PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization

Mamshad Nayeem Rizve*‡ Gaurav Mittal*† Ye Yu† Matthew Hall† Sandra Sajeev†
Mubarak Shah‡ Mei Chen†
†Microsoft ‡University of Central Florida
{gaurav.mittal, yu.ye, mathall, ssajeev, mei.chen}@microsoft.com
nayeemrizve@knights.ucf.edu shah@crcv.ucf.edu

Abstract

*Weakly-supervised Temporal Action Localization (WTAL) attempts to localize the actions in untrimmed videos using only video-level supervision. Most recent works approach WTAL from a localization-by-classification perspective where these methods try to classify each video frame followed by a manually-designed post-processing pipeline to aggregate these per-frame action predictions into action snippets. Due to this perspective, the model lacks any explicit understanding of action boundaries and tends to focus only on the most discriminative parts of the video resulting in incomplete action localization. To address this, we present PivoTAL, **Piv**or-driven Super**v**isi**o**n for Weakly-supervised **T**emporal **A**ction **L**ocalization, to approach WTAL from a localization-by-localization perspective by learning to localize the action snippets directly. To this end, PivoTAL leverages the underlying spatio-temporal regularities in videos in the form of action-specific scene prior, action snippet generation prior, and learnable Gaussian prior to supervise the localization-based training. PivoTAL shows significant improvement (of at least 3% avg mAP) over all existing methods on the benchmark datasets, THUMOS-14 and ActivitNet-v1.3.*

1. Introduction

Temporal action localization (TAL) [5,24,43,46,47,56] refers to the task of predicting where and what category of action happens in an arbitrarily long untrimmed video. While TAL is crucial in a wide variety of applications ranging from sports, robotics, and safety, it is challenging as it requires the model to develop a strong temporal and spatial understanding of the video scene and events for effective localization. Furthermore, fully-supervised TAL relies on the availability of expensive dense annotations in terms of the start and end of each action snippet in the training videos.

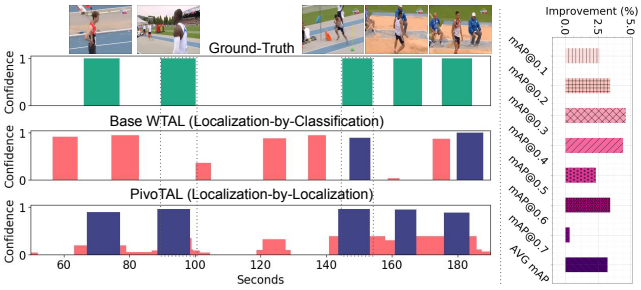

Figure 1. Left: Green denotes ground truth action snippets, Purple denotes true positive action snippets, and Pink denotes false positive action snippets. We observe that PivoTAL significantly outperforms Base WTAL by detecting all the ground-truth instances correctly. Right: Improvement of PivoTAL over the previous state-of-the-art on THUMOS'14 at different IoU thresholds.

Weakly-supervised Temporal Action Localization (WTAL) serves to mitigate this dependency on dense annotations by operating only on video-level annotations (*i.e.*, knowing which actions occur without knowing their precise locations in a video) during training while still being able to predict the start and end of the action snippet in test videos. Several methods [20, 32, 33, 38, 45, 50, 53] have attempted to perform WTAL by employing different techniques which include Multiple Instance Learning (MIL) [20, 33] and attention mechanism [45, 50]. However, to the best of our knowledge, all of these previous works approach WTAL from a *localization-by-classification* perspective where the underlying method tries to classify each video frame into zero or more action categories followed by a manually-designed post-processing pipeline to aggregate these per-frame action predictions into action snippets with explicit boundaries.

Fig. 1 shows the final action snippet predictions on a video with *Long Jump* action from a typical *localization-by-classification* method which we refer to as *Base WTAL* in the figure. We observe that *Base WTAL* suffers from some

*Authors with equal contribution.
This work was done as Mamshad's internship project at Microsoft.

challenges. First, the localization-by-classification training is performed only with the coarse video-level labels, which encourages the model to focus on the most discriminative parts of the video, resulting in incomplete and fragmented action snippets (140-160s for Base WTAL in Fig. 1). There is also a higher rate of false positives due to the misclassification of background that closely resembles foreground (at $\sim$60s for Base WTAL in Fig. 1). Second, since the model is trained to perform per-frame prediction, it lacks any explicit notion of action boundaries, thus resulting in a discrepancy between the classification-based training and localization-based test objectives. This is generally addressed by incorporating carefully-designed post-processing algorithms. Even though such post-hoc transformations can encode crucial prior knowledge of temporal structure of videos, they cannot influence the model training for improving the localization performance directly.

To resolve this discrepancy, we propose PivoTAL, **Pri**or-driven Super**vis**i**o**n for weakly-supervised **T**emporal **A**ction **L**ocalization. PivoTAL approaches WTAL from *localization-by-localization* perspective by learning to localize the action snippets directly. To this end, PivoTAL introduces a novel algorithm that exploits the inherent spatio-temporal structure of the video data in the form of *action-specific scene prior*, *action snippet generation prior*, and *learnable Gaussian prior* to derive pseudo-action snippets. These pseudo-action snippets act as an additional source of supervision in PivoTAL to complement the under-constrained video-level weak-supervision to perform the localization task.

PivoTAL first employs a *Base WTAL Head* to perform weakly-supervised temporal action localization using video-level supervision. While doing so, PivoTAL employs a novel *action-specific scene prior* in the background MIL loss to inject action-specific bias into the background frames to improve action boundaries (at 180s for PivoTAL in Fig. 1). PivoTAL also complements the per-frame actionness scores learned by the model with *learnable Gaussian prior*-based actionness scores to incorporate context from nearby frames and to improve the smoothness of predicted action snippets (140-160s for PivoTAL in Fig. 1). Next, PivoTAL creates pseudo-action snippets by employing *action snippet generation prior* and makes them confidence-aware using the confidence predictions of the *Base WTAL Head*. Finally, these pseudo-action snippets are used to train the *Prior-driven Localization Head* of the model to predict the action snippets directly. We conduct extensive experiments on the standard WTAL datasets, THUMOS'14 and ActivityNet-v1.3, achieving 3.2% and 3.0% absolute improvement on the average mAP respectively over all previous methods. This demonstrates PivoTAL's advantage in effectively utilizing the priors, leading to a significant improvement in the localization performance.

Our work makes the following major contributions,

1. We introduce PivoTAL, the first method to approach WTAL from a *localization-by-localization* perspective by generating pseudo-action snippets as supervision to localize action snippets directly.

2. In the process, PivoTAL exploits the underlying spatio-temporal regularities in videos in the form of *action-specific scene prior*, *action snippet generation prior*, and *learnable Gaussian prior* to complement the available weak video-level supervision.

3. PivoTAL significantly outperforms all previous methods on WTAL benchmarks THUMOS'14 and ActivityNet, with 3% or higher absolute increase on average mAP metric.

2. Related Work

Temporal Action Localization. The fully-supervised methods for temporal action localization can be broadly divided into two categories: anchor-based methods [3, 46, 49, 59] and anchor-free methods [23–26, 56]. The anchor-based methods learn the action boundaries by performing regression based on a pre-defined set of action proposals. Because of relying on pre-defined anchors, these methods tend to perform poorly on actions which are extremely short or long. The anchor-free methods mitigate this by explicitly predicting the action offset and probability for each clip. The primary difference between these methods and PivoTAL is that these methods require expensive per-clip annotations while we only use video-level labels.

Weakly-Supervised Temporal Action Localization. The recent weakly-supervised temporal action localization (WTAL) methods can be broadly classified into two categories: single-stage, and multi-stage methods. The single-stage WTAL methods can be further divided into three main categories: MIL-based [20, 33, 35, 40], attention-based [11, 31, 36, 41, 45, 50], and erasing-based [19, 34, 55, 61] methods. The MIL-based methods are the simplest of these single-stage methods which treat a video as a bag consisting of positive samples (i.e., clips corresponding to foreground actions) and negative samples (i.e., background clips). For training, MIL-based methods perform top-k positive sample selection and aggregate their prediction to train with a video-level label. Attention-based methods try to avoid top-k based hard selection and perform class-agnostic foreground actionness based attentional pooling to aggregate the clip-level scores to obtain video-level predictions for training. The erasing-based methods take an adversarial complementary learning approach [57] to address the WTAL methods' tendency to focus on the most discriminative parts. To

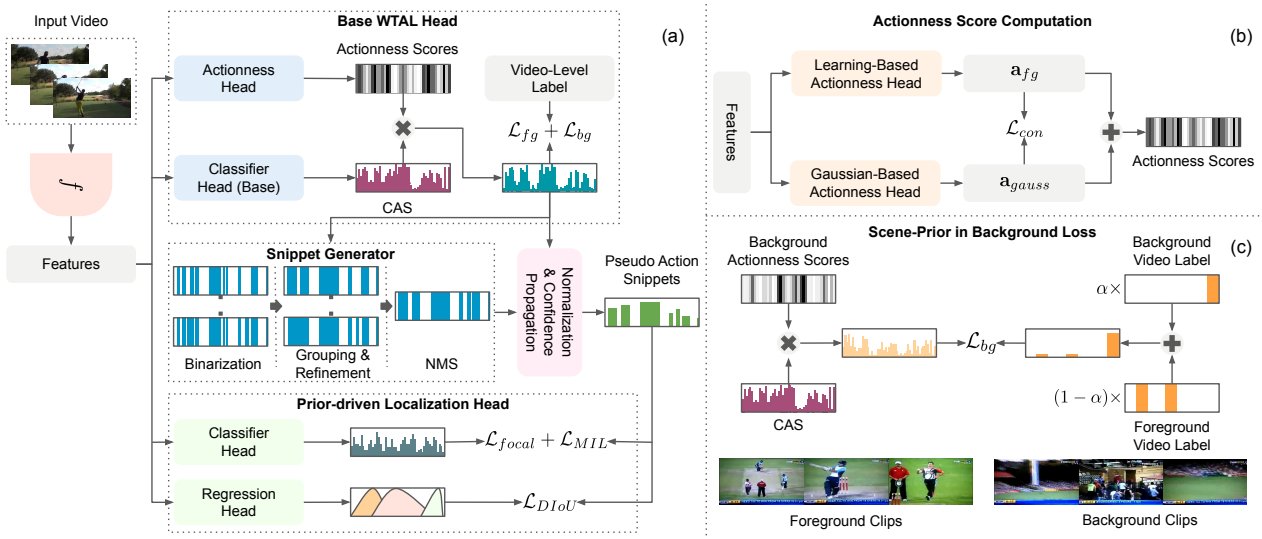

Figure 2. PivoTAL Overview: (a) **Training pipeline:** After processing a video through feature extractor f, we obtain class-agnostic Actionness and class activation sequences (CAS) scores from the features via Actionness Head and Classifier Head (Base) respectively. We enhance the CAS scores further using the Actionness scores via Hadamard product and apply foreground, $\mathcal{L}_{fg}$, and background, $\mathcal{L}_{bg}$, MIL losses w.r.t video-level weak label. Once we finish training this *Base WTAL Head*, we apply Snippet Generator SG, including Binarization, Grouping & Refinement and NMS, on the *Base WTAL Head* confidence predictions to obtain hard action snippets. Then we convert them to soft pseudo-action snippets using confidence propagation from *Base WTAL Head* confidence predictions and perform per-class confidence normalization. We finally train the *Prior-driven Localization Head* using the pseudo-action snippets as ground-truth to predict and localize the action snippets directly. (b) **Actionness Score Computation:** Our Actionness Head consists of a Learning-based and a Gaussian prior-based Actionness Head which both process input features to obtain corresponding actionness scores that are averaged to obtain the final Actionness Scores. We minimize a consistency loss as a regularizer to reduce disagreement between output of the two Actionness heads. (c) **Scene prior in Background MIL loss:** We minimize the Background MIL loss between the background-specific CAS scores (yellow plot, center-left) and the composite background label created by combining foreground and background video label (center-right) to inject action-specific scene prior. From the bottom frames, we observe that even background frames (photos of the stadium) are relevant to the foreground actions of *Cricket Bowling* and *Cricket Shot*.

this end, these methods try to increase the weight of less discriminative parts of the video. Our proposed solution is complementary to all these approaches since we primarily focus on incorporating existing human priors into training to perform *localization-by-localization*.

The multi-stage training-based methods [9,32,38,51,53] generally take a self-training approach. The primary objective of these methods is to generate per-clip pseudo-labels from an initial WTAL model and then perform further training with those generated pseudo-labels. Technically such pseudo-label-based self-training can be repeated for multiple iterations [9, 38]. The primary difference between our work and these methods is that these methods do not utilize the action snippets explicitly and therefore, do not have an explicit notion of action boundaries. Therefore, this per-clip pseudo-label-based self-training still falls into the *localization-by-classification* category.

3. Method

PivoTAL attempts to solve the under-constrained weakly supervised temporal action localization (WTAL) task from a *localization-by-localization* perspective. Figure 2 provides an overview of our method.

3.1. Preliminaries

In the WTAL setting, we assume that we have access to a set of weakly labeled videos $\mathbb{V} = \{\mathbf{v}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, where N represents the total number of samples, $\mathbf{v}^{(i)}$ represents an untrimmed input video, and $\mathbf{y}^{(i)}$ represents the set of action classes present in video $\mathbf{v}^{(i)}$ with no information about their precise locations in the video. Specifically, we represent $\mathbf{y}^{(i)}$ with a multi-label one-hot encoding such that $\mathbf{y}^{(i)} \in \{0,1\}^{C+1}$, where C is the total number of action classes present in the dataset, we add an additional class to model the background. During inference, the objective is to predict a set of action snippets $\mathbb{A}^{(i)} = \{c_j, s_j, e_j\}_{j=1}^{M}$ for video $\mathbf{v}^{(i)}$ where M is the total number of action snippets, c_j is the predicted class, s_j is the start time, and e_j is the end time of a particular action snippet j. We denote $\mathbf{v}^{(i)}$ and $\mathbf{y}^{(i)}$ with $\mathbf{v}$ and $\mathbf{y}$ respectively in the subsequent text for simplicity.

Baseline Approach for WTAL. As shown in Figure 2a, PivoTAL comprises a *Base WTAL Head* to perform the MIL-based WTAL using video-level supervision. Since we are working with arbitrarily long untrimmed video for the WTAL task, it is computationally prohibitive to encode the entire video in a single forward pass through a feature

encoder. Therefore, following prior works [9, 14, 40, 54], we split a given video into multiple small clips i.e. $\mathbf{v} = \{\mathbf{g}_k\}_{k=1}^{T}$ where T is the total number of clips present in that video. As shown in Figure 2a, we then process these clips using a feature extractor, f, to obtain the feature embeddings s.t. $f : \mathbf{g} \mapsto \mathbf{z}_g$, where $\mathbf{z}_g \in \mathbb{R}^d$. We obtain the video-level feature, $\mathbf{z}_v$, by concatenating all the clip level features such that $\mathbf{z}_v \in \mathbb{R}^{T\times d}$. Next, we process these video-level features, $\mathbf{z}_v$, using a classifier to project them into the output space to obtain the class activation sequence (CAS) scores, $\mathbf{q}$, s.t. $\mathbf{q} \in \mathbb{R}^{T\times(C+1)}$. In parallel, following prior art [9, 38, 41], we employ an actionness score generator modeled using a linear layer to obtain class-agnostic actionness scores, $\mathbf{a}$, s.t. $\mathbf{a} \in \mathbb{R}^{T\times 2}$, where we use a 2-dimensional output vector for each clip to model both foreground and background actionness scores (Figure 2a).

For training, we only have access to the video-level labels, $\mathbf{y}$. Therefore, following prior work [16, 20, 40, 41], we train the network using a multiple instance learning (MIL) based classification loss. To this end, first, we enhance the CAS scores, $\mathbf{q}$, by taking a Hadamard product with the class-agnostic foreground actionness scores $\mathbf{a}_{fg}$. Next, we perform top-k selection followed by average pooling across the temporal dimension to obtain the video-level foreground classification logits, $\hat{\mathbf{y}}_{fg}$ s.t. $\hat{\mathbf{y}}_{fg} = 1/K \sum_{k=1}^{K} \text{topK}(\mathbf{a}_{fg} \odot \mathbf{q})$, where $\odot$ is the Hadamard product operator. Finally, we use cross-entropy loss to optimize the network parameters as,

$$\mathcal{L}_{fg} = -\sum_{c=1}^{C+1} \mathbf{y}(c) \log \hat{\mathbf{y}}_{fg}(c). \tag{1}$$

While the foreground loss in equation 1 can help to localize the action snippets, the model still underperforms due to the absence of any explicit loss to reduce false positives. Therefore, following prior work [37, 41, 41], we generate complimentary labels, $\mathbf{y}_{bg}$, by setting the background class to 1 and all other action classes to 0 in $\mathbf{y}$. Next, we obtain video-level background logits, $\hat{\mathbf{y}}_{bg}$ s.t. $\hat{\mathbf{y}}_{bg} = 1/K \sum_{k=1}^{K} \text{topK}(\mathbf{a}_{bg} \odot \mathbf{q})$, where $\mathbf{a}_{bg}$ is the background actionness score. After that, we compute a background loss $\mathcal{L}_{bg}$ in the following manner,

$$\mathcal{L}_{bg} = -\sum_{c=1}^{C+1} \mathbf{y_{bg}}(c) \log \hat{\mathbf{y}}_{bg}(c). \tag{2}$$

We finally optimize the *Base WTAL Head*, as shown in Figure 2a, using the combined loss, $\mathcal{L}_{base} = \mathcal{L}_{fg} + \mathcal{L}_{bg}$.

CAS to Action Snippet Generation. As we can see from Equation 1 and 2, the WTAL objective does not train the model for localization but instead trains for classifying the clips, $\mathbf{g}$, to predict the CAS scores. Since the final objective is to generate action snippets, $\mathbb{A}$, containing explicit start and end times along with the action label, we need to convert CAS to $\mathbb{A}$. However, this is a non-trivial task since the network lacks any explicit notion of action boundaries and the means to aggregate the CAS score to form $\mathbb{A}$.

To address this discrepancy, it is common to introduce manual priors as post processing to transform CAS scores, $\mathbf{q}$, into action snippets, $\mathbb{A}$. The transformation involves multiple steps. First, the CAS scores are binarized with a broad range of thresholds. This is followed by generating connected components from the binarized CAS scores to form the initial set of action snippets. Next, some morphological operations (such as erosion and dilation) are applied to refine the action snippet boundaries. Finally, non-maximum suppression (NMS) operation is performed to obtain the best candidate action snippets, $\mathbb{A}$.

As is evident from the above, generating action snippets from CAS scores involves injecting a series of manual priors in the WTAL task. We empirically find each of these operations playing a significant role in the downstream performance, but at the same time, are rarely being mentioned in text in existing approaches and are only found in their code repositories. These priors make it possible to deal with the under-constrained task of action localization with weak supervision. Therefore, in our work, we incorporate them into our training pipeline and holistically refer to them as the snippet generator function, SG s.t. SG : $\mathbf{q} \mapsto \mathbb{A}$ (Fig. 2a).

3.2. Prior-driven Weak Localization

While SG allows for action snippet, $\mathbb{A}$, generation from model outputs, $\mathbf{q}$, the *Base WTAL Head* alone is not optimal since it cannot directly output action snippets. Moreover, the priors are introduced after training and cannot influence model optimization. To address this, PivoTAL introduces prior-driven weak localization by integrating the action snippet generation prior into WTAL training. We discuss the potential approaches to do so in the following.

Self-Training with Hard Pseudo-Action Snippets. One straightforward way to incorporate the priors encoded in SG into a WTAL method is to perform self-training with the generated hard pseudo-action snippets so that the network can explicitly learn the action boundaries. However, this strategy is also not optimal since some of the pseudo-action snippets generated using manual priors will be noisy. One way to deal with this noise would be to incorporate a denoising mechanism into training [8, 17, 22, 42]. However, this will require incorporating additional components into the design, like multiple networks, sample selection, and subsequent semi-supervised training, etc.

Self-Training with Soft Pseudo-Action Snippets. In contrast, in this work, we propose a simple solution that leverages both human priors and the distilled knowledge from the available video-level weak annotations. To this end, we perform self-training with the pseudo-action snippets, $\mathbb{A}$, generated with SG, and also utilize the confidence of these action snippets obtained from the *Base WTAL Head*.

For this, we propagate the *Base WTAL Head*'s confidence predictions by setting the confidence of an action snippet as the average of the confidence scores within the span of the action snippet. We argue that this confidence-aware self-training strategy based on pseudo-action snippets, $\mathbb{A}$, takes advantage of both weak annotations and human priors. However, this strategy still has one remaining challenge, neural network predictions are not well calibrated and tend to be overconfident [7, 44], especially on the easy classes [18,39]. Therefore, a vanilla confidence propagation from the *Base WTAL Head* to the action snippets will not be optimal for the relatively harder and underrepresented classes. To address this issue, we propose to normalize the predictive confidence scores for each class independently.

To train based on pseudo-action snippets, we introduce *Prior-driven Localization Head*, as shown in Figure 2a. To train this head, we incorporate two more loss terms besides the MIL-based classification loss, $\mathcal{L}_{MIL}$. Following prior work on supervised temporal action localization [56], we use a focal loss [27], $\mathcal{L}_{focal}$, for per-clip action classification where the target is derived from the pseudo-action snippets. Next, to determine the action boundaries in an anchor-free manner, we predict the action offsets from each time step, and to achieve this objective, we use a DIoU based [60] regression loss, $\mathcal{L}_{DIoU}$. The target for the regression loss is also derived from the pseudo-action snippets. Therefore, our overall *localization-by-localization* training objective with the pseudo-action snippets is as follows:

$$\mathcal{L}_{loc} = \mathcal{L}_{focal} + \mathcal{L}_{DIoU} + \mathcal{L}_{MIL}. \tag{3}$$

Even though the proposed solution of learning from both human priors encoded in SG and weak supervision enables us to solve the WTAL task in a *localization-by-localization* manner, we expect that a WTAL method can benefit further from additional priors. The primary intuition behind this is that the pseudo-action snippets are generated from a base model that focuses on discriminative parts of actions due to localization-by-classification (Figure 1), and tends to err especially when the visual information around action boundaries is ambiguous. Therefore, injecting priors which can address such failure cases should improve the quality of the extracted pseudo-action snippets and the final localization.

3.3. Prior-driven Base WTAL

We inspect the design of the *Base WTAL Head* to inject additional priors into the learning process. From a high level, the *Base WTAL Head* has three main components: (i) class-agnostic actionness score generation, (ii) foreground MIL loss ($\mathcal{L}_{fg}$), and (iii) background MIL loss ($\mathcal{L}_{bg}$). We find that the actionness score generation and the background MIL loss are added to complement the foreground MIL loss. Therefore, we propose to add additional prior to these components.

Scene Prior in Background MIL Loss. Understanding temporal dynamics is essential for temporal action localization. Meanwhile, it is well established that spatial information can be a strong cue for recognizing actions [10,48,58]. We utilize this observation in the form of an action-specific scene prior into our objective function. In particular, we modify our background MIL loss to incorporate a foreground action specific prior. In typical MIL-based WTAL systems, the background loss is computed on the least probable video clips. However, we expect that even the least probable video clips contain foreground-related information. Therefore, instead of encouraging the model to predict only background class on the least probable video clips, we encourage the model to also predict the appropriate foreground class as the second most dominant class. For this, we generate a composite background label, $\tilde{\mathbf{y}}_{bg}$, which contains foreground-specific information (Fig. 2c). We formulate our modified background loss with scene prior as,

$$\begin{aligned} \tilde{\mathbf{y}}_{bg} &= \alpha \mathbf{y}_{bg} + (1-\alpha)\mathbf{y}, \\ \mathcal{L}_{bg} &= -\sum_{c=1}^{C+1} \tilde{\mathbf{y}}_{bg}(c) \log \hat{\mathbf{y}}_{bg}(c), \end{aligned} \tag{4}$$

where, α controls the strength of the background label.

Gaussian Prior for Actionness Prediction. Following prior works [9,38,41], in the Base WTAL approach, we predict the class agnostic actionness scores using a linear layer. The actionness score is generally determined on a per-clip basis without using any context of nearby clips. One naive way to improve this would be to incorporate more local context by utilizing long temporal convolutional kernels or attention mechanisms. However, effective optimization of such an actionness predictor is non-trivial since the available video-level labels lack any clip-level local information.

One of the primary motivations behind adding local context is that the predictions have to be locally consistent. To enforce this prior explicitly, we model the foreground actionness scores with learnable Gaussian masks. In particular, we obtain a second foreground actionness score, $\mathbf{a}_{gauss}$, by introducing a Gaussian mask prediction branch which predicts Gaussian kernels $\{\sigma_i, \mu_i\}_{i=1}^{T}$ for each clip to flexibly model the actionness scores (Fig. 2b). We generate clip-specific local Gaussian masks, $\mathbf{G}_i$, from the predicted parameter $\{\sigma_i, \mu_i\}$. To effectively preserve the local context in the final actionness scores, $\mathbf{a}_{gauss}$, we perform local selection from the clip-specific local masks $\mathbf{G}$ to generate $\mathbf{a}_{gauss}(i)$ by selecting the value at corresponding i^{th} temporal position from the i^{th} Gaussian mask, $\mathbf{G}_i$ as,

$$\begin{aligned} \mathbf{G}_i &= \exp\Big(-\frac{\beta(j/T-\mu_i)^2}{\sigma_i^2}\Big)_{j=1}^{T}, \\ \mathbf{a}_{gauss} &= \{\mathbf{G}_i(i)\}_{i=1}^{T}, \end{aligned} \tag{5}$$

where, β controls the variance of the Gaussian mask, $\mathbf{G}$.

Supervision	Method	mAP@IoU (%)							AVG		
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.5)	(0.3:0.7)	(0.1:0.7)
Full	SSN ICCV'17 [59]	60.3	56.2	50.6	40.8	29.1	-	-	49.6	-	-
	BSN ECCV'18 [26]	-	-	53.5	45.0	36.9	28.4	20.0	-	36.8	-
	GTAN CVPR'19 [30]	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-	-
Weak	CleanNet ICCV'19 [29]	-	-	37.0	30.9	23.9	13.9	7.1	-	22.6	-
	RPN AAAI'20 [12]	62.3	57.0	48.2	37.2	27.9	16.7	8.1	46.5	27.6	36.8
	TSCN ECCV'20 [53]	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	28.8	37.8
	EM-MIL ECCV'20 [32]	59.1	52.7	45.5	36.8	30.5	22.7	16.4	45.0	30.4	37.7
	A2CL-PT ECCV'20 [34]	61.2	56.1	48.1	39.0	30.1	19.2	10.6	46.9	29.4	37.8
	HAM-Net AAAI'21 [16]	65.4	59.0	50.3	41.1	31.0	20.7	11.1	49.4	30.8	39.8
	WUM AAAI'21 [21]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	32.9	41.9
	AUMN CVPR'21 [31]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4	41.5
	CoLA CVPR'21 [54]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	TS-PCA CVPR'21 [28]	67.6	61.1	53.4	43.4	34.3	24.7	13.7	52.0	33.9	42.6
	UGCT CVPR'21 [51]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	ASL CVPR'21 [33]	67.0	-	51.8	-	31.1	-	11.4	-	-	-
	CO2-Net MM'21 [11]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.6	44.6
	D2-Net ICCV'21 [36]	65.7	60.2	52.3	43.4	36.0	-	-	51.5	-	-
	FAC-Net ICCV'21 [13]	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	33.1	42.2
	ACG-Net AAAI'22 [52]	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	33.4	42.5
	ASM-Loc CVPR'22 [9]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	RSKP CVPR'22 [14]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	DELU ECCV'22 [4]	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.5	37.4	46.4
	PivoTAL (Ours)	74.1	69.6	61.7	52.1	42.8	30.6	16.7	**60.1**$_{\uparrow 3.6}$	**40.8**$_{\uparrow 3.4}$	**49.6**$_{\uparrow 3.2}$

Table 1. Temporal action localization performance comparison with state-of-the-art methods on the **THUMOS-14** dataset. PivoTAL outperforms all existing methods on all different IoU thresholds and achieves at least 3.2% better average mAP than all existing methods.

Even though generating foreground actionness scores based on Gaussian priors can generate locally smooth actionness scores, we observe that integrating it with the learning-based actionness scores is not straightforward. We experiment with different aggregation strategies and observe that the performance deteriorates with any of these aggregation strategies. We hypothesize that this happens because of the inconsistency/disagreement between these two actionness scores. To resolve this issue, we introduce an actionness consistency loss as defined below.

$$\mathcal{L}_{con} = \sum_{i} (\mathbf{a}_{fg}(i) - \mathbf{a}_{gauss}(i))^2. \tag{6}$$

Therefore, the overall loss to train the *Base WTAL Head* is $\mathcal{L}_{base} = \mathcal{L}_{fg} + \mathcal{L}_{bg} + \mathcal{L}_{con}$.

4. Experimental Evaluation

Datasets. We evaluate our method on the two standard datasets for weakly-supervised action localization: THUMOS-14 [15] and ActivityNet-v1.3 [1]. THUMOS-14 contains 20 action classes. We use the 200 untrimmed videos in the validation set as our training set and test the model on a set of 212 test videos. ActivityNet-v1.3 contains 200 action classes. We use the 10,024 videos from the training set to train our model and use the 4,926 videos from the validation set to test our model.

Implementation Details. To extract clip level features, $\mathbf{z}_g$, following prior works [9, 51, 54], we use a I3D network [2] pretrained on the Kinetics-400 [2] dataset. We use both RGB and optical flow [6] features. We train the WTAL head for 150 and 50 epochs on THUMOS-14 and ActivityNet-v1.3 datasets, respectively. We use Adam optimizer with a learning rate of $1e-4$. We set the value of α to 0.8, set β to 0.1, and set the coefficient of all loss terms to 1.0. For the prior-driven localization head training, we use a similar network architecture as [56]. Finally, for a fair comparison, we do not optimize the parameters of the snippet generator, SG, but rather use the same parameters as [9].

4.1. Comparison with State-of-the-Art

We compare the performance of PivoTAL with the existing state-of-the-art methods on the THUMOS-14 dataset in Table 1. Following prior works [9, 14], we also report results from a few representative fully-supervised TAL methods for reference. Table 1 shows that PivoTAL outperforms all the previous methods by establishing a new state-of-the-art with 49.6% avg mAP over IoU thresholds 0.1:0.7. We can observe that PivoTAL outperforms the existing best method DELU [4] by more than 3%. Similar improvements are observed for other avg mAP scores. Note that PivoTAL is the only method that outperforms all the previous methods at each individual IoU threshold ranging from 0.1 to 0.7. PivoTAL also significantly outperforms the multi-stage self-training based methods (ASM-Loc [9], UGCT [51]).

Supervision	Method	mAP@IoU (%)			
		0.5	0.75	0.95	AVG
Full	TAL-Net CVPR'18 [3]	38.2	18.3	1.3	20.2
	BSN ECCV'18 [26]	46.5	30.0	8.0	30.0
	GTAN CVPR'19 [30]	52.6	34.1	8.9	34.3
Weak	TSCN ECCV'20 [53]	35.3	21.4	5.3	21.7
	BaS-Net AAAI'20 [20]	34.5	22.5	4.9	22.2
	A2CL-PT ECCV'20 [34]	36.8	22.0	5.2	22.5
	ACM-BANet MM'20 [35]	37.6	24.7	6.5	24.4
	WUM AAAI'21 [21]	37.0	23.9	5.7	23.7
	AUMN CVPR'21 [31]	38.3	23.5	5.2	23.5
	TS-PCA CVPR'21 [28]	37.4	23.5	5.9	23.7
	UGCT CVPR'21 [51]	39.1	22.4	5.8	23.8
	FAC-Net ICCV'21 [13]	37.6	24.2	6.0	24.0
	RSKP CVPR'22 [14]	40.6	24.6	5.9	25.0
	ASM-Loc CVPR'22 [9]	41.0	24.9	6.2	25.1
	PivoTAL (Ours)	45.1	28.2	5.0	**28.1**$_{\uparrow 3.0}$

Table 2. Temporal action localization performance comparison with existing methods on the **ActivityNet-v1.3** dataset. PivoTAL outperforms all existing methods on avg mAP by at least 3.0%.

We compare PivoTAL with existing methods on the more challenging AcitivityNet-v1.3 dataset in Table 2 and observe a similar trend where PivoTAL outperforms the existing best method ASM-Loc [9] by more than 3%. We believe this is particularly significant given the relatively smaller improvements reported by other recent methods.

4.2. PivoTAL: Ablation Study

We conducted a comprehensive set of ablation experiments to empirically validate the effectiveness of the different components of PivoTAL, and we report the results of these experiments in Table 3. Row 1 shows the performance of the MIL-based Base WTAL head without any of our priors. Row 2 demonstrates that adding our Gaussian prior improves the average mAP by 2.6%. Row 3 shows the effectiveness of our proposed scene prior, which further improves the performance by 1.3%. Row 4 shows the results with the incorporation of a prior-driven localization head with hard pseudo-action snippets. We observed no performance improvement over Row 3, which validates our hypothesis that training with hard pseudo-action snippets is suboptimal due to the noise present in the pseudo-action snippets generated using manual priors. Row 5 demonstrates that the incorporation of soft pseudo-action snippets and per-class normalization in the prior-driven localization training significantly improves the performance, with a 6.9% improvement in average mAP. This validates our hypothesis that confidence propagation (soft pseudo-action snippets) and per-class normalization are effective in minimizing the influence of any errors in pseudo-labels. Finally, Row 6 demonstrates that incorporating MIL loss further improves the performance by 3%. We believe that MIL loss is critical in training the prior-driven localization head since it further aids in mitigating the potential negative effects of dense, albeit noisy supervision (soft pseudo-action snippets) via weak, albeit noise-free supervision. These ablation experiments demonstrate that each design component has a noticeable impact on the overall performance.

Base WTAL	Gaussian Prior	Scene Prior	Localization Head			AVG mAP
			Hard Pseudo-Action Snippets	Soft Pseudo-Action Snippets	MIL	
✓						35.8
✓	✓					38.4
✓	✓	✓				39.7
✓	✓	✓	✓			39.0
✓	✓	✓		✓		46.6
✓	✓	✓		✓	✓	**49.6**

Table 3. Ablation studies on the **THUMOS-14** dataset showing the effectiveness of each component of PivoTAL.

4.3. Discussion

Gaussian Prior. We discuss various design aspects of our Gaussian prior-based actionness score generation. In PivoTAL, we predict one Gaussian mask for each clip and sample one value from each mask. We experiment with other designs to achieve the same objective and report the results in Table. 4a. Row 1 in Table. 4a shows the result for the case when we predict a single Gaussian mask (global mask) over the entire video. For global mask, we generate a single $\{\mu, \sigma\}$ for the entire video, making $\{\mu_i, \sigma_i\}_{i=1}^{T}$ in Eq. 5 $\{\mu, \sigma\}$. We observe that the performance goes down by 4.4%. This is to be expected since a video will have multiple actions happening at different times. Row 2 and 3 expand on this idea and try to predict multiple global Gaussian masks (multiple $\{\mu, \sigma\}$) and we aggregate the actionness scores of these multiple global Gaussian masks via averaging. We observe a further performance drop that we believe is due to the lack of diversity between multiple global masks and the inadequacy of the global mask predictions in modeling fine local variations. In Row 4, we model the Gaussian masks locally and average their contributions to obtain the video-level actionness scores. Here, local Gaussian mask refers to generating clip specific (hence, local) T Gaussian parameters $\{\mu_i, \sigma_i\}_{i=1}^{T}$ and averaging over T such that $\mathbf{a}_{gauss}(i) = 1/T \sum_{j=1}^{T} \mathbf{G}_j(i)$. We find this design to be sub-optimal validating the design decision made in PivoTAL that the Gaussian masks need to be modeled *locally* and their contribution will have to be selected *locally* to have enough flexibility in modeling. Finally in Row 5, we replace the Gaussian-prior head with a *second* learning-based head (similar to $\mathbf{a}_{fg}$) and observe a 1.4% drop in avg mAP, further validating the advantage of our Gaussian prior.

Actionness Aggregation. Next, we discuss how to aggregate the two sources of actionness scores (learning based, gaussian prior based) in PivoTAL. We experiment with different aggregation strategies and report the results in Table 4b. The first row shows that the performance goes down by 4.1% if we take the Hadamard product between these two actionness scores. This is to be expected since such an aggregation strategy will only focus on areas where both

Method	AVG mAP
PivoTAL w. global mask (m=1)	45.2
PivoTAL w. global mask (m=5)	41.3
PivoTAL w. global mask (m=10)	44.0
PivoTAL w. local mask centered at each clip	43.0
PivoTAL w. a second learning-based head	48.2
PivoTAL	**49.6**

(a)

Method	AVG mAP
PivoTAL w. Product	45.5
PivoTAL w. Max	48.0
PivoTAL w. Mean	**49.6**

(b)

Method	AVG mAP
PivoTAL w/o consistency loss	47.0
PivoTAL w. consistency loss	**49.6**

(c)

Table 4. Analysis on different (a) Gaussian prior implementation techniques, (b) actionness mask aggregation techniques, and (c) actionness mask consistency setups on the **THUMOS-14** dataset. Our Gaussian prior implementation outperforms all the other baseline solutions.

actionness scores agree. In the second row, we report results by taking element-wise max and observe that performance improves over the Hadamard product-based aggregation strategy but is still 1.6% lower than the mean-based aggregation strategy used in PivoTAL. We hypothesize that the max-based aggregation strategy underperforms because of overpredicting the foreground actionness scores, especially in the cases where one actionness score is significantly higher than the other.

Consistency Loss. Finally, in Table 4c, we investigate the effectiveness of the proposed consistency loss between the two actionness scores. We observe that the performance goes down by 2.6% without any consistency loss. This empirically validates our hypothesis that the two sources of actionness scores need to be aligned for the best performance.

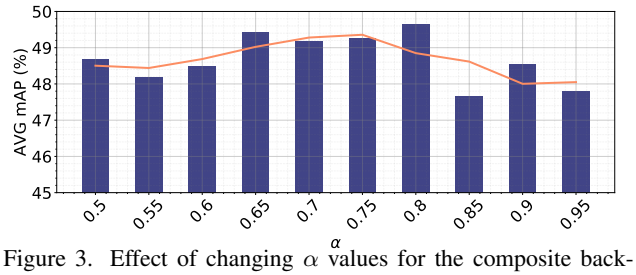

Figure 3. Effect of changing α values for the composite background labels on the **THUMOS-14** dataset. We observe that the performance predictably deteriorates at large and small α values.

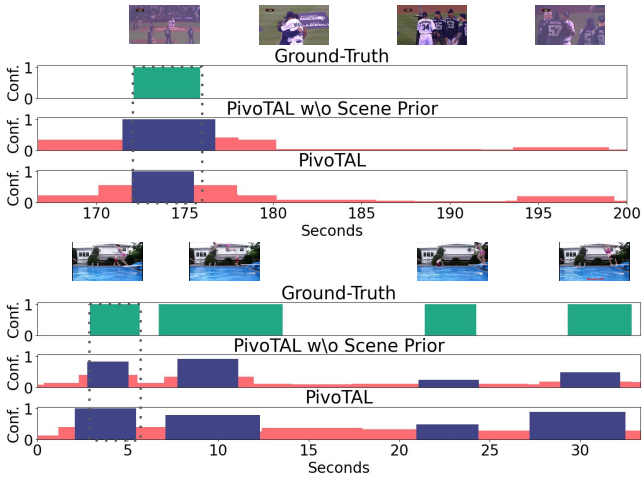

Figure 4. PivoTAL's prediction with and without scene prior. From the dotted boxes, we can observe that PivoTAL with scene prior outputs better-aligned predictions with better action boundaries.

Scene Prior. To enforce the scene prior, we use α to control the strength of action-specific information in the composite background labels. Even though the ablation experiments (Table 3) validated the importance of this prior, here, we conduct a more fine-grained analysis. We vary the value of α and report the results on the THUMOS'14 dataset in Fig 3. We report the performance by varying the value of α from 0.5 to 0.95. We do not use a lower value since we want the background label to be dominant. We notice that the performance deteriorates at large and small values of α. This is to be expected since in both cases either we will have very little action-specific information or we will suppress the background too much and end up with more false positives. We also observe that a wide range of values (0.65 to 0.80) yield similar results. This validates that PivoTAL is robust and not sensitive to a particular choice of α.

In Fig 4, we present some qualitative results to further demonstrate the effectiveness of our action-specific scene priors. In the first video, we see a *Baseball Pitch* action happening. If we look at the foreground and background frames shown in the top row, we can see that the background (stadium scene) carries a meaningful cue for the foreground action. We report the results of PivoTAL without scene priors in the third row and the last row shows results with the scene prior. We observe that even though the overall performance of these two variations is not drastically different, PivoTAL with scene prior can detect the action boundaries better. We notice the same trend in the second video where a *Diving* action is happening, and the pool serves as an informative background. Overall, PivoTAL with scene prior outputs more confident predictions with better-aligned action boundaries. Please refer to Supplementary Materials for more visual results from PivoTAL.

5. Conclusion

PivoTAL is a novel approach that tackles weakly-supervised temporal action localization in a *localization-by-localization* manner. PivoTAL complements the available video-level weak supervision with learned priors to capture the underlying spatio-temporal structure in video data. Empirical studies validated the effectiveness of each component in PivoTAL, and results on WTAL benchmark datasets establish PivoTAL as new state-of-the-art in this area.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 6

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

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018. 2, 7

[4] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision (ECCV)*, 2022. 6

[5] Ishan Dave, Zacchaeus Scheffer, Akash Kumar, Sarah Shiraz, Yogesh Singh Rawat, and Mubarak Shah. Gabriellav2: Towards better generalization in surveillance videos for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 122–132, 2022. 1

[6] Vincent Duval, Jean-François Aujol, and Yann Gousseau. The tvl1 model: a geometric point of view. *Multiscale Modeling & Simulation*, 8(1):154–189, 2009. 6

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 5

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 4

[9] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022. 3, 4, 5, 6, 7

[10] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. Human action recognition without human. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 11–17, Cham, 2016. Springer International Publishing. 5

[11] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021. 2, 6

[12] Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Relational prototypical network for weakly supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11053–11060, 2020. 6

[13] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8002–8011, 2021. 6, 7

[14] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022. 4, 6, 7

[15] 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

[16] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1637–1645, 2021. 4, 6

[17] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 4

[18] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9686, 2022. 5

[19] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017. 2

[20] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11320–11327, 2020. 1, 2, 4, 7

[21] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1854–1862, 2021. 6, 7

[22] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. 4

[23] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11499–11506, 2020. 2

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2021. 1, 2

[25] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 2

[26] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2, 6, 7

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[28] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6176–6185, 2021. 6, 7

[29] Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3899–3908, 2019. 6

[30] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019. 6, 7

[31] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9969–9979, 2021. 2, 6, 7

[32] 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 *European conference on computer vision*, pages 729–745. Springer, 2020. 1, 3, 6

[33] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7596, 2021. 1, 2, 6

[34] Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *European conference on computer vision*, pages 283–299. Springer, 2020. 2, 6, 7

[35] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2166–2174, 2020. 2, 7

[36] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13608–13617, 2021. 2, 6

[37] Phuc Xuan Nguyen, Deva Ramanan, and Charless C Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5502–5511, 2019. 4

[38] Alejandro Pardo, Humam Alwassel, Fabian Caba, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3319–3328, 2021. 1, 3, 4, 5

[39] Alejandro Pardo, Mengmeng Xu, Ali Thabet, Pablo Arbeláez, and Bernard Ghanem. Baod: budget-aware object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1256, 2021. 5

[40] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 2, 4

[41] Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021. 2, 4, 5

[42] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018. 4

[43] Mamshad Nayeem Rizve, Ugur Demir, Praveen Tirupattur, Aayush Jung Rana, Kevin Duarte, Ishan R Dave, Yogesh S Rawat, and Mubarak Shah. Gabriella: An online system for real-time activity detection in untrimmed security videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4237–4244. IEEE, 2021. 1

[44] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. 5

[45] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1019, 2020. 1, 2

[46] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058, 2016. 1, 2

[47] Praveen Tirupattur, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Modeling multi-label action dependencies for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1460–1470, 2021. 1

[48] Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, and Josef Sivic. Predicting actions from static scenes. In *European Conference on Computer Vision*, pages 421–436. Springer, 2014. 5

[49] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017. 2

[50] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9070–9078, 2019. 1, 2

[51] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021. 3, 6, 7

[52] Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3090–3098, 2022. 6

[53] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *European conference on computer vision*, pages 37–54. Springer, 2020. 1, 3, 6, 7

[54] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 4, 6

[55] 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 *Proceedings of the 27th ACM international conference on multimedia*, pages 738–746, 2019. 2

[56] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2, 5, 6

[57] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018. 2

[58] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6566–6575, 2018. 5

[59] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017. 2, 6

[60] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020. 5

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