# Fine-grained Temporal Contrastive Learning for Weakly-supervised Temporal Action Localization

Junyu Gao[1,2], Mengyuan Chen[1,2], and Changsheng Xu[1,2,3]
[1] National Lab of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Peng Cheng Laboratory, ShenZhen, China
{junyu.gao, csxu}@nlpr.ia.ac.cn; chenmengyuan2021@ia.ac.cn

## Abstract

*We target at the task of weakly-supervised action localization (WSAL), where only video-level action labels are available during model training. Despite the recent progress, existing methods mainly embrace a localization-by-classification paradigm and overlook the fruitful fine-grained temporal distinctions between video sequences, thus suffering from severe ambiguity in classification learning and classification-to-localization adaption. This paper argues that learning by contextually comparing sequence-to-sequence distinctions offers an essential inductive bias in WSAL and helps identify coherent action instances. Specifically, under a differentiable dynamic programming formulation, two complementary contrastive objectives are designed, including Fine-grained Sequence Distance (FSD) contrasting and Longest Common Subsequence (LCS) contrasting, where the first one considers the relations of various action/background proposals by using match, insert, and delete operators and the second one mines the longest common subsequences between two videos. Both contrasting modules can enhance each other and jointly enjoy the merits of discriminative action-background separation and alleviated task gap between classification and localization. Extensive experiments show that our method achieves state-of-the-art performance on two popular benchmarks. Our code is available at* https://github.com/MengyuanChen21/CVPR2022-FTCL.

## 1. Introduction

Action localization is one of the most fundamental tasks in computer vision, which aims to localize the start and end timestamps of different actions in an untrimmed video [41, 63, 67, 75]. In the past few years, the performance has gone through a phenomenal surge under the fully-supervised setting. However, collecting and annotating precise frame-wise information is a bottleneck and consequently limits the scalability of a fully supervised framework for real-world scenarios. Therefore, weakly-supervised action localization (WSAL) has been explored [26, 27, 56, 69], where only video-level category labels are available.

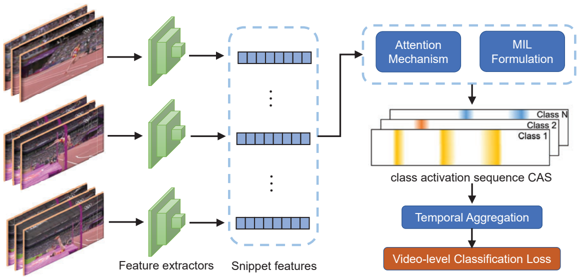

Figure 1. Pipeline of the localization-by-classification paradigm. It first extracts snippet-level features and adopts attention/MIL mechanisms for learning CAS under video-level supervisions.

To date in the literature, current approaches mainly embrace a localization-by-classification paradigm [54, 57, 65, 68], which divides each input video into a series of fixed-size non-overlapping snippets and aims for generating the temporal Class Activation Sequences (CAS) [56, 71]. Specifically, as shown in Figure 1, by optimizing a video-level classification loss, most existing WSAL approaches adopt the multiple instance learning (MIL) formulation [45] and attention mechanism [56] to train models to assign snippets with different class activations. The final action localization results are inferred by thresholding and merging these activations. To improve the accuracy of learned CAS, various strategies have been proposed, such as uncertainty modeling [69], collaborative learning [26, 27], action unit memory [42], and causal analysis [37], which have obtained promising performance.

Despite achieving significant progress, the above learning pipelines still suffer from severe localization ambi-

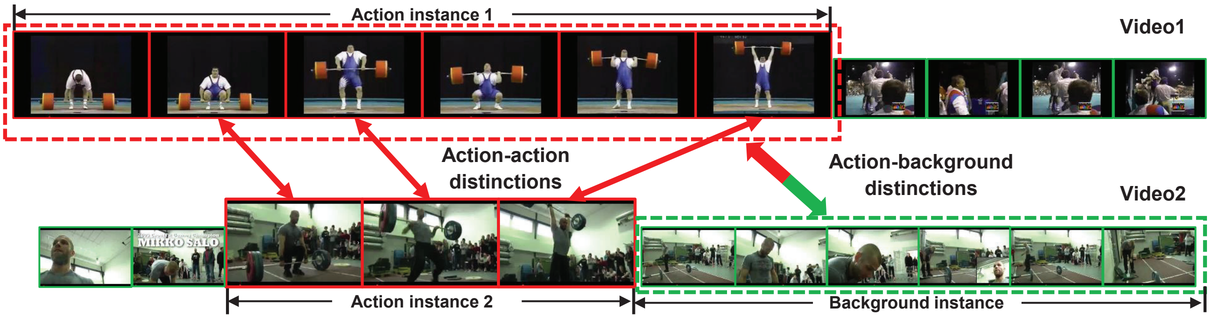

Figure 2. Fine-grained temporal distinctions between two videos. Here, the two untrimmed videos are from the same action category *CleanAndJerk*. Note that the distinctions are derived from two aspects: (1) Fine-grained action-background distinctions. The snippets in the action instances and background subsequences are semantically different, which should be effectively separated in a robust WSAL model. (2) Fine-grained distinctions between action instances. In this example, the three snippets of the action instance in Video2 can be aligned with the partial action instance in Video1. In addition, we can observe that the three snippets linked by the red arrows are the longest common sequences of both videos. We argue that considering the above fine-grained distinctions can benefit WSAL learning.

guity due to the lack of fine-grained frame-wise annotations in the temporal dimension, which dramatically hinders the WSAL performance of the localization-by-classification paradigm. Specifically, the ambiguity is two-fold: **(1)** Without sufficient annotations in the weakly-supervised setting, the learned classifier itself is not discriminative and robust enough, causing difficulties in action-background separation. **(2)** Since there exists a large task gap between classification and localization, the learned classifiers usually focus on the easy-to-distinguish snippets while ignoring those that are not prominent in localization. As a result, the localized temporal sequences are often incomplete and inexact.

To alleviate the above ambiguity, we argue that videos naturally provide a rich source of temporal structures and additional constraints for improving weakly-supervised learning. As outlined in Figure 2, an action video generally includes a series of fine-grained snippets, while different action/background instances possess correlative and fine-grained temporal distinctions. For example, given a pair of videos from the same action category but captured in varied scenes, there exists a latent temporal association between both videos. With this in mind, a key consideration is to leverage such temporal distinctions for improving representation learning in WSAL. However, when elaborately comparing two videos, no guarantee ensures that they can be aligned directly. Recently, dynamic time warping (DTW) [2, 55] was proposed to tackle the misalignment issue in various video analysis tasks such as action classification [25], few-shot learning [7], action segmentation and video summarization [9, 10]. DTW computes the discrepancy between two videos based on their optimal alignment from dynamic programming. However, the above approaches either assume the video is trimmed [7, 25] or require additional supervision [9, 10] such as action orders, which impedes the direct use of DTW in WSAL.

In this paper, to address the above issues, we propose a novel Fine-grained Temporal Contrastive Learning (FTCL) framework for weakly-supervised temporal action localization. By capturing the distinctive temporal dynamics of different video sequences, FTCL focuses on optimizing the structural and fine-grained snippet-wise relations between videos by leveraging end-to-end differentiable dynamic programming goals, with loss that is informed from the structural relations. Specifically, **(1)** To improve the robustness of action-background separation, we contrast the fine-grained sequence distance (FSD) calculated from different action/background instance pairs by designing an improved and differentiable edit distance measurement. The measurement can evaluate whether two sequences are structurally analogous by calculating the minimum cost required to transform one to the other. **(2)** To alleviate the task gap between classification and localization, we aim at contrasting the mined Longest Common Subsequence (LCS) between two untrimmed videos that contain the same action. Different video sequences from the same category can provide complementary clues for exploring the complete action instance by optimizing the LCS. Therefore, LCS learning between different video sequences improves the coherence in a predicted action instance. Finally, with FSD and LCS contrasting, a unified framework is constructed in an end-to-end manner, while the proposed FTCL strategy can be seamlessly integrated into any existing WSAL approach.

The main contributions of this paper are three-fold:

- In light of the above analysis, we contend that localizing action by contextually contrasting fine-grained temporal distinctions offers an essential inductive bias in WSAL. We thus introduce the first discriminative sequence-to-sequence comparing framework for robust WSAL to address the lack of frame-wise annotations, capable of leveraging fine-grained temporal distinctions.
- A unified and differentiable dynamic programming formulation, including fine-grained sequence distance

learning and longest common subsequence mining, is designed, which jointly enjoys the merits of (1) discriminative action-background separation and (2) alleviated task gap between classification and localization.

- Extensive experimental results on two popular benchmarks demonstrate that the proposed FTCL algorithm performs favorably. Note that the proposed strategy is model-agnostic and non-intrusive, and hence can play a complementary role over existing methods to promote the action localization performance consistently.

## 2. Related Work

**Fully-supervised Temporal Action Localization (TAL).** Compared with traditional video understanding tasks [8,17, 19,20,23], TAL aims to classify every activity instance in an untrimmed video and predict their accurate temporal locations. Existing TAL approaches can be roughly divided into two categories: two-stage methods [11,13,61,63,66,73,75] and one-stage methods [4,34,35,41,58,64,67]. For the former one, action proposals are firstly generated and then fed into a classifier. This pipeline mainly focuses on improving the quality of proposals [11, 61, 75] and the robustness of classifiers [63, 73]. One-stage methods instead predict action location and category simultaneously. SS-TAD [4] utilizes recurrent neural networks to regress the temporal boundaries and action labels jointly. Lin *et al.* [34] introduces an anchor-free framework in a coarse-to-fine manner. Although the above model achieves significant performance, the fully-supervised setting limits their scalability and practicability in the real world [18, 21, 22].

**Weakly-supervised Action Localization.** To overcome the above limitation, WSAL has drawn significant attention in recent years by leveraging different types of supervisions, *e.g.*, web videos [16], action orders [3], single-frame annotation [31, 44], and video-level category labels [36, 52, 65]. Among these weak supervisions, the last one is the most commonly used due to the low cost. UntrimmedNet [65] is the first work that uses video-level category labels for WSAL via a relevant segment selection module. Currently, most existing approaches can be roughly divided into three groups, namely attention-based methods [26, 26, 39, 42, 49, 56, 57, 68], MIL-based methods [32, 43, 45, 48, 54], and erasing-based methods [62, 72, 74]. Attention-based approaches aim at selecting snippets of high activation scores and suppressing background snippets. ACM-Net [56] investigates a three-branch attention module by simultaneously and effectively considering action instances, context, and background information. MIL-based pipeline treats the entire video as a bag and utilizes a top-$k$ operation to select positive instances. W-TALC [54] introduces a co-activity relation loss to model inter- and intra-class information. The erasing-based methods, *e.g.*, Hide-and-Seek [62], typically attempt to erase input segments during training for highlighting less discriminative snippets.

Note that most existing methods only consider the video-level supervision but ignore the fine-grained temporal distinctions between videos, and can hardly benefit from discriminative learning of snippet-wise contrasting. Although some approaches have investigated different types of contrastive regularization, *e.g.*, hard snippet contrasting in CoLA [71], they perform contrasting by only considering video-level information [30, 50, 54] or neglecting the fine-grained temporal structures [49, 53, 71]. To the best of our knowledge, we are the first to introduce the contrastive learning of fine-grained temporal distinctions to the WSAL task. Experimental results demonstrate that the proposed FTCL learns discriminative representations, thus facilitating the action localization.

**Dynamic Programming for Video Understanding.** Recent progress has shown that learning continuous relaxation of discrete operations (*e.g.*, dynamic programming) can benefit video representation learning [7,9,10,25]. A popular framework is to adopt sequence alignment as a proxy task and then uses dynamic time warping (DTW) to find the optimal alignment [2,6,12,14,15,46,55]. For example, based on a novel probabilistic path finding view, Hadji *et al.* [25] design contrastive and cycle-consistency objectives for video representation learning by leveraging differentiable DTW. Chang *et al.* [10] propose discriminative prototype DTW to learn class-specific prototypes for temporal action recognition. However, the above dynamic programming strategies either assume the video is trimmed [7, 25] or require additional supervision [9, 10] such as action orders, thus cannot be applied to the WSAL task. Different from the above approaches, this paper proposes to leverage fine-grained sequence distance and longest common subsequence contrasting for discriminative foreground-background separation and robust classification-to-localization adaption.

## 3. Our Approach

In this work, we describe our WSAL approach based on Fine-grained Temporal Contrastive Learning (FTCL). As shown in Figure 3, given a set of video sequence pairs, our training objective is the learning of an embedding function applied to each snippet. We firstly adopt feature extractors to obtain the appearance (RGB) and motion (optical flow) features of each snippet (Section 3.1). Then, under a differentiable dynamic programming formulation, two complementary contrastive objectives are designed for learning fine-grained temporal distinctions including Fine-grained Sequence Distance (FSD) contrasting (Section 3.2) and Longest Common Subsequence (LCS) contrasting (Section 3.3). Finally, the whole framework is end-to-end learned (Section 3.4), which can jointly achieve discriminative action-background separation and alleviated task gap between classification and localization.

### 3.1. Notations and Preliminaries

Given an untrimmed video $\mathbf{X}$ with its groundtruth label $\mathbf{y} \in \mathbb{R}^C$, where $C$ is the number of action categories. $\mathbf{y}_i = 1$ if the $i$-th action class is present in the video and $\mathbf{y}_i = 0$ otherwise. For the video, we divide it into non-overlapping $T$ snippets and apply feature extractors to obtain snippet-wise features $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_T] \in \mathbb{R}^{D \times T}$, where $D$ is the feature dimension and each snippet has 16 frames. In this paper, for a fair comparison, we follow previous approaches [50, 54, 56, 71] to extract features from both RGB and optical flow streams by using the I3D network [8] pre-trained on the Kinetics dataset. After that, the two types of features are concatenated together and then input into an embedding module, *e.g.*, convolutional layers [56], for generating $\mathbf{X}$. The goal of WSAL is to learn a model that simultaneously localizes and classifies all action instances in a video with timestamps as $(t_s, t_e, c, \phi)$, where $t_s, t_e, c$, and $\phi$ denote the start time, the end time, the predicted action category and the confidence score of the action proposal, respectively.

Currently, existing dominant approaches mainly embrace a localization-by-classification framework, which first learns importance scores for aggregating snippet-level features into a video-level embedding and then perform action classification by using the video-level labels:

$$\begin{aligned} \overline{\mathbf{x}} &= \sum_{t=1}^{T} \alpha_t * \mathbf{x}_t \\ \mathcal{L}_{cls} &= -\sum_{i=1}^{C} \mathbf{y}_i \log \widetilde{\mathbf{y}}_i \end{aligned} \tag{1}$$

where $\alpha_t = f_\alpha(\mathbf{x}_t)$ is the learned importance score. The generated video-level feature is further fed into a classifer to obtain the prediction results $\widetilde{\mathbf{y}} = f_{cls}(\overline{\mathbf{x}})$. After model training, $f_\alpha(\cdot)$ and $f_{cls}(\cdot)$ is used for inferring the snippet-level Class Activation Sequences (CAS) of a test video. To learn the two functions, various strategies can be applied such as multiple attention learning [56] and modality collaborative learning [26].

### 3.2. Discriminative Aciton-Background Separation via FSD Contrasting

To learn discriminative action-background separation in the above localization-by-classification framework, a few existing methods resort to performing contrastive learning by either using global video features [30, 50, 54] or only considering intra-video contrast without temporal modeling [49, 53, 71]. However, these models ignore the fine-grained temporal distinctions between videos, resulting in the insufficient discriminative ability for classification.

In this work, we propose to contrast two video sequences temporally in a fine-grained manner. Existing methods usually calculate the similarity of two sequences by measuring the vector distance between their global feature representations. Different from this matching strategy, as shown in the left of Figure 3, we would like to determine whether two sequences are structurally analogous by evaluating the minimum cost required to transform one sequence to the other. The naive idea is to exhaustively compare all the possible transformations, which is NP-hard. A fast solution is to utilize solvable dynamic programming techniques, where sub-problems can be nested recursively inside larger problems. Here, motivated by the widely used edit distance[1] [51] in computational linguistics and computer science, we design differentiable *Match*, *Insert*, and *Delete* operators for sequence-to-sequence similarity calculation. Specifically, with the learned CAS, we can generate various action/background proposals, where an action proposal $\mathbf{U}$ contains snippets with high action activations and a background proposal $\mathbf{V}$ is just the opposite. For the two proposal sequences with lengths of $M$ and $N$, $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_i, ..., \mathbf{u}_M] \in \mathbb{R}^{D \times M}$ and $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_i, ..., \mathbf{v}_M] \in \mathbb{R}^{D \times N}$, their similarity is evaluated with the following recursion:

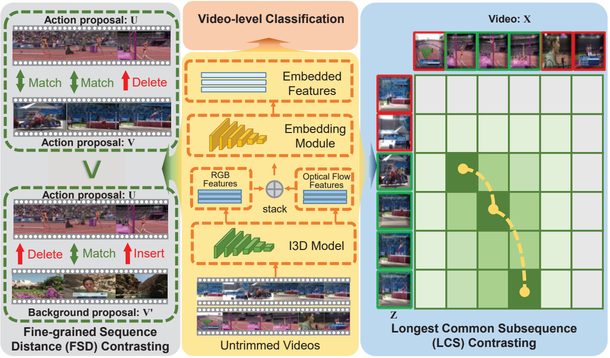

Figure 3. Our proposed FTCL architecture and toy example. The pretrained I3D model is first adopted for the input videos to obtain RGB and optical flow features. Then an embedding module is further applied to extract snippet-wise features under video-level supervisions. To achieve discriminative action-background separation, FSD contrasting is designed to consider the relations of different action/background proposals using Match, Insert, and Delete operators. For classification-to-localization adaption, we employ LCS contrasting to find the longest common subsequences between two videos. Both contrasting strategies are implemented via differentiable dynamic programming.

$$\mathbf{S}(i,j) = \mu_{i,j} + \max \begin{cases} \mathbf{S}(i-1, j-1) & (\text{ Match }) \\ g_{i,j} + \mathbf{S}(i-1, j) & (\text{ Insert }) \\ h_{i,j} + \mathbf{S}(i, j-1) & (\text{ Delete }) \end{cases} \tag{2}$$

[1] Edit distance is a way of quantifying how dissimilar two strings are to one another by counting the minimum number of operations required to transform one string into the other.

where the sub-sequence similarity score $\mathbf{S}(i,j)$ is evaluated on position $i$ in the first sequence $\mathbf{U}$ and on position $j$ in the second sequence $\mathbf{V}$. $\mathbf{S}(0,:)$ and $\mathbf{S}(:,0)$ are initialized to zeros. Intuitively, in position $(i,j)$, if $\mathbf{u}_i$ and $\mathbf{v}_j$ are matched, the sequence similarity score should be increased. If the insert or delete operation is conducted, there should be a penalty on the similarity score. To this end, we learn three types of residual values (scalars), $\mu_{i,j}$, $g_{i,j}$, and $h_{i,j}$ for these operations. Taking $\mu_{i,j}$ and $g_{i,j}$ as an example, which can be calculated as follows:

$$\mu_{i,j} = \sigma_\mu(\cos(\mathbf{\Delta}^\mu_{i,j})), \quad g_{i,j} = \sigma_g(\cos(\mathbf{\Delta}^g_{i,j})) \tag{3}$$

where $\mathbf{\Delta}^\mu_{i,j} = [f_\mu(\mathbf{u}_i), f_\mu(\mathbf{v}_j)]$ and $\mathbf{\Delta}^g_{i,j}$ is defined similarly. $f_\mu(\cdot)$, $f_g(\cdot)$, and $f_h(\cdot)$ are three fully-connected layers. We utilize these functions to simulate different operations including match, insert, and delete. $\sigma_\mu$ and $\sigma_g$ are activation functions for obtaining the residual values.

After conducting the above recursive calculation, $\mathbf{S}(i,j)$ is guaranteed to be the optimal similarity score between the two sequences. It is evident that the similarity between two action proposals from the same category should be larger than it between an action proposal and a background proposal. By leveraging this relation, we design the FSD contrasting loss as follows:

$$\mathcal{L}_{\text{FSD}} = \ell\left(s_{[\mathbf{UV}']} - s_{[\mathbf{UV}]}\right) + \ell\left(s_{[\mathbf{U}'\mathbf{V}]} - s_{[\mathbf{UV}]}\right) \tag{4}$$

where $\ell(x)$ denotes the ranking loss. The subscript $[\mathbf{UV}]$ indicates the two action proposals from the same category for calculating the sequence-to-sequence similarity $s = \mathbf{S}(M,N)$. $\mathbf{U}'$ and $\mathbf{V}'$ represents the background proposals. In our implementation, we utilize the learned importance score $\alpha$ [56] to select action and background proposals.

**Smooth Max Operation.** As the max operation in Eq. (2) is not differentiable, the recursive matrices and the traceback cannot be differentiated in current formulation. Therefore, we are motivated to utilize a standard smooth approximation for the max operator [46]:

$$\text{smoothMax}(\mathbf{a};\gamma) = \log(\sum_i \exp(\gamma \mathbf{a}_i)) \tag{5}$$

where $\mathbf{a} = [\mathbf{a}_1, ..., \mathbf{a}_i, ...]$ is a vector for max operator. $\gamma$ represents the temperature hyper-parameter. Note that other types of smooth approximation [6, 12, 25] can also be applied for differentiating while designing a novel smooth max operation is not the goal of our paper.

### 3.3. Robust Classification-to-Localization Adaption via LCS Contrasting

In the above section, action-background separation is considered, which improves the discriminative ability of the learned action classifiers. However, the goal of WSAL task is to localize action instances temporally with precise timestamps, resulting in a large task gap between classification and localization. To alleviate this gap, we attempt to mine the longest common subsequence (LCS) between two untrimmed videos $\mathbf{X}$ and $\mathbf{Z}$ thus improve the coherence in the learned action proposals. The intuition behind this idea is two-fold: (1) If the two videos do not share the same actions, the length of LCS between $\mathbf{X}$ and $\mathbf{Z}$ should be small. Obviously, due to the diverse background and substantial difference between the two types of actions, snippets from the two individual videos are likely to be highly inconsistent, resulting in short LCS. (2) Similarly, if two videos share the same action, their LCS is prone to be long since action instances from the same category are composed of similar temporal action snippets. Ideally, the LCS in this situation is as long as the shorter action instance. For example, as shown in Figure 2, the action *CleanAndJerk* consists of several sequential sub-actions like *squat*, *grasp*, and *lift*.

Based on the above observation, as shown in the right of Figure 3, we propose to model the LCS between $\mathbf{X}$ and $\mathbf{Z}$ by designing a differentiable dynamic programming strategy. Specifically, we maintain a recursive matrix $\mathbf{R} \in \mathbb{R}^{(T+1)\times(T+1)}$, with elements $\mathbf{R}(i,j)$ stores the length of longest common subsequence of prefixes $\mathbf{X}_i$ and $\mathbf{Z}_j$. To find the LCS of prefixes $\mathbf{X}_i$ and $\mathbf{Z}_j$, we first compare $\mathbf{x}_i$ and $\mathbf{z}_j$. If they are equal, then the calculated common subsequence is extended by that element and thus $\mathbf{R}(i,j) = \mathbf{R}(i-1,j-1)+1$. If they are not equal, the largest length calculated before is retained for $\mathbf{R}(i,j)$. In the WSAL task, since a pair of snippets cannot be exactly the same even they depict the same action, we adopt their similarities to calculate the accumulated soft length of two sequences. As a result, we design the recursion formula of LCS modeling:

$$\mathbf{R}(i,j) = \begin{cases} 0, \quad i=0 \text{ or } j=0 \\ \mathbf{R}(i-1,j-1)+c_{i,j}, \quad c_{i,j} \geqslant \tau \\ \max\{\mathbf{R}(i-1,j), \mathbf{R}(i,j-1)\}, c_{i,j} < \tau \end{cases} \tag{6}$$

where $\tau$ is a threshold that determines whether the $i$-th snippet of video $\mathbf{X}$ and the $j$-th snippet of video $\mathbf{Z}$ is matched. $c_{i,j} = \cos(\mathbf{x}_i, \mathbf{z}_j)$ is the cosine similarity of snippets $\mathbf{x}_i$ and $\mathbf{z}_j$. Note that by using the equation above, we can seek the longest common subsequence between two videos. Although not used here, the mined subsequence can qualitatively demonstrate the effectiveness and improve the interpretability of our approach (Section 4.3).

With the above dynamic programming, the resulting values $r = \mathbf{R}(T,T)$ represents the soft length of the longest common subsequence between the two videos. We utilize a cross-entropy loss to serve as a constraint for LCS learning:

$$\mathcal{L}_{\text{LCS}} = \delta_{xz}\log(r_{[\mathbf{XZ}]}) + (1-\delta_{xz})\log(1-r_{[\mathbf{XZ}]}) \tag{7}$$

where $\delta_{xz}$ is the groundtruth indicating whether the two videos $\mathbf{X}$ and $\mathbf{Z}$ have the same action categories.

**Discussion.** In this work, FSD and LCS learning strategies are proposed via differentiable dynamic programming, while both are designed for sequence-to-sequence contrasting. However, the two modules are not redundant and have substantial difference: (1) They have different goals by considering different types of sequences. We utilize FSD to learn robust action-background separation while different action and background proposals are employed. While LCS contrasting is designed to find coherent action instances in two untrimmed videos, thus achieving classification-to-localization adaption. (2) They have different contrasting levels. In FSD contrasting, the relations between different action/background pairs are considered (Eq. (4)), whereas in LCS, the contrasting is conducted in a pair of untrimmed videos (Eq. (7)). We also demonstrate that jointly learning FSD and LCS can enhance and complement each other for pursuing effective WSAL in Section 4.3.

### 3.4. Learning and Inference

**Training.** The above two objectives can be seamlessly integrated into existing WSAL frameworks and collaborate with each other. For optimizing the whole model, we compose the classification loss and the two contrastive losses:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{\mathrm{FSD}} + \mathcal{L}_{\mathrm{LCS}} \tag{8}$$

Since our proposed method is model-agnostic and non-intrusive, the two contrastive losses can well cooperate with any other weakly-supervised action localization objectives by replacing $\mathcal{L}_{cls}$ with different types of loss functions and backbones (Please refer to Section 4.3).

**Inference.** Given a test video, we first predict the snippet-level CAS and then apply a threshold strategy to obtain action snippet candidates following the standard process [56]. Finally, continuous snippets are grouped into action proposals, and then we perform non-maximum-suppression (NMS) to remove duplicated proposals.

## 4. Experimental Results

We evaluate the proposed FTCL on two popular datasets: THUMOS14 [28] and ActivityNet1.3 [5]. Extensive experimental results demonstrate the effectiveness of our proposed method.

### 4.1. Experimental Setup

**THUMOS14.** It contains 200 validation videos and 213 test videos annotated with temporal action boundaries from 20 action categories. Each video contains 15.4 action instances on average, making this dataset challenging for weakly-supervised temporal action localization. Following previews works [26, 37, 56, 69, 71], we apply the validation set for training and the test set for evaluation.

**ActivityNet1.3.** ActivityNet1.3 contains 10,024 training videos and 4,926 validation videos from 200 action categories, and each video contains 1.6 action instances on average. Following the standard protocol in previous work [26, 37, 56, 69, 71], we train on the training set and test on the validation set.

**Evaluation Metrics.** Following previous models [38, 54, 65], we use mean Average Precision (mAP) under different temporal Intersection over Union (t-IoU) thresholds as evaluation metrics. The t-IoU thresholds for THUMOS14 is [0.1:0.1:0.7] and for ActivityNet is [0.5:0.05:0.95].

**Implementation Details.** Following existing methods, we use I3D [8] model pretrained on Kinetics dataset as the RGB and optical flow feature extractors. The dimension of the output feature is 2048. Note that no fine-tuning operations are applied to the I3D feature extractor for a fair comparison. The number of sampled snippets $T$ for THUMOS14 and ActivityNet is set to 750 and 75, respectively. To implement $f_{\alpha}(\cdot)$ and $f_{cls}(\cdot)$, we adopt the pre-trained ACM-Net [56] as the backbone for video-level classification. For FSD contrasting, we select action/background proposals by using the learned CAS. For LCS contrasting, to save the computational cost, we do not use the entire untrimmed video but select the top-$J$ activated snippets for contrasting, $J$ is set to 30 and 10 for THUMOS14 and ActivityNet, respectively. The output dimension of $f_{\mu}(\cdot)$ and $f_{g}(\cdot)$ is 1024. For simplicity, $f_{h}(\cdot)$ is the same with $f_{g}(\cdot)$. The temperature hyper-parameter $\gamma$ and threshold $\tau$ in Eq. (5) and Eq. (6) are 10 and 0.92. Our model is implemented with PyTorch 1.9.0, and we utilize Adam with a learning rate of $10^{-4}$ and a batch size of 16 for optimization. We train our model until the training loss is smooth.

### 4.2. Comparison with State-of-the-art Methods

**Evaluation on THUMOS14.** As shown in Table 1, FTCL outperforms previous weakly supervised methods in almost all IoU metrics on the THUMOS14 dataset. Specifically, our method achieves favorable performance of 35.6% mAP@0.5 and 43.6% mAP@Avg. And an absolute gain of 1.4% and 1.0% is obtained in terms of the average mAP when compared to the SOTA approaches ACM-Net [56] and FAC-Net [27]. Furthermore, we observe that our methods can even achieve comparable performance with several fully-supervised methods, although we utilize much less supervision during training. Note that CoLA [71] gets a higher mAP@0.7 than ours. However, we get 2.7% absolute gains at average mAP. CoLA adopts a hard snippet mining strategy to pursue action completeness, which can be further equipped with our FTCL for more effective WSAL.

**Evaluation on ActivityNet1.3.** As in Table 2, our method also achieves state-of-the-art performance on the ActivityNet1.3 datasets. Specifically, compared with state-of-the-art ACM-Net [56], we obtain the relative gain of 0.8%. Note

Table 1. Temporal action localization performance comparison with state-of-the-art methods on the THUMOS14 dataset. Note that weak$^+$ represents methods that utilize external supervision information besides video labels.

| Supervision | Method | mAP@t-IoU(%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | [0.1:0.5] | [0.3:0.7] | Avg |
| Fully | S-CNN [60], CVPR2016 | 47.7 | 43.5 | 36.3 | 28.7 | 19.0 | - | - | 35.0 | - | - |
| | CDC [58], CVPR2017 | - | - | 40.1 | 29.4 | 23.3 | 13.1 | 7.9 | - | - | - |
| | R-C3D [66], ICCV2017 | 54.5 | 51.5 | 44.8 | 35.6 | 28.9 | - | - | 43.1 | - | - |
| | SSN [73], ICCV2017 | 66.0 | 59.4 | 51.9 | 41.0 | 29.8 | - | - | 49.6 | - | - |
| | TAL-Net [11], CVPR2018 | 59.8 | 57.1 | 53.2 | 48.5 | 42.8 | 33.8 | 20.8 | 52.3 | 39.8 | 45.1 |
| | GTAN [41], CVPR2019 | 69.1 | 63.7 | 57.8 | 47.2 | 38.8 | - | - | 55.3 | - | - |
| Weakly$^+$ | STAR [68], AAAI2019 | 68.8 | 60.0 | 48.7 | 34.7 | 23.0 | - | - | 47.0 | - | - |
| | 3C-Net [50], ICCV2019 | 59.1 | 53.5 | 44.2 | 34.1 | 26.6 | - | 8.1 | 43.5 | - | 37.6 |
| Weakly | UntrimmedNet [65], CVPR2017 | 44.4 | 37.7 | 28.2 | 21.1 | 13.7 | - | - | 29.0 | - | - |
| | Hide-and-Seek [62], ICCV2017 | 36.4 | 27.8 | 19.5 | 12.7 | 6.8 | - | - | 20.6 | - | - |
| | AutoLoc [59], ECCV2018 | - | - | 35.8 | 29.0 | 21.2 | 13.4 | 5.8 | - | - | - |
| | STPN [52], CVPR2018 | 52.0 | 44.7 | 35.5 | 25.8 | 16.9 | 9.9 | 4.3 | 35.0 | 18.5 | 27.0 |
| | W-TALC [54], ECCV2018 | 55.2 | 49.6 | 40.1 | 31.1 | 22.8 | - | 7.6 | 39.8 | 25.4 | 34.4 |
| | CMCS [36], CVPR2019 | 57.4 | 50.8 | 41.2 | 32.1 | 23.1 | 15.0 | 7.0 | 40.9 | 23.7 | 32.4 |
| | WSAL-BM [53], ICCV2019 | 60.4 | 56.0 | 46.6 | 37.5 | 26.8 | 19.6 | 9.0 | 45.5 | 27.9 | 36.6 |
| | DGAM [57], CVPR2020 | 60.0 | 54.2 | 46.8 | 38.2 | 28.8 | 19.8 | 11.4 | 45.6 | 29.0 | 37.0 |
| | TCAM [24], CVPR2020 | - | - | 46.9 | 38.9 | 30.1 | 19.8 | 10.4 | - | 29.2 | - |
| | Bas-Net [32], AAAI2020 | 58.2 | 52.3 | 44.6 | 36.0 | 27.0 | 18.6 | 10.4 | 43.6 | 27.3 | 35.3 |
| | A2CL-PT [47], ECCV2020 | 61.2 | 56.1 | 48.1 | 39.0 | 30.1 | 19.2 | 10.6 | 46.9 | 29.4 | 37.8 |
| | RefineLoc [1], WACV2021 | - | - | 40.8 | 32.7 | 23.1 | 13.3 | 5.3 | - | 23.0 | - |
| | Liu et al [38], AAAI2021 | - | - | 50.8 | 41.7 | 29.6 | 20.1 | 10.7 | - | 30.6 | - |
| | ACSNet [40], AAAI2021 | - | - | 51.4 | 42.7 | 32.4 | 22.0 | 11.7 | - | 32.0 | - |
| | HAM-Net [29], AAAI2021 | 65.9 | 59.6 | 52.2 | 43.1 | 32.6 | 21.9 | 12.5 | 50.7 | 32.5 | 41.1 |
| | Lee et al [33], AAAI2021 | 67.5 | 61.2 | 52.3 | 43.4 | 33.7 | 22.9 | 12.1 | 51.6 | 32.9 | 41.9 |
| | ASL [45], CVPR2021 | 67.0 | - | 51.8 | - | 31.1 | - | 11.4 | - | - | 40.3 |
| | CoLA [71], CVPR2021 | 66.2 | 59.5 | 51.5 | 41.9 | 32.2 | 22.0 | **13.1** | 50.3 | 32.1 | 40.9 |
| | D2-Net [49], ICCV2021 | 65.7 | 60.2 | 52.3 | 43.4 | **36.0** | - | - | 51.5 | - | - |
| | FAC-Net [27], ICCV2021 | 67.6 | 62.1 | 52.6 | 44.3 | 33.4 | 22.5 | 12.7 | 52.0 | 33.1 | 42.2 |
| | ACM-Net [56], arXiv2021 | 68.9 | 62.7 | 55.0 | 44.6 | 34.6 | 21.8 | 10.8 | 53.2 | 33.4 | 42.6 |
| | **FTCL(Ours)** | **69.6** | **63.4** | **55.2** | **45.2** | 35.6 | **23.7** | 12.2 | **53.8** | **34.4** | **43.6** |

Table 2. Comparison results on ActivityNet1.3 dataset.

| Method | mAP@t-IoU(%) | | | |
|---|---|---|---|---|
| | 0.5 | 0.75 | 0.95 | Avg |
| STPN [52], CVPR2018 | 26.3 | 16.9 | 2.6 | 16.3 |
| MAAN [70], ICLR2019 | 33.7 | 21.9 | 5.5 | - |
| Bas-Net [32], AAAI2020 | 34.5 | 22.5 | 4.9 | 22.2 |
| A2CL-PT [47], ECCV2020 | 36.8 | 22.0 | 5.2 | 22.5 |
| Lee et al [31], AAAI2021 | 37.0 | 23.9 | 5.7 | 23.7 |
| FAC-Net [27], ICCV2021 | 37.6 | 24.2 | 6.0 | 24.0 |
| ACM-Net [56], arXiv2021 | **40.1** | 24.2 | 6.2 | 24.6 |
| **FTCL(Ours)** | 40.0 | **24.3** | **6.4** | **24.8** |

that the performance improvement on this dataset is not as significant as it on the THUMOS14 dataset; the reason may lie in that videos in ActivityNet are much shorter than those in THUMOS14. ActivityNet only contains 1.6 instances per video on average, while the number in THUMOS14 is 15.6. Obviously, sufficient temporal information can facilitate the fine-grained temporal contrasting.

### 4.3. Further Remarks

To better understand our algorithm, we conduct ablation studies and in-depth analysis on the THUMOS14 dataset.

**Effectiveness of FSD Contrasting.** We utilize FSD contrasting for discriminative foreground-background separation. To evaluate the effectiveness of this contrasting, we wipe out this module (denoted as FTCL(w/o FSD)) from the full model and observe a significant decrease in performance, as shown in Table 3. Specifically, our full model FTCL outperforms the baseline by relative gains of (0.8%, 1.7%, 2.9%, 6.1%) mAP on t-IoU thresholds of [0.10, 0.30, 0.50, 0.70]. Without the FSD contrasting, fine-grained foreground-background distinctions can not be well handled, leading to insufficient classifier learning.

**Effectiveness of LCS Contrasting.** We also remove LCS contrasting from the full model (FTCL(w/o LCS)) to evaluate its contribution to the overall performance, and the corresponding performance consistently drops as shown in Table 3, proving the positive impact for robust classification-

Table 3. Ablation study of module effectiveness on THUMOS14.

| | mAP@t-IoU(%) | | | | |
|---|---|---|---|---|---|
| | 0.1 | 0.3 | 0.5 | 0.7 | Avg |
| ACM-Net | 68.9 | 55.0 | 34.6 | 10.8 | 42.6 |
| FTCL(w/o FSD) | 69.0 | 54.3 | 34.6 | 11.5 | 42.8 |
| FTCL(w/o LCS) | 69.3 | 55.0 | 34.8 | 11.4 | 43.0 |
| FTCL(both-FSD) | 69.6 | 55.0 | 35.3 | 11.8 | 43.2 |
| FTCL(both-LCS) | 69.4 | 55.1 | 34.8 | 11.5 | 43.1 |
| FTCL | **69.6** | **55.2** | **35.6** | **12.2** | **43.6** |

Table 4. Comparison with DTW-based methods on THUMOS14.

| | mAP@t-IoU(%) | | | | |
|---|---|---|---|---|---|
| | 0.1 | 0.3 | 0.5 | 0.7 | Avg |
| CC-DTW [25] | 69.1 | 54.9 | 34.8 | 11.2 | 42.9 |
| Drop-DTW [14] | 69.5 | 55.2 | 35.4 | 11.3 | 43.2 |
| DTW [25] | 69.2 | 55.1 | 35.0 | 11.7 | 43.1 |
| FTCL | **69.6** | **55.2** | **35.6** | **12.2** | **43.6** |

Table 5. Evaluation of the complementary role of FTCL.

| | mAP@t-IoU(%) | | | | |
|---|---|---|---|---|---|
| | 0.1 | 0.3 | 0.5 | 0.7 | Avg |
| STPN [52](reproduced) | 52.2 | 35.6 | 16.8 | 4.1 | 27.2 |
| **STPN+FTCL** | **54.1(+1.9)** | **38.4(+2.8)** | **18.2(+1.4)** | **4.8(+0.7)** | **29.0(+1.8)** |
| W-TALC [54](reproduced) | 55.7 | 40.0 | 22.7 | 7.7 | 34.5 |
| **W-TALC+FTCL** | **57.5(+1.8)** | **40.9(+0.9)** | **23.8(+1.1)** | **8.4(+0.7)** | **35.7(+1.2)** |
| CoLA [71](reproduced) | 66.1 | 52.1 | 34.3 | 13.1 | 41.7 |
| **CoLA+FTCL** | **67.1(+1.0)** | **52.9(+0.8)** | **34.8(+0.5)** | **13.2(+0.1)** | **42.3(+0.6)** |

to-localization adaption. Mining LCS for untrimmed videos enables the model to discover coherent snippets in an action instance, thus facilitating localization performance.

**Are the Above Two Modules Redundant?** Both the FSD and LCS objectives are adopted for sequence-to-sequence contrasting but with different goals. Astute readers may be curious about whether the FSD and LCS learning strategies are redundant, *i.e.*, can we adopt either FSD or LCS for jointly modeling the foreground-background separation and classification-to-localization adaption? To answer this question, we conduct experiments with only FSD or LCS contrasting for tackling both the separation and adaption objectives, namely FTCL(both-FSD) and FTCL(both-LCS) in Table 3. We observe that our full model outperforms both variants, proving that the above two modules are not redundant. Another observation is that the two variants achieve better performance than FTCL(w/o FSD)) and FTCL(w/o LCS)). The reason lies in that both FSD and LCS belong to the sequence-to-sequence measurement, which can promote the separation and adaption objectives solely. However, since the two objectives have their unique properties, we design the FSD and LCS contrasting strategies to address them, which obtains the best performance.

**Why Not Resort to other Dynamic Programming Strategies like DTW?** We observe that some recent works are pursuing the video sequence alignment based on dynamic time warping (DTW) [7, 14, 25]. However, DTW assumes that the two sequences can be fully aligned, thus requiring trimmed videos. To validate the effectiveness of our FTCL, as shown in Table 4, we compare our proposed method with the current state-of-the-art DTW-based approaches, Cycle-Consistency DTW (CC-DTW) [25] and Drop-DTW [14]. The results consistently demonstrate the superiority of our framework. We also replace our FSD and LCS strategies (Eq. (2) and Eq. (6)) with the standard differential DTW operator [25] (denoted as DTW), which obtains inferior results as we analyzed above.

**Complementary Role of the proposed FTCL.** It is obvious that the proposed strategy is model-agnostic and non-intrusive, and hence can play a complementary role over existing methods. In Table 5, we plug our FSD and LCS contrasting into three WSAL approaches including STPN [52], W-TALC [54], and CoLA [71]. The results show that our proposed learning strategies can consistently improve their performance. In addition, our method does not introduce computational cost during model inference. Note that CoLA also adopts contrastive learning in snippet-level, while our proposed method can further boost its performance by additionally considering the fine-grained temporal distinctions.

## 5. Conclusions

This paper proposes a fine-grained temporal contrastive learning framework for WSAL, which jointly enjoys the merits of discriminative action-background separation and alleviated task gap between classification and localization. Specifically, two types of contrasting strategies, including FSD and LCS contrasting, are designed via differentiable dynamic programming, capable of making fine-grained temporal distinctions. The encouraging performance is demonstrated in extensive experiments.

**Limitations.** In this work, similar to existing WSAL models, we equally employ a fixed snippet division strategy for all videos. However, since different videos have different duration and shots, the simple and fixed way may hinder the fine-grained temporal contrastive learning. In the future, we plan to conduct FTCL in an adaptive manner, *e.g.*, considering hierarchical temporal structures or performing shot detection and action localization in a unified framework.

## Acknowledgements

This work was supported by the National Key Research & Development Plan of China under Grant 2020AAA0106200, in part by the National Natural Science Foundation of China under Grants 62036012, 61721004, U21B2044, 62102415, 62072286, 61720106006, 61832002, 62072455, 62002355, and U1836220, in part by the Key Research Program of Frontier Sciences of CAS under Grant QYZDJSSWJSC039, in part by Beijing Natural Science Foundation (L201001), in part by Open Research Projects of Zhejiang Lab (NO.2022RC0AB02).

## References

[1] Humam Alwassel, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In *ACM*, 2019. 7

[2] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370, 1994. 2, 3

[3] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 3

[4] 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 *BMVC*, 2017. 3

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

[6] Xingyu Cai and Tingyang Xu. Dtwnet: a dynamic timewarping network. *NeurIPS*, 2019. 3, 5

[7] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *CVPR*, 2020. 2, 3, 8

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*. 3, 4, 6

[9] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*, 2019. 2, 3

[10] Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In *CVPR*, 2021. 2, 3

[11] 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 *CVPR*, 2018. 3, 7

[12] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017. 3, 5

[13] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, 2017. 3

[14] Nikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan D Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. *arXiv:2108.11996*, 2021. 3, 8

[15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, 2019. 3

[16] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, 2016. 3

[17] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 3

[18] Junyu Gao and Changsheng Xu. Learning video moment retrieval without a single annotated video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 3

[19] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 690–699. ACM, 2018. 3

[20] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019. 3

[21] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 3

[22] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3476–3491, 2021. 3

[23] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu. Deep relative tracking. *IEEE Transactions on Image Processing*, 26(4):1845–1858, 2017. 3

[24] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian. Learning temporal co-attention models for unsupervised video action localization. In *CVPR*. 7

[25] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *CVPR*, 2021. 2, 3, 5, 8

[26] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *ACM MM*, 2021. 1, 3, 4, 6

[27] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *ICCV*, 2021. 1, 6, 7

[28] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 6

[29] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *AAAI*, 2021. 7

[30] Ashraful Islam and Richard Radke. Weakly supervised temporal action localization using deep metric learning. In *WACV*, 2020. 3, 4

[31] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *ICCV*, 2021. 3, 7

[32] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI*. 3, 7

[33] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, 2021. 7

[34] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, 2021. 3

[35] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, 2017. 3

[36] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019. 3, 7

[37] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In *CVPR*, 2021. 6

[38] Ziyi Liu, Le Wang, Wei Tang, Junsong Yuan, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through learning explicit subspaces for action and context. In *AAAI*, 2021. 6, 7

[39] 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 *ICCV*, 2019. 3

[40] Ziyi Liu, Le Wang, Qilin Zhang, Wei Tang, Junsong Yuan, Nanning Zheng, and Gang Hua. Acsnet: Action-context separation network for weakly supervised temporal action localization. In *AAAI*, 2021. 7

[41] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, 2019. 1, 3, 7

[42] 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. 1, 3

[43] 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 *ECCV*, 2020. 3

[44] 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. 3

[45] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *CVPR*, 2021. 1, 3, 7

[46] Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *ICML*, 2018. 3, 5

[47] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *ECCV*. 7

[48] 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 *ACM MM*, 2020. 3

[49] 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. 3, 4, 7

[50] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In *ICCV*, 2019. 3, 4, 7

[51] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001. 4

[52] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 2018. 3, 7, 8

[53] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *ICCV*, 2019. 3, 4, 7

[54] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018. 1, 3, 4, 6, 7, 8

[55] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978. 2, 3

[56] Zhijun Li Lijun Zhang Fan Lu Alois Knoll Sanqing Qu, Guang Chen. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv:2104.02967*, 2021. 1, 3, 4, 5, 6, 7

[57] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*, 2020. 1, 3, 7

[58] 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 *CVPR*, 2017. 3, 7

[59] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*. 7

[60] Zheng Shou, Dongang Wang, and S Chang. Action temporal localization in untrimmed videos via multi-stage cnns. In *CVPR*. 7

[61] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 3

[62] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 3, 7

[63] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *ICCV*, 2021. 1, 3

[64] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. *arXiv:2102.01894*, 2021. 3

[65] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 1, 3, 6, 7

[66] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 3, 7

[67] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020. 1, 3

[68] 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 *AAAI*, 2019. 1, 3, 7

[69] 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, 6

[70] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *ICLR*, 2019. 7

[71] 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, 4, 6, 7, 8

[72] Chengwei Zhang, Yunlu Xu, Zhanzhan Cheng, Yi Niu, Shiliang Pu, Fei Wu, and Futai Zou. Adversarial seeded sequence growing for weakly-supervised temporal action localization. In *ACM MM*, 2019. 3

[73] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 3, 7

[74] 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 *ACM MM*, 2018. 3

[75] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *ICCV*, 2021. 1, 3