Slow Motion Matters: A Slow Motion Enhanced Network for Weakly Supervised Temporal Action Localization

Weiqi Sun, Rui Su, Qian Yu* *Member, IEEE* and Dong Xu, *Fellow, IEEE*

arXiv:2211.11324v1 [cs.CV] 21 Nov 2022

Abstract—Weakly supervised temporal action localization (WTAL) aims to localize actions in untrimmed videos with only weak supervision information (*e.g.*, video-level labels). Most existing models handle all input videos with a fixed temporal scale. However, such models are not sensitive to actions whose pace of the movements is different from the "normal" speed, especially slow-motion action instances, which complete the movements with a much slower speed than their counterparts with a "normal" speed. Here arises the slow-motion blurred issue: It is hard to explore salient slow-motion information from videos at normal speed. In this paper, we propose a novel framework termed Slow Motion Enhanced Network (SMEN) to improve the ability of a WTAL network by compensating its sensitivity on slow-motion action segments. The proposed SMEN comprises a Mining module and a Localization module. The mining module generates mask to mine slow-motion-related features by utilizing the relationships between the normal motion and slow motion; while the localization module leverages the mined slow-motion features as complementary information to improve the temporal action localization results. Our proposed framework can be easily adapted by existing WTAL networks and enable them be more sensitive to slow-motion actions. Extensive experiments on three benchmarks are conducted, which demonstrate the high performance of our proposed framework.

Index Terms—Weakly-supervised learning, temporal action localization, slow motion.

I. INTRODUCTION

TEMPORAL action localization (TAL) is an important yet challenging task for video understanding. It aims at localizing the temporal boundaries (*i.e.*, the starting and ending frames) of the actions of interest and recognizing their action categories in untrimmed videos [1], [2]. TAL has wide real-world applications such as video surveillance [3] and abnormality alarm [4] for aged care. Therefore, this task has attracted increasing attention from the research community and embraced great improvements [5]–[14] in recent years. However, these fully supervised methods require extensive manual frame-level annotations, which is labor-consuming and time-costing.

To address this problem, researchers introduced the task of weakly supervised temporal action localization (WTAL). Specifically, instead of using the frame-level temporal annotation, WTAL leverages weaker but cheaper annotations, *e.g.*, video-level labels, during the training process. There are

Weiqi Sun and Qian Yu are with Beihang University. Rui Su is with Shanghai Artificial Intelligence Laboratory. Dong Xu is with the Department of Computer Science, The University of Hong Kong. The corresponding author is Qian Yu (e-mail:qianyu@buaa.edu.cn).

several solutions to the WTAL task. For example, the works in [15]–[17] formulate this problem as a multiple instance learning (MIL) task [18] and treat the entire untrimmed video as a bag containing both positive and negative instances, generating class activation sequence (CAS) to obtain the localization results.

Although the existing WTAL works [15]–[17] have achieved great progress, they overlooked the fact that action instances could exhibit different paces of movement, especially *Slow Motion*, which is defined as an action taking place at a slower than normal speed. Slow motion actions are common, such as the playback in sports videos: In THUMOS'14 [1] dataset, there are more than 64.0% videos and 26.4% action instances containing slow motion. As shown in Fig. 1a, the two clips are from the same video. The first clip displays the action "Javelin Throw" at normal speed, and the second clip shows a slow-motion replay of the action. It is clear to see that a slow-motion action is drastically different from the same action at normal speed. A commonly used pipeline of the existing works is to extract features from video frames sampled at a *fixed* rate and then to process the extracted features to get the final predictions. When determining the sampling rate, researchers mainly consider actions occurring at a normal rate, *i.e.*, normal motion, and ignore the action segments with slow motion. As a result, it is hard for the WTAL frameworks following this pipeline to localize slow motion actions.

Localizing slow motion is not easy. First, unlike normal motion which has salient characteristics, slow motion only has subtle changes in consecutive frames within a temporal period, as shown in Fig. 1a. As a result, slow motion is not easy to be detected. Second, slow motion is easy to be confused with background instances as they share similar characteristics that most of the contents are unchanged in adjacent frames. Therefore, to promote the development of WTAL, a model that is not only sensitive to normal-motion actions but slow-motion actions is expected.

This paper addresses the above problems by introducing a novel framework called Slow Motion Enhanced Network (SMEN). The idea is straightforward: To make a model be sensitive to slow motion, the model should see sufficient slow motion samples during the training process. We thus propose a data mining strategy to find slow-motion-related features. SMEN consists of two modules: 1) a Mining module to generate masks for filtering out slow-motion-related features from the whole video features, and 2) a Localization module to leverage both the whole video features and minded slow-

(a) Comparison of a normal motion and a slow motion. These two video segments are sampled from the same video at the same sampling rate.

(b) Comparison of CAS value produced by two models respectively trained on the original video features and sub-sampled features. Baseline network is the same.

Fig. 1. An example of action instance of the category "Javelin Throw" in the THUMOS'14 dataset. The barcode is the ground truth (GT). Specifically, the green barcode is the action instance with *normal motion*, while the red is the action instance with *slow motion*. The following line charts are CAS value of baseline (ACM-NET [19]) using normal-sampled feature and sub-sampled feature, respectively (*i.e.*, CAS_{normal} and CAS_{sub}). The baseline network trained on sub-sampled feature can better handle action instances with slow motion.

motion features to predict temporal boundaries of actions.

Considering that only video-level labels are available under the weakly-supervised setting, the conventional mining strategy which relies on loss is not applicable. Therefore, our proposed mining module mines slow motion actions by using the prior that the action at normal speed is essentially a "speed-up" version of its slow motion. In other words, a slow motion action can be converted, or speed-up, to normal motion through sub-sampling, so that their action characteristics become more salient and they are easier to be detected by the mining module. The proposed localization module then respectively processes the mined slow-motion features and the original video features with a two-branch architecture, and produces the final temporal action localization results by combining predictions of these two branches.

The contributions of this work are three-fold:

(1) We propose a novel weakly supervised temporal action localization (WTAL) framework called Slow Motion Enhanced Network (SMEN). To the best of our knowledge, this is the first work to explore the salient slow-motion information in the WTAL task.

(2) We introduce a novel slow-motion Mining strategy, which utilizes a prior of slow-motion actions to improve their "actionness". A two-branch localization network is proposed to handle slow- and normal-motion actions simultaneously.

(3) Comprehensive experiments conducted on two benchmark datasets, THUMOS'14 and ActivityNet v1.3, demonstrate the effectiveness of our framework for the WTAL task. Our proposed SMEN outperforms all state-of-the-art methods by a significant margin. Furthermore, our proposed framework can be easily adapted to different base networks and effectively improve their performance in the WTAL task.

II. RELATED WORKS

Weakly-supervised Temporal Action Localization. More and more studies draw increasing attention to WTAL due to the time-consuming and error-prone manual labeling in a fully-supervised setting. UntrimmedNet [20] introduced a classification module and a selection module for predicting a classification score and selecting relevant video segments, respectively. On top of that, STPN [21] introduces a sparse

loss to enforce the sparsity of selected segments. W-TALC [15], CoLA [22] and FTCL [23] employ deep metric learning to force features from different classes to get farther distance than those from the same classes. Nguyen et al. [21] and BaS-Net [17] introduce an auxiliary background class to model background activity. ACM-NET [19] introduces a class-agnostic action-context branch to tackle the action-context confusion issue. Wang et al. [24] and Li et al. [25] utilize temporal consistency of video to refine localization results especially for less discriminative action segments. ACN [26] features a new coherence loss that better supervises action boundary learning and facilitates proposal regression.

TSCN uses a self-training strategy with pseudo labels produced by two stream branches to improve the performance. UGCT [27] and Huang et al [28] focuses on improving the quality of pseudo labels. CO2-NET [29] utilizes a cross-modal consensus module to filter out the information redundancy in the main modality with the help of information from different perspectives of the auxiliary modality. STAR [30], 3C-NET [31], SF-NET [32], SODA [33], and BackTAL [34] use additional weak information during training and obtain huge improvement.

Slow Motion in Video. Slow-motion action instance often occurs in sports video [35]–[38] in the form of replays. Many existing work [39]–[42] focus on detecting slow motion in sports by playback speed classification. However, these works usually employ a specific domain analysis. Differently, SpeedNet [43] works on any videos and proposes a novel network to automatically predict the "speediness" of moving objects in videos, *i.e.*, whether they are at or slower than their "natural" speed. To the best of our knowledge, most existing WTAL works do not consider that most videos contain action instances with different movements and handle all input videos with a fixed temporal scale. In addition, Zhu et al. [44] also observed that the speed of motion varies constantly. They proposed a Multirate Visual Recurrent Model (MVRM) by encoding frames of a video clip with different intervals and apply it in self-supervised learning in video domain. Unlike previous WTAL works, we introduce a data mining module to explore slow-motion information in a video and design a two-branch network for learning from original video features and the enhanced slow-motion features.

Masking Mechanism. Most WTAL methods tend to focus on the most discriminative action segments but ignore trivial action segments, e.g., the beginning or the end of an action, which results in incomplete action localization. Therefore, masking mechanism is proposed to highlight less discriminative segments. For example, Hide-and-Seek [45] proposes to randomly erase input segments during training, which can force the model to discover less discriminative segments. CleanNet [46] utilizes mask mechanism to help improve boundary regression. And more sophisticated masking mechanisms are used in later work such as Zhong et al. [47], ASSG [48] and A2CL-PT [49].

Although our proposed Mining module also employs a masking mechanism, it is distinguished from others as it specifically focuses on mining slow-motion-related features. Furthermore, the proposed Mining module works in a novel

way by utilizing the prior that the normal-motion action can be treated as the accelerated version of the slow-motion action. Thus we can convert a slow-motion action to the corresponding normal-motion action by sub-sampling without using any additional annotations or loss feedback.

III. METHODOLOGY

In this section, we first introduce basic notations and preliminaries in Section III-A. In Section III-B, we briefly review ACM-NET [19], which is used as the CAS generation backbone in our Slow Motion Enhanced Network (SMEN). We then introduce the proposed SMEN in Section III-C.

A. Notations and Preliminaries

Given an untrimmed video V, we first divide it into non-overlapping 16-frame segments $V = \left\{ v_i \in \mathbb{R}^{16 \times H \times W \times 3} \right\}_{i=1}^{T}$ as in previous methods [15], [17], [21], where T denotes the total number of segments. Each segment v_i is then fed into a pre-trained feature extraction network (e.g. I3D [50]) to generate a d dimension feature vector $x_i \in \mathbb{R}^d$, and feature vectors of T segments are stacked together to form a feature sequence $\mathbf{X} = [x_1, x_2, \cdots, x_T]^\top \in \mathbb{R}^{T \times d}$ as the video representation. Each video has a ground-truth video-level action class label, *i.e.*, a multi-hot vector $\mathbf{Y} = [y_1, y_2, \cdots, y_C, y_{C+1}]^\top$, where C is the number of action classes. $y_c = 1, c \in [1, 2, \cdots, C]$ indicates that the c-th action happens in the input video and $y_{C+1} = 1$ indicates that the input video contains non-action background class.

B. Review of ACM-NET

Our proposed SMEN uses ACM-NET [19] as our CAS generation backbone to produce CAS in order to obtain the sub-sampled mask and generate the temporal action localization results. Note that our proposed SMEN can use any CAS generation methods, and we take ACM-NET as an example in this work. We also evaluate our proposed SMEN on different CAS generation backbones in Section IV.

ACM-NET [19] first utilizes a classification branch to generate the initial CAS (Eq. 1). Then, a three-branch class-agnostic attention module is used to generate three sets of attention weights A for discriminating *action instance*, *action context*, and *non-action background*, respectively (Eq. 2).

$$CAS = \Phi_{cls}(X) \tag{1}$$

$$A = \Phi_{attn}(X) \tag{2}$$

where $CAS \in \mathbb{R}^{T \times (C+1)}$ denotes the classification logit of each action class over time, $A = \left\{ (attn_{ins}(t), attn_{con}(t), attn_{bac}(t)) \right\}_{t=1}^{T} \in \mathbb{R}^{T \times 3}$ indicates the likelihood of t-th snippet being an *action instance*, an *action context*, and a *non-action background*, respectively.

Based on the attention weights generated by these three branches, ACM-NET constructs three sets of CAS as follow:

$$CAS_* = attn_* \times CAS, * = \{ins, con, bac\} \tag{3}$$

Fig. 2. Overview of our proposed Slow Motion Enhanced Network (SMEN). Our proposed SMEN first extracts the video feature $\mathbf{X}$ from the original videos. (a) The Mining module applies sub-sampling operation on the original video feature $\mathbf{X}$ and uses the CAS generation backbone to generate CAS_{sub}, which is then used to produce the mask to generate the enhanced slow-motion feature $\mathbf{X}_{slow}$. (b) The Localization module consists of two branches with one (*i.e.*, the Normal Motion Centric Branch) taking the original video feature $\mathbf{X}$ as the input while the other one (*i.e.*, the Slow Motion Centric Branch) taking the enhanced slow-motion feature $\mathbf{X}_{slow}$ as the input. The CAS and the attention weights (*i.e.*, CAS_{normal}, $attn_{normal}$, CAS_{slow}, $attn_{slow}$) generated from these two branches are combined via Fusion module to output the final predictions.

which are then aggregated to compute video-level classification scores for *action instance*, *action context*, and *non-action background*.

These video-level classification scores are supervised by the pre-defined labels $\mathbf{Y}_{ins} = [y_c = 1, y_{C+1} = 0]$, $\mathbf{Y}_{con} = [y_c = 1, y_{C+1} = 1]$ and $\mathbf{Y}_{bac} = [y_c = 0, y_{C+1} = 1]$ for action instance, action context, and non-action background, respectively.

To train the ACM-NET [19], three binary cross-entropy loss are used in the objective functions for these three branches, which are denoted as L_{ins}, L_{con} and L_{bac}, respectively. The whole ACM-NET is trained by jointly minimizing the overall loss function as follow:

$$L_{acm-net} = L_{cls} + L_{add} \qquad (4)$$

$$L_{cls} = L_{ins} + L_{con} + L_{bac} \qquad (5)$$

$$L_{add} = \lambda_1 L_{gui} + \lambda_2 L_{feat} + \lambda_3 L_{spa} \qquad (6)$$

where L_{gui}, L_{feat} and L_{spa} are the attention guide loss, the action feature separation loss and the sparse attention loss, respectively. We kindly refer readers to [19] for more details about ACM-Net.

C. Slow Motion Enhanced Network

As shown in Fig. 2, our Slow Motion Enhanced Network (SMEN) consists of a Mining module and a Localization module. The Mining module takes the original video feature as input and outputs masks used to filter slow-motion-related features. The Localization module takes both the mined slow-motion features and the original video feature to produce temporal action localization results. We will introduce the details in the following sections.

1) Slow-motion Mining: The proposed Mining module aims to mine slow-motion-related features. However, slow-motion actions do not have characteristics as salient as normal motions. Therefore, we first improve the "actionness" of slow-motion actions by sub-sampling the original video feature $\mathbf{X} \in \mathbb{R}^{T \times d}$ with ratio τ, which results in a speed-up video features $\mathbf{X}_{sub} \in \mathbb{R}^{(T//\tau) \times d}$.

Specifically, we sample one snippet for every τ snippets from the original video feature $\mathbf{X}$ to produce the sub-sampled feature $\mathbf{X}_{sub}$. After sub-sampling, the characteristics of slow motion are enhanced, so that become easier to be detected.

We use ACM-NET as a backbone of the Mining module which generates a mask M to localize slow-motion features within sub-sampled feature $\mathbf{X}_{sub}$. The CAS predicted by the backbone network are used for mask generation.

To obtain high-quality masks, we apply a smooth mask generation mechanism on the predicted CAS. In the smooth mask generation mechanism, we use the maximum values of the CAS across all action classes to represent the action activations, which are then normalized by using the Min-Max Normalization, i.e., $M^{norm} \in \mathbb{R}^{(T//\tau) \times d}$.

Furthermore, inspired by [51], we use the Coefficient of Variation Smoothing to smooth the normalized action activations M^{norm} based on the variation in temporal domain to remove the short temporal segments which are usually considered as noise. Specifically, we define the Coefficient of Variation c_v as in Eq. 7:

$$c_v = \frac{\sqrt{\mathbb{D}\left(M^{norm}\right)}}{\mathbb{E}\left(M^{norm}\right)} \qquad (7)$$

where $\mathbb{D}(\cdot)$ and $\mathbb{E}(\cdot)$ are the functions to calculate the deviation and the mean, respectively. We then use the Coefficient of Variation c_v together with a scale factor s to smooth the normalized action activations M^{norm} as follows:

$$M_i^{smooth} = \left(M_i^{norm}\right)^{\alpha} \qquad (8)$$

$$\alpha = 1 - s \times c_v \tag{9}$$

where M_i^{smooth} and M_i^{norm} are the i-th element in the smoothed action activations $M^{smooth} \in \mathbb{R}^{(T//\tau) \times d}$ and the normalized action activations M^{norm}. Then we apply a binary function to generate the mask M by using a pre-defined threshold θ as following:

$$M_i = \begin{cases} 1, & M_i^{smooth} \geq \theta \\ 0, & M_i^{smooth} < \theta \end{cases} \tag{10}$$

The mask M is then up-sampled to the temporal length of the original video feature $\mathbf{X}$ by nearest neighbor interpolation, finally get mask $M \in \mathbb{R}^{T \times d}$.

As the mask is produced based on the high action activation values generated by using the sub-sampled input feature $\mathbf{X}_{sub}$, it can be used to filter the slow-motion feature from the original video feature $\mathbf{X}$.

2) Temporal Localization: Once slow-motion-related features are obtained, they can be used as complementary for original video features for action localization. Precisely, the Localization module consists of two branches, a Normal-motion branch (N-branch) and a Slow-motion branch (S-branch), which both are built upon our CAS generation backbone (*i.e.*, ACM-NET [19]). Similar to ACM-NET [19], the N-branch takes the original video feature $\mathbf{X}$ as input and produces the CAS and attention weights. While the S-branch receives the enhanced slow-motion feature $\mathbf{X}_{slow}$ as input which are generated by filtering the original video feature with the mask produced by the Mining module. We then apply max pooling operation to combine the CAS and the attention weights generated from these two branches and use the same post-processing method and inference method as in [19] to produce the final weakly supervised temporal action localization results.

3) Training Details: For demonstration, our proposed SMEN framework uses the ACM-NET as the backbone. We use Eq. 4-6 as loss functions to optimize the Mining module and Localization module separately. Note that the Mining module is first trained and then fixed to generate mask M. Besides, as the Localization module consists of two branches, there are two action feature separation loss L_{feat}^{normal} and L_{feat}^{slow}. We combine them as follows:

$$L_{feat}^{fuse} = (1 - \beta)L_{feat}^{normal} + \beta L_{feat}^{slow} \tag{11}$$

In this work, we empirically set the hyper-parameter $\beta = 0.5$. During the inference stage, the Mining module is discarded as the Slow-motion branch of the Localization module is well trained to focus on the salient slow-motion features. Both branches in the Localization module take the original video feature $\mathbf{X}$ as the input.

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

1) Datasets.: We perform extensive experiments on two temporal action localization benchmark datasets THU-MOS'14 [1] and ActivityNet1.3 [61].

THMOUS14 [1] contains 200 and 213 untrimmed videos for validation and testing sets, respectively, and the action instances are annotated with precise temporal action boundaries and action classes from 20 different action categories. On average, each video contains 15.4 action instances, and more than 60% videos included slow motion. In addition, the video length varies from a few seconds to more than one hour, which makes it very challenging, especially for weakly-supervised temporal action localization. Following previous works [17], [19], we use the videos in the validation and the testing sets for training and testing, respectively.

ActivityNet1.3 [61] contains 19,994 untrimmed videos with 10,024, 4,926, and 5,044 videos for training, validation, and testing sets, respectively. All action instances in the training and validation sets are labeled from 200 different action categories. On average, each video contains 1.6 action instances.

HACS [62] dataset contains 50k videos spanning 200 classes, where the training/validation/testing set consists of 38k/6k/6k videos, respectively. Compared with existing benchmarks, HACS contains large-scale videos and action instances, serving as a more realistic and challenging benchmark.

2) Evaluation Metrics.: We use the mean Average Precision (mAP) with different temporal Intersection over Union (t-IoU) thresholds to evaluate our weakly-supervised temporal action localization performance, which denotes as mAP@t-IoU. Specifically, the t-IoU thresholds used to calculate the average mAP is [0.1:0.1:0.7] for THUMOS'14 [1], [0.5:0.05:0.95] for ActivityNet v1.3 [61] and HACS [62].

B. Implementation Details

We use the I3D network [50], which is pre-trained based on the Kinetics dataset [63], as our feature extractor to extract both the RGB and optical flow features. The optical flow maps are generated by using the TVL1 algorithm [64]. Instead of extracting the features for each frame, we divide the video into a set of non-overlapping 16-frame segments and extract the RGB and flow features for each segment.

During the training process, the batch size is set to be 16, 512, and 64 for the THUMOS'14 dataset [1], the ActivityNet v1.3 dataset [61], and the HACS dataset [62], respectively. We train the proposed Mining module for 1500 iterations on the THUMOS'14 and the ActivityNet v1.3 datasets and 10 iterations on the HACS dataset. For the optimizer, we choose the Adam optimizer [65] for all three datasets. The learning rate is set to be 5×10^{-5}, 1×10^{-4}, and 1×10^{-4}, respectively. We increased the learning rate to ten times its value to train the Localization module. The hyperparameter r, θ and s are set to be 4, 0.4 and 0.3 for all datasets.

We kindly refer readers to [19] for other parameters of backbones. All the experiments are conducted with PyTorch [66] on a single GTX 2080Ti GPU.

C. Comparison with the State-of-the-arts Methods

We compare our proposed network with existing fully supervised and weakly supervised temporal action localization methods on two benchmark datasets THUMOS'14 [1], ActivityNet v1.3 [61] and HACS [62].

TABLE I
TEMPORAL LOCALIZATION PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE THUMOS'14 TEST SET [1]. NOTE THAT †
REPRESENTS METHODS THAT UTILIZE EXTERNAL SUPERVISION INFORMATION BESIDES FROM VIDEO LABELS.

Supervision	Year	Methods	mAP@t-IoU(%)							AVG[0.1-0.5]	AVG[0.3-0.7]	AVG[0.1-0.7]
			0.10	0.20	0.30	0.40	0.50	0.60	0.70			
Full	2017	SSN [52]	66.0	59.4	51.9	41.0	29.8	-	-	49.6	-	-
	2018	BSN [53]	-	-	53.5	45.0	36.9	28.4	20.0	-	36.8	-
	2019	G-TAD [54]	-	-	54.5	47.6	40.2	30.8	23.4	-	39.3	-
	2019	GTAN [5]	**69.1**	**63.7**	57.8	47.2	38.8	-	-	55.3	-	-
	2021	BSN++ [55]	-	-	**59.9**	49.5	41.3	31.9	22.8	-	41.1	-
Weak†	2019	3C-NET [31]	59.1	53.5	44.2	34.1	26.6	-	8.1	43.5	-	-
	2020	SF-NET [32]	71.0	63.4	53.2	40.7	29.3	18.4	8.6	51.5	30.2	40.8
	2021	SODA [33]	-	-	53.1	44.9	35.6	26.4	15.8	-	35.2	-
	2021	BackTAL [34]	-	-	54.4	45.5	36.3	26.2	14.7	-	35.4	-
	2021	LACP [56]	**75.7**	**71.4**	**64.6**	**56.5**	**45.3**	**34.5**	**21.8**	**62.7**	**44.5**	**52.8**
Weak	2017	Hide-and-seek [45]	36.4	27.8	19.5	12.7	6.8	-	-	20.6	-	-
	2018	STPN [21]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	18.5	27.0
	2018	W-TALC [15]	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-	-
	2018	Zhong et al. [47]	45.8	39.0	31.1	22.5	15.9	-	-	30.9	-	-
	2018	STAR [30]†	68.8	60.0	48.7	34.7	23.0	-	-	47.0	-	-
	2019	ASSG [48]	65.6	59.4	50.4	38.7	25.4	15	6.6	47.9	27.2	37.3
	2019	CleanNet [46]	-	-	37.0	30.9	23.9	13.9	7.1	-	22.6	-
	2019	ACN [26]	-	-	35.9	30.7	24.2	15.7	7.4	-	22.8	-
	2020	BaS-Net [17]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	43.6	27.3	35.3
	2020	A2CL-PT [49]	61.2	56.1	48.1	39.0	30.1	19.2	10.6	46.9	29.4	37.8
	2020	TSCN [57]	63.4	57.6	47.8	37.7	28.7	19.4	10.2	47.0	28.8	37.8
	2021	Wang et al. [24]	66.1	60.0	52.3	43.2	32.9	-	-	50.9	-	-
	2021	Li et al. [25]	67.8	61.9	54.1	43.7	32.7	22.1	12.3	52.0	33.0	42.1
	2021	ASL [58]	67.0	-	51.8	-	31.1	-	11.4	-	-	-
	2021	CoLA [22]	66.2	59.5	51.5	41.9	32.2	22	13.1	50.3	32.1	40.9
	2021	UGCT [27]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	2021	CO2-Net [29]	70.1	63.6	54.5	45.7	**38.3**	**26.4**	**13.4**	54.0	35.7	44.6
	2021	ACGNet [59]	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	33.4	42.5
	2022	FTCL [23]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
	2022	Huang et al. [28]	<u>71.3</u>	65.3	55.8	<u>47.5</u>	<u>38.2</u>	<u>25.4</u>	12.5	<u>55.6</u>	<u>35.9</u>	<u>45.1</u>
	2022	ASM-Loc [60]	71.2	<u>65.5</u>	<u>57.1</u>	46.8	36.6	25.2	**13.4**	55.4	35.8	<u>45.1</u>
	2021	ACM-NET [19]	68.9	62.7	55.0	44.6	34.6	21.8	10.8	53.2	33.4	42.6
	-	SMEN (Ours)	**74.0**	**68.5**	**60.1**	**49.4**	36.9	23.6	<u>12.9</u>	**57.8**	**36.6**	**46.5**

TABLE II
TEMPORAL LOCALIZATION PERFORMANCE COMPARISON WITH
STATE-OF-THE-ART METHODS ON THE ACTIVITY-NET V1.3 [61]
VALIDATION SET. NOTE THAT † REPRESENTS METHODS THAT UTILIZE
EXTERNAL SUPERVISION INFORMATION BESIDES FROM VIDEO LABELS.

Supervision	Methods	mAP@t-IoU(%)			AVG
		0.50	0.75	0.95	
Full	SSN [52]	39.1	23.5	5.5	24.0
	BSN [53]	46.5	30.0	8.0	30.0
	BSN++ [55]	**51.3**	**35.7**	**8.3**	**34.9**
Weak	STPN [21]	29.3	16.9	2.6	-
	STAR [30]†	31.1	18.8	4.7	-
	ASSG [48]	32.3	20.1	4.0	-
	TSM [67]	30.3	19.0	4.5	-
	BaS-Net [17]	34.5	22.5	4.9	22.2
	TSCN [57]	35.3	21.4	5.3	21.7
	ACSNet [68]	36.3	24.2	5.8	23.9
	Wang et al. [24]	37.1	24.1	5.8	24.1
	UGCT [27]	39.1	22.4	5.8	23.8
	FAC-Net [69]	37.6	24.2	6.0	24.0
	FTCL [23]	40.0	24.3	<u>6.4</u>	24.8
	Huang et al. [28]	40.6	24.6	5.9	25.0
	ASM-Loc [60]	<u>41.0</u>	24.9	6.2	25.1
	Li et al. [25]	40.9	**25.7**	5.6	<u>25.6</u>
	ACM-NET [19]	40.1	24.2	6.2	24.6
	SMEN (Ours)	**41.7**	<u>25.6</u>	**6.6**	**26.0**

1) Results on the THUMOS'14 Dataset: We report the mAP results on the THUMOS'14 dataset [1] in Table I. We observe that our proposed SMEN can achieve the average mAP of 46.5%, which is 3.9% higher than that of the baseline method ACM-NET [19]. We believe this improvement is mainly brought by the salient slow-motion information exploitation in our proposed modules. We also observe that our proposed SMEN can achieve the best *Average mAP* results and outperforms baseline models ACM-NET [19] by 7%, 9%, 9%, 11%, 7%, 8%, 19% at t-IoU=[0.1,:0.1,0.7] in terms of *relative improvement*, indicating that our proposed SMEN have consistent improvement across all t-IoU thresholds. It is worthy to note that the methods like STAR [30], 3C-NET [31], SF-NET [32], SODA [33], and BackTAL [34], use additional information as supervision during training instead of using the video-level annotations as the only supervision. However, our proposed SMEN outperforms all these methods in terms of all evaluation metrics. Furthermore, our proposed SMEN even outperforms the fully supervised method SSN.

2) Results on the Activity-Net v1.3: In Table II, we report the mAPs of our proposed SMEN and compare state-of-the-art approaches at different t-IoU thresholds (*i.e.*, t-IoU=0.5, 0.75, 0.95) on the validation set of the Activity-Net v1.3 dataset. Besides, we also report the average mAP, calculated by

TABLE III
TEMPORAL LOCALIZATION PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS ON THE
HACS [62] VALIDATION SET. NOTE THAT + DENOTES THE METHODS USING FULLY SUPERVISION, † REPRESENTS THE METHODS THAT UTILIZE
EXTERNAL SUPERVISION INFORMATION IN ADDITION TO THE VIDEO LABELS. # DENOTES THE RESULTS ARE BASED ON THE IMPLEMENTATION IN
BACKTAL [34], THE OTHER RESULTS ARE FROM OUR IMPLEMENTATION.

Methods	mAP@tIoU										AVG[0.5-0.95]
	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	
SSN+# [52]	28.8	-	-	-	-	18.8	-	-	-	5.3	19.0
BaS-NET# [17]	30.6	27.7	25.1	22.6	20.0	17.4	14.8	12.0	9.2	5.7	18.5
BackTAL†# [34]	31.5	29.1	26.8	24.5	22.0	19.5	17.0	14.2	10.8	4.7	20.0
ACM-NET [19]	25.4	23.2	21.0	18.6	16.5	14.4	12.5	10.1	7.5	4.7	15.4
SMEN(w/ ACM-NET)	**27.6**	**25.2**	**23.1**	**20.9**	**18.7**	**16.4**	**14.1**	**11.4**	**8.5**	**5.4**	**17.1**
BaS-NET [17]	30.8	28.0	25.7	23.5	21.0	18.7	16.2	13.5	10.5	**6.4**	19.4
SMEN(w/ BaS-NET)	**32.1**	**29.4**	**26.9**	**24.3**	**21.8**	**19.4**	**16.7**	**13.9**	10.5	5.9	**20.1**

TABLE IV
PERFORMANCE COMPARISON OF SEVERAL VARIATIONS FOR OUR
PROPOSED METHOD ON THE THUMOS'14 TEST SET [1]. THE AVERAGE
MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7]. "TB"
DENOTES TWO BRANCHES; "MM" DENOTES MINING MODULE; "SSF"
DENOTES SUB-SAMPLED FEATURE.

TB	MM	SSF	mAP@tIoU							AVG[0.1-0.7]
			0.10	0.20	0.30	0.40	0.50	0.60	0.70	
✗	✗	✗	68.9	62.7	55.0	44.6	34.6	21.8	10.8	42.6
✓	✗	✗	69.9	63.1	54.7	44.2	32.2	22.3	11.7	42.6
✓	✗	✓	71.4	64.8	56.0	45.9	33.9	21.5	11.0	43.5
✓	✓	✗	73.4	66.4	57.8	46.6	34.2	21.7	11.4	44.5
✓	✓	✓	**74.0**	**68.5**	**60.1**	**49.4**	**36.9**	**23.6**	**12.9**	**46.5**

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT β FOR OUR PROPOSED
METHOD ON THE THUEMOS'14 TEST SET. THE AVERAGE MAP (%) IS
COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7].

β	mAP@t-IoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
0.00	71.3	65.0	56.8	47.0	35.1	21.8	11.7	44.1
0.25	72.6	66.8	58.7	47.7	35.8	23.3	12.3	45.3
0.50	**74.0**	**68.5**	**60.1**	**49.4**	**36.9**	**23.6**	**12.9**	**46.5**
0.75	73.6	68.2	59.4	49.0	36.4	23.9	13.1	46.2
1.00	55.1	47.1	37.6	26.5	16.4	9.0	2.9	27.8

TABLE VI
PERFORMANCE COMPARISON BETWEEN DIFFERENT SOTA BACKBONES
AND OUR PROPOSED METHOD ON THE THUEMOS'14 SLOW-MOTION
TESTING SET. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU
THRESHOLDS [0.1:0.1:0.7].

Methods	mAP@tIoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
BaS-NET	43.5	40.4	35.4	31.5	22.7	18.7	14.2	29.5
SMEN(w/ BaS-NET)	**49.0**	**45.3**	**41.1**	**35.9**	**27.5**	**20.9**	**14.3**	**33.4**
ASL	48.3	43.9	37.5	33.2	26.1	20.1	14.1	31.9
SMEN(w/ ASL)	**50.9**	**47.9**	**43.3**	**38.5**	**32.7**	**26.2**	**19.4**	**37.0**
ACM-NET	55.8	52.1	47.2	40.4	31.9	23.3	13.4	37.7
SMEN(w/ ACM-NET)	**58.7**	**54.5**	**48.5**	**42.6**	**35.2**	**24.8**	**16.0**	**40.0**

averaging all mAPs across multiple t-IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05. We observe that our proposed SMEN outperforms all state-of-the-art WTAL approaches in terms of all evaluation metrics, including its CAS generation backbone ACM-NET. Similar to our observations on the THUMOS'14 dataset, it can be seen that the average mAP of our proposed SMEN is higher than that of some fully supervised temporal action localization approaches (*e.g.*, SSN [52]) and STAR [30], which takes advantage of extra supervision information other than video-level labels.

3) Results on the HACS: We conduct experiments on HACS dataset with the baseline methods ACM-NET [19] and BaS-NET [17]. The experimental results are shown in Table III. Our method can achieve the average mAP of 17.1% and 20.1%, which is 1.7% and 0.7% higher than baseline methods ACM-NET and BaS-NET, respectively. The improvements mainly come from the salient slow-motion information exploitation in our proposed modules. It is worthy mentioning that our proposed SMEN, which uses the video-level annotation as the only supervision, outperforms the SOTA method BackTAL [34] slightly, which uses additional weak information during training. Furthermore, our proposed SMEN even outperforms the state-of-the-art fully supervised approach SSN [52].

D. Model Analysis

In this section, we perform a series of experiments on the THUMOS'14 dataset [1] to further demonstrate the contribution of individual components of our proposed SMEN, and

provide more insights about this task. Unless specified, all the reported results are achieved by using ACM-NET [19] as the CAS generation backbone.

1) Ablation Studies: Given that our proposed Localization module consists of two branches, we start from a plain two-branch baseline model, which is essentially an ensemble model of two ACM-NET models. As shown in the first and second row of Table IV, the temporal action localization performance cannot be improved by simply combining two ACM-NET models, i.e., both the baseline model and the two-branch variant achieve the mAP of 42.6%.

Sub-sampled feature. On top of the baseline model, we replace the input of one branch with the sub-sampled features in which the slow-motion-related features are enhanced. We can see that by using sub-sampled features, it can bring about 1% improvement (43.5% vs. 42.6%), which verifies the effectiveness of the prior used in our method, i.e., normal motion actions can be treated as the speed-up version of slow

motion, and the characteristics of the slow motion will become more salient after speeding up.

Mining module. The key of our proposed SMEN is the slow-motion mining module. In our proposed SMEN, the mining module takes the sub-sampled feature as the input during the training process and outputs a mask to be used for filtering slow-motion-related features. As shown in the third and fifth rows of Table IV, the mining module significantly increases the performance from 43.5% to 46.5%, i.e., 3% improvement.

It is interesting to see that even only with the original video features, the mining module can increase the average mAP by 1.9% (44.5% vs. 42.6%) as shown in the second and fourth row. A possible explanation is that the mining module learns to filter out less discriminative features from the original video features for the action-instances with normal motion. This is because a pre-defined threshold (Eq. 10) is used to generate masks by preserving areas with high activation scores, which helps to suppress false positive predictions.

Balance Coefficient β. We conduct an ablation study on β and report the results in Table V. A bigger β means the action feature separation loss generated from the S-branch produces a larger weight. When $\beta = 0.5$, the model achieves the best performance.

Performance on Slow Motion. To test the performance of the proposed method on handling slow-motion actions, a slow-motion testing set is formed by manually annotating the slow-motion action instances in THUMOS'14 testing set. Three volunteers are involved. This slow-motion testing set consists of 126 videos with 839 slow-motion action instances from all 20 action categories. That means there are more than 60% videos and 25% action instances of the THUMOS'14 testing set contain slow motion.

We compare the WTAL performance on slow-motion action instances of three baseline method, i.e. BaS-NET [17], ASL [58] and ACM-NET [19] and our proposed SMEN. Note that we replaced the feature corresponding to normal-motion action with $\mathbf{0} \in \mathbb{R}^d$ during testing. The results are reported in Table VI. We can see that our proposed SMEN achieves the average mAP of 33.4%, 37.0%, 40.0%, surpassing the baseline methods BaS-NET(29.5%), ASL(31.9%), ACM-NET (37.7%) by 3.9%, 5.1%, 2.3%, respectively. This margin demonstrates the superiority of our proposed SMEN in localizing slow-motion actions. All results listed in Table IV and Table VI show that our proposed mining module and the two-branch localization module can work cooperatively to enhance the model's localization ability.

2) The Role of Each Branch in Localization Module: We additionally conduct experiments to show the role of individual branch, i.e., Normal-motion branch (N-branch) and Slow-motion branch (S-branch), of the localization module. We first *directly* evaluate their performance respectively, and the results are shown in Table VII. It is not surprisingly to see that the S-branch performs poorly on this task since this branch only "see" slow-motion-related features during training.

To reveal the role of the S-branch, we provide the performance of the variant that using the CAS predictions of the N-branch and the attention predictions of the S-branch, i.e.,

TABLE VII
PERFORMANCE COMPARISON OF SEVERAL VARIANTS FOR OUR PROPOSED METHOD ON THE THUMOS'14 TESTING SET. *N*-BRANCH DENOTES THE RESULTS FROM *N*ORMAL-BRANCH, WHILE *S*-BRANCH DENOTES THE RESULTS FROM *S*LOW-BRANCH. SMEN-COMBO DONATES THE VARIANT WHERE CAS IS FROM THE *N*ORMAL-BRANCH WHILE ATTENTION VALUES FROM THE *S*LOW-BRANCH. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7].

sub-sampled feature	mAP@tIoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
N-branch	70.8	63.3	53.2	41	28.9	18.1	9.7	40.7
S-branch	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.1
SMEN	**74.0**	**68.5**	**60.1**	**49.4**	**36.9**	**23.6**	**12.9**	**46.5**
SMEN-combo	73.5	67.6	59.2	48.0	34.8	22.7	12.1	45.4

TABLE VIII
PERFORMANCE COMPARISON OF OUR PROPOSED METHOD WHEN USING DIFFERENT BACKBONES ON THE THUMOS'14 TEST SET [1]. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.9].

Methods	mAP@tIoU									AVG[0.1-0.9]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
BaS-NET	58.2	52.3	44.6	36.0	27.0	18.6	10.4	3.9	0.5	27.9
SMEN(w/ BaS-NET)	60.7	55.3	47.9	38.5	28.5	21.4	11.8	4.5	0.7	29.9
ASL	67.0	-	51.8	-	31.1	-	11.4	-	0.7	32.2
SMEN(w/ ASL)	69.0	63.7	54.9	43.7	32.9	22.7	14.3	6.5	1.0	34.2

SMEN-combo in Table VII. In other words, SMEN-combo differs from the N-branch in that it uses the attention values predicted by the S-branch to compute the final CAS (Eq. 3). As shown in Table VII, SMEN-combo outperforms *N*-branch by 4.7%, suggesting that the *S*-branch alone is incapable of localizing a specific action, but it helps *N*-branch by providing more accurate attention weights. To be more specific, originally the slow-motion actions are easily confused with action context or non-action background by the N-branch. Since the S-branch learns from the enhanced slow-motion features, it becomes more sensitive than the N-branch in discriminating action-instance, action-context, and non-action background.

3) Generalization: Our proposed SMEN framework can be easily adapted to different baseline methods. We conduct experiments by changing the CAS generation backbone with different existing networks, such as BaS-NET [17] and ASL [58]. The average mAPs (at t-IoU thresholds [0.1:0.1:0.9]) of our proposed SMEN with different CAS generation backbones are reported in Table VIII. We observe that SMEN (w/ BaS-NET) and SMEN (w/ ASL) can achieve the average mAP of 29.9% and 34.2%, which are 2.0% higher than those of the baseline methods BaS-NET [17] and ASL [58], respectively. This indicates that our proposed SMEN can generalize to different CAS generation methods and consistently improve their performance in the task of weakly-supervised temporal action localization.

4) Smooth Mask Generation: In order to verify the effectiveness of the proposed Smooth Mask Generation strategy used in the Mining module, we conduct experiments by comparing two different methods: 1) generating mask with a fixed threshold, i.e., *normal mask*; 2) generating mask with the proposed Coefficient of Variation Smoothing mechanism which is inspired by [51], i.e., *smooth mask*. Our proposed SMEN adopts the latter option. Table IX compares the average mAPs achieved by using these two mask generation methods

(a) An example of action class "SoccerPenalty".

(b) An example of action class "Blizzards".

(c) An example of action class "LongJump".

(d) An example of action class "Diving".

Fig. 3. Qualitative comparisons with baseline on THUMOS'14 dataset. The barcode is the ground-truth (GT). Specifically, the green barcode is the action instance with *normal motion* while the red barcode is the action instance with *slow motion*. The following line charts are CAS value of baseline model (ACM-NET [19]) and our proposed SMEN, respectively. For beteer clarity, the frames with green bounding boxes refer to ground-truth actions with *normal motion*, the frames with red bounding boxes refer to ground-truth actions with *slow motion*, while those in blue refer to ground-truth *backgrounds*. The red dotted boxes show that our proposed SMEN performs better in localizing slow-motion action instances. The yellow dotted box shows that SMEN can suppress the false positive instance, *i.e.*, *Background→Action*, incorrectly predicted by the baseline model (ACM-NET [19]).

(a) A failure example of action class "JavelinThrow".

(b) A failure example of action class "Diving".

Fig. 4. Failure case of our proposed SMEN. The barcode is the ground-truth (GT). Specifically, the green barcode is the action instance. The following line charts are CAS value of baseline model (ACM-NET [19]) and our proposed SMEN, respectively. For better clarity, the frames with green bounding boxes refer to ground-truth actions with *normal motion*, while those in blue refer to ground-truth *backgrounds*. The purple dotted boxes show that our proposed SMEN can not well localize the action instances of short duration.

TABLE IX
PERFORMANCE COMPARISON OF OUR PROPOSED METHOD WHEN USING DIFFERENT MASK GENERATION METHODS ON THE THUMOS'14 TEST SET [1]. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7].

mask generation methods	mAP@tIoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
normal	73.4	67.6	59.4	48.9	36.5	23.7	13.3	46.1
smooth	74.0	68.5	60.1	49.4	36.9	23.6	12.9	46.5

TABLE X
PERFORMANCE COMPARISON OF OUR PROPOSED METHOD WHEN USING DIFFERENT SUB-SAMPLING METHODS ON THE THUMOS14 TEST SET [1]. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7].

sub-sampled methods	mAP@tIoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
frame	74.1	68.1	58.9	47.9	36.0	22.9	12.1	45.8
feature	74.0	68.5	60.1	49.4	36.9	23.6	12.9	46.5

TABLE XI
PERFORMANCE COMPARISON OF OUR PROPOSED METHOD WHEN USING DIFFERENT SUB-SAMPLING RATE ON THE THUMOS14 TEST SET [1]. THE AVERAGE MAP (%) IS COMPUTED AT T-IOU THRESHOLDS [0.1:0.1:0.7].

τ	mAP@tIoU							AVG[0.1-0.7]
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	
1	73.4	66.4	57.8	46.6	34.2	21.7	11.4	44.5
2	73.5	67.7	58.8	47.8	35.4	22.9	12.0	45.5
4	**74.0**	**68.5**	**60.1**	**49.4**	**36.9**	**23.6**	**12.9**	**46.5**
8	72.5	66.5	58.1	48.0	35.4	23.4	12.4	45.2

TABLE XII
COMPARISON BETWEEN OUR PROPOSED METHOD AND THE BASELINE METHODS IN TERMS OF MEMORY USAGE, TRAINING TIME AND INFERENCE TIME ON THE THUMOS'14 DATASET [1].

Methods	Memory	traing time (s/epoch)	inference time (s/video)
BaS-NET	3536MB	6.64	0.02
SMEN(w/ BaS-NET)	5430MB	11.27	0.03
ASL	1743MB	3.87	0.03
SMEN(w/ ASL)	2239MB	4.64	0.04
ACM-NET	2335MB	4.92	0.07
SMEN(w/ ACM-NET)	3201MB	7.93	0.10

on the THUMOS'14 dataset. We can observe that using the smooth mask generation mechanism can achieve a 0.4% higher average mAP than using the normal mask. A possible explanation for this is that the smooth mask generation mechanism can de-noise CAS values and produce more robust masks for selecting slow-motion-related features.

5) Sub-sampling Strategy: **Feature-level vs. Frame-level.** We sub-sample on video features in our proposed framework to speed-up slow motion to normal motion. However, we can

alternatively speed up the motions by sub-sampling on *video frames*. Therefore, we compare the performance of using these two sub-sampling strategies.

Table X shows the average mAPs achieved by using different sub-sampling methods on the THUMOS'14 dataset. We can see that sub-sampling on video features can achieve 0.7% higher average mAP than sub-sampling on input video frames. Apart from superior performance, sub-sampling on video features is more efficient as it does not require re-extracting video features by a pre-trained feature extractor.

Sub-sampling Rates. We also investigate the influence of different sub-sampling rates for generating sub-sampled features. The results are reported in Table XI. It can be seen that when the sub-sampling rate $\tau = 4$, our proposed SMEN can achieve the best temporal action localization performance. A possible explanation is that a lower sub-sampling rate may not speed up the slow motion actions enough to stand out their salient characteristics, while a larger sub-sampling rate may lead to losing too much temporal information, making it hard to discriminate from background instances.

6) Efficiency: We compare the efficiency of several baseline methods and our proposed SMEN on the machine with a single GTX 2080Ti GPU. The results are shown in Table XII. The inference time of our proposed SMEN is slightly higher than the baseline methods.

E. Qualitative Results

We visualize the localized regions and the CAS results for four actions on the THUMOS'14 dataset in Fig. 3. Our proposed SMEN has a more informative CAS distribution compared to baseline method, thus leading to more accurate localization for slow-motion segments.

Figure 3a depicts a typical case ("SoccerPenalty") that the video contains two segments, one is normal motion and the other is slow motion. Specifically, the second action segment is the slow-motion replay of the first. As shown in Fig. 3a, both segments begin from preparing Soccer Penalty. After 3-second interval, the normal-motion segment turns to finish, but the slow-motion segment just turns to torch the ball. With the Mining module, the Localization module can leverage the generated mask to filter slow-motion-related features as complementary information to improve the temporal action localization results. As a result, while the backbone model fails to localize the slow-motion action segments, our proposed SMEN can accurately localize the action instance at a much slower speed. Another example in shown in Fig. 3b.

On the other hand, by introducing the Mining module, our proposed SMEN can filter out less discriminative features, thereby avoiding many false positives produced by a single branch (backbone). Fig. 3c demonstrates an example of "LongJump" action in which the backbone model incorrectly localizes a background segment (highlighted in blue box). This background segment belongs to the class of "action context" which is not the target action but contains similar scenes and slow movements. Such action context segments are easily confused with actions, especially slow motion actions. Another example in shown in Fig. 3d.

By considering the slow motion, our proposed SMEN can distinguish action-instances from slow-motion-alike action-contexts, and therefore, improve the WTAL results.

Failure Cases. As shown in Fig. 4, the baseline method and our proposed SMEN cannot well handle the action instances with very short duration well (highlighted in purple dotted boxes). A possible explanation is both the baseline and our method do not have any specific designs for such cases. We will leave this problem in our future work.

V. CONCLUSION

This paper has proposed a novel framework, SMEN, to address the problem that slow motion provides little information for the WTAL frameworks to understand the content and distinguish them from action instances. The new framework consists of two modules, a Mining module and a Localization module. The Mining module is proposed to generate slow-motion-related mask and the Localization module is designed to leverage the generated mask to select enhanced slow-motion features as complementary information to improve the temporal action localization results. Experiments conducted on three benchmarks, including THUMOS'14, ActivityNet v1.3, and HACS, have validated the state-of-the-art performance of SMEN.

REFERENCES

[1] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, "The thumos challenge on action recognition for videos "in the wild"," *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.

[2] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2782–2795, 2013.

[3] K. Yang, D. Liu, Z. Chen, F. Wu, and W. Li, "Spatiotemporal generative adversarial network-based dynamic texture synthesis for surveillance video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 359–373, 2022.

[4] Y. Li, W. Cao, W. Hu, and M. Wu, "Abnormality detection for drilling processes based on jensen–shannon divergence and adaptive alarm limits," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 6104–6113, 2020.

[5] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[6] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[7] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[8] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[10] Y. Liu, J. Chen, X. Chen, B. Deng, J. Huang, and X.-S. Hua, "Centerness-aware network for temporal action proposal," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 5–16, 2022.

[11] L. Shao, S. Jones, and X. Li, "Efficient search and localization of human actions in video databases," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 3, pp. 504–512, 2014.

[12] H. Eun, S. Lee, J. Moon, J. Park, C. Jung, and C. Kim, "Srg: Snippet relatedness-based temporal action proposal generator," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4232–4244, 2020.

[13] L. Xu, X. Wang, W. Liu, and B. Feng, "Cascaded boundary network for high-quality temporal action proposal generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3702–3713, 2020.

[14] Y. Chen, B. Guo, Y. Shen, W. Wang, W. Lu, and X. Suo, "Capsule boundary network with 3d convolutional dynamic routing for temporal action detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[15] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *European conference on computer vision*, 2018.

[16] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," *arXiv preprint arXiv:1905.08586*, 2019.

[17] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[18] Z.-H. Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, 2004.

[19] S. Qu, G. Chen, Z. Li, L. Zhang, F. Lu, and A. Knoll, "Acm-net: Action context modeling network for weakly-supervised temporal action localization," *arXiv preprint arXiv:2104.02967*, 2021.

[20] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[21] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[22] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, "Cola: Weakly-supervised temporal action localization with snippet contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[23] J. Gao, M. Chen, and C. Xu, "Fine-grained temporal contrastive learning for weakly-supervised temporal action localization," *arXiv preprint arXiv:2203.16800*, 2022.

[24] B. Wang, X. Zhang, and Y. Zhao, "Exploring sub-action granularity for weakly supervised temporal action localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[25] B. Li, R. Liu, T. Chen, and Y. Zhu, "Weakly supervised temporal action detection with temporal dependency learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[26] Y. Zhai, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, "Action coherence network for weakly supervised temporal action localization," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3696–3700.

[27] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[28] L. Huang, L. Wang, and H. Li, "Weakly supervised temporal action localization via representative snippet knowledge propagation," *arXiv preprint arXiv:2203.02925*, 2022.

[29] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng, "Cross-modal consensus network for weakly supervised temporal action localization," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[30] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proceedings of the AAAI conference on artificial intelligence*, 2019.

[31] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[32] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou, "Sf-net: Single-frame supervision for temporal action localization," in *European conference on computer vision*, 2020.

[33] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "Soda: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2474–2498, 2021.

[34] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-click supervision for temporal action localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[35] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4578–4590, 2020.

[36] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2018.

[37] L. Kong, D. Pei, R. He, D. Huang, and Y. Wang, "Spatio-temporal player relation modeling for tactic recognition in sports videos," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[38] L. Kong, D. Huang, J. Qin, and Y. Wang, "A joint framework for athlete tracking and action recognition in sports videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 532–548, 2020.

[39] L. Wang, X. Liu, S. Lin, G. Xu, and H.-Y. Shum, "Generic slow-motion replay detection in sports video," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, 2004.

[40] C.-M. Chen and L.-H. Chen, "A novel method for slow motion replay detection in broadcast basketball video," *Multimedia Tools and Applications*, vol. 74, no. 21, pp. 9573–9593, 2015.

[41] A. Javed, K. B. Bajwa, H. Malik, and A. Irtaza, "An efficient framework for automatic highlights generation from sports videos," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 954–958, 2016.

[42] V. Kiani and H. R. Pourreza, "An effective slow-motion detection approach for compressed soccer videos," *International Scholarly Research Notices*, vol. 2012, 2012.

[43] S. Benaim, A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, and T. Dekel, "Speednet: Learning the speediness in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9922–9931.

[44] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2653–2662.

[45] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[46] Z. Liu, L. Wang, Q. Zhang, Z. Gao, Z. Niu, N. Zheng, and G. Hua, "Weakly supervised temporal action localization through contrast based evaluation networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[47] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[48] C. Zhang, Y. Xu, Z. Cheng, Y. Niu, S. Pu, F. Wu, and F. Zou, "Adversarial seeded sequence growing for weakly-supervised temporal action localization," in *Proceedings of the 27th ACM international conference on multimedia*, 2019.

[49] K. Min and J. J. Corso, "Adversarial background-aware loss for weakly-supervised temporal activity localization," in *European conference on computer vision*, 2020.

[50] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[51] Y. Li, Z. Kuang, L. Liu, Y. Chen, and W. Zhang, "Pseudo-mask matters in weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[52] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[53] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *European conference on computer vision*, 2018.

[54] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[55] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, "Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation," *arXiv preprint arXiv:2009.07641*, 2020.

[56] P. Lee and H. Byun, "Learning action completeness from points for weakly-supervised temporal action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 648–13 657.

[57] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, and G. Hua, "Two-stream consensus network for weakly-supervised temporal action localization," in *European conference on computer vision*, 2020.

[58] J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, "Weakly supervised action selection learning in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[59] Z. Yang, J. Qin, and D. Huang, "Acgnet: Action complement graph network for weakly-supervised temporal action localization," *arXiv preprint arXiv:2112.10977*, 2021.

[60] B. He, X. Yang, L. Kang, Z. Cheng, X. Zhou, and A. Shrivastava, "Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization," *arXiv preprint arXiv:2203.15187*, 2022.

[61] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[62] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8668–8678.

[63] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijaya-narasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[64] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*, 2007.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[66] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[67] T. Yu, Z. Ren, Y. Li, E. Yan, N. Xu, and J. Yuan, "Temporal structure mining for weakly supervised action detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[68] Z. Liu, L. Wang, Q. Zhang, W. Tang, J. Yuan, N. Zheng, and G. Hua, "Acsnet: Action-context separation network for weakly supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2233–2241.

[69] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Weiqi Sun is currently a master candidate in School of Software, Beihang University. His current research interests include weakly-supervised temporal action localization and video retrieval.

Rui Su received the B.S. degree in school of Electronic and Information Engineering from South China University of Technology in 2013, the MPhil. degree in school of Information Technology and Electrical Engineering from the University of Queensland in 2016, and the PhD. degree in the school of Electrical and Information Engineering, the University of Sydney. His current research interests include action detection and its applications in computer vision.

Qian Yu an associate professor at Beihang University. Before joining Beihang, she was a post-doctoral research scientist with UC Berkeley and International Computer Science Institute, working with Professor Stella Yu. She received her doctorate from the Queen Mary University of London, co-supervised by Professor Yi-Zhe Song and Professor Tao Xiang. Her research interest is sketch understanding and cross-modality modeling. Her research works have been published in top-tier journals and conferences including IJCV, CVPR, ECCV and ICCV. She was the recipient of Best Scientific Paper of BMVC 2015. She is one of the main organizers of ICCV 2021 workshop *Sketching for Human Expressivity*.

Dong Xu received the BE and PhD degrees from University of Science and Technology of China, in 2001 and 2005, respectively. While pursuing the PhD degree, he was an intern with Microsoft Research Asia, and a research assistant with the Chinese University of Hong Kong, for more than two years. He was a post-doctoral research scientist with Columbia University, for one year. He also worked as a Faculty Member at Nanyang Technological University, and the Chair of computer engineering at The University of Sydney. He is currently a Professor with the Department of Computer Science, The University of Hong Kong. His current research interests include computer vision, statistical learning, and multimedia content analysis. He was the co-author of a paper that won the Best Student Paper award in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2010, and a paper that won the Prize Paper award in IEEE Transactions on Multimedia (T-MM) in 2014. He is a fellow of the IEEE.