D2-Net: Weakly-Supervised Action Localization via Discriminative Embeddings and Denoised Activations

Sanath Narayan[1] Hisham Cholakkal[2] Munawar Hayat[3] Fahad Shahbaz Khan[2,4]
Ming-Hsuan Yang[5,6,7] Ling Shao[1]

[1]Inception Institute of Artificial Intelligence [2]Mohamed Bin Zayed University of AI [3]Monash University
[4]Linköping University [5]University of California, Merced [6]Google Research [7]Yonsei University

Abstract

This work proposes a weakly-supervised temporal action localization framework, called D2-Net, which strives to temporally localize actions using video-level supervision. Our main contribution is the introduction of a novel loss formulation, which jointly enhances the discriminability of latent embeddings and robustness of the output temporal class activations with respect to foreground-background noise caused by weak supervision. The proposed formulation comprises a discriminative and a denoising loss term for enhancing temporal action localization. The discriminative term incorporates a classification loss and utilizes a top-down attention mechanism to enhance the separability of latent foreground-background embeddings. The denoising loss term explicitly addresses the foreground-background noise in class activations by simultaneously maximizing intra-video and inter-video mutual information using a bottom-up attention mechanism. As a result, activations in the foreground regions are emphasized whereas those in the background regions are suppressed, thereby leading to more robust predictions. Comprehensive experiments are performed on multiple benchmarks, including THUMOS14 and ActivityNet1.2. Our D2-Net performs favorably in comparison to the existing methods on all datasets, achieving gains as high as 2.3% in terms of mAP at IoU=0.5 on THUMOS14. Source code is available at https://github.com/naraysa/D2-Net.

1. Introduction

Temporal action localization is a challenging problem, which aims to jointly classify and localize the temporal boundaries of actions in videos. Most existing approaches [37, 5, 36, 30, 43, 32] are based on strong supervision, requiring manually annotated temporal boundaries of actions during training. In contrast to these strong frame-level supervision based methods, weakly-supervised action localization learns to localize actions in videos, leveraging only video-level supervision. Weakly-supervised action localization is therefore of greater importance since the manual annotation of temporal boundaries in videos is laborious, expensive and prone to large variations [28, 27].

Existing methods [33, 34, 23, 25, 31] for weakly-supervised action localization typically use video-level annotations in the form of action classes and learn a sequence of class-specific scores, called temporal class activation maps (TCAMs). In general, a classification loss is used to obtain the discriminative foreground regions in TCAMs. Some approaches [23, 25, 22, 24] learn TCAMs using action labels and obtain temporal boundaries via a post-processing step, while others [31, 15] use a TCAM-generating video classification branch along with an explicit localization branch to directly regress action boundaries. Nevertheless, the localization performance is heavily dependent on the quality of the TCAMs. The quality of TCAMs is likely to improve in fully-supervised settings where frame-level annotations are available. Such frame-level information (true foreground and background regions) are unavailable in the weakly-supervised paradigm. In such a paradigm, the predicted foreground regions often overlap with the ground-truth background regions, while predicted background regions are likely to overlap with the ground-truth foreground regions. This leads to noisy activations, *i.e.*, false positives and false negatives, in the learned TCAMs. Most existing weakly-supervised action localization methods that learn TCAMs typically rely on separating foreground and background regions (foreground-background separation) and do not explicitly handle its noisy outputs.

In this work, we address the problem of foreground-background separation along with explicit tackling of noise in TCAMs for weakly-supervised action localization. We propose a unified loss formulation that is jointly optimized to classify and temporally localize action snippets (group of frames) in videos. Our loss formulation comprises a discriminative and a denoising loss term. The discriminative loss seeks to maximally separate backgrounds from actions

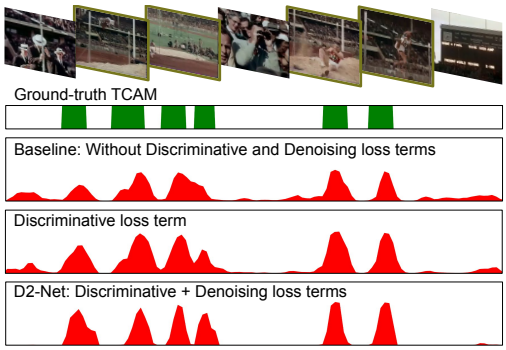

Figure 1. **Impact of our proposed loss formulation** on the quality of the output TCAMs. Compared to the baseline (without our discriminative and denoising loss terms), the introduction of the discriminative loss term improves the separation between foreground and background activations (*e.g.*, third and fourth ground-truth action instance from the left). Furthermore, our final `D2-Net` comprising both the discriminative and the denoising loss terms reduces the noise in the TCAMs, leading to more robust TCAMs.

(foregrounds) via interlinked classification and localization learning objectives (Sec. 3.1). The denoising loss (Sec. 3.2) complements the discriminative term by explicitly addressing the foreground-background noise in activations, thereby producing robust TCAMs (see Fig. 1).

In our loss formulation, we learn distinct latent embeddings such that their foreground-background separation is maximized based upon the corresponding top-down attention generated from the output TCAMs. Furthermore, the embeddings are employed to generate pseudo-labels based on their foreground scores (bottom-up attention). These pseudo-labels are utilized to explicitly handle the noise by emphasizing the corresponding output activations in pseudo-foreground regions, while suppressing the activations in pseudo-background regions. This pseudo-background suppression and pseudo-foreground enhancement is achieved by maximizing the mutual information (MI) between activations and generated pseudo-labels within an action video (intra-video). Maximizing MI between predicted activations and labels decreases the uncertainty of predictions, leading to more robust predictions. In addition to capturing intra-video MI, our formulation also strives to maximize MI between the action class predictions and video-level ground-truth labels, across videos in a mini-batch (inter-video).

Contributions: We introduce a weakly-supervised action localization framework, `D2-Net`, which incorporates a novel loss formulation that jointly enhances the foreground-background separability and explicitly tackles the noise to robustify the output TCAMs. Our main contributions are:

- We introduce a discriminative loss term, which simultaneously aims at video categorization and enhanced foreground-background separation.
- We introduce a denoising loss term to improve the robustness of TCAMs. Our denoising loss explicitly addresses noise in TCAMs by maximizing the MI between activations and labels within a video (intra-video) *and* across videos (inter-video). To the best of our knowledge, we are the first to introduce a loss term that simultaneously captures MI across multiple snippets within a video and across all videos in a batch for weakly-supervised action localization.
- Experiments are performed on multiple benchmarks, including THUMOS14 [6] and ActivityNet1.2 [3]. Our `D2-Net` performs favorably against existing weakly-supervised methods on all datasets, achieving gains as high as 2.3% mAP at IoU=0.5 on THUMOS14.

2. Related Work

Several weak supervision strategies have been explored in the context of action localization, including category labels [33, 23, 34, 25, 31, 41], sparse temporal points [19], order of actions [26, 2], instance count [22, 39] and single-frame annotations [17]. Most existing weakly-supervised action localization methods employ category labels as weak supervision and typically utilize features extracted from backbone networks [35, 4] trained on the action recognition task. The work of [34] proposes a selection module for detecting the relevant temporal segments and employs a classification loss for training. The Autoloc method [31] extends [34] by adding an explicit localization branch and utilizes an outer-inner contrastive loss for its training. In contrast, [25, 8] match similar segments of actions in paired videos by employing classification and similarity-based losses that require multiple videos of same actions in a mini-batch. Different from these works, our approach explicitly addresses the issue of large number of easy negatives overwhelming a smaller number of hard positives via sample re-weighting and performs foreground-background separation by inter-linking classification and localization objectives.

Snippet-level loss: While the work of [24] employs a background-aware loss along with a self-guided loss for modeling the background, [21] additionally utilizes an iterative multi-pass erasing step for discovering different action segments in TCAMs. Differently, the training in [16] alternates between updating a key-instance assignment branch and a classification branch via Expectation Maximization. In contrast, the recent work of [12] classifies the foreground/background snippets as in/out-of-distribution based on the feature magnitude and entropy over foreground classes. However, all these approaches aggregate per-snippet losses for training and do not explicitly capture the mutual information (MI) between the activations and labels, which is likely to be more beneficial due to the absence of snippet-level labels in a weakly-supervised setting. Different from existing methods [24, 21, 16, 12, 22, 23, 1, 8], our approach addresses the problem of foreground-background noise by exploiting both inter- *and* intra-video MI between class acti-

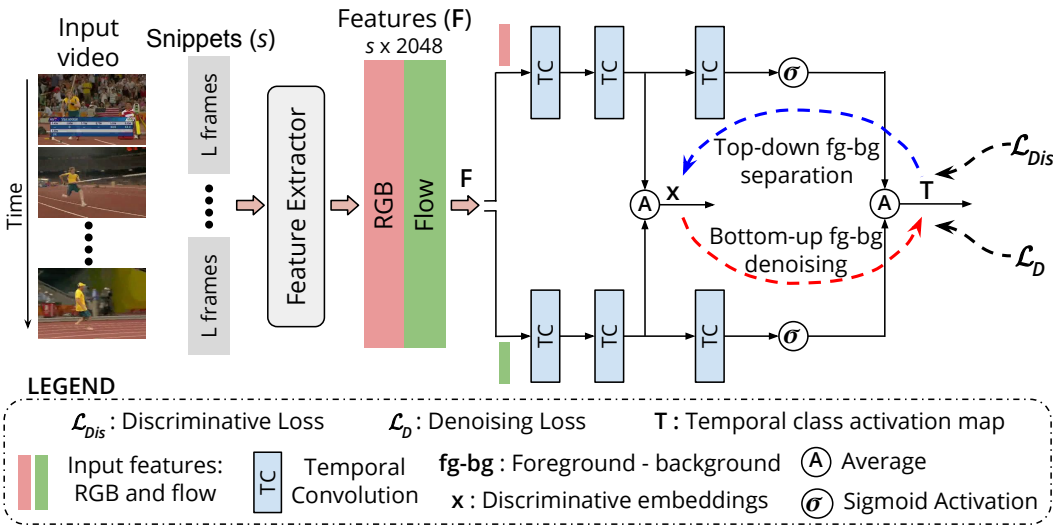

Figure 2. **Overall architecture** of our D2-Net. The focus of our design is the introduction of a novel loss formulation that jointly enhances the discriminability of latent embeddings and explicitly addresses the foreground-background noise in the output class activations. The network comprises two identical parallel streams (RGB and flow) consisting of three temporal convolutional TC layers. The second TC layer activations from both streams are averaged to obtain latent embeddings $\mathbf{x}$. The final outputs of both streams are then averaged to obtain the temporal class activation maps (TCAMs) $\mathbf{T}$ of untrimmed input videos. A discriminative loss $\mathcal{L}_{Dis}$ (Sec. 3.1) is introduced to enhance the foreground-background separability (←--) of embeddings $\mathbf{x}$ by utlizing a top-down attention mechanism, in addition to achieving video classification. Furthermore, a denoising loss $\mathcal{L}_D$ (Sec. 3.2) is introduced to explicitly address the foreground-background noise (--→) in the class activations of $\mathbf{T}$, by utilizing a bottom-up attention. The network is trained jointly using both loss terms $\mathcal{L}_{Dis}$ and $\mathcal{L}_D$.

vations and corresponding labels, resulting in robust TCAMs. To the best of our knowledge, we are the first to propose a weakly-supervised action localization approach that simultaneously captures MI across multiple snippets within a video and across videos in a mini-batch (see also Fig. 4).

3. Proposed Method

Our D2-Net strives to improve the separation of foreground-background feature representations in videos, while jointly enhancing the robustness of output TCAMs w.r.t. foreground-background noise. This leads to better differentiation between foreground actions and surrounding background regions, resulting in enhanced action localization in the challenging weakly-supervised setting. Here, we first present our overall architecture, followed by a detailed description of our proposed losses for training D2-Net.

Overall architecture of D2-Net is illustrated in Fig. 2. Given a video v, we divide it into non-overlapping snippets of $L = 16$ frames each. Features are then extracted to encode appearance (RGB) and motion (optical flow) information. Similar to [23, 25, 22], we use the Inflated 3D (I3D) [4] to obtain $d = 2048$ dimensional features for each 16-frame snippet. Let $\mathbf{F} \in \mathbb{R}^{s \times d}$ denote features for a video, where s is the number of snippets. The extracted features become the inputs to our D2-Net, which comprises two parallel streams for RGB and optical flow. Each stream consists of three temporal convolutional (TC) layers. The first two layers learn latent discriminative embeddings $\mathbf{x}(t) \in \mathbb{R}^{d/2}$ (with time $t \in [1, s]$), from the input features $\mathbf{F}$. The output of the final TC layer is passed through a *sigmoid* activation. Subsequently, the outputs from both streams are averaged to obtain TCAMs $\mathbf{T} \in \mathbb{R}^{s \times C}$ representing a sequence of class-specific scores over time for C action classes. The main contribution of our work is the introduction of a novel loss formulation to train the proposed D2-Net. Our training objective combines a discriminative ($\mathcal{L}_{Dis}$) and a denoising term ($\mathcal{L}_D$), with a balancing weight α,

$$\mathcal{L} = \mathcal{L}_{Dis} + \alpha \mathcal{L}_D. \tag{1}$$

These two loss terms utilize foreground-background attention sequences computed in opposite directions: (*i*) the discriminative loss $\mathcal{L}_{Dis}$ utilizes a top-down attention, which is computed from the output TCAMs (the top-most layer) and (*ii*) the denoising loss $\mathcal{L}_D$ utilizes a bottom-up attention, which is derived from the foreground scores of the latent embeddings (intermediate layer features). We describe these losses in detail in Sec. 3.1 and 3.2.

3.1. Foreground-Background Discriminability: $\mathcal{L}_{Dis}$

In this work, we introduce a discriminative loss ($\mathcal{L}_{Dis}$) to learn separable class-agnostic foreground and action-free background feature representations, in terms of latent embeddings, using a top-down attention from the TCAMs. The embedding of a video with s snippets is defined by a weighted temporal pooling based on the class activations $\mathbf{T} \in \mathbb{R}^{s \times C}$. Let the top-down foreground attention $\boldsymbol{\lambda}(t) = \max_c \mathbf{T}[t, c]$

denote the maximum foreground activation across all action classes $c \in \{1, \ldots, C\}$, where $t \in [1, s]$ and C is the number of classes. Then, the class-agnostic foreground and background embeddings are:

$$\mathbf{x}_{fg} = \sum_{\boldsymbol{\lambda}(t)>\tau} \boldsymbol{\lambda}(t)\mathbf{x}(t), \quad \mathbf{x}_{bg} = \sum_{\boldsymbol{\lambda}^b(t)>\tau} \boldsymbol{\lambda}^b(t)\mathbf{x}(t), \quad (2)$$

where τ=0.5 and $\boldsymbol{\lambda}^b(t)$=1−$\boldsymbol{\lambda}(t)$ is the background attention. Maximizing the distance between foreground and background embeddings enhances the separability of the corresponding output activations, leading to improved localization. In addition, different sets of action classes are likely to share certain characteristics among them *e.g.*, *Hammer Throw* and *Discus Throw* have similar spatial context and motion. Hence, clustering foreground embeddings amongst themselves at a coarse level is likely to aid "coarse-to-fine" snippet-level classification. Similarly, clustering background embeddings helps in learning an approximate universal background embedding, which is likely to aid in generalization at test time to new backgrounds. Hence, three weight terms, w_{fb}, w_{fg} and w_{bg}, are introduced in our $\mathcal{L}_{Dis}$, targeting foreground-background separation, foreground grouping and background grouping, respectively. They are defined as:

$$\begin{aligned} w_{fb} &= \max(0, \cos(\mathbf{x}_{fg}, \tilde{\mathbf{x}}_{bg})), \\ w_{fg} &= \gamma(1 - \cos(\mathbf{x}_{fg}, \tilde{\mathbf{x}}_{fg})), \\ w_{bg} &= \gamma(1 - \cos(\mathbf{x}_{bg}, \tilde{\mathbf{x}}_{bg})), \end{aligned} \quad (3)$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ denote embeddings from different videos in a mini-batch. Here, γ denotes the intra-class compactness weight used for grouping same class (foreground *vs.* background) embeddings. Alongside robust localization, our other objective is the multi-label classification of action categories. A major challenge is introduced by the class-imbalance problem, where easy background snippets overwhelmingly outnumber the hard foregrounds. To address this, inspired by the focal loss for object detection [13], we propose to include penalty terms based on the weights (Eq. 3), in our $\mathcal{L}_{Dis}$. To this end, a video-level prediction $\mathbf{p} \in \mathbb{R}^C$ is obtained by performing a temporal *top-k* pooling on $\mathbf{T}$. Our $\mathcal{L}_{Dis}$ term, which jointly addresses the class-imbalance and enhances foreground-background separation, is defined by

$$\begin{aligned} \mathcal{L}_{Dis} = &- \sum_{c:\mathbf{y}[c]=1} (1 - \mathbf{p}[c] + w_{fg} + w_{fb})^{\beta} \log(\mathbf{p}[c]) \\ &- \sum_{c:\mathbf{y}[c]=0} (\mathbf{p}[c] + w_{bg} + w_{fb})^{\beta} \log(1 - \mathbf{p}[c]), \end{aligned} \quad (4)$$

where $\mathbf{y} \in \{0, 1\}^C$ denotes the video-level label and β is the focusing parameter. The first term in Eq. 4 denotes the loss for a positive action class, while the second term incorporates the loss for a negative class. The weight term w_{fb} (see Eq. 3) is added for both positive action classes and background classes since it represents the foreground-background separation. The terms w_{fg} and w_{bg} enhance intra-class compactness for the positive and background classes, respectively. The first term in Eq. 4 indicates that the loss due to a positive action class c is low only when (i) its predicted probability $\mathbf{p}[c]$ is high, and (ii) the foreground grouping w_{fg} and foreground-background separation w_{fb} for the corresponding video are both simultaneously low. A similar observation holds in the second term for the negative class. Thus, $\mathcal{L}_{Dis}$ enhances the discriminability of embeddings $\mathbf{x}(t)$ by encouraging foreground-background separation while simultaneously achieving classification.

3.2. Robust Temporal Class Activation Maps: $\mathcal{L}_D$

Our discriminative loss $\mathcal{L}_{Dis}$ improves action localization by enhancing the distinctiveness of latent embeddings. However, the temporal locations of true foreground regions are unknown under weak supervision, resulting in noisy output temporal class activations (and noisy top-down attention) learned from video-level labels. Consequently, the foreground and background embeddings ($\mathbf{x}_{fg}$ and $\mathbf{x}_{bg}$), learned from the top-down attention $\boldsymbol{\lambda}(t)$, are likely to be noisy. Our goal is to explicitly reduce this foreground-background noise caused by the absence of snippet-level labels and improve the robustness of the output class activations. To this end, we introduce a denoising loss $\mathcal{L}_D$ comprising a novel pseudo-Determinant based Mutual Information (pDMI) loss. Our $\mathcal{L}_D$ exploits both intra- and inter-video mutual information (MI) between the class activations and corresponding labels.

Our pseudo-Determinant based Mutual Information (pDMI) loss is inspired by the Determinant based Mutual Information (DMI) [38]. The original DMI, proposed for multi-class classification, is computed as the determinant of a joint distribution matrix, *i.e.*, DMI($\mathbf{P}, \mathbf{Y}$)=$|\det(\mathbf{U})|$. Here, $\mathbf{U} = {}^1\!/\!_n\mathbf{PY}$ is the joint distribution over the predicted posterior probabilities $\mathbf{P}$ and the ground-truth (noisy) labels $\mathbf{Y}$. The matrices $\mathbf{P}$ and $\mathbf{Y}$ are of sizes $C \times n$ and $n \times C$, where n denotes the mini-batch size and C the number of classes. The DMI loss $\mathcal{L}_{dmi}$ is defined as

$$\mathcal{L}_{dmi} = -\mathbb{E}[\log(|\det(\mathbf{U})|)], \quad (5)$$

where $\mathbb{E}$ denotes Expectation. Note that $\mathcal{L}_{dmi}$ depends on the determinant of $\mathbf{U}$. To ensure a non-zero $\det(\mathbf{U})$, the label matrix $\mathbf{Y}$ must be full-rank, *i.e.*, a mini-batch must contain instances from all classes. This is prohibitive for a large number of classes. Such a mini-batch sampling for action localization also leads to memory issues in GPUs due to the long duration of untrimmed videos in the dataset, especially when capturing inter-video MI.

Our pDMI loss overcomes these limitations and ensures a non-degenerate value of DMI by avoiding an explicit computation of the determinant. To this end, we observe that for

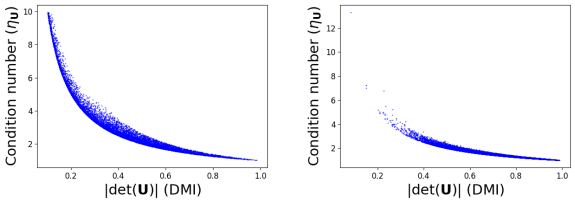

Figure 3. **Condition number ($\eta_{\mathbf{U}}$) *vs*. Determinant ($|\det(\mathbf{U})|$)** for joint distribution matrices $\mathbf{U}$. On the left: 25k randomly sampled $\mathbf{U}$. On the right: $\mathbf{U}$ obtained during our snippet-level training. In both cases, minimizing $\eta_{\mathbf{U}}$ leads to maximizing $|\det(\mathbf{U})|$ (DMI).

the DMI loss to tend to zero, the determinant of the joint distribution $|\det(\mathbf{U})|$ must tend to one. Formally,

$$\mathcal{L}_{dmi} \to 0 \implies |\det(\mathbf{U})| \to 1 \implies \mathbf{U} \to \mathbf{I}. \quad (6)$$

As a result, DMI is maximum when $|\det(\mathbf{U})|=1$, with the identity matrix $\mathbf{I}$ as an optima for $\mathbf{U}$ of size $C \times C$ (since elements of $\mathbf{U}\in[0,1]$). Furthermore, the condition number η for the optimal solution $\mathbf{I}$ is minimum, *i.e.*, $\eta=1$. Hence, instead of maximizing $|\det(\mathbf{U})|$, we can alternatively minimize its η. In effect, $\mathbf{U}$ becomes better-conditioned and this improves the robustness of the activations towards label noise. The proposed pDMI loss $\mathcal{L}_{pdmi}$ is then given by

$$\mathcal{L}_{pdmi} = \mathbb{E}[\log(\text{pDMI}(\mathbf{P},\mathbf{Y}))] = \mathbb{E}[\log(\eta_{\mathbf{U}})], \quad (7)$$

where $\eta_{\mathbf{U}}$ denotes the condition number of $\mathbf{U}$. Since the rank of $\mathbf{U}$ is $r \leq C$, $\eta_{\mathbf{U}}$ is computed as σ_1/σ_r, where $\{\sigma_1,\ldots,\sigma_r\}$ are non-zero singular values of $\mathbf{U}$. Thus, our pDMI loss avoids an explicit computation of the determinant and overcomes the limitations of the standard DMI. Fig. 3 shows plots of $\eta_{\mathbf{U}}$ *vs*. $|\det(\mathbf{U})|$ for joint distribution matrices $\mathbf{U}$ that are randomly sampled (left) and encountered during intra-video MI training (right, described in Sec. 3.2.1). It can be observed that minimizing $\eta_{\mathbf{U}}$ indeed maximizes $|\det(\mathbf{U})|$, *i.e.*, DMI, in turn maximizing MI. Consequently, our pDMI serves as a promising alternative to the original DMI when optimizing with noisy temporal action labels.

3.2.1 Snippet-level and Video-level Noise Removal

To robustify the TCAMs, we employ our $\mathcal{L}_{pdmi}$ at two levels: (*i*) snippet-level to exploit intra-video MI, and (*ii*) video-level to exploit inter-video MI. Snippet-level denoising incorporates a bottom-up attention to emphasize the foreground activations, while suppressing the background ones by capturing the MI between the temporal activations and corresponding foreground labels within a video. On the other hand, the video-level denoising step exploits MI between the video representations and corresponding labels, across videos, to achieve the same objective. Fig. 4 shows a conceptual illustration of loss computation with and without capturing MI.

Snippet-level joint distribution: It captures the MI between the foreground-background activations and the

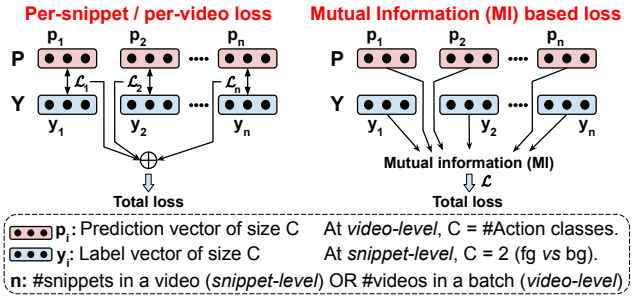

Figure 4. **A conceptual illustration of loss computation** with (on the right) and without (on the left) capturing mutual information (MI). Typically, existing methods compute the loss without MI (*e.g.*, cross-entropy loss) by aggregating individual losses ($\mathcal{L}_i$) between prediction $\mathbf{p}_i$ and labels $\mathbf{y}_i$ either at a per-video or per-snippet level. Instead, we compute a collective loss across (i) all snippets within a video (snippet-level) *and* (ii) all videos in a batch (video-level), by capturing the MI between predictions ($\mathbf{P}$) and labels ($\mathbf{Y}$).

snippet-level pseudo-labels within a video. For this, we utilize a bottom-up attention mechanism, which encodes the foreground scores $\boldsymbol{\lambda}'(t)$ of latent embeddings $\mathbf{x}(t)$ for the corresponding snippets. The scores $\boldsymbol{\lambda}'(t)$ are computed w.r.t. a reference background embedding $\mathbf{x}_{ref}$ and are given by

$$\boldsymbol{\lambda}'(t) = 0.5(1-\cos(\mathbf{x}(t),\mathbf{x}_{ref})), \quad t\in[1,s], \quad (8)$$

where $\mathbf{x}_{ref}^{[m]} = 0.9\mathbf{x}_{ref}^{[m-1]} + 0.1\mathbf{x}_{bg}^{\mu,[m]}$ is progressively computed as a running mean of $\mathbf{x}_{bg}$ over m iterations. Here, $\mathbf{x}_{bg}^{\mu,[m]}$ denotes the mean of the background embeddings in a mini-batch at iteration m. Let $t_f=\{t:\boldsymbol{\lambda}'(t)>0.5\}$ and $t_b=\{t:\boldsymbol{\lambda}'(t)<0.5\}$ denote the time instants for selecting the foreground and background activations w.r.t. $\boldsymbol{\lambda}'(t)$. Using the pseudo-foreground temporal locations t_f, a row matrix $\boldsymbol{\lambda}_f$ of width $n_f=|t_f|$ is constructed using top-down attention $\boldsymbol{\lambda}(t), t\in t_f$. Similarly, $\boldsymbol{\lambda}_b$ of width $n_b=|t_b|$ is constructed for the pseudo-background snippets. Then, the prediction matrix $\mathbf{P}_1$ and pseudo-label matrix $\mathbf{Y}_1$ are given by

$$\mathbf{P}_1 = \begin{bmatrix} \boldsymbol{\lambda}_f & \boldsymbol{\lambda}_b \\ 1-\boldsymbol{\lambda}_f & 1-\boldsymbol{\lambda}_b \end{bmatrix}, \quad \mathbf{Y}_1 = {}^1\!/_z \begin{bmatrix} \mathbf{1}_{n_f} & \mathbf{0}_{n_f} \\ \mathbf{0}_{n_b} & \mathbf{1}_{n_b} \end{bmatrix}, \quad (9)$$

where $z=n_f+n_b$, $\mathbf{P}_1\in\mathbb{R}^{2\times z}$, $\mathbf{Y}_1\in\mathbb{R}^{z\times 2}$, $\mathbf{1}_k$ and $\mathbf{0}_k$ are k dimensional column vectors of ones and zeros. The snippet-level joint distribution is then defined as $\mathbf{U}_1 = \mathbf{P}_1\mathbf{Y}_1$.

Video-level joint distribution: Here, the noise stems from the video-level prediction $\mathbf{p}\in\mathbb{R}^C$ and is predominantly caused by the temporal *top-k* pooling. Under the weakly-supervised setting, all the *top-k* locations predicted for an action class need not necessarily belong to that class. Moreover, actions in untrimmed videos may not span $k=\lceil s/8\rceil$ snippets. Hence, denoising the video-level prediction $\mathbf{p}$ eventually robustifies the output class activations at the snippet-level. Let the prediction $\mathbf{P}_2$ and label $\mathbf{Y}_2$ be

$$\mathbf{P}_2 = \left[\mathbf{p}_1,\ldots,\mathbf{p}_n\right] \text{ and } \mathbf{Y}_2 = {}^1\!/_n \left[\mathbf{y}_1,\ldots,\mathbf{y}_n\right]^\top, \quad (10)$$

where $\mathbf{p}_i \in \mathbb{R}^C$ and $\mathbf{y}_i \in \{0,1\}^C$ denote the video-level prediction and associated label of i-th video in a mini-batch. Then, the video-level joint distribution that captures the MI between class activations and action classes across videos is $\mathbf{U}_2 = \mathbf{P}_2\mathbf{Y}_2$. We finally define our denoising loss as

$$\begin{aligned}\mathcal{L}_D &= \mathcal{L}_{DS} + \mathcal{L}_{DV} \qquad (11)\\ &= \mathbb{E}[\log(\text{pDMI}(\mathbf{P}_1, \mathbf{Y}_1))] + \mathbb{E}[\log(\text{pDMI}(\mathbf{P}_2, \mathbf{Y}_2))],\end{aligned}$$

where the pDMI loss is given by Eq. 7. Here, $\mathcal{L}_{DS}$ and $\mathcal{L}_{DV}$ denote the snippet-level and video-level losses. Thus, our denoising loss improves the TCAMs, at the snippet-level and video-level, by making them robust to the foreground-background noise under the weakly-supervised setting.

3.3. Inference: Action Localization from TCAMs

At inference, given a video, D2-Net outputs a bottom-up attention sequence $\boldsymbol{\lambda}'$ (Eq. 8) of length s and a class activation map $\mathbf{T}$ of size $s \times C$. We perform *top-k* pooling to obtain the predicted class probabilities $\mathbf{p} \in \mathbb{R}^C$, which are then used to find the relevant action classes above a threshold $p_{th} = 0.5 \max(\mathbf{p})$. For every relevant class c, its corresponding class activations $\mathbf{T}_c \in \mathbb{R}^s$ are multiplied element-wise with $\boldsymbol{\lambda}' \in \mathbb{R}^s$ to obtain a refined sequence $\mathbf{r}_c = \boldsymbol{\lambda}' \mathbf{T}_c$. The snippets with activations above a threshold are retained and a 1-D connected component is used to obtain segment proposals. Multiple thresholds are used to obtain a larger pool of proposals. Each proposal is then scored using the contrast between the mean activation of the proposal itself and its surrounding areas [31], $S = S_i - S_o$, where S_i and S_o respectively denote the mean activation of the proposal and its neighboring background. The neighboring background is obtained by inflating the proposal on either side by 25% of its width, as in [31]. Proposals with high overlap are removed using class-wise NMS. Only high-scoring proposals (*i.e.*, $S > S_{th}$) are retained as final detections.

4. Experiments

Datasets: We evaluate D2-Net on multiple challenging temporal action localization benchmarks. The **THUMOS14** [6] dataset contains temporal annotations for 200 validation and 212 test videos from 20 action categories. The dataset is challenging since each video contains 15 action instances on an average. As in [25, 1], the validation and test set are used for training and evaluating, respectively. The **ActivityNet1.2** [3] dataset has annotations of 100 categories in 4819 training and 2383 validation videos, with 1.5 activity instances per video on an average. As in [31, 25], we use the training and validation sets to respectively train and evaluate.
Implementation details: For each snippet, 2048-d features are extracted from RGB and Flow I3D models pre-trained on Kinetics [4]. The kernel size and dilation rate of the temporal convolutional layers are: (3, 1) for THUMOS14 and (5, 2) for ActivityNet1.2. The first two convolutions in each stream are followed by a leaky ReLU with 0.2 negative slope. Our D2-Net is trained with a mini-batch size of 10 for 20K iterations, using the Adam [9] optimizer with a 10^{-4} learning rate and 0.005 weight decay. The k for *top-k* is set to $\lceil s/8 \rceil$, as in [25, 22]. All the hyperparameters are chosen via cross-validation. The balancing parameter α is set to 0.2 and 10^{-3} for THUMOS14 and ActivityNet1.2. The intra-class compactness weight γ and focusing parameter β are set to 0.01 and 2 for both datasets. Multiple thresholds from 0.025 to 0.5 with increments of 0.025 are used for proposal generation. The NMS threshold is set to 0.5 while the score threshold S_{th} for retaining detections in a video is set to 10% of the maximum proposal score in that video.

Table 1. **State-of-the-art comparison** on the THUMOS14 dataset. Methods with superscript '+' require strong frame-level supervision for training. Our D2-Net performs favorably in comparison to existing weakly-supervised methods and achieves consistent improvements, in terms of mean average precision (mAP).

Approach	mAP @ IoU				
	0.1	0.2	0.3	0.4	0.5
R-C3D [36]$^+$	54.5	51.5	44.8	35.6	28.9
GTAD [37]$^+$	-	-	54.5	47.6	40.2
TAL-Net [5]$^+$	59.8	57.1	53.2	48.5	42.8
P-GCN [40]$^+$	**69.5**	**67.8**	**63.6**	**57.8**	**49.1**
Autoloc [31]	-	-	35.8	29.0	21.2
W-TALC [25]	53.7	48.5	39.2	29.9	22.0
CMCS [14]	57.4	50.8	41.2	32.1	23.1
BM [24]	64.2	59.5	49.1	38.4	27.5
3C-Net [22]	59.1	53.5	44.2	34.1	26.6
BaS-Net [11]	58.2	52.3	44.6	36.0	27.0
DGAM [29]	60.0	54.2	46.8	38.2	28.8
DML [8]	62.3	-	46.8	-	29.6
A2CL-PT [20]	61.2	56.1	48.1	39.0	30.1
EM-MIL [16]	59.1	52.7	45.5	36.8	30.5
ACM-BANet [21]	64.6	57.7	48.9	40.9	32.3
HAM-Net [7]	65.4	59.0	50.3	41.1	31.0
UM [12]	**67.5**	**61.2**	**52.3**	**43.4**	33.7
ASL [18]	67.0	-	51.8	-	31.1
CoLA [42]	66.2	59.5	51.5	41.9	32.2
Ours: D2-Net	65.7	60.2	**52.3**	**43.4**	**36.0**

4.1. State-of-the-art Comparison

Tab. 1 and 2 compare D2-Net with state-of-the-art methods on THUMOS14 and ActivityNet1.2, respectively. Methods with '+' require strong supervision for training.
THUMOS14: Similar to ours, all weakly-supervised methods in Tab. 1 use an I3D backbone, except Autoloc [31], which uses TSN [35]. While BM [24] considers an additional background class, DGAM [29] extends BM using a VAE [10]. Although DML [8] and EM-MIL [16] achieve a promising mAP of 29.6 and 30.5 at IoU=0.5, they do not generalize well to ActivityNet1.2 (see Tab. 2). As discussed earlier, the recent work of UM [12] employs out-of-distribution detection of background snippets. We also empirically

Table 2. **State-of-the-art comparison** on the ActivityNet1.2 dataset. Our D2-Net performs favorably compared to existing weakly-supervised approaches. Furthermore, our D2-Net performs comparably to SSN [43], which is trained with strong supervision (denoted with superscript '+'). AVG denotes the mean of the mAP values for IoU in $[0.5, 0.95]$ with steps of 0.05.

Approach	mAP @ IoU			AVG
	0.5	0.75	0.95	
SSN [43]$^+$	**41.3**	**27.0**	**6.1**	**26.6**
DML [8]	35.2	-	-	-
EM-MIL [16]	37.4	-	-	20.3
CMCS [14]	36.8	22.0	5.6	22.4
3C-Net [22]	37.2	-	-	21.7
BaS-Net [11]	38.5	24.2	5.6	24.3
DGAM [29]	41.0	23.5	5.3	24.4
UM [12]	41.2	**25.6**	**6.0**	25.9
ASL [18]	40.2	-	-	25.8
Ours: D2-Net	**42.3**	25.5	5.8	**26.0**

validate the complementarity of our approach with UM by intergrating the loss terms and observe an average gain of 1% mAP across different IoUs. Our D2-Net performs well against existing weakly-supervised approaches, including the recent CoLA [42] and ASL [18]. Our approach achieves an absolute gain of 2.3% at IoU=0.5 over the best existing method (UM). Moreover, promising localization performance is obtained at other IoU thresholds.

ActivityNet1.2: Similar to our D2-Net, all weakly-supervised methods in Tab. 2 use I3D backbone. Following standard evaluation protocol [3], we report the mean of the mAP scores (denoted as AVG) at different IoU thresholds ($[0.5, 0.95]$ in steps of 0.05). The generative modeling based approach DGAM [29] and background suppression based BaS-Net [11] perform comparably, achieving mean mAP scores of 24.4 and 24.3, respectively. In comparison, the recent approaches such as UM [12] and ASL [18] achieve localization performances of 25.9 and 25.8, respectively, in terms of mean mAP. Our proposed D2-Net performs comparably against these existing approaches and achieves a promising localization performance of 26.0 mean mAP. Additional results are provided in the supplementary.

4.2. Ablation Study

As discussed earlier, our D2-Net comprises a discriminative $\mathcal{L}_{Dis}$ and a denoising loss $\mathcal{L}_D$. Here, we perform comparisons by replacing the two proposed loss terms ($\mathcal{L}_{Dis}$ and $\mathcal{L}_D$) in our framework with either the standard cross-entropy loss $\mathcal{L}_{CE}$ or the focal loss $\mathcal{L}_F$. In addition, we also show the performance of our D2-Net with only $\mathcal{L}_{Dis}$. Tab. 3 presents these performance comparisons, in terms of mAP and F1, on THUMOS14. Employing a standard cross-entropy loss ($\mathcal{L}_{CE}$ in Tab. 3) in our framework results in an mAP score of 23.0 at IoU=0.5. We observe that

Table 3. **Performance comparison** by replacing our two loss terms ($\mathcal{L}_{Dis}$ and $\mathcal{L}_D$) in the proposed D2-Net with either the standard cross-entropy loss ($\mathcal{L}_{CE}$) or the focal loss ($\mathcal{L}_F$). In addition, we also show the performance of our D2-Net with only $\mathcal{L}_{Dis}$. Results are shown in terms of mAP and F1 score at IoU=0.5, on THUMOS14. Replacing the proposed loss terms in our framework with $\mathcal{L}_{CE}$ and $\mathcal{L}_F$ results in mAP scores at IoU=0.5 of 23.0 and 26.7, respectively. Our D2-Net with the discriminative loss term $\mathcal{L}_{Dis}$ achieves consistent improvement in performance over $\mathcal{L}_F$ with an absolute gain of 5.5% in terms of mAP at IoU=0.5. Furthermore, our final D2-Net comprising both loss terms ($\mathcal{L}_{Dis}$ and $\mathcal{L}_D$) achieves the best performance with absolute gains of 12.9% and 9.2% in terms of mAP at IoU=0.5 over $\mathcal{L}_{CE}$ and $\mathcal{L}_F$, respectively.

Loss term	mAP @ IoU					F1
	0.1	0.2	0.3	0.4	0.5	
$\mathcal{L}_{CE}$	55.0	47.6	38.7	30.7	23.0	23.5
$\mathcal{L}_F$	58.8	52.4	44.3	35.7	26.7	27.2
$\mathcal{L}_{Dis}$	65.4	59.7	50.1	40.4	32.2	30.7
D2-Net: $\mathcal{L}_{Dis} + \mathcal{L}_D$	**65.7**	**60.2**	**52.3**	**43.4**	**36.0**	**36.7**

Table 4. **Impact of MI-based denoising** on THUMOS14. Our D2-Net, employing MI-based pDMI loss in $\mathcal{L}_D$ performs favorably compared to utilizing standard losses (L1 and BCE) in $\mathcal{L}_D$.

	L1	BCE	**Ours: D2-Net**
mAP at IoU=0.5	32.9	33.5	**36.0**

training with the standard focal loss (obtained by zeroing the weights w in Eq. 4) helps alleviate the issue of a large number of easy samples overwhelming hard samples. This setting, $\mathcal{L}_F$ in Tab. 3, gains 3.7% mAP at IoU=0.5 over $\mathcal{L}_{CE}$, thereby highlighting the need to tackle imbalance between easy backgrounds and hard foregrounds. To the best of our knowledge, we are the first to evaluate the standard focal loss, $\mathcal{L}_F$, in weakly-supervised action localization setting. Our D2-Net with the discriminative loss term $\mathcal{L}_{Dis}$, which jointly addresses class-imbalance and enhances background-foreground separation, provides consistent improvements over $\mathcal{L}_F$ and achieves 32.2% mAP at IoU=0.5. An absolute gain of 5.5% in terms of mAP at IoU=0.5 is obtained by the introduction of our proposed $\mathcal{L}_{Dis}$ in place of $\mathcal{L}_F$. Furthermore, our D2-Net comprising both $\mathcal{L}_{Dis}$ and $\mathcal{L}_D$ obtains the best results with an mAP score of 36.0% at IoU=0.5. Our D2-Net achieves absolute gains of 12.9% and 9.2% in terms of mAP at IoU=0.5, over $\mathcal{L}_{CE}$ and $\mathcal{L}_F$, respectively. It is noteworthy that our final D2-Net, containing both $\mathcal{L}_{Dis}$ and $\mathcal{L}_D$, obtains a significant gain of 5.9% in terms of F1 score over $\mathcal{L}_{Dis}$ alone. This improvement over $\mathcal{L}_{Dis}$ alone is obtained due to explicitly addressing the noise in TCAMs by our $\mathcal{L}_D$, leading to a substantial reduction (28%) in the number of false positives without affecting the recall.

Impact of MI-based denoising: We also perform an experiment by replacing the proposed pDMI loss in our $\mathcal{L}_D$ with the standard L1 and BCE losses for denoising the snippet-level activations. The L1 and BCE losses, which do not

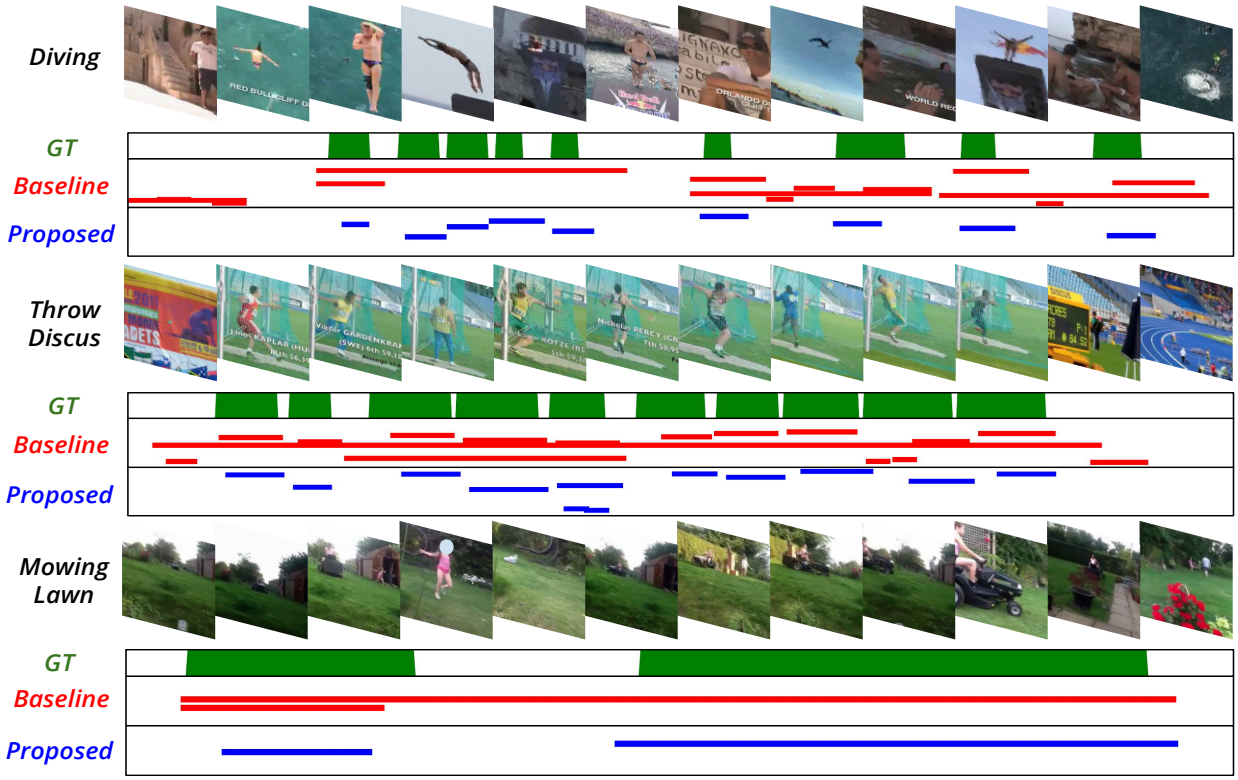

Figure 5. **Qualitative temporal action localization results** of our proposed D2-Net on example test videos, with *Diving*, *Throw Discus* actions from THUMOS14, and *Mowing Lawn* activity from ActivityNet1.2. For each video, example frames (top row), ground-truth GT segments (green), baseline detections (red) and D2-Net detections (blue) are shown. The height of a detection is indicative of its score. The Baseline incorrectly merges multiple GT instances, has false positives in background regions and falsely detects the presence of the activity over the entire video length. Our D2-Net correctly detects multiple instances (*e.g.*, 1 to 5 GT in *Diving*, 3 to 5 in *Throw Discus*) and suppresses most false positives in the background regions, achieving promising localization performance.

explicitly capture MI, achieve mAP scores of 32.9% and 33.5% at IoU=0.5, respectively, on THUMOS14 (see Tab. 4). Our D2-Net, which employs MI-based pDMI loss in $\mathcal{L}_D$, achieves improved results with an mAP score at IoU=0.5 of 36.0%. These results suggest that our MI-based denoising is able to robustify the TCAMs in a weakly-supervised setting.

Qualitative results: Fig. 5 shows a qualitative comparison between the baseline (red) and D2-Net (blue), along with the ground-truth (GT) action segments (green). The baseline employs only $\mathcal{L}_F$ and is the same as the one used in Fig. 1. Example test videos with *Diving* and *Throw Discus* actions from THUMOS14 are shown in the first two rows. The baseline incorrectly merges multiple GT instances (*e.g.*, 1 to 5 GT in *Diving*) and produces false positives in background regions (*e.g.*, towards the beginning of *Diving* video). Our D2-Net correctly detects these multiple action instances and suppresses most false positives in the background regions. The third row shows an example test video with *Mowing Lawn* activity from ActivityNet1.2. The baseline incorrectly detects the presence of the activity over the entire video length. In contrast, our D2-Net improves the detection of multiple activity instances, leading to promising localization performance. Additional results and discussions are provided in the supplementary.

5. Conclusion

We propose a weakly-supervised action localization approach, called D2-Net, that comprises a discriminative and a denoising loss. The discriminative loss term strives for improved foreground-background separability through interlinked classification and localization objectives. The denoising loss term complements the discriminative term by tackling the foreground-background noise in the activations. This is achieved by maximizing the mutual information between activations and labels within a video (intra-video) and across videos (inter-video). Comprehensive experiments performed on multiple benchmarks show that our D2-Net performs favorably against existing methods on all datasets.

Acknowledgements

This work is partially supported by ARC DECRA Fellowship DE200101100, NSF CAREER Grant #1149783 and VR starting grant 2016-05543.

References

[1] Humam Alwassel, Alejandro Pardo, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. *arXiv preprint arXiv:1904.00227*, 2019. 2, 6

[2] 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. 2

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 6, 7

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 3, 6

[5] 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. 1, 6

[6] 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". *CVIU*, 2017. 2, 6

[7] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. *arXiv preprint arXiv:2101.00545*, 2021. 6

[8] Ashraful Islam and Richard Radke. Weakly supervised temporal action localization using deep metric learning. In *WACV*, 2020. 2, 6, 7

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

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014. 6

[11] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *AAAI*, 2020. 6, 7

[12] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. *arXiv preprint arXiv:2006.07006*, 2020. 2, 6, 7

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4

[14] Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019. 6, 7

[15] 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. 1

[16] 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. *arXiv preprint arXiv:2004.00163*, 2020. 2, 6, 7

[17] 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. 2

[18] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *CVPR*, 2021. 6, 7

[19] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 2

[20] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. *arXiv preprint arXiv:2007.06643*, 2020. 6

[21] Md Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C Leu. Action completeness modeling with background aware networks for weakly-supervised temporal action localization. In *ACMMM*, 2020. 2, 6

[22] 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. 1, 2, 3, 6, 7

[23] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 2018. 1, 2, 3

[24] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *ICCV*, 2019. 1, 2, 6

[25] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018. 1, 2, 3, 6

[26] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *CVPR*, 2017. 2

[27] Scott Satkin and Martial Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010. 1

[28] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? In *CVPR*, 2008. 1

[29] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*, 2020. 6, 7

[30] 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. 1

[31] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, 2018. 1, 2, 6

[32] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 1

[33] 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. 1, 2

[34] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 1, 2

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2, 6

[36] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 1, 6

[37] 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, 6

[38] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, 2019. 4

[39] 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. 2

[40] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019. 6

[41] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2010.11594*, 2020. 2

[42] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, 2021. 6, 7

[43] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 1, 7