Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization

Daochang Liu[1], Tingting Jiang[1], Yizhou Wang[1,2,3]
[1]NELVT, Cooperative Medianet Innovation Center, School of EECS, Peking University
[2]Peng Cheng Lab, [3]Deepwise AI Lab
{daochang, ttjiang, yizhou.wang}@pku.edu.cn

Abstract

Temporal action localization is crucial for understanding untrimmed videos. In this work, we first identify two underexplored problems posed by the weak supervision for temporal action localization, namely action completeness modeling and action-context separation. Then by presenting a novel network architecture and its training strategy, the two problems are explicitly looked into. Specifically, to model the completeness of actions, we propose a multi-branch neural network in which branches are enforced to discover distinctive action parts. Complete actions can be therefore localized by fusing activations from different branches. And to separate action instances from their surrounding context, we generate hard negative data for training using the prior that motionless video clips are unlikely to be actions. Experiments performed on datasets THUMOS'14 and ActivityNet show that our framework outperforms state-of-the-art methods. In particular, the average mAP on ActivityNet v1.2 is significantly improved from 18.0% to 22.4%. Our code will be released soon.

1. Introduction

Temporal action localization is an important visual task with potential applications in video surveillance [42], video summarization [28], skill assessment [16], and others. The goal is to predict not only the action label but also the start and end times of each action instance from untrimmed videos. Fully supervised temporal action localization has witnessed remarkable progress recently [39, 48, 15, 9, 37, 51, 47, 8, 2, 31]. However, precisely annotating the temporal extent of action instances is labor-intensive and time-consuming, which undermines fully supervised approaches in real-world large-scale scenarios. Therefore the weakly supervised setting, where only video-level category labels are available during training, is more practical and draws increasing attention from the community. This paper works on temporal action localization with such weak labels.

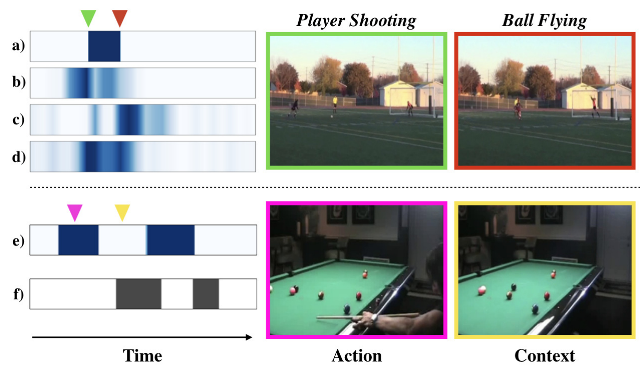

Figure 1. Illustration of the two issues. **Top:** Completeness modeling with the proposed multi-branch network. Branches are trained to discover different action parts, such that complete action can be localized using the averaged activations over branches. a) Ground truth of an action instance of *Soccer Penalty*. b) Class Activation Sequence (CAS) of one branch from our trained model, localizing the *Player Shooting* part. c) CAS of another branch, localizing the *Ball Flying* part. d) Average CAS, localizing the instance completely. **Bottom:** Context separation with the hard negative video generation. e) Ground truth of an action instance of *Billiards*. f) Background clips obtained, discriminating the co-occurring context. On the right are four frames taken from the time locations of correspondingly colored arrows. Best viewed in color.

Most of existing weakly supervised methods [45, 33, 38, 36, 52] fall into the framework of Multiple Instance Learning (MIL) [54]. In this framework, a video is treated as a bag of sampled frames or snippets and fed into video-level classification networks. Action instances are then localized using the Class Activation Sequence (CAS) [38], a 1D temporal classification score sequence of each action.

Compared with its fully-supervised counterpart, weakly supervised temporal action localization introduces two new challenges, dubbed as *action completeness modeling* and *action-context separation*. The two issues have not been well considered before and have notably limited the performance. **The first challenge** is how to detect action instances in their entirety without full annotations. An action is intrinsically a temporal composition of elementary sub-actions [20], which are supposed to be wholly included in the prediction without omission. In the fully supervised setting, whether an action is complete is learned directly from the ground truth of temporal boundaries. In contrast, when weakly supervised, the lack of fine-grained annotations complicates the completeness modeling since the localization task is now formulated as the classification on video level. Identifying one fragment of an action is sufficient for video-level classification but not for segment-level localization. For example, the action *Soccer Penalty* can be roughly divided into two sub-actions, *i.e.*, *Player Shooting* and *Ball Flying*. The activation only on the more discriminative *Player Shooting* part is adequate to classify the video, but leaving the *Ball Flying* part false negative for localization. **The second challenge** of action-context separation is how to distinguish action instances from their context with weak labels. Action instances of the same class are usually surrounded by visually similar clips, such as action *Billiards* commonly enclosed by commentary clips with a static pool table in the screen. Such clips appear together with the true actions in most videos, thus termed as *context* in this paper. Context clips are different from ordinary background clips in terms of distributions in videos. Context clips co-occur with the true action in most cases and are not involved in videos of other action categories, while background clips are class-independent and distributed randomly. For this reason, context clips can be regarded as hard negatives. Video-level classifiers learn the correlation between videos with the same tag and discover their common contents, which unfortunately include not only the common action (*e.g.*, *Billiards*) but also the common context (*e.g.*, the static pool table). We argue that action-context separation is inherently difficult with weak supervision, unless employing the prior knowledge about actions.

To tackle the two issues respectively, we propose a multi-branch network architecture and a hard negative data generation scheme. To model action completeness, feature sequences extracted from input videos are fed into a network with multiple classification branches in parallel. A diversity loss is devised to ensure the dissimilarity between the Class Activation Sequences output by different branches, such that each branch is trained to locate distinct fractions of an action. Thereupon, as the example in Fig. 1, complete actions can be retrieved by aggregating activations from multiple branches. The class activation is then pooled over time with temporal attention, producing a video-level category distribution. We compute its cross-entropy with the ground truth, *i.e.*, the standard MIL loss, which is minimized along with the diversity loss to learn the network parameters. As for action-context separation, we develop a simple yet effective strategy for mining hard negatives using the prior that actions should be of motions. We search for stationary clips in the training videos, as illustrated in Fig. 1. Then pseudo videos are generated using static clips and labeled with a new *background* class. Such a strategy can assist the model in rejecting the common context, as long as some hard negatives are included in the generated pseudo videos.

On two benchmark datasets THUMOS'14 [21] and ActivityNet [6], the proposed method outperforms the state-of-the-art methods, demonstrating the effectiveness of handling the two problems. In summary, our contributions are three-fold: 1) A multi-branch network with diversity loss is proposed to model action completeness. 2) A hard negative video generation scheme is devised to separate common context. 3) Our method achieves superior results on two benchmark datasets.

2. Related Works

Action recognition on trimmed videos has been extensively studied in the past. Early methods were mainly based on hand-crafted features [26, 43, 34]. In recent years, various deep networks have been proposed, such as two-stream networks [40, 46], LSTM [12], 3D ConvNets [41], I3D [7], and others [23, 44, 53]. Please refer to recent surveys [1, 3, 22, 19] for a detailed review.

Fully supervised temporal action localization approaches have been largely based on the proposal-plus-classification paradigm [39, 37, 51, 15, 5, 9, 47, 8, 31], where temporal proposals are generated first and then classified. Other categories of methods have also been studied, such as those based on single-shot detectors [4, 30] or sequential decision-making process [48, 2]. Given full annotations, the proposal-plus-classification methods usually filter out the common context at the proposal stage via a binary actionness classifier. As for the completeness modeling, Zhao *et al.* [51] used a structural temporal pyramid pooling followed by an explicit binary classifier to determine whether an instance is complete. Hou *et al.* [20] clustered video segments of an action into different sub-actions and then detected the whole action as an ordered sequence of sub-actions. Yuan *et al.* [49] structured an action into three components, *i.e.* the start, middle and end, to model its temporal evolution. But they all require full annotations. Other works on spatial-temporal action detection [17] and video temporal segmentation [27] are beyond our scope.

Weakly supervised temporal action localization algorithms mostly belong to the Multiple Instance Learning (MIL) [54]. Wang *et al.* [45] proposed a framework called UntrimmedNet composed of a classification module and a selection module, based on which a sparsity regularization was later introduced in [33]. Paul *et al.* [36] used a co-activity similarity loss to enforce the feature similarity between the localized instances of the same class. Instead of

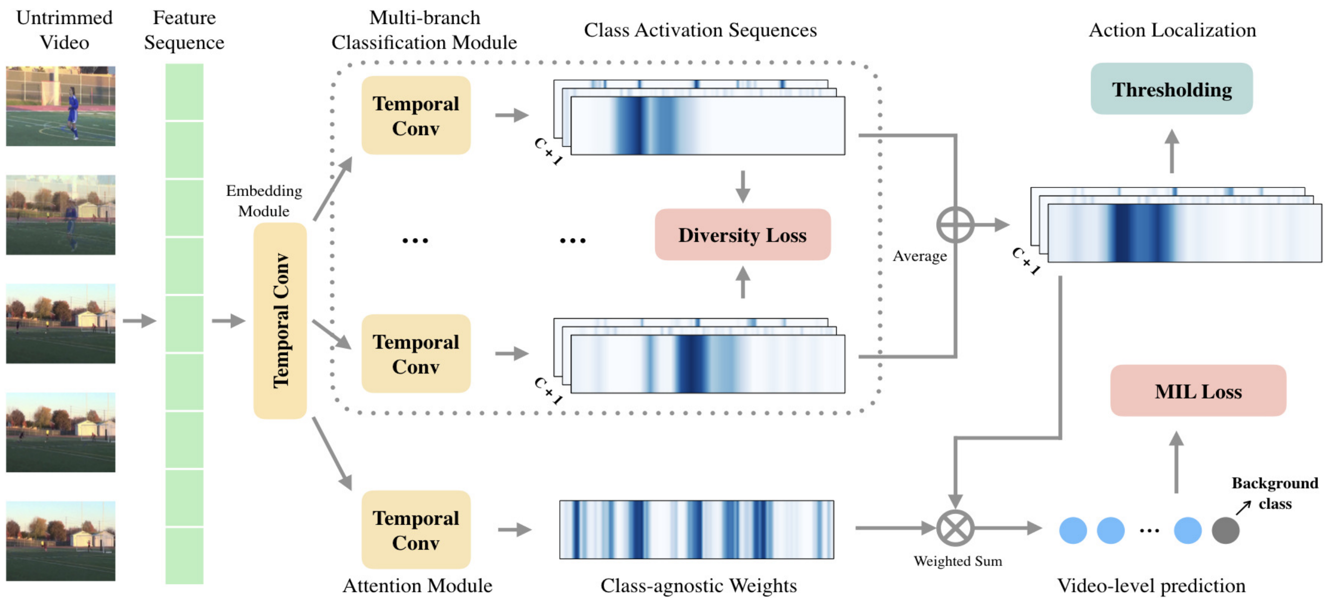

Figure 2. **Overview architecture.** The proposed multi-branch network consists of a feature extraction module, a feature embedding module, a multi-branch classification module, and a temporal attention module. In the classification module, multiple branches are trained with the diversity loss to discover different action parts.

thresholding on the CAS, AutoLoc [38] directly predicted the temporal boundaries to detect actions. Regarding the first challenge, there are two prior works that attempt to model action completeness. Hide-and-Seek [25] hid random frame sequences while training to force the network responsive to multiple relevant parts. However, randomly hiding frames does not always guarantee the discovery of new parts and also disrupts the training process. Recently, Zhong *et al*. [52] trained a series of classifiers iteratively to find complementary pieces, by erasing the predictions of predecessor classifiers from input videos. The major drawback with this approach is the extra time cost and computational expense to train multiple classifiers. The other challenge of action-context separation is inherently tricky and remains unexplored in the literature. The selection module in UntrimmedNet [45] is intended for eliminating irrelevant background clips rather than semantically related context. Researchers have also studied action localization with other types of weak supervision, such as movies scripts [13], ordered action lists [11], and web images [14].

Diversity Loss was initially introduced for text embedding [32] to extract different aspects of a sentence. Recently, Li *et al*. [29] utilized the diversity loss to deal with occlusions in person re-identification. Unlike previous works, we use diversity loss to model the action completeness, which is of different specification and motivation.

3. Proposed Method

In this section, we present the proposed methodology for weakly supervised temporal action localization. The input is an untrimmed video with varying frames in length. Let a one-hot vector $y \in \{0,1\}^{C+1}$ denotes the ground truth video-level category label, where C is the number of action classes and $C+1$ represents the newly added background class. During the test time, the output for each test video is a set of localized action instances $\{(s_i, e_i, c_i, q_i)\}$, where s_i and e_i denote the start time and end time of the i^{th} detection, c_i represents the predicted category, and q_i denotes the confidence score.

3.1. Hard Negative Video Generation

Weakly supervised models are inclined to confuse the true action with its surrounding context, *i.e*. the hard negatives, especially when the context appears in a majority of videos of that class. We observe that it is the motion that makes an action different from its context. An action must involve the movement of human or other subjects, while the context clips are allowed to stay static (*e.g*., the static pool table). Therefore, we generate hard negative training data using stationary video clips, labeling them with a new *background* class. Concretely, for each video in the training set, we compute its optical flow using the TV-L1 algorithm [50] and average the intensity in every frame. Since the motion magnitude differs among action categories and even some actions exhibit minor motion, a small predefined percentage ρ of video frames with the lowest optical flow intensity are picked out from each video individually. Frames picked from the same video are then concatenated into a pseudo video, which is labeled with the background class and added to the training set. We expect that the generated videos partially include the hard negatives and drop hints for the proposed network to deal with the challenge of action-context separation. Details and the generated video examples are provided in the supplementary material.

3.2. Multi-branch Network

To model action completeness, a multi-branch network is designed such that each branch focuses on different action parts. As shown in Fig. 2, the proposed multi-branch network consists of a feature extraction module, an embedding module, a multi-branch classification module, and a temporal attention module, which are detailed as follows.

Feature extraction module. Given an input video, a snippet-wise feature sequence $X \in \mathbb{R}^{T\times D}$ is first extracted by pre-trained deep networks, where T denotes the number of snippets and D denotes the feature dimensions. The extracted feature sequence provides a high-level representation of the appearance and motion of the input video and is fed into the next layers in the network. Note that T and D depend on the choice of feature extraction network. In experiments, we focus on two off-the-shelf models, namely UntrimmedNet [45] and I3D [7].

Embedding module. The feature extraction module is followed by an embedding module. Since the features may not be originally trained for weakly supervised action localization, a task-specific embedding of the features is desired. We utilize a temporal convolutional layer followed by a ReLU activation layer to embed the features:

$$\phi(X) = \max\left(W_{emb} \circledast X + b_{emb}, 0\right) \tag{1}$$

where $\circledast$ represents the convolution operation, W_{emb} and b_{emb} are the weights and biases of temporal filters, $\phi(X) \in \mathbb{R}^{T\times F}$ denotes the learned embedding, and F is the number of filters. Temporal convolutions integrate the information from neighboring time locations, enabling the network to capture the temporal structure. The embedded feature sequence is then passed to subsequent layers.

Multi-branch classification module. In this module, K classification branches are organized in parallel to discover complementary pieces of an action. Each branch inputs the embedded feature sequence into a temporal convolutional layer and outputs a sequence of classification scores:

$$A^k = W^k_{cls} \circledast \phi(X) + b^k_{cls} \tag{2}$$

where $A^k \in \mathbb{R}^{T\times(C+1)}$, W^k_{cls} and b^k_{cls} are respectively the classification scores, the filter weights, and filter biases in the k^{th} branch. Then each A^k is passed through a softmax along the category dimension, yielding class distributions at each time location:

$$\overline{A^k} = \mathrm{softmax}(A^k) \tag{3}$$

where $\overline{A^k}$ is referred to as Class Activation Sequence (CAS). For clarity, we use the bar notation in this paper to indicate it has undergone a softmax. For action completeness modeling, we expect the CASes from multiple branches differ from each other. However, without constraint, the branches could lazily concentrate on a single same action part. To avoid such degenerate cases where branches give identical results, a diversity loss based on cosine similarity is imposed on the CASes:

$$\mathcal{L}_{div} = \frac{1}{Z}\sum_{c=1}^{C+1}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\overline{A^i}_{*,c}\cdot\overline{A^j}_{*,c}}{\|\overline{A^i}_{*,c}\|\|\overline{A^j}_{*,c}\|} \tag{4}$$

which is the cosine similarities between the CASes from every two branches, averaged over all branch pairs and action categories. $\overline{A^i}_{*,c} \in \mathbb{R}^T$ means the activation sequence for class c from the i^{th} branch and $Z = \frac{1}{2}K(K-1)(C+1)$ is a normalization factor. By minimizing such a diversity loss, branches are encouraged to produce activations on different action parts. Then CASes from multiple branches are averaged and passed through a softmax along the category dimension:

$$A^{avg} = \frac{1}{K}\sum_{k=1}^{K} A^k \tag{5}$$

$$\overline{A^{avg}} = \mathrm{softmax}(A^{avg}) \tag{6}$$

where $\overline{A^{avg}} \in \mathbb{R}^{T\times(C+1)}$ is referred to as average CAS, which combines all part activations and encodes the full action. Moreover, the softmax operation inhibits activations of action classes when the score of the background class is large, thus reducing false positives on context clips.

We empirically notice that A^k from some branches tend to be nearly all zeros while those from other branches explode, which may corrupt the training process. More importantly, if one branch dominates, the average CAS is effectively responsive to single action part instead of the whole action. From another perspective, these parallel branches can be considered to be in an adversarial relationship, competing with each other to find different discriminative action segments. It is expected that the branches are balanced to have comparable strength. Similar ideas can be seen in the training strategy of Generative Adversarial Networks [18]. Therefore we introduce another regularization term on the norms of the original score sequences without softmax:

$$\mathcal{L}_{norm} = \frac{1}{K(C+1)}\sum_{c=1}^{C+1}\sum_{i=1}^{K}\left|\|A^i{}_{*,c}\| - \|A^{avg}{}_{*,c}\|\right| \tag{7}$$

which is the deviations from the norm of A^{avg}, averaged over branches and categories. Equipped with the diversity loss and norm regularization, the multi-branch design is capable of discovering diverse action parts without full supervision and therefore modeling the action completeness.

Temporal attention module. Since the input video is untrimmed and contains irrelevant backgrounds, we utilize a temporal attention module to learn the importance of video snippets. The attention module feeds the embedded feature sequence into a temporal convolutional layer followed by a softmax along the *temporal* dimension:

$$\overline{U} = \text{softmax}(W_{att} \circledast \phi(X) + b_{att}) \tag{8}$$

where W_{att} and b_{att} are the weight parameter and bias of the temporal filter, and $\overline{U} \in \mathbb{R}^T$ represents the sequence of learned class-agnostic attention. To get the video-level classification prediction, we perform a softmax along the category dimension on the summation of A^{avg} weighted by the attention:

$$\overline{p} = \text{softmax}(\sum_{t=1}^{T} \overline{U}_t {A^{avg}}_{t,*}) \tag{9}$$

where $\overline{p} \in \mathbb{R}^{C+1}$ is a probabilistic distribution over action classes, including the background class. Then its cross-entropy with the ground truth, *i.e.*, the standard MIL loss, is computed:

$$\mathcal{L}_{mil} = -\sum_{c=1}^{C+1} y_c \log \overline{p}_c \tag{10}$$

Finally, we combine the MIL loss with the diversity loss and norm regularization:

$$\mathcal{L}_{sum} = \mathcal{L}_{mil} + \alpha \mathcal{L}_{div} + \beta \mathcal{L}_{norm} \tag{11}$$

where α and β are coefficients. All the three components have sub-gradients at least and can be minimized using gradient descent.

3.3. Action Localization

During the test time, we leverage the trained multi-branch network to classify test videos and localize actions. Since multiple categories of actions can occur in one video, we first threshold on the video-level classification score. Given a test video, we detect action instances for each non-background category c with $\overline{p}_c$ larger than 0.1. Then we threshold on the average CAS of class c, *i.e.*, $\overline{A^{avg}}_{*,c}$, to localize action instances. Let $\{(s_i, e_i, c, q_i)\}$ denotes the corresponding output detections. Similar to the Outer-Inner-Contrastive loss proposed in [38], we score each localized instance using the contrast between the mean activation of the instance itself and its surrounding areas:

$$\begin{aligned} q_i &= m_{inner} - m_{outer} + \gamma \overline{p}_c \\ m_{inner} &= \text{mean}(\overline{A^{avg}}_{s_i:e_i,c}) \\ m_{outer} &= \text{mean}([\overline{A^{avg}}_{s_i-l_i:s_i,c}, \overline{A^{avg}}_{e_i:e_i+l_i,c}]) \end{aligned} \tag{12}$$

where $[\cdot]$ denotes concatenation and $l_i = (e_i - s_i)/4$ is the inflation length. The video-level score $\overline{p}_c$ is combined as well with the coefficient γ.

4. Experiments

In this section, we first discuss the datasets and our implementation details. Then comparisons between the proposed method and state-of-the-art approaches are presented. At last, we examine the impact of each model component by ablation studies. In the supplementary material, more experiment results are reported.

4.1. Datasets

Extensive experiments are conducted on two large-scale benchmarks: THUMOS'14 [21] and ActivityNet [6]. Videos in both datasets are untrimmed and only video-level category labels are used for training.

THUMOS'14. A subset of THUMOS'14 including 20 action classes is provided with temporal annotations and used for the localization task. Following the previous convention, we use the validation set of 200 videos for training, and the test set of 213 videos for evaluation. From the training data, 152 hard negative videos[1] are generated. This dataset has a large amount of action instances per video and the length of videos varies widely.

ActivityNet. Experiments are performed on both two release versions of ActivityNet. ActivityNet1.3 covers 200 action classes and consists of 10,024 training videos, 4,926 validation videos and 5,044 testing videos, with 7323 hard negative videos generated using the training videos. We train on the training set and report results on the validation set as well as the testing set. To facilitate comparisons, we also evaluate on ActivityNet1.2, a subset of version 1.3, which has 4,819 training videos, 2,383 validation videos, 2,480 testing videos, and 100 classes. 3469 hard negative videos are generated on ActivityNet1.2. We employ the training set for training and the validation set for evaluation.

Evaluation Metrics. We follow the standard evaluation protocol and report mean average precision (mAP) at different thresholds of temporal intersection over union (IoU). The mAP values are computed using the evaluation codes provided by the datasets. All results on THUMOS'14 are averaged over three runs. The performance on ActivityNet1.3 testing set is obtained by submitting results to the evaluation server.

4.2. Implementation Details

Two deep networks with two-stream architecture are tried out for feature extraction, namely UntrimmedNet [45] and I3D [7], which are pre-trained and fixed during training. UntrimmedNet is pre-trained on ImageNet [10] and takes video snippets of 1 RGB frame and 5 stacked optical flow frames as input. I3D is pre-trained on Kinetics [7] and takes non-overlapping 16-frame chunks as input for both two streams. Video snippets are sampled every 15 frames

[1]Please refer to the supplementary for details.

Supervision	Methods ↓ IoU threshold →	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG (0.1:0.5)
Full	S-CNN [39], *CVPR 2016*	47.7	43.5	36.3	28.7	19.0	-	5.3	35.0
Full	R-C3D [47], *ICCV 2017*	54.5	51.5	44.8	35.6	28.9	-	-	43.1
Full	SSN [51], *ICCV 2017*	**60.3**	56.2	50.6	40.8	29.1	-	-	47.4
Full	Chao *et al*. [8], *CVPR 2018*	59.8	**57.1**	**53.2**	**48.5**	**42.8**	**33.8**	**20.8**	**52.3**
Weak	Hide-and-Seek [25], *ICCV 2017*	36.4	27.8	19.5	12.7	6.8	-	-	20.6
Weak	UntrimmedNet [45], *CVPR 2017*	44.4	37.7	28.2	21.1	13.7	-	-	29.0
Weak	Zhong *et al*. [52], *ACM MM 2018*	45.8	39.0	31.1	22.5	15.9	-	-	30.9
Weak	STPN (UNT) [33], *CVPR 2018*	45.3	38.8	31.1	23.5	16.2	9.8	5.1	31.0
Weak	W-TALC (UNT) [36], *ECCV 2018*	49.0	42.8	32.0	26.0	18.8	-	**6.2**	33.7
Weak	AutoLoc (UNT) [38], *ECCV 2018*	-	-	35.8	29.0	**21.2**	**13.4**	5.8	-
Weak	Ours (UNT), Full	**53.5**	**46.8**	**37.5**	**29.1**	19.9	12.3	6.0	**37.4**
Weak	STPN (I3D) [33], *CVPR 2018*	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0
Weak	W-TALC (I3D) [36], *ECCV 2018*	55.2	49.6	40.1	31.1	22.8	-	**7.6**	39.8
Weak	Ours (I3D), Full	**57.4**	**50.8**	**41.2**	**32.1**	**23.1**	**15.0**	7.0	**40.9**

Table 1. Results on THUMOS'14 testing set. The mAP values at different IoU thresholds are reported, and the column AVG indicates the average mAP at IoU thresholds from 0.1 to 0.5. UNT and I3D are abbreviations for UntrimmedNet features and I3D features respectively. With both UntrimmedNet and I3D features, our full model outperforms state-of-the-art methods at most IoUs.

Methods IoU→	0.5	0.75	0.95	AVG
Zhong *et al*. [52]	27.3	14.7	2.9	15.6
AutoLoc (UNT) [38]	27.3	15.1	3.3	16.0
Ours (UNT)	**33.9**	**19.9**	**5.1**	**20.5**
W-TALC (I3D) [36]	**37.0**	-	-	18.0
Ours (I3D)	36.8	**22.0**	**5.6**	**22.4**

Table 2. Results on ActivityNet1.2 validation set. The column AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95. The proposed method exceeds previous ones by a large margin.

	Validation				Testing
Methods IoU→	0.5	0.75	0.95	AVG	AVG
STPN (I3D) [33]	29.3	16.9	2.6	-	20.1
Ours (I3D)	**34.0**	**20.9**	**5.7**	**21.2**	**23.1**

Table 3. Results on ActivityNet1.3. The column AVG indicates the average mAP at IoU thresholds 0.5:0.05:0.95. Our method also achieves superior performance.

for UntrimmedNet and every 16 frames for I3D. The feature dimension is 1024 in each stream for both two networks. We adopt the early-fusion of the RGB and optical flow streams for UntrimmedNet and late-fusion for I3D. The proposed method is implemented with PyTorch [35]. Network parameters are learned using mini-batch stochastic gradient descent with Adam optimizer [24]. Branch number is set as $K = 4$ in the multi-branch classification module. The kernel size of temporal convolutions is set as 3 in the classification module and 1 in both embedding module and attention module. The dimension of embedded features is set as $F = 32$. The coefficients α and β in Eq.(11) are both set as 0.2 and γ in Eq.(12) is set as 0.25. The selection ratio ρ for hard negative mining is chosen as 25%. Other details are provided in the supplementary material.

4.3. Comparisons with the State-of-the-art

Experimental results on THUMOS'14 testing set are shown in Table 1. Our proposed multi-branch network along with hard negative mining is compared to existing methods for weakly supervised temporal action localization, as well as several fully supervised ones. Our model outperforms previous weakly supervised methods at most IoU thresholds regardless of the choice of feature extraction network. The gain is not as substantial at higher IoU due to the observation that our model sometimes produces *overly complete* instances which lead to false positives. Note that AutoLoc [38] regresses temporal action boundaries for localization and therefore obtain high mAP at higher IoU thresholds, while we simply threshold on the CAS and still achieve comparable results. We argue that their method and ours can promote the performance further if combined.

Table 2 presents the results on ActivityNet1.2 validation set, and results on the validation and testing sets of ActivityNet1.3 are reported in Table 3. On both versions of this large dataset, the proposed method outperforms the state-of-the-art significantly, verifying the effectiveness of handling action completeness modeling and context separation.

4.4. Ablation Studies

To analyze the contribution of each model component, we perform a set of ablation studies, with results on THUMOS'14 testing set shown in Table 4. Our best model is compared to a baseline and other configurations with each of the following components removed: 1) multi-branch design 2) hard negative generation 3) both diversity loss and norm regularization 4) only norm regularization 5) the tem-

Methods	AVG (0.1:0.5)
Ours (UNT), Single + $\mathcal{L}_{mil}$ (Baseline)	28.8
Ours (UNT), Single + $\mathcal{L}_{mil}$ + HN	32.7
Ours (UNT), Multiple + $\mathcal{L}_{sum}$	34.8
Ours (UNT), Multiple + $\mathcal{L}_{mil}$ + HN	34.7
Ours (UNT), Multiple + $\mathcal{L}_{mil}$ + $\mathcal{L}_{div}$ + HN	35.6
Ours (UNT), Multiple + $\mathcal{L}_{sum}$ + HN (No TA)	36.3
Ours (UNT), Multiple + $\mathcal{L}_{sum}$ + HN (Full)	**37.4**

Table 4. Ablation study results on THUMOS'14 testing set. 'Single' and 'Multiple' indicate the number of branches in the classification module, and 'HN' denotes that hard negative videos are used when training. 'TA' denotes the temporal attention module.

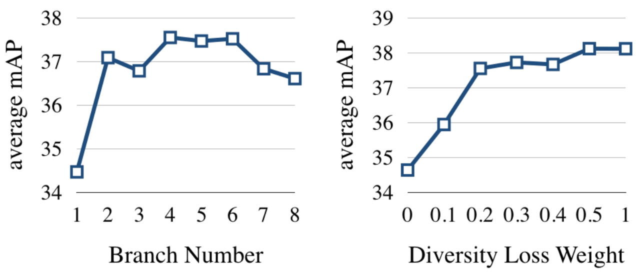

Figure 3. **Left:** Experiments on the branch number. **Right:** Experiments on the diversity loss weight. The average mAP at IoU thresholds from 0.1 to 0.5 is reported.

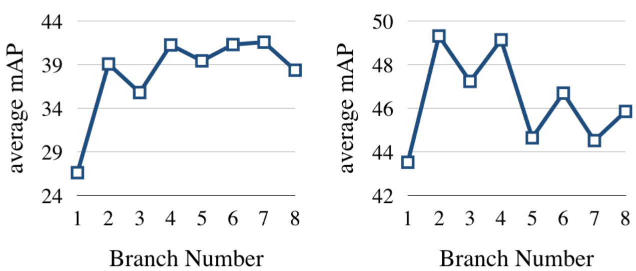

Figure 4. Class-specific results with different branch number. The optimal branch number depends on the complexity of the action. **Left:** Results of *Shotput*. **Right:** Results of *Cliff Diving*. The average mAP at IoU thresholds from 0.1 to 0.5 is reported.

Ratio	15%	20%	25%	30%
AVG	32.4	33.3	32.7	32.8

Table 5. Impact of the selection ratio ρ in hard negative mining. AVG indicates the average mAP at IoU thresholds from 0.1 to 0.5.

poral attention module. Results show that all these components are required to achieve the best performance, and the multi-branch design is especially important. Besides, we conduct experiments with UntrimmedNet features to investigate the impact of the branch number, the diversity loss weight as well as the selection ratio in hard negative mining.

Branch Number. A comparative experiment is performed on the branch number, in which the branch number K is altered from one to eight. To avoid tuning parameters on test data, we conduct this experiment on THUMOS'14 validation set, *i.e.*, the one used for training. As the results shown in Fig. 3, all models with multiple branches exceed the one with single branch evidently, and the difference among models with two to eight branches is insignificant. Since the complexity of actions varies among categories, the optimal branch number could be different for each action class. As the example given in Fig. 4, complex actions such as *Shotput* are comprised of more parts and thus need a larger branch number, while the structure of simpler actions like *Cliff Diving* can be captured with only two branches.

Weight of Diversity Loss. Another comparative experiment on the weight of diversity loss is conducted on THUMOS'14 validation set, with results posted in Fig. 3. Since the norm regularization and diversity loss are intended for constraining the multiple branches jointly, coefficients α and β are set to a same changing value. The experiment shows that our model is insensitive to the diversity loss weight when it is larger than 0.2, demonstrating the robustness of the proposed method.

Selection Ratio in Hard Negative Mining. Several different percentages are tried out when selecting the static frames. Since the hard negative videos are generated from the validation set, this experiment is performed on THUMOS'14 testing set. Results produced by the single branch model are presented in Table 5, which are stable across different ratios.

4.5. Qualitative Results

We plot several interesting examples of localized actions and corresponding CASes in Fig. 5 to show the effectiveness of tackling the two challenges qualitatively. Examples are from THUMOS'14 testing set using UntrimmedNet features. In the first example of *Diving*, incomplete actions displayed in red bounding boxes such as the one with only *Entry into the Water* part and the other one with only *Standing on the Platform* part are discovered by single branch but excluded from final predictions due to incompleteness. In the second example of *Billiards*, commentary clips (yellow boxes) semantically similar to the true actions (pink boxes) are effectively filtered out using hard negative generation. In the third example of *High Jump*, CASes from multiple branches are very diverse, localizing different action parts.

5. Discussion and Future Work

We devise a simple yet effective data generation scheme to separate action context, while the assumption behind might not hold in all cases. We find its effect closely related to the action class, with details in the supplementary. In future, more advanced techniques such as Generative Adversarial Networks can be applied to mine the hard negatives deeper. As for action completeness modeling, distinctive

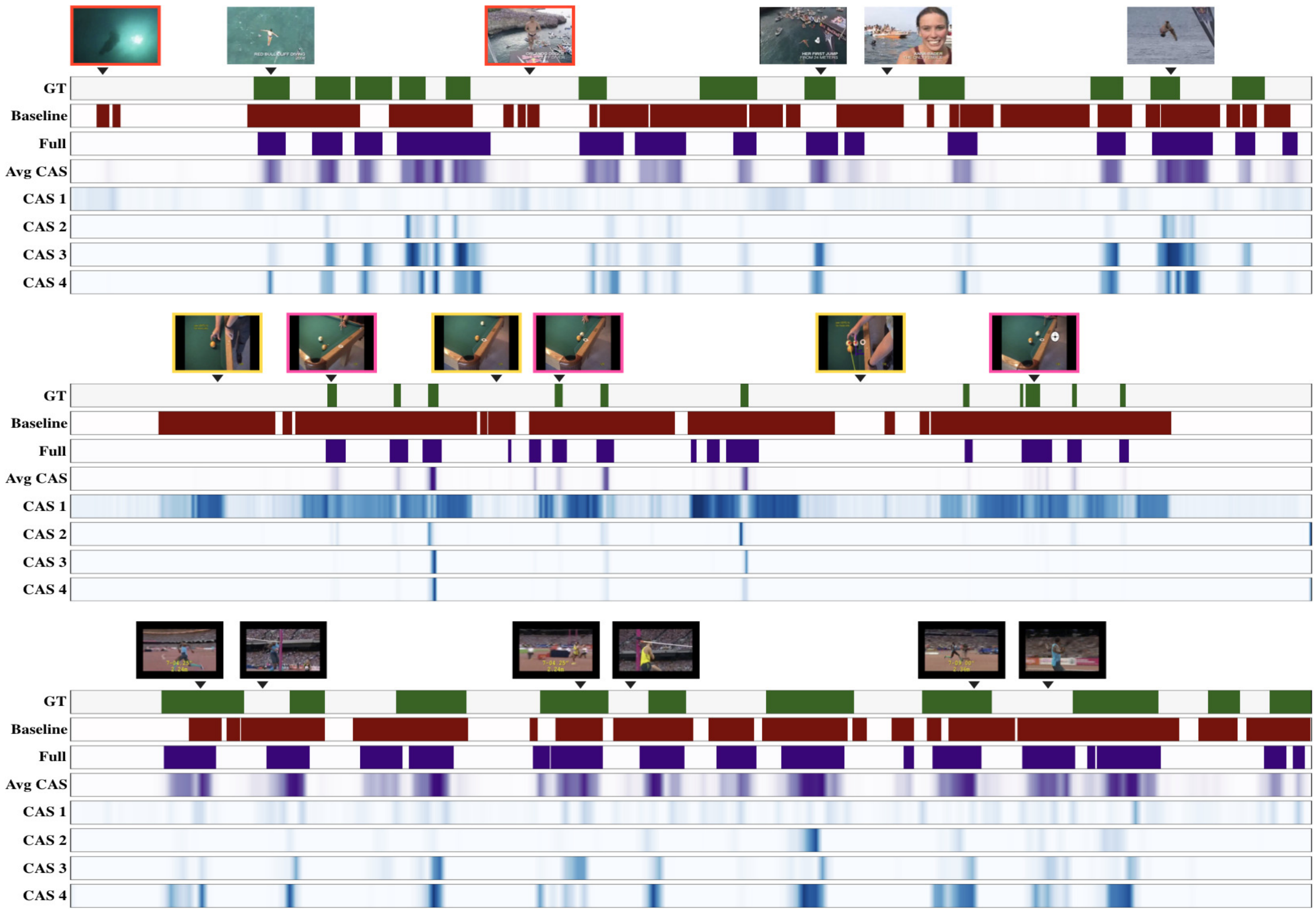

Figure 5. Three prediction examples. The eight barcodes in each case are respectively: 1) ground truth of action instances 2) predictions of the baseline model, *i.e.*, the single-branch network without hard negative generation 3) predictions of our full model 4) the average CAS of our full model 5-8) CASes from the four branches. **Top:** *Diving*. Incomplete actions (red boxes) only activate single branch and are thus ruled out by the average CAS. **Middle:** *Billiards*. The proposed method focuses on the true actions (pink boxes) and the false positives of semantically similar context (yellow boxes) are reduced. **Bottom:** *High Jump*. In all examples, each branch outputs a distinct CAS.

action parts are discovered automatically by the proposed multi-branch module in an unsupervised manner. In practice, the learned action parts may not exactly correspond to semantically meaningful sub-actions. Instead, the model may capture different action modes, aspects, stages, or other underlying structures, depending on which kind of representation benefits the learning target most. Along the key idea of dividing an action into parts, there can be many potential future directions for weakly supervised temporal action localization, including but not limited to 1) using the learned part representation to understand actions or measure action complexity 2) modeling the temporal configuration of action parts 3) representing actions hierarchically to handle ambiguities or subjective annotation biases.

6. Conclusion

In this work, we identified two challenges posed by the weak supervision for temporal action localization, namely action completeness modeling and action-context separation. To handle the first challenge, a multi-branch network was proposed to find different action parts and therefore locate action instances in their integrity. Meanwhile, we mined hard negatives to handle the second issue of action-context separation. Experiments on two benchmarks showed that our framework tackles the two problems effectively and outperforms state-of-the-art methods.

Acknowledgment. This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61527804, 61625201. We acknowledge the high-performance computing platform of Peking University for providing computational resources. Thanks to Jingjia Huang for providing the optical flow data.

References

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[3] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 476–483. IEEE, 2017.

[4] S Buch, V Escorcia, B Ghanem, L Fei-Fei, and JC Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[5] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. SCC: Semantic context cascade for efficient action detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[11] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[13] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1498. IEEE, 2009.

[14] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *The European Conference on Computer Vision (ECCV)*, pages 849–866. Springer, 2016.

[15] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[16] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014.

[17] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[19] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4–21, 2017.

[20] Rui Hou, Rahul Sukthankar, and Mubarak Shah. Real-time temporal action localization in untrimmed videos by subaction discovery. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, page 7, 2017.

[21] 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.

[22] Soo Min Kang and Richard P Wildes. Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906*, 2016.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[26] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[27] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *The IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), July 2017.

[28] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353. IEEE, 2012.

[29] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 988–996. ACM, 2017.

[31] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[32] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[33] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.

[36] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[37] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[38] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[40] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

[43] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[44] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[46] 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 *The European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016.

[47] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region convolutional 3D network for temporal activity detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[48] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[49] Zehuan Yuan, Jonathan C. Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[50] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[51] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[52] 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 2018 ACM on Multimedia Conference*. ACM, 2018.

[53] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[54] Zhi-Hua Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004.