Temporal Feature Enhancement Dilated Convolution Network for Weakly-supervised Temporal Action Localization

Jianxiong Zhou, Ying Wu
jianxiongzhou2026@u.northwestern.edu, yingwu@northwestern.edu

INTRODUCTION

Weakly-supervised temporal action localization (WTAL) aims to classify and locate action instances in untrimmed videos with only video-level labels. We argue there are two limitations that hinder the performance of WTAL.

Limitations:

- The short temporal span of snippets. A complete action instance usually covers a relatively long temporal span, while a snippet is unable to observe the full dynamics of the action instance.
- The inappropriate initial features. Most WTAL methods directly use the RGB and optical flow features extracted by pre-trained models, e.g., I3D, which are customized and trained for trimmed video action classification rather than WTAL.

Motivations:

- **Temporal Feature Enhancement Dilated Convolution Module (TFE-DC)** enlarges the receptive field, enabling the model to obtain temporal information of complete action instances and eliminating incoherence of temporal information caused by the short temporal span of snippets.
- **Modality Enhancement Module** keeps the consistency between the two modalities and introduces improved optical flow features to enhance RGB features.

Contributions:

- A novel architecture to solve two limitations of WTAL
- The TFE-DC Module that reflects the influence of temporal information at different receptive scales on final attention weights
- A Modality Enhancement Module that keeps the consistency between two modalities and enhances RGB features
- Our method outperforms all state-of-the-art WTAL methods on THUMOS'14 and ActivityNet v1.3

PROBLEM FORMULATION

Given a set of N untrimmed videos $\{v_n\}_{n=1}^{N}$ and the video-level categorical labels $\{y_n\}_{n=1}^{N}$, where $y_n \in \mathbb{R}^C$ is a normalized multi-hot vector and C is the number of action categories, the goal of WTAL is to generate classification and temporal localization results of all action instances as action proposals for each video.

We divide each video v_n into 16-frame non-overlapping snippets and sample a fixed number of T snippets to represent the video. The RGB features $X_n^{RGB} = \{x_{n,i}^{RGB}\}_{i=1}^{T}$ and the optical flow features $X_n^{Flow} = \{x_{n,i}^{Flow}\}_{i=1}^{T}$ are extracted from the sampled RGB snippets and optical flow snippets respectively with the pre-trained feature extractor, i.e., I3D. $x_{n,i}^{RGB}, x_{n,i}^{Flow} \in \mathbb{R}^D$ are features of the i-th RGB snippet and optical flow snippet, and D is the feature dimension.

METHOD

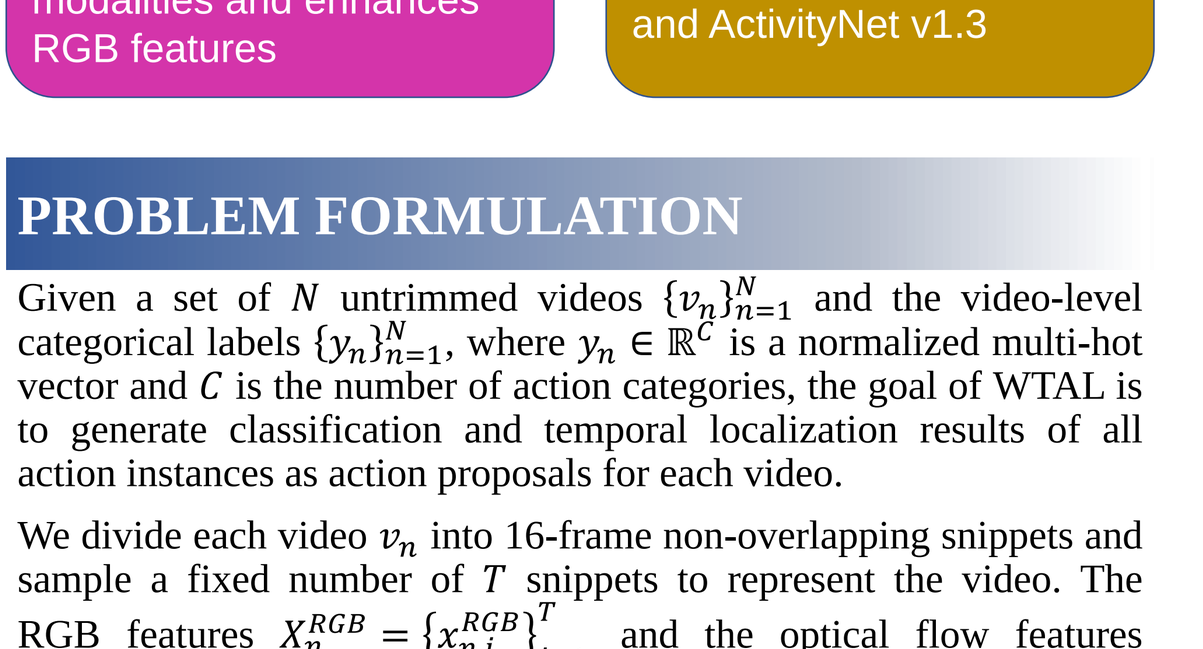

Figure 1: An overview of the Temporal Feature Enhancement Dilated Convolution Network, which consists of four parts: (1) pre-trained feature extractor that outputs RGB features and optical flow features; (2)Temporal Feature Enhancement Dilated Convolution Module (TFE-DC Module) that generates enhanced optical flow features and temporal attention weights; (3) Modality Enhancement Module that generates enhanced RGB features and spatial attention weights; (4) classifier and element-wise multiplication that generate the Temporal Class Activation Sequence (TCAS) and suppressed TCAS.

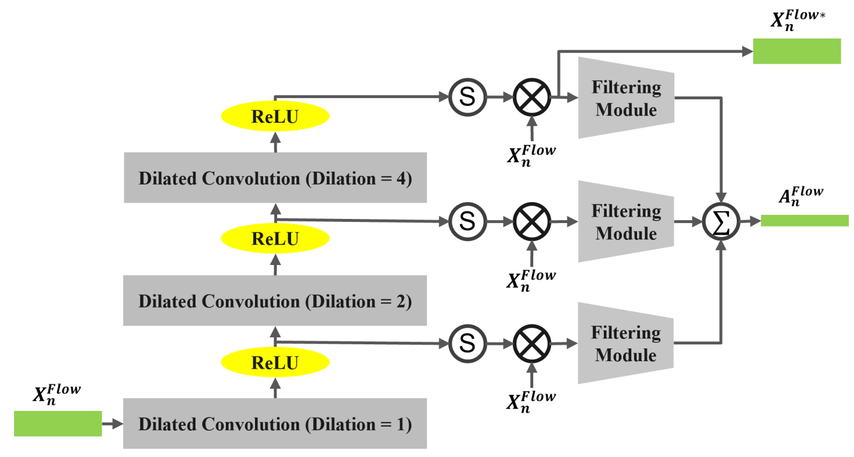

Figure 2: An overview of the proposed TFE-DC Module. The module contains a K-layer dilated convolution network (K=3 in this figure) to enlarge the receptive field and capture dependencies between snippets with different temporal scales. It also has an attention weights generation mechanism that averages the attention weights obtained from the outputs of each layer. This makes the final attention weights A_n^{Flow} can cover temporal information of receptive fields with different sizes.

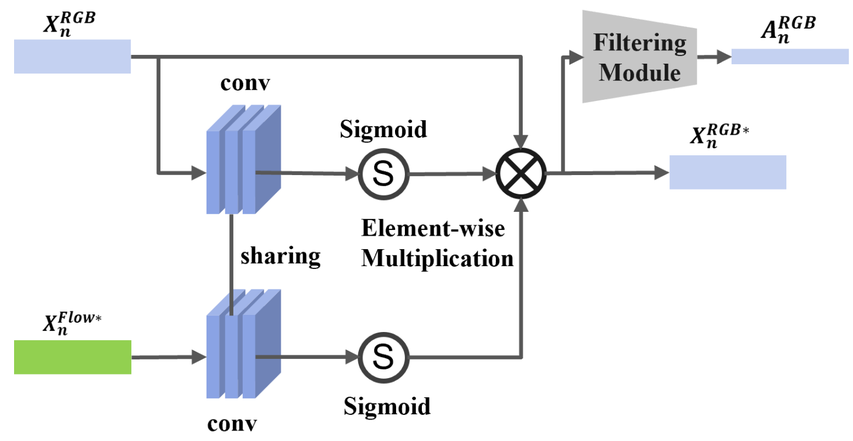

Figure 3: An overview of the proposed Modality Enhancement Module. This module aims to enhance RGB features X_n^{RGB} with the help of enhanced optical flow features X_n^{Flow*}. The sharing convolution layer is beneficial to make weights distributions of the two modalities approached. The enhanced RGB features X_n^{Flow*} are fed into the filtering module to obtain spatial attention weights A_n^{RGB}.

RESULTS AND ANALYSIS

Supervision (Feature)	Method	Publication	mAP@IoU (%)							AVG	
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1:0.5	0.1:0.7
Fully (-)	SSN [42]	ICCV'17	60.3	56.2	50.6	40.8	29.1	-	-	47.4	-
	TAL-Net [2]	CVPR'18	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	45.1
	GTAN [24]	CVPR'19	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-
	P-GCN [39]	ICCV'19	69.5	67.5	63.6	57.8	49.1	-	-	61.5	-
Weakly (UNT)	Liu *et al.* [21]	CVPR'19	53.5	46.8	37.5	29.1	19.9	12.3	6.0	37.4	29.3
	BaS-Net [16]	AAAI'20	56.2	50.3	42.8	34.7	25.1	17.1	9.3	41.8	33.6
	TSCN [40]	ECCV'20	58.9	52.9	45.0	36.6	27.6	18.8	10.2	44.2	35.7
Weakly (I3D)	Lee *et al.* [17]	AAAI'21	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	41.9
	CoLA [41]	CVPR'21	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	40.9
	AUMN [25]	CVPR'21	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	41.5
	TS-PCA [22]	CVPR'21	67.6	61.1	53.4	43.4	34.3	24.7	**13.7**	52.0	42.6
	UGCT [37]	CVPR'21	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	43.6
	FAC-Net [10]	ICCV'21	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	42.2
	CO_2-Net [8]	MM'21	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	44.6
	ACGNET [38]	AAAI'22	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	42.5
	FTCL [5]	CVPR'22	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	43.6
	DCC [19]	CVPR'22	69.0	63.8	55.9	45.9	35.7	24.3	**13.7**	54.1	44.0
	Huang *et al.* [11]	CVPR'22	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	45.1
	ASM-Loc [6]	CVPR'22	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	45.1
	TFE-DCN	WACV'23	**72.3**	**66.5**	**58.6**	**49.5**	**40.7**	**27.1**	**13.7**	**57.5**	**46.9**

Table 1. Comparisons of our method with state-of-the-art fully-supervised and weakly-supervised TAL methods on the THUMOS'14 testing set. UNT and I3D are abbreviations for UntrimmedNet features and I3D features, respectively. AVG is the average mAP at multiple IoU thresholds, i.e., 0.1:0.1:0.5 and 0.1:0.1:0.7.

Table 1 and Table 2 show the results on THUMOS'14 and ActivityNet v1.3. The evaluation metric is the mean Average Precision (mAP) under different Intersection-over-Union (IoU) thresholds. Our method reaches 57.5% average mAP (0.1:0.5) on THUMOS'14 and 25.3% average mAP on ActivityNet v1.3.

Method	mAP@IoU (%)			
	0.5	0.75	0.95	AVG
BaS-Net [16], AAAI'20	34.5	22.5	4.9	22.2
TSCN [40], ECCV'20	35.3	21.4	5.3	21.7
ACSNet [23], AAAI'21	36.3	24.2	5.8	23.9
AUMN [25], CVPR'21	38.3	23.5	5.2	23.5
TS-PCA [22], CVPR'21	37.4	23.5	5.9	23.7
UGCT [37], CVPR'21	39.1	22.4	5.8	23.8
FAC-Net [10], ICCV'21	37.6	24.2	6.0	24.0
FTCL [5], CVPR'22	40.0	24.3	**6.4**	24.8
DCC [19], CVPR'22	38.8	24.2	5.7	24.3
Huang *et al.* [11], CVPR'22	40.6	24.6	5.9	25.0
ASM-Loc [6], CVPR'22	41.0	**24.9**	6.2	25.1
TFE-DCN, WACV'23	**41.4**	24.8	**6.4**	**25.3**

Table 2. Comparison of our method with state-of-the-art WTAL methods on the ActivityNet v1.3 validation set. AVG is the average mAP at the IoU threshold 0.5:0.05:0.95.

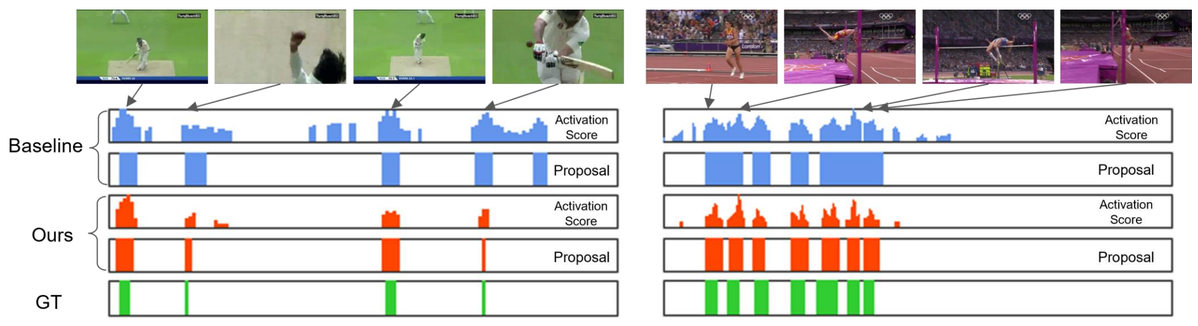

Figure 4: Qualitative visualization of two typical video examples from THUMOS'14. The results of BaS-Net (baseline), our method, and ground truth (GT) are shown in blue, red, and green, respectively.

In Figure 4, we demonstrate the results of two typical video samples. The first example contains category 'Cricket Bowling' and 'Cricket Shot' and each action instance of these two categories is extremely short (about 0.6 sec). While the second example contains the category 'High Jump' and each action instance of this class is relatively long (about 6.1 sec). Our method has more accurate localization results than the baseline (BaS-Net), indicating our method effectively utilizes temporal information. For instance, in the second example, the baseline method incorrectly combines several action instances into one. While our method can locate each action instance correctly.