007. 6

[62] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7094–7103, 2019. 2

[63] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 6, 8

[64] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5, 6, 8

[65] Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. S3d: single shot multi-span detector via fully 3d convolutional networks. *arXiv preprint arXiv:1807.08069*, 2018. 2

[66] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 539–555, 2020. 1, 2

[67] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10623–10633, 2021. 3

[68] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017. 1, 2, 8

[69] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *Proceedings of the 26th ACM International Conference on Multimedia (ACM MM)*, pages 35–44, 2018. 2