Temporal RPN Learning for Weakly-Supervised Temporal Action Localization

Jing Huang HUANGJIN9@ZJU.EDU.CN
Zhejiang University, Hangzhou 310058, China

Ming Kong ZJUKONGMING@ZJU.EDU.CN
College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
Hikvision Research Institute, Hangzhou 310051, China

Luyuan Chen CHENLY@BISTU.EDU.CN
Beijing Information Science and Technology University, Beijing 100101, China

Tian Liang LIANGTIAN2022@ZJU.EDU.CN
Zhejiang University, Hangzhou 310058, China

Qiang Zhu(✉) ZHUQ@ZJU.EDU.CN
College of Computer Science, Zhejiang University, Hangzhou 310058, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Weakly-Supervised Temporal Action Localization (WSTAL) aims to train an action instance localization model from untrimmed videos with only video-level labels, similar to the Object Detection (OD) task. Existing Top-k MIL-based WSTAL methods cannot flexibly define the learning space, which limits the model's learning efficiency and performance. Faster R-CNN is a classic two-stage object detection architecture with an efficient Region Proposal Network. This paper successfully migrates the Faster R-CNN liked two-stage architecture to the WSTAL task: first to build a T-RPN and integrate it with the traditional WSTAL framework; and then to propose a pseudo label generation mechanism to enable the T-RPN learning without temporal annotations. Our new framework has achieved breakthrough performances on THUMOS-14 and ActivityNet-v1.2 datasets, and comprehensive ablation experiments have verified the effectiveness of the innovations. Code will be available at: https://github.com/ZJUHJ/TRPN.

Keywords: Weakly-Supervised Learning; Action Localization; Temporal Region Proposal.

1. Introduction

Temporal Action Localization (TAL) Karaman et al. (2014); Wang et al. (2014), which aims to locate the temporal position of the start and end of the action instances from an untrimmed video and identify the action category effectively, has become one of the significant branches of computer vision. Considering the tremendous data labeling costs, Weakly-Supervised Temporal Action Localization (WSTAL) Wang et al. (2017), i.e., to achieve the action localization model learning with only video-level labels, is considered to be more in line with actual needs and has become mainstream of TAL research.

Existing WSTAL methods mainly use multi-instance learning (MIL) supervision paradigm since temporal annotations cannot be accessed. They predict and pool the class distribu-

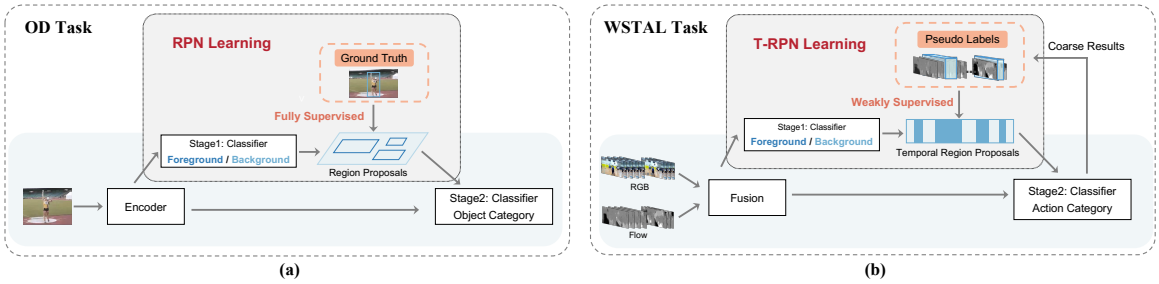

Figure 1: Migration of Two-Stage Faster R-CNN-like Framework.

tion of each snippet to obtain the video-level prediction. However, the sparsity of action instances in videos becomes a challenge, which weakens the loss from action salient parts and hinders training. To mitigate this, most WSTAL methods use top-k pooling to aggregate the k most confident snippets. But the action occurrence parts vary across videos, so even the k highest-scoring snippets cannot fully represent the video. This may cause false negatives when k snippets miss some action instances, or false positives otherwise. Moreover, this design is hard to transfer across domains.

WSTAL is very similar to OD task, for both tasks are dedicated to accurately locating the subspace containing target instances from inputs and completing further recognition. Faster R-CNN Ren et al. (2015) is a classic two-stage object detection framework. As shown at the left of Figure 1: Stage-1 aims to learn a Region Proposal Network (RPN) for foreground/background classification; Stage-2 aims to learn a classifier for object classification and boundary optimization with image features and region proposals. Such an architecture design constrains the learning space of stage-2 within region proposals, which makes the model very effective.

Correspondingly, we can migrate the two-stage detection framework to the WSTAL task. As shown in the right of Figure 1. Stage-1 formulates a temporal region proposals network (T-RPN) to coarsely localize the video foreground parts in a video. Stage-2 further classifies and refines the action category and boundary of the extracted region proposals. This design enables the model to focus on the regions relevant for action localization and ignore the irrelevant background, thus significantly improving the model's efficiency. Given the difference between the two tasks, the migration should consider the following aspects:

1) **How to generate the temporal region proposals for the WSTAL task**: The main challenge is how to integrate a temporal region proposal network (T-RPN) with the WSTAL framework, which generates proposals by scoring each snippet, rather than the proposal coordinates as the OD method does. Moreover, we need to avoid the T-RPN from over-focusing on the salient parts, which impairs its background recognition.

2) **How to achieve the weakly-supervised training process**: The core difference between the OD and WSTAL is that the supervision signals of RPN training related to the OD task come from the ground-truth labeling of the training set, while the videos of the WSTAL task lack the temporal annotation of the action instance. Hence, a well-designed pseudo label derived from the coarse classification results of stage 2 is essential for enabling the learning of T-RPN without relying on temporal annotation.

In this paper, we present a *Dual-Modality Temporal Region Proposal Localization* (DTRP-Loc), a novel two-stage detection framework for WSTAL. We formulate T-RPN by treating each video snippet as an anchor and sample temporal proposals based on their foreground probabilities. Then, we achieve the refinement of the proposals by optimizing the scores assigned to each snippet. To alleviate the imbalance of foreground and background learning caused by the introduction of the T-RPN, we propose to adopt a co-learning mechanism during the training phase. Finally, we optimize the T-RPN learning without temporal annotations by generating high-quality pseudo labels based on the classification results of stage-2. Our method achieves an average mAP of 47.0% on THUMOS-14 with t-IoU thresholds ranging from [0.1, 0.7], and an average mAP of 27.0% on ActivityNet-v1.2 with t-IoU thresholds ranging from [0.5, 0.95], both achieve the current state-of-the-art performance.

The core innovations and main contributions of this work are summarized as follows:

- Migrate a Faster R-CNN-like two-stage object detection architecture to the WSTAL task for the first time by formulating a T-RPN and integrating it with the conventional MIL-based WSTAL framework;
- Design a robust pipeline for generating pseudo labels that offer high-quality supervision for enabling T-RPN learning under the weakly-supervised paradigm;
- Promote the SOTA results of WSTAL on two major benchmarks, and sufficient ablation experiments also prove the universality and effectiveness of T-RPN learning for the WSTAL task.

2. Related Work

Considering the TAL task as a similar classic vision task of Object Detection (OD), and drawing inspiration from the OD research, there are already a lot of existing works to refer to. Zheng *et al.* followed the design of the multi-scale sliding window in the OD task to locate Temporal Region Proposals in the time dimension Shou et al. (2016). However, similar to the OD problem, encoding the features of all candidate windows will result in huge computational overhead. SPPNet He et al. (2015) and Fast R-CNN improve this problem by using feature pyramid and RoI pooling respectively. By mapping feature maps of different scales into fixed-dimension vectors, image encoding is required only once. MSCA Liu et al. (2019) used this idea to improve the feature mapping and aggregation of candidates in the TAL task, which greatly improves computational efficiency. The RPN learning proposed by Faster R-CNN realizes a more efficient Region Proposals algorithm, and R-C3D Xu et al. (2017) drew on similar ideas and can accept video input of any length. In addition, there was also an anchor-free design Lin et al. (2018, 2019) that can not use a predefined anchor, which can directly predict the boundary probability and action score of each video clip to obtain reliable proposals. However, the above techniques all require fine-grained labels for each action instance that occurs in the video to achieve fully-supervised learning, which cannot be directly transferred to the WSTAL task.

Weakly-Supervised Temporal Action Localization (WSTAL) does not need to label fine-grained action instances, only the category labels of the entire video, which makes the WSTAL task more similar to video multi-label classification problems. It should be pointed

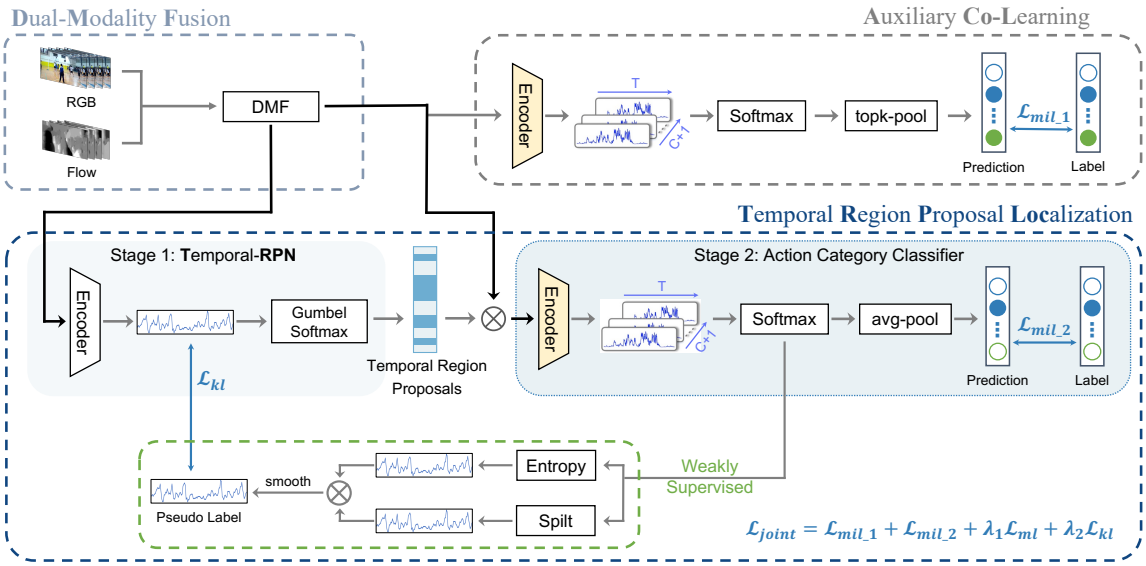

Figure 2: Our WSTAL Solution via Weakly-Supervised T-RPN Learning.

out that the accurate evaluation of WSTAL needs to predict the start and end times of actions, so it can still be regarded as a TAL task that lacks supervisory signals in the training phase. Wang *et al.* first proposed to divide the video evenly into several snippets, then classify each snippet, and aggregate to obtain class activation sequence (CAS) via the MIL architecture, and further proposed to use an attention network to eliminate the interference of background snippets Wang et al. (2017). *Hong et al.* proposed the Cross-Modal Consensus Module, which enables the two modalities of RGB and optical flow to assist each other by activating weights to remove their respective redundant information Hong et al. (2021). Subsequently, *Min et al.* proposed a two-stream adversarial network Min and Corso (2020), which is used to locate salient parts and provide supplementary information respectively, which improves the problem of CAS-based MIL method focusing too much on high activation fragments Choe and Shim (2019); Feng et al. (2021). Furthermore, some research works introduce pseudo-labels to guide model training Paul et al. (2018); Min and Corso (2020); Zeng et al. (2020); Xu et al. (2017). However, because the WSTAL task lacks fine-grained labels, it cannot use the ground truth to train RPN like Faster R-CNN, making it difficult to benefit from the progress of Object Detection research.

3. Methods

As shown in Figure 2, our proposed WSTAL solution can be divided into the following main components:

- **Dual-Modality Fusion:** To encode, enhance and fuse the RGB and optical flow information of the input video to realize information interaction and fusion between different modalities

- **Temporal Region Proposal Localization:** Using the Faster R-CNN-like framework, the T-RPN of Stage-1 learns to generate temporal region proposals, then performs action categorization in Stage-2 based on the generated proposals;
- **Auxiliary Co-Learning:** A standard WSTAL pipeline is introduced as an auxiliary co-learning branch to improve background recognition during the training phase.

3.1. Dual-Modality Fusion

Since WSTAL is a task that requires the cooperation of optical flow and RGB modalities, we begin by presenting the Dual-Modality Fusion module adopted by our method. First, we refer to the work Wang et al. (2017) to uniformly sample the video into T snippets and encode the RGB and optical flow features to obtain the visual and optical flow features X_{rgb}, $X_{flow} \in \mathbb{R}^{T\times D}$, where X_{rgb}, X_{flow} represent the appearance feature and the motion feature of the video, respectively. The salient area of the appearance feature often represents the foreground part of the static image, while the motion feature represents the area where actions occur. In the context of Action Recognition, the overlapping part of the salient area and the motion area is often the focus of the entire TAL network learning. From the perspective of modal complementary, we can imagine the following two scenarios: 1) A complete action may have a pause moment, at which appearance dominates the action recognition; 2) The static image contains multiple foreground objects, which indicates moving objects are more valuable. At this time, motion dominates action recognition.

Following the above discussion, we propose a simple and effective Dual-Modality Fusion (DMF) for cross-modality learning. The RGB modality generates channel attention to make a channel-wise enhancement to the optical flow feature to obtain the enhanced feature X_{flow^+}. Correspondingly, the optical flow modality generates temporal attention to make a temporal-wise enhancement to the RGB modality and obtain X_{rgb^+}. The DMF module can be regarded as strengthening the specific dimension of the dominant modality with the auxiliary modality and enhancing the superposition of the two modalities. For more detail about DMF, please refer to the supplementary material.

3.2. Temporal Region Proposal Localization

Next, we elaborate on the technical details of each stage of our *Temporal Region Proposal Localization*, which reveals how we formulate the T-RPN module and integrate it with the conventional WSTAL framework. Then, we explain how we design a *Pseudo-Label Generation* mechanism to supervise the learning of T-RPN without temporal annotations.

3.2.1. T-RPN Learning

Similar to the Region Proposal Network under the OD task, the T-RPN we propose first generates a Foreground Score Sequence (FSS) through a foreground classifier and then selects temporal feature segments according to the FSS to generate temporal region proposals.

For foreground extraction, unlike previous attention-based methods, we employ a hard-selection strategy that explicitly separates localization and classification tasks. Specifically, given video features $X \in \mathbb{R}^{T\times D}$, the foreground classifier g_θ obtains the foreground score of the input segment according to the input features, denoted as:

$$FSS_{t\text{-}rpn} = \sigma(g_\theta(X)) \tag{1}$$

where σ is the *sigmoid* activation function, which is for mapping the output of g_θ to $[0, 1]$ so that it corresponds to the probability of action occurrence. In addition, $FSS_{t\text{-}rpn} \in \mathbb{R}^{T\times 1}$, which means that each snippet of the input video is assigned a foreground score.

After obtaining the foreground probability distribution of the entire video, the intuitive idea is to set a hard threshold to filter out temporal region proposals. However, in the WSTAL task, the model tends to continuously enhance more discriminative segments, while ignoring the learning of relatively vague action boundaries. This directly causes the model to generate fragmented temporal region proposals, which cannot completely cover action instances. Therefore, we propose to use Gumbel-Softmax Jang et al. (2016), which introduces some randomness while achieving hard sampling, thus allowing the network to find "overlooked foregrounds". For sampling, we can model it as a binary classification problem, using arg max to align the model's predictions with discrete labels, i.e:

$$TP_{mask} = \textit{one-hot}(\arg\max(\log p)) \tag{2}$$

where p is the probability of the corresponding snippet belonging to the foreground. Considering arg max is a non-derivable operation, we use Softmax to solve the problem and obtain the features of the generated temporal region proposals, denoted as:

$$\begin{aligned} TP_{mask} &= softmax\left(\frac{log(\ p + \mathcal{N})}{\tau}\right) \\ TP_{rgb} &= X_{rgb} \otimes TP_{mask} \\ TP_{flow} &= X_{flow} \otimes TP_{mask} \end{aligned} \tag{3}$$

where τ is the temperature coefficient, which is used to adjust the smoothness of softmax, and $\otimes$ means element-wise multiplication. In this paper, τ is set to a very small value to achieve a hard sampling of temporal region proposals. $\mathcal{N}$ is a standard normal distribution, which introduces some randomness into our sampling process.

It should be pointed out that due to the weakly supervised nature of WSTAL (no action instance-level labels are provided), T-RPN learning cannot directly obtain supervision signals from ground-truth labeling, and a self-supervised task must be constructed to form closed-loop learning. We describe a Pseudo-Label Generation mechanism to alleviate the missing supervisory signal problem in section 3.2.3.

3.2.2. Action Categorization

Given the features sampled according to temporal region proposals, we apply an action classifier to obtain the snippet-level action classification results. Finally, these classification results are aggregated into video-level predictions by averaging pooling and calculating multiple instance learning loss with video labels.

Specifically, we first concatenate the different modality features sampled by the T-RPN module to achieve the cross-modality feature integration:

$$X_{fuse} = concat(TP_{rgb}, TP_{flow}) \tag{4}$$

where $X_{fuse} \in \mathbb{R}^{T \times 2D}$. Next, X_{fuse} is fed into the action classifier for action classification. The action classifier used in our method consists of a 3×1 convolution f_θ and a 1×1 convolution $\mathcal{H}_\theta$. The convolution is performed in the temporal dimension of X_{fuse} using f_θ to capture the temporal dependencies between neighboring snippets, while $\mathcal{H}_\theta$ is responsible for establishing the mapping between intermediate features and action labels.

$$T\text{-}CAS = \mathcal{H}_\theta \left(\phi \left(f_\theta \left(X_{fuse} \right) + b \right) \right) \tag{5}$$

where ϕ represents $relu$ activation function, b is the bias and $T\text{-}CAS \in \mathbb{R}^{c+1}$. It should be noticed that we follow the setup of Lee et al. (2020), using $\{c+1\}$-th class to represent the background. Since the temporal region proposal filtering is performed in the previous stage, the label of the background class is set to 0, i.e $y = \{y_1, y_2, \ldots, y_c, 0\}$. After obtaining the $T\text{-}CAS$, we use average pooling to aggregate it into video-level prediction:

$$p_i^{ac} = \frac{1}{N} \sum_{j \subset \mathcal{T}}^{N} T\text{-}CAS_j^{(i)} \tag{6}$$

where p_i^{ac} denotes the score of i-th action category in the video-level prediction, $T\text{-}CAS_j^{(i)}$ denotes the probability of i-th action category occurring in the j-th snippet, and $\mathcal{T}$ denotes the set of snippets covered by temporal region proposals. The loss of multiple-instance learning can be represented as follows:

$$\mathcal{L}_{mil_1} = - \left(\sum_{i}^{C_+} \frac{y_i}{Y_+} \log p_i^{ac} + \sum_{i}^{C_-} \frac{1 - y_i}{Y_-} \log \left(1 - p_i^{ac} \right) \right) \tag{7}$$

where C_+ and C_- represent positive categories and negative categories, respectively, while Y_+ and Y_- represent the sum of positive categories and negative categories correspondingly. We perform the aforementioned regularization on the labels to balance the contribution of positive and negative categories to the loss.

3.2.3. Pseudo Label Generation

As shown in the module indicated by the weakly-supervised arrow in Figure 2, the pseudo-label comes from the snippet-level score sequence $T\text{-}CAS$ output in the Action Categorization stage. After curve smoothing and confidence-weighted supervision loss, high-quality fine-grained learning of T-RPN under the weak supervision paradigm is achieved.

First, we obtain the foreground score curve from $T\text{-}CAS$ as follows:

$$FSS_{cls} = 1 - split(T\text{-}CAS) \tag{8}$$

where $split(\cdot)$ represents the separation of the score sequences of the background class of $T\text{-}CAS$, from which we can obtain a foreground score sequence $FSS_{cls} \in \mathbb{R}^T$ at the snippet-level. since the video has coherence in the time dimension, there should be similar score distributions between neighboring snippets, and we achieve this constraint using Gaussian filtering $\mathcal{G}$:

$$FSS_{smooth} = Softmax(\mathcal{G}(FSS_{cls})) \tag{9}$$

Here, FSS_{smooth} is a rough pseudo-label that can be used to supervise T-RPN. However, the pseudo-labels face the problem of lower quality in the early stage of training. Supervising

T-RPN with low-quality pseudo-labels can easily cause instability in the model training process, thus affecting the model performance. Inspired by Lee et al. (2021), we propose a dynamic adjustment strategy for loss weights based on the uncertainty of the model, making Kullback-Leibler Divergence the loss of supervised T-RPN learning:

$$\begin{aligned} \varepsilon^{(i)} &= -\sum_{j}^{C+1} log\ T\text{-}CAS_i^{(j)} \\ \mathcal{L}_{kl} &= -\frac{1}{T}\sum_{i}^{T} -\varepsilon^{(i)} FSS_{t\text{-}rpn}^{(i)} log \frac{FSS_{smooth}^{(i)}}{FSS_{t\text{-}rpn}^{(i)}} \end{aligned} \tag{10}$$

Through fine pseudo-label, we can improve the T-RPN learning and output high-quality temporal region proposals, so as to achieve the performance improvement of WSTAL.

3.3. Auxiliary Co-Learning

The *Action Categorization* module in Section 3.2.2 performs action classification on the foreground segment selected by T-RPN. However, since this process excludes background segments without action occurring, the learning of background segment recognition ability is neglected. To alleviate this problem, we introduce an **A**uxiliary **Co**-**L**earning (ACoL) branch to construct a collaborative learning mechanism with the above-mentioned Two-Stage Localization framework. We aggregate the losses of both by:

$$\mathcal{L}_{joint} = \mathcal{L}_{mil_1} + \mathcal{L}_{mil_2} \tag{11}$$

where $\mathcal{L}_{mil_1}$ represents the learning loss of Stage-2, and $\mathcal{L}_{mil_2}$ represents the multiple instance learning loss of the ACoL branch. By co-optimizing the two losses of $\mathcal{L}_{mil_1}$ and $\mathcal{L}_{mil_2}$, we can better guide the main task learning to improve the Action Category Classifier in Stage-2. Next, we describe the ACoL branch in detail.

Specifically, the ACoL supplements the consideration of the background to enhance the background recognition ability. The input of the ACoL is the complete temporal feature sequence X without being selected by T-RPN, and the snippet-level score sequence of $T\text{-}CAS_{full}$ is obtained by the same processing steps as *Action Categorization* in Section 3.2.2. In addition, the video label corresponding to the ACoL is $y = \{y_1, y_2, \ldots, y_c, 1\}$, i.e., it always contains background snippets. This branch shares the parameters of the action classifier with the Action Categorization of Stage-2 and thus can obtain improvement for background recognition during the co-training process.

Following the setting of the standard WSTAL framework, we hereby apply *topk-pooling* to aggregate $T\text{-}CAS_{full}$ into a video-level prediction:

$$p_i^s = \frac{1}{K}\sum_{j \subset top(k)}^{K} T\text{-}CAS_{full}{}_j^{(i)} \tag{12}$$

where p_i^s denotes the predicted score of the *i-th* class of actions after aggregation of the standard WSTAL branch, and $top(K)$ denotes the K highest scoring snippets. Similar to Section 3.2.2, the multiple instance learning loss for the ACoL branch is calculated as follows:

$$\mathcal{L}_{mil_2} = -\left(\sum_{i}^{C_+} \frac{y_i}{Y_+} \log p_i^s + \sum_{i}^{C_-} \frac{1-y_i}{Y_-} \log\left(1-p_i^s\right)\right) \tag{13}$$

3.4. Learning and Inference

To sum up, in the training phase, we will jointly optimize all the above training objectives, that is, the final loss function can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{joint} + \lambda_1 \mathcal{L}_{ml} + \lambda_2 \mathcal{L}_{kl} \tag{14}$$

where λ_1 and λ_2 are hyper-parameters that adjust the importance of $\mathcal{L}_{ml}$ and $\mathcal{L}_{kl}$.

The inference of our method follows the two-stage process. We first use the action classifier in Stage-2 to generate the $T\text{-}CAS$ of the test video and follow the standard post-processing process Hong et al. (2021) to obtain temporal region proposals, then group the continuous classification scores of snippets to produce action proposals, and finally to remove redundant outputs with non-maximum suppression (NMS).

4. Experiments

4.1. Datasets

To evaluate the effectiveness of our model, we validate our method on two well-known WSTAL benchmarks:

THUMOS-14 Idrees et al. (2017) is a well-known temporal action localization dataset with 101 action categories from daily scenes and sports scenes. The dataset contains 1,010 untrimmed videos as the validation set and 1,547 as the testing set. For a fair comparison, we follow the setting of previous work Zhang et al. (2021) and use 200 validation videos with temporal labeling for model training and 213 test videos for evaluation.

ActivityNet-v1.2 Heilbron et al. (2015) is a larger TAL dataset with 100 complex daily action categories, which contains 4,819 videos for the training set, 2,383 videos for the validation set, and 2,480 videos for the testing set, with an average of 1.5 actions per video. Since the annotations for the testing are not provided, we evaluate the model performance on the validation set, as in previous work Islam et al. (2021).

4.2. Implementation Details and Metrics

Implementation Details: For a fair comparison, we follow the previous work Chen et al. (2022) to calculate optical flow from RGB video frames based on the TVL1 Pérez et al. (2013) algorithm and use the kinetics-400 dataset pre-trained I3D Carreira and Zisserman (2017) model to extract snippet-level features. We sample 16 consecutive and non-overlapping frames from the video as a snippet, and the number of snippet sampling is set to 320 and 60 for THUMOS-14 and ActivityNet-v1.2 datasets, respectively. We use Adam as the optimizer, and for both datasets, the initialized learning rate and weight decay are set to 1e-4 and 1e-3, respectively. The size of the mini-batch is set to 16. For the hyperparameter α, we set it to 0.9 and 0.1 in Stage-1 and Stage-2, respectively. And λ_1 and λ_2 are

set to 1 and 0.8 respectively. All the experiments are implemented on PyTorch 1.8.1, and the training and testing processes are operated on 1× Nvidia Tesla A100 GPU.

Evaluation Metrics: Like the previous work Chen et al. (2022), we use the mAPs under different intersection-over-union (IoU) thresholds as the metric to evaluate the performance of the model, where the threshold is set to [0.1:0.1:0.7] on THUMOS-14, and [0.50:0.05:0.95] on ActivityNet-v1.2. We adopted the ActivityNet-v1.2 provided evaluation toolkit for the evaluation calculation.

Supervision	Method	mAP@t-IoU(%)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	[0.1:0.5]	[0.3:0.7]	AVG
Fully	SSN	60.3	56.2	50.6	40.8	29.1	-	-	47.4	-	-
	TAL-Net	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	45.1	
	P-GCN	69.5	67.5	63.6	57.8	49.1	-	-	61.5	-	-
	BU-TAL	-	-	53.9	50.7	45.4	38.0	28.5	-	43.3	-
	VSGN	-	-	66.7	60.4	52.4	41.0	30.4	-	50.2	-
Weakly	TSCN	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	28.8	37.8
	CoLA	66.2	59.2	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	AUMN	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4	41.5
	FAC-Net	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	33.0	42.2
	CO2Netv	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
	ASM-Loc	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	FTCL	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
	RSKP	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	DCC	69.0	63.8	55.9	45.9	35.7	24.3	13.7	54.1	35.1	44.0
	DELU	**71.5**	**66.2**	56.5	47.7	40.5	27.2	**15.3**	56.5	37.4	46.4
	DTRP-Loc (Ours)	71.4	66.0	**58.0**	**49.3**	**41.4**	**28.0**	15.0	**57.2**	**38.3**	**47.0**

Table 1: Performance comparison on THUMOS-14 dataset

4.3. Comparison with SOTAs

As shown in Table 1, we compare our proposed method with SOTA results on the THUMOS-14 dataset. In addition to the recent WSTAL work, we also list some fully supervised TAL methods for reference. The SOTA methods we compare are list as follows: SSN Zhao et al. (2017), TAL-Net Chao et al. (2018), P-GCN Zeng et al. (2019), BU-TAL Zhao et al. (2020), VSGN Zhao et al. (2021), TSCN Zhai et al. (2020), CoLA Zhang et al. (2021), AUMN Luo et al. (2021), FAC-Net Huang et al. (2021), CO2Net Hong et al. (2021), ASM-Loc He et al. (2022), FTCL Gao et al. (2022), RSKP Huang et al. (2022), DCC Li et al. (2022), DELU Chen et al. (2022). The results illustrate that on the three most significant indicators of mAP@[0.1:0.5], mAP@[0.3:0.7], and mAP@AVG, our method surpasses the most advanced DELU by 0.7%, 0.9%and 0.6%. Table 2 shows the performance comparison on the ActivityNet-v1.2 dataset. The results illustrated that our proposed method still obtains SOTA results on most indicators.

The experimental results on both datasets reveal an interesting "coincidence" that our methods perform especially outstanding in the case of high t-IoU for both datasets. We infer that the T-RPN learning outputs tend to generate high-quality temporal region proposals that highly overlap with the ground-truth action fragments. Of course, this may decrease accuracy on low t-IoU ranges because T-RPN may tend to discard incomplete

Method	mAP@t-IoU(%)			
	0.5	0.75	0.95	AVG
UGCT	41.8	25.3	5.9	25.8
CoLA	42.7	25.7	5.8	26.1
D2-Net	42.3	25.5	5.8	26.0
CO2-Net	43.3	26.3	5.2	26.4
ACGNet	41.8	26.0	5.9	26.1
DELU	**44.2**	26.7	5.4	26.9
DTRP-Loc (Ours)	43.7	**26.9**	**6.1**	**27.0**

Table 2: Performance comparison on ActivityNet-v1.2 dataset

DMF	T-RPN	ACoL	PLG	mAP@AVG
				42.6
✓				43.5
	✓			44.0
✓	✓			44.8
	✓	✓		45.6
✓	✓	✓		46.5
	✓		✓	44.7
✓	✓		✓	45.3
	✓	✓	✓	46.0
✓	✓	✓	✓	**47.0**

Table 3: Validation of proposed components, where DMF, ACoL and PLG represent Dual-Modality Fusion, Auxiliary Co-Learning branch and Pseudo-Label Generation, respectively.

temporal region proposals. We also find that the fully-supervised TAL methods are with a similar trend (perform better on high t-IoU threshold ranges than low ranges), proving that our method provides high-quality supervision (pseudo-labels) that improves the weakly supervised learning.

4.4. Ablation Study

Effectiveness of Proposed Component: Table 3 presents the ablation experiments on the significant parts of the proposed model to explore their contribution. The results observed that each proposed module achieves a gain in model performance. Firstly, Dual-Modality Fusion (DMF) can significantly enhance the base model, which shows that our proposed cross-modal interaction can make the semantic information of the input features more explicit. The introduction of T-RPN further improves the model performance (+1.4 and +1.3 w/wo DMF), which can provide a better learning space for the action classifier and effectively alleviate the semantic ambiguity problem of the Topk-MIL learning paradigm, thereby improving the learning quality. Introducing the Auxiliary Co-Learning Branch (ACoL) alleviates the balance of foreground and foreground learning problems, resulting

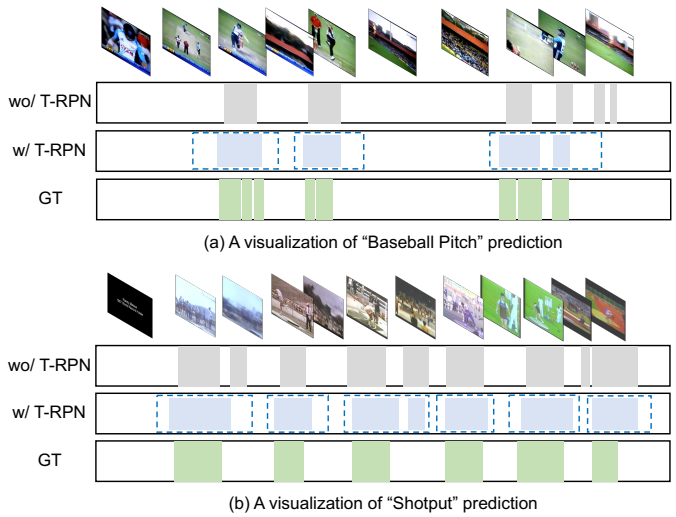

Figure 3: Visualization of the comparison between model with and without T-RPN. Gray regions and blue regions represent proposals generated by models w/wo T-RPN respectively, blue dashed regions are temporal region proposals generated by T-RPN, and green regions are ground truth.

Methods	w/o Proposal	T-RPN	T-RPN*	GT Label
CO2Net	44.4	45.2	45.5	46.0
ASM-Loc	**45.1**	45.8	46.0	46.8
DTRP-Loc	45.0	**46.5**	**47.0**	**47.7**

Table 4: Performance Comparison on different temporal region proposal generation strategies. T-RPN* represents the T-RPN trained with the pseudo labels.

in a considerable gain in the model performance. Finally, although T-RPN can be learned end-to-end, our Pseudo-Label Generation (PLG) strategy supervises the learning process of T-RPN, which can further improve the model performance by 0.5 to 47.0, which shows that the refined pseudo-label can realize the high-quality supervision of T-RPN process.

Impact of Temporal Region Proposals: Figure 3 shows the visualization of the comparison between model with and without T-RPN. First, we show the proposals generated by the model wo/ T-RPN (i.e., the gray regions). Then, we show the proposals and temporal region proposals generated by the model w/ T-RPN (the blue regions and the dashed regions). By comparing them with the ground truth (the green regions), we have the following observations: 1) the model can suppress the false positive predictions after integrating T-RPN. This benefits from the fact that T-RPN confines the learning space of stage-2 to the temporal region proposals, thus filtering out the background; 2) the temporal region proposals generated by T-RPN can effectively cover the regions where the action instances occur, and this region range is much smaller than the whole video, so the proposed method can reduce the task learning difficulty compared with traditional methods.

The above ablation experiment has proved the effectiveness of T-RPN learning. We can further verify the universality and effectiveness of temporal region proposals for all existing WSTAL solutions in this experiment. As shown in Table 4, we demonstrate the model performance on CO2Net Hong et al. (2021), ASM-Loc He et al. (2022), and our method under different strategies of TP generation: without TP, TP generated using T-RPN, TP generated using pseudo-label optimization (i.e., T-RPN*) and with ground-truth labeling. It can be seen that temporal region proposals and their qualities (T-RPN <T-RPN* <Ground Truth) have a significant impact on model performance, which fully proves the importance and effectiveness of the T-RPN learning strategy proposed in this paper.

5. Conclusion

Aiming at the similarities and subtle differences between OD and WSTAL tasks, we focus on several technical challenges in migrating the two-stage OD framework to WSTAL: firstly, we formulate a T-RPN and integrate it with the conventional WSTAL framework; secondly, we construct Pseudo-Labels to complete the self-supervised learning closed-loop of T-RPN; finally, we design a co-learning mechanism to prevent the T-RPN from collapsing under weakly-supervised paradigm.

The core logic of object detection is to first look for "locations where objects may exist (Proposals)" in the source, and then judge "what is the object in this location (Classification)". Therefore, due to its advantages in learning efficiency and computational resource consumption, we believe that building a WSTAL model resembling the OD framework could potentially emerge as a pivotal research direction for addressing long video action localization tasks in the future, warranting continued exploration

Acknowledgments

This work was supported by the National Key RD Program of China (2022ZD0115904).

References

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.

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.

Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 192–208. Springer, 2022.

Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 961–970. IEEE, 2015.

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.

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.

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.

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.

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.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, volume 1, page 5, 2014.

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.

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.

Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19914–19924, 2022.

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.

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.

Haijun Liu, Shiguang Wang, Wen Wang, and Jian Cheng. Multi-scale based context-aware net for action detection. *IEEE Transactions on Multimedia*, 22(2):337–348, 2019.

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.

Kyle Min and Jason J Corso. Adversarial background-aware loss for weakly-supervised temporal activity localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 283–299. Springer, 2020.

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.

Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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.

Limin Wang, Yu Qiao, Xiaoou Tang, et al. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017.

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.

Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2526–2534, 2020.

Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7094–7103, 2019.

Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 37–54. Springer, 2020.

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.

Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021.

Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer, 2020.

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.