Similar Modality Enhancement and Action Consistency Learning for Weakly Supervised Temporal Action Localization

Maodong Li[1], Chao Zheng[1], Jian Wang[1*], Bing Li[1,2]

[1]School of Computer Science, Wuhan University
[2]Hubei Luojia Laboratory, Wuhan
{limaodong2023, chaozheng, jianwang, bingli}@whu.edu.cn

Abstract

Weakly-supervised temporal action localization (WTAL) aims to identify and localize action instances in untrimmed videos using only video-level labels. Existing methods typically rely on original features from frozen pre-trained encoders designed for trimmed action classification (TAC) tasks, which inevitably introduces task discrepancy. Additionally, these methods often overlook the importance of considering action consistency from multiple perspectives, specifically the consistency in action processes and action semantics, both of which are crucial for the model's understanding of actions. To address these issues, we propose a novel WTAL method based on **s**imilar modality **e**nhancement and **a**ction consistency **l**earning (SEAL). First, we construct global descriptors for each action category, and use the pseudo-labels generated based on these descriptors to guide the model in learning more consistent representations, thereby mitigating task discrepancy. Second, we design two types of losses to achieve action consistency learning: process consistency loss, which penalizes candidate proposals that deviate from the action center to ensure the completeness of the action process, and semantic consistency loss, which employs local descriptors to help proposals of the same action category (especially those with apparent semantic confusion) learn similar feature distributions. Extensive experiments on the THUMOS14 and ActivityNet datasets demonstrate the superior performance of the proposed method compared to state-of-the-art methods.

Code — https://github.com/Lkydong2020/SEAL_WTAL

Introduction

Temporal Action Localization (TAL) aims to accurately locate action instances in untrimmed videos (Wang et al. 2023), enabling broad applications in video editing (Tejero-de Pablos et al. 2018), industrial video analysis (Li and Zhao 2021), intelligent surveillance (Yun et al. 2019), human-computer interaction (Vignolo et al. 2017), and more.

To date, most research efforts in TAL have been based on fully supervised methods (Lin et al. 2019; Zhao et al. 2022; Zhang, Wu, and Li 2022). However, fully supervised TAL (FTAL) relies on meticulously annotated frame-level labels, which are costly and time-consuming in real-world applications (Li et al. 2023). In contrast, Weakly Supervised Temporal Action Localization (WTAL) (Li et al. 2023; Wang et al. 2017) only requires video-level labels, which has therefore attracted increasing attention from researchers.

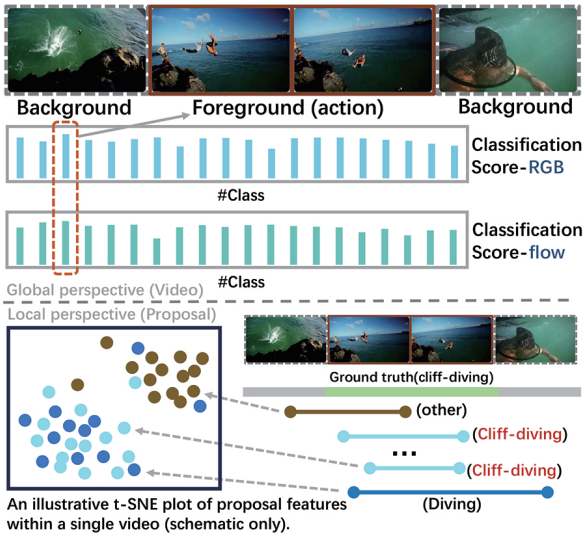

Figure 1: Conceptual illustration of consistency in learning. At a global level, enforcing cross-modal consistency—aligning RGB and flow representations toward shared target categories—reduces task discrepancies and mitigates the impact of noisy or modality-specific artifacts. At a local level, promoting semantic similarity among proposals from the same action class fosters more coherent feature distributions, thereby enhancing the model's discriminative capability and improving localization accuracy.

Recent WTAL methods predominantly adopt a segment-level multiple instance learning framework (S-MIL) (Wang et al. 2017; Huang, Wang, and Li 2021; Luo et al. 2020a) within the pre-classification pipeline. Based on a finely designed two-branch architecture, S-MIL is able to parallelize the processing of original RGB features and optical flow features extracted from frozen pre-trained encoders and transform the localization task into a classification task. In this framework, the class-aware branch is used to generate segment-level class activation sequences (CAS), which pro-

*Corresponding author.

vide segment-level category information, while the class-agnostic branch is used to generate segment-level foreground scores. By integrating the output sequences of both branches and applying manually designed multi-threshold strategies, the model generates numerous high-quality action proposals for effective temporal region detection.

However, most architectures, including S-MIL, typically rely on frozen pre-trained encoders designed for trimmed video action classification tasks. Due to the differing objectives of these tasks, these segment-level video encoders often overlook the handling of action boundaries and background frames, inevitably introducing task discrepancy (Xu et al. 2021). To address this issue, some studies (such as (Hong et al. 2021)) have encouraged cross-modal (or cross-branch) mutual supervision and collaboration to mitigate task discrepancy and enhance localization performance. Although the specific implementations vary, there is an underlying consensus: maintaining consistency in the learning process can lead to higher-quality localization results. Specifically, leveraging the complementarity of different modalities helps disregard task-irrelevant information, as illustrated in Figure 1. However, achieving this typically necessitates the incorporation of additional learning layers for feature refinement or the development of sophisticated pseudo-label strategies. In contrast to these schemes, we designed the relatively simple and lightweight Similar Modality Enhancement (SME) Module, which corrects the learning direction of the model solely through the use of simple pseudo-labels. Specifically, we fully leverage the original input features to construct hard-global descriptors (Appearance and Motion Descriptors) for each action category, preserving a feature distribution similar to the original. By utilizing the similarity scores calculated between these descriptors and the original video features, we generate high/moderate-confidence pseudo-labels. These pseudo-labels guide the classifier to learn more consistent representations, further mitigating task discrepancy.

Moreover, the attention mechanism in the dual-branch S-MIL architecture disproportionately emphasizes the most discriminative segments of an action, hindering the model's ability to capture the complete temporal dynamics of action instances. Additionally, as shown in Figure 1, since the main goal of existing methods is to precisely distinguish between foreground and background using confidence scores rather than thoroughly exploring the semantics of each frame, the model often neglects the semantic consistency of segments within the same action category. This may cause the model to confuse semantics when handling segments of the same action category and lead to the accumulation of noise and uncertainty when aggregating proposal scores, making it difficult to accurately regress temporal boundaries. To address these issues, we observed actions from multiple perspectives and proposed Action Consistency Learning (ACL). This mechanism is implemented through two newly designed loss functions: Process Consistency (PC) Loss and Semantic Consistency (SC) Loss. The process consistency loss treats the central region of each action instance as the action center of the action process. Given the same Intersection over Union (IoU), candidate proposals that capture the action center are considered to have better process consistency, thereby encouraging the model to generate more complete localization results. The semantic consistency loss assumes that features of the same action category share a consistent semantic space. It aggregates proposal-level features that yield coherent classification outcomes under different classification methods (basic/background suppression) into soft-local descriptors. These descriptors update dynamically to provide supervisory signals aligned with the model's training progress, thereby enhancing semantic consistency among features associated with the same action category. The synergy between these two losses enables the model to comprehensively learn more consistent action representations, achieving more accurate localization.

Overall, our contributions can be summarized as:

- We propose a SME module deriving two sets of pseudo-labels from original features, guiding the model toward balanced, consistent cross-modal representations, thereby alleviating task discrepancies.
- We propose an ACL mechanism that includes a PC Loss, which ensures action completeness by penalizing proposals failing to capture the action center, and a SC Loss, which refines the model's semantic understanding of proposals through soft-local descriptors, thereby improving localization accuracy.
- Extensive experiments on the THUMOS14 and ActivityNet datasets demonstrate that SEAL outperforms several state-of-the-art methods in terms of performance.

Related Work

Fully Supervised Temporal Action Localization. FTAL relies on frame-level annotations to localize actions in untrimmed videos. Existing methods are categorized into anchor-based (Gao et al. 2017; Lin, Zhao, and Shou 2017) and anchor-free approaches (Yang et al. 2020; Lin et al. 2021). Anchor-based methods include single-stage (Yang et al. 2022) and two-stage approaches (Kang et al. 2023): the former predicts boundaries and categories simultaneously, while the latter first generates proposals using predefined anchors, followed by boundary regression and category prediction. In contrast, anchor-free methods directly regress instance boundaries without anchors, offering simplicity and robustness but facing temporal center uncertainty issues (Wang et al. 2023). Overall, the labor-intensive precise annotations, while bringing higher performance to FTAL, also represent its primary bottleneck for expansion into the real-world scenarios.

Weakly Supervised Temporal Action Localization. In contrast to FTAL, WTAL requires only video-level labels. WTAL methods generally fall into two categories: pre-classification (Lee, Uh, and Byun 2020; Ma et al. 2021) and post-classification pipelines (Xu et al. 2019). Pre-classification pipelines apply temporal convolution to classify each frame, then aggregate these predictions (e.g., via Top-K pooling) to derive video-level labels. Among these, the Multiple Instance Learning (MIL) framework (Luo et al. 2020b) is highly influential, with S-MIL emerging as one

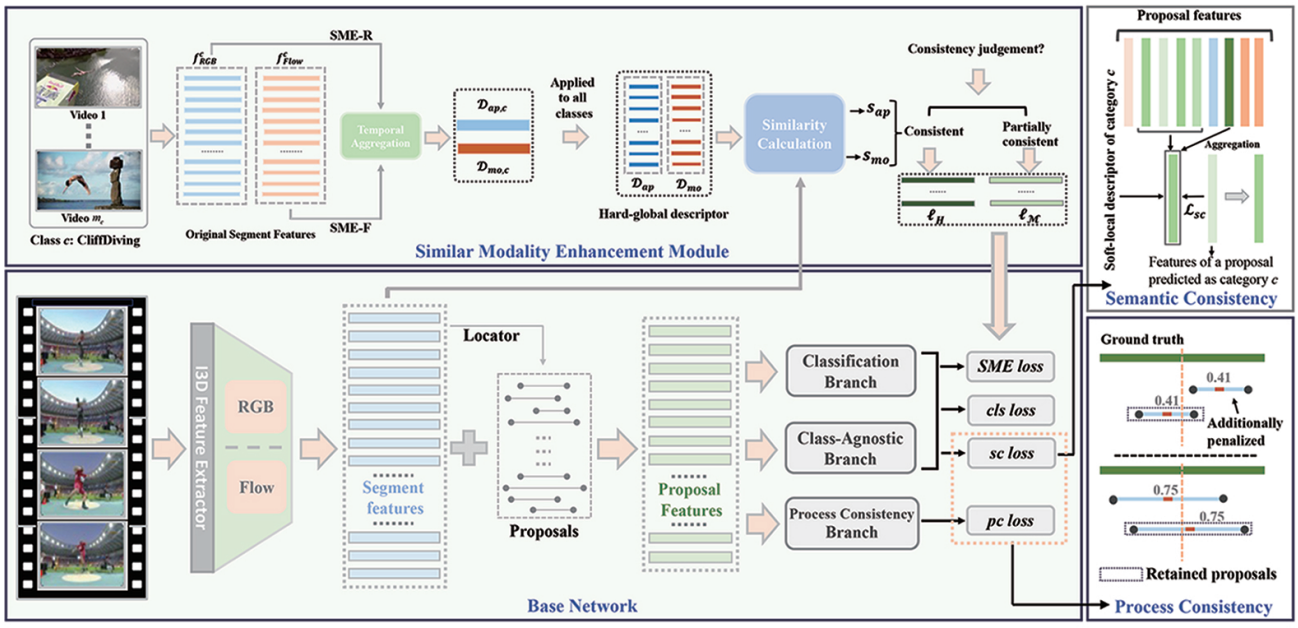

Figure 2: Overview of the Proposed Framework. Our method integrates two key modules—SME and ACL—that collaboratively refine representations and enhance localization. The SME module leverages original input features to construct hard-global descriptors for each action category, generating two types of pseudo-labels that guide the model toward more consistent representations. Meanwhile, the ACL module applies two specialized loss functions, PC loss and SC loss, ensuring complete proposals and aligning semantic features among proposals of the same category.

of its most popular implementations. In contrast, post-classification pipelines first evaluate frame-level task relevance to assign attention weights, then aggregate weighted features into video-level predictions. Attention-based methods (Islam, Long, and Radke 2021) are a well-known example, focusing on distinguishing action frames from background frames through generated attention maps.

Facing the task discrepancy problem addressed in this paper, early approaches often pursued cross-modal collaboration. For example, (Hong et al. 2021) proposed a cross-modal consensus network that partitions modalities into primary and auxiliary streams to filter out task-irrelevant information. More recent efforts emphasize pseudo-label-based methods (Huang, Wang, and Li 2022): ASMLOC (He et al. 2022) employs MIL to generate action proposal pseudo-labels, while (Yun et al. 2024) construct high-fidelity pseudo-labels by inferring salient segment features. In contrast, we derive pseudo-labels directly from original features, focusing on how the model learns from both modalities, resulting in a lighter and simpler approach. Regarding action consistency, the most relevant work (Wang, Li, and Wang 2023) integrates a learnable dictionary with static class centers to capture semantic correlations among segments. Our method, however, constructs soft-local descriptors from predicted proposal-level features, enabling dynamic and flexible semantic consistency learning. Additionally, we consider action completeness from a motion-process perspective, ensuring more accurate and robust localization. These two aspects jointly form our action consistency learning.

Method

Problem Definition

Given a set of N untrimmed videos $V = \{v_i\}_{i=1}^{N}$ containing C different action categories, with only video-level multi-hot encoded category labels $Y = \{y_i\}_{i=1}^{N}$ provided for training, the task is to predict a set of action instances $\Lambda_i = \{(s_j, e_j, c_j, q_j)\}_{j=1}^{M}$ for each test video v_i during the inference phase. Here, $y_i \in \mathbb{R}^{C+1}$, and the additional category is the background category. If action category c appears in the i-th video, then $y_i^c = 1$; otherwise, it is 0. M represents the number of proposals. s_j, e_j, c_j, and q_j correspond to the start time, end time, category, and confidence of the j-th action segment in the i-th video, respectively.

Pipeline Overview

The overall framework of our SEAL is shown in Figure 2.

Segment-level Feature Extraction. Following most previous works, each untrimmed video is divided into non-overlapping 16-frame segments, yielding T total segments. A pretrained extractor is then used to obtain segment-level RGB stream features $f_{RGB} \in \mathbb{R}^{T\times D}$ and optical flow features $f_{flow} \in \mathbb{R}^{T\times D}$, where D denotes the feature dimension. These two features are concatenated to form the original input feature $f \in \mathbb{R}^{T\times 2D}$.

Proposal-based Framework. In this work, we adopt a proposal-based learning framework (Ren et al. 2023) to construct the base model, ensuring scoring consistency during training and inference. The SCFE (Ren et al. 2023) module

is also introduced to preserve the spatiotemporal continuity of video actions. Using proposal-level features $f_{prop} = SCFE(f) \in \mathbb{R}^{M\times 2D}$ and candidate proposals generated by the pretrained S-MIL model without post-processing, the framework supports flexible training and inference.

Similarity Modality Enhancement. As shown in Figure 2, the SME module consists of SME-R and SME-F, corresponding to inputs from two modalities. SME-R constructs appearance descriptors for all action categories, while SME-F builds action descriptors. Together, these form the Hard-global Descriptors. Using the category scores from the similarity calculation between Hard-global Descriptors and video features, the model generates high/moderate-confidence pseudo-labels to supervise consistent action representation learning.

Action Consistency Learning. ACL consists of process consistency loss and semantic consistency loss. In terms of implementation, the semantic consistency loss relies solely on the basic classification score s_{base} and background suppression score s_{supp} from the base model's dual-branch head. To enable process consistency loss, we introduce a dedicated process consistency branch to predict corresponding scores. Finally, the model fuses these three types of scores to produce the final localization.

Similar Modality Enhancement Module

The SME module generates two types of pseudo-labels to guide the learning process, enabling the model to produce action representations better suited for the localization task. Specifically, it consists of two key components:

Hard-global Descriptors. The Hard-global Descriptors are composed of an Appearance Descriptor, which suggests visual recognition information, and a Motion Descriptor, which indicates progressive motion information. Taking the construction of the Appearance Descriptor as an example, for the c-th action category, assuming there are m_c videos of the same category, the RGB features can be denoted as $f^c_{RGB} \in \mathbb{R}^{m_c\times T\times D}$. Based on the assumption that video features of the same action category should have similar semantics, we can obtain two different scales of aggregated features:

$$\mathcal{S}^c_{RGB} = \frac{1}{\sum_{i=1}^{m_c} T_i} \sum_{i=1}^{m_c}\sum_{t=1}^{T_i} f^c_{RGB}\left[i,t,:\right] \in \mathbb{R}^D \quad (1)$$

$$\mathcal{V}^c_{RGB} = \frac{1}{m_c} \sum_{i=1}^{m_c} \left(\frac{1}{T_i} \sum_{t=1}^{T_i} f^c_{RGB}\left[i,t,:\right] \right) \in \mathbb{R}^D \quad (2)$$

where $\mathcal{S}^c_{RGB}$ represents the segment-level appearance descriptor for action category c, emphasizing the frame-level semantic correlation of the action. On the other hand, $\mathcal{V}^c_{RGB}$ represents the video-level Appearance Descriptor for category c, which focuses more on intra-class variations across different videos, highlighting the semantic relationships between actions at the video level. For the sake of simplicity, we assume that one of these methods is chosen to construct the Appearance Descriptor, denoted as $\mathcal{D}_{ap,c}$. Then, the set of Appearance Descriptors for all C action categories can be represented as:

$$\mathcal{D}_{ap} = [\mathcal{D}_{ap,1}, \cdots \mathcal{D}_{ap,c}, \cdots \mathcal{D}_{ap,C}] \in \mathbb{R}^{C\times D} \quad (3)$$

Similarly, the Motion Descriptor set $\mathcal{D}_{mo} \in \mathbb{R}^{C\times D}$ is defined in the same manner as $\mathcal{D}_{ap}$.

By concatenating the two descriptor sets, the Hard-global Descriptor can be obtained as:

$$\mathcal{D}_{hard} = cat[\mathcal{D}_{ap}, \mathcal{D}_{mo}] \in \mathbb{R}^{C\times 2D} \quad (4)$$

Compared to the original features, the Hard-Descriptors highly encapsulate the core semantic information of different action categories, including visual characteristics, motion dynamics, and overall representation, with reduced noise and enhanced robustness. The frame-level descriptors focus on the refinement of local features, while the video-level descriptors provide a unified expression from a global perspective. Both types, to varying degrees, reflect the essential characteristics of actions and the underlying semantic correlations, which form the foundation of our research.

High/Moderate-Confidence Pseudo-labels. Based on the aforementioned design, pseudo-labels are constructed to guide proposal learning. To effectively leverage the complementarity between the two modalities, the model aggregates latent information from both streams as supervisory signals. Specifically, given the feature $f_n \in \mathbb{R}^{T\times 2D}$ of the current n-th video, cosine similarity is used to calculate the similarity scores $s_{ap,n}$ and $s_{mo,n}$ between its features and the two descriptor sets:

$$s_{ap,n} = norm\left(\cos\left(\hat{f}_{RGB,n}, \mathcal{D}_{ap}\right)\right) \in \mathbb{R}^C, \quad (5)$$

$$s_{mo,n} = norm\left(\cos\left(\hat{f}_{flow,n}, \mathcal{D}_{mo}\right)\right) \in \mathbb{R}^C, \quad (6)$$

where $\hat{f}_{RGB,n}$ and $\hat{f}_{flow,n}$ represent the video-level RGB and optical flow features obtained by pooling along the temporal dimension of the n-th video, respectively. For effective localization, the cues provided by features from both streams should exhibit high consistency. Thus, the action category determined by the maximum of the two similarity scores is expected to be the same. Therefore, we classify the scoring scenarios into three categories:

- The categories determined by both scores belong to the same class.
- The category determined by one of the scores matches the category determined by the average of the two scores.
- The categories determined by both scores and the category determined by the average similarity score are all different.

We consider that the video features in the first category exhibit strong cross-modal consistency. The pseudo-labels constructed using the average similarity score in this case are referred to as high-confidence pseudo-labels, represented as follows:

$$\ell_{H,n} = onehot(\arg\max((s_{ap,n} + s_{mo,n})/2)) \quad (7)$$

In the second category, the video features have the potential to be further explored. Thus, the pseudo-labels constructed using the average similarity score are referred to as

moderate-confidence pseudo-labels, denoted similarly to ℓ_H and represented as $\ell_{\mathcal{M}}$. For video features in the third category, we consider them to potentially contain substantial noise or multiple actions, so no additional labels are constructed in this case to avoid disrupting the stability of the learning process. Based on this, the loss function for the SME module can be expressed as:

$$\mathcal{L}_{SME} = \frac{1}{m_H}\sum_{i=1}^{m_H}\varphi(p_{H,i}, \ell_{H,i}) + \frac{\lambda(t)}{m_{\mathcal{M}}}\sum_{i=1}^{m_{\mathcal{M}}}\varphi(p_{\mathcal{M},i}, \ell_{\mathcal{M},i}) \tag{8}$$

where m_H and $m_{\mathcal{M}}$ represent the number of videos in the first and second categories, respectively. $\varphi(\cdot)$ denotes the mean squared error function, which provides smoother gradient information. $\lambda(t)$ is an auxiliary coefficient decaying from 1 to 0 (t denotes the current epoch iteration), reflecting that moderate-confidence pseudo-labels may embed unforeseen uncertainties. To prevent accumulating latent noise, we progressively reduce reliance on these labels as training advances. Here, p_H and $p_{\mathcal{M}}$ denote the video-level proposal classification scores for the first and second categories, respectively, obtained by averaging the proposal-level background suppression scores weighted by their attention scores across the video. For example,

$$p_H = \frac{1}{M}\sum_{i=1}^{M} s_{base,i} \times A_{attn,i}. \tag{9}$$

Action Consistency Learning

Process Consistency Loss. Candidate proposals significantly impact the model's performance, but relying solely on Non-Maximum Suppression (NMS) may activate low-quality proposals. Our observations indicate that the main content of an action typically occurs in the central region of the action instance. Thus, for proposals with the same IoU, greater deviation from the action center leads to poorer consistency and ambiguity in the model's understanding. To address this, we propose an process consistency loss to enhance candidate proposal quality:

$$\mathcal{L}_{pc} = \frac{1}{N}\sum_{i=1}^{N}\left(\varphi(C_{P,i}, C_{G,i}) + \frac{d_i}{I_i + \varepsilon} + \delta(\mathcal{C}(i))\right), \tag{10}$$

where $C_{P,i}$ and $C_{G,i}$ represent the centers of the i-th candidate proposal and the pseudo-true action instance with the highest IoU with this proposal, respectively. The pseudo-true action instance set is constructed by selecting proposals within the same video whose attention scores satisfy $A_{attn} > 0.9 \cdot \max(A_{attn})$. d_i denotes the distance between these two centers, I_i is the intersection area of the two proposals, and ε is a small constant to prevent division by zero. The penalty function $\delta(\mathcal{C}(i))$ penalizes proposals that overlap with pseudo-true instances but fail to sufficiently capture the action center, where $\mathcal{C}(i) = (I_i < 0.5L_{G,i})$ represents the condition for insufficient overlap. Here, $L_{G,i}$ denotes the length of the pseudo-true action instance corresponding to the i-th proposal. Based on these definitions, $\mathcal{L}_{pc}$ encourages the model to align proposal centers with the action center. When the centers are sufficiently close, a larger I_i results in a smaller loss, promoting proposal completeness.

Semantic Consistency Loss. Based on our initial consideration, we believe that proposals classified into the same category within the same video should share consistent semantic representations. To achieve this, features of proposals classified into the same category by both the basic classification score s_{base} and the background suppression score s_{supp} are aggregated into a Soft-local Descriptor. The semantic consistency loss enforces features of other proposals classified as the same category under a single score to align with this Soft-local Descriptor, promoting more accurate semantic representations of actions. For a video with category c, this loss is expressed as:

$$\mathcal{L}_{st} = \frac{1}{N_c}\sum_{i=1}^{N_c}\left\| f_{prop,i} - \frac{1}{|S_c|}\sum_{j\in S_c} f_{prop,j}\right\|_2 \tag{11}$$

where N_c represents the number of proposals consistently classified as action category c by either score. $f_{prop,i}$ denotes the feature of the i-th proposal in the current video, and S_c represents the set of proposals used to construct the soft-local descriptor. Since the soft-local descriptor relies heavily on proposals with consistent classification results from both scores, we further introduce a consistency loss to ensure an adequate number of such proposals, maintaining training quality:

$$\mathcal{L}_c = \varphi(s_{base}, s_{supp}) \tag{12}$$

Thus, the overall action semantic consistency loss is:

$$\mathcal{L}_{sc} = \alpha(t)\mathcal{L}_{st} + \mathcal{L}_c \tag{13}$$

where $\alpha(t)$ is a coefficient function that increases during training (from 0.3 to 0.5). This design reflects the assumption that semantic representation of features in the early stages of the model may not be reliable, and as learning progresses, the model's action representation becomes more suited to focused learning.

Overall Loss

Reflecting on the previous content, the overall loss used for training is:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{SME} + \lambda_2\mathcal{L}_{pc} + \lambda_3\mathcal{L}_{sc} \tag{14}$$

where λ_1, λ_2, and λ_3 are hyperparameters used to control the influence of the different losses. $\mathcal{L}_{cls}$ is a commonly used classification loss (Ren et al. 2023) in the field.

Experiments

Datasets and Evaluation Metrics

Datasets. **THUMOS14** (Yun et al. 2024; Idrees et al. 2017) includes 200 test videos and 213 validation videos, covering 20 action categories. For a fair comparison, we followed the standard practice in the field, training on the training videos and testing on the validation videos; **ActivityNet1.3** (Caba Heilbron et al. 2015) contains coarser annotations but a much larger scale. It includes 200 action categories, 10,024 training videos, and 4,926 validation videos. Since the dataset's annotations do not publicly provide information about the test set, we trained on the training videos and tested on the validation videos.

Sup	Method	mAP@IoU(%)							AVG	AVG	AVG
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	(0.1:0.5)	(0.3:0.7)	(0.1:0.7)
Fully	S-CNN(Shou, Wang, and Chang 2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3	35.0	19.9	27.3
	TAL-Net(Chao et al. 2018)	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
	GCM(Zeng et al. 2021)	72.5	70.9	66.5	60.8	51.9	-	-	64.5	-	-
	RefactorNet(Xia et al. 2022)	-	-	**70.7**	**65.4**	**58.6**	**47.0**	**32.1**	-	**54.8**	-
Weak	STPN(Nguyen et al. 2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	18.5	27.0
	DGAM(Shi et al. 2020)	60.0	54.2	46.8	38.2	28.8	19.8	11.4	45.6	29.0	37.0
	UM(Lee et al. 2021)	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	32.9	41.9
	CoLA(Zhang et al. 2021)	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	UGCT(Yang et al. 2021)	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	CO2-Net(Hong et al. 2021)	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
	ASM-Loc(He et al. 2022)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	DELU(Chen et al. 2022)	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.5	37.4	46.4
	P-MIL(Ren et al. 2023)	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
	PivoTAL(Rizve et al. 2023)	74.1	69.6	61.7	52.1	42.8	30.6	16.7	60.1	40.8	49.6
	AICL(Li, Wang, and Liu 2023)	73.1	67.8	58.2	48.7	36.9	25.3	14.9	-	-	46.4
	ISSF(Yun et al. 2024)	72.4	66.9	58.4	49.7	41.8	25.5	12.8	57.8	37.6	46.8
	SEAL(Ours)	**78.4**	**73.6**	**66.1**	**55.8**	**44.5**	**31.2**	**19.9**	**63.7**	**43.5**	**52.8**

Table 1: Comparison on the THUMOS14 dataset.

Method	mAP@IoU(%)			
	0.5	0.75	0.95	AVG
STPN(Nguyen et al. 2018)	29.3	16.9	2.6	16.3
TSCN(Zhai et al. 2020)	35.3	21.4	5.3	21.7
TS-PCA(Liu et al. 2021)	37.4	23.5	5.9	23.7
UGCT(Yang et al. 2021)	39.1	22.4	5.8	23.8
AUMN(Luo et al. 2021)	38.3	23.5	5.2	23.5
CO2-Net(Hong et al. 2021)	35.4	20.9	5.6	22.9
FAC-Net(Huang, Wang, and Li 2021)	37.6	24.2	6.0	24.0
RSKP(Huang, Wang, and Li 2022)	40.6	24.6	5.9	25.0
ASM-Loc(He et al. 2022)	41.0	24.9	**6.2**	25.1
P-MIL(Ren et al. 2023)	**41.8**	25.4	5.2	25.5
STCL-Net(Fu, Gao, and Xu 2023)	40.6	24.0	6.0	24.7
SEAL(Ours)	41.6	**25.8**	5.2	**25.6**

Table 2: Comparison on the ActivityNet1.3 dataset. AVG means average mAP from IoU 0.5 to 0.95 with a 0.05 increment.

Evaluation Metrics. Following standard evaluation protocols, we report performance results using mean Average Precision (mAP) under different temporal Intersection over Union (t-IOU) thresholds. For THUMOS14, the t-IOU thresholds are set at $[0.1 : 0.1 : 0.7]$; for ActivityNet1.3, the thresholds are set at $[0.5 : 0.05 : 0.95]$.

Supplementary Details. The input features are extracted using a pre-trained I3D model for RGB and the TV-L1 algorithm for optical flow. All three learnable branches consist of a 1D convolutional layer with a kernel size of 1, followed by a ReLU activation layer. Additionally, the heads of the three branches are simple 1D convolutional layers, with outputs corresponding to $s_{base} \in \mathbb{R}^{M\times(C+1)}$ (class-aware branch), $A_{attn} \in \mathbb{R}^{M}$ (class-agnostic branch), and $s_{pc} \in \mathbb{R}^{M}$ (process consistency branch). For model learning, we used the Adam optimizer with a learning rate of 0.00005 and a regularization rate of 0.001. In terms of proposal generation, we utilized raw proposals generated by the classical point-level weakly supervised temporal action localization method LAC (Lee and Byun 2021) without any post-processing. Our motivation for this approach stems from the observation that P-WTAL tasks, due to more refined labels, often produce more promising proposals. However, as previous research has shown (Xia et al. 2024), an average of 65.2% of proposals generated by LAC can be replaced by better alternatives, indicating that the model's learning capacity is less constrained. Therefore, this approach allows for a more accurate assessment of our framework's learning ability and generalization performance. Additionally, the batch size is set to 10, the training runs for 200 epochs, λ_1 decays linearly from 1 to 0 over 30 epochs, λ_2 is set to 1.5, and λ_3 decays non-linearly from 1 to 0 over 50 epochs.

Comparison with State-of-the-Art Methods

Results on THUMOS14. Table 1 presents a numerical comparison of our SEAL with fully-supervised TAL methods and state-of-the-art WTAL methods on the THUMOS14 dataset. Compared to the strong performance of FTAL methods, our SEAL often achieves results that are comparable to or even better than these methods. When compared with the state-of-the-art WTAL methods, our method demonstrates a significant performance lead. Specifically, our approach improves the average mAP@0.1:0.7 by 3.2% over the best existing method. This improvement is attributed to our SME module's focus on cross-modal consistency and action consistency learning, which further enhances the semantics and completeness of action segments.

Losses of Different Modules			mAP@IoU (%)			
$\mathcal{L}_{SME}$	$\mathcal{L}_{pc}$	$\mathcal{L}_{sc}$	0.3	0.5	0.7	AVG
			54.0	36.8	14.6	43.1
✓			64.2	43.0	17.2	50.9
	✓		64.5	43.1	18.4	51.2
		✓	59.6	40.6	16.0	47.2
✓	✓		65.3	44.4	18.9	52.1
	✓	✓	64.9	44.2	18.7	51.7
✓	✓	✓	66.0	44.5	19.9	52.8

Table 3: Ablation study of three losses in SEAL.

Aggregation Method	mAP@IoU (%)			
	0.3	0.5	0.7	AVG
Segment level	64.3	42.2	18.0	50.8
Video level	64.2	43.0	17.2	50.9

Table 4: Comparison of the similar modality enhancement module under different aggregation methods.

Results on ActivityNet1.3. Table 2 compares our SEAL with state-of-the-art WTAL methods on ActivityNet1.3. As shown in the table, SEAL demonstrates strong competitiveness. On the average mAP@0.5:0.95, our results outperform all the comparison methods.

Ablation Study

To evaluate the effectiveness and underlying considerations of our design, we conduct ablation studies on the THUMOS14 dataset.

Impact of Each Module. We isolate each component's corresponding loss to analyze its individual contribution. As shown in Table 3, a detailed ablation study evaluates these losses' effectiveness. The results show that every loss enhances overall performance, and no conflicts emerge when combined, confirming the soundness of our design.

Impact of Aggregation Methods for Hard-global Descriptors. We consider two approaches for aggregating hard-global descriptors in the SME module: segment-level and video-level aggregation. The segment-level approach generalizes action semantics in a flattened manner, while the video-level approach integrates contextual information from a broader perspective. Since the chosen aggregation method can profoundly affect SME performance, we compare the two under conditions using only the SME module. As shown in Table 4, the video-level aggregation method yields better results, likely due to its incorporation of contextual cues.

Impact of High-confidence and Moderate-confidence Labels. In the SME module, pseudo-labels are constructed using cosine similarity scores derived from two modality features. High/moderate-confidence pseudo-labels are further generated based on the degree of consensus between the scores on action categories. To investigate whether both types of labels contribute to the model's learning, we conducted ablation experiments on the SME module, with the numerical results reported in Table 5. Experimental results indicate that the combined use of both types of pseudo-labels effectively steers the model toward superior performance, thereby verifying the soundness of our proposed design.

Method	mAP@IoU (%)			
	0.3	0.5	0.7	AVG
Base	54.0	36.8	14.6	43.1
Base+ℓ_H	61.3	41.0	17.1	48.7
Base+ℓ_H+$\ell_{\mathcal{M}}$	64.2	43.0	17.2	50.9

Table 5: Ablation study of high confidence and moderate confidence pseudo-labels in SME.

SEAL with different proposals	mAP@IoU (%)			
	0.3	0.5	0.7	AVG
Base + P_{pw}	54.0	36.8	14.6	43.1
Ours + P_{pw}	66.0	44.5	19.9	52.8
Base + P_w	49.5	33.8	11.4	39.7
Ours + P_w	59.6	39.8	16.2	47.1

Table 6: Comparison under different proposals.

Impact of Candidate Proposals. Our candidate proposals P_{pw} are based on the classic P-WTAL method. While this method is both challenging and promising, it remains unclear whether SEAL can be effectively applied to candidate proposals generated by other methods. To test the generalization capability of SEAL, we applied it to candidate proposals P_w generated by the classic method CO2-Net (Hong et al. 2021), with the results shown in Table 6. As can be seen, the change in candidate proposal sets does not affect the superiority of our framework. Specifically, the newly proposed modules achieve a 7.4% improvement in the average mAP@0.1:0.7, indicating that our method demonstrates reliable generalization performance.

Conclusion

This work proposes SEAL, a novel WTAL method that includes two key new modules. The SME module leverages the aggregation of appearance and motion descriptors derived from original features to construct high/moderate-confidence pseudo-labels, guiding the model to learn more robust action representations and mitigating task discrepancies. The ACL module focuses on the spatiotemporal process and feature semantics of actions, encouraging the model to maintain semantic consistency among similar actions while ensuring the completeness of candidate proposals, thus achieving more accurate localization. Extensive experiments on two datasets demonstrate the effectiveness of our SEAL.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant (No. 62032016).

References

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Chao, Y.-W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1130–1139.

Chen, M.; Gao, J.; Yang, S.; and Xu, C. 2022. Dual-evidential learning for weakly-supervised temporal action localization. In *European conference on computer vision*, 192–208. Springer.

Fu, J.; Gao, J.; and Xu, C. 2023. Semantic and temporal contextual correlation learning for weakly-supervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12427–12443.

Gao, J.; Yang, Z.; Chen, K.; Sun, C.; and Nevatia, R. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE international conference on computer vision*, 3628–3636.

He, B.; Yang, X.; Kang, L.; Cheng, Z.; Zhou, X.; and Shrivastava, A. 2022. 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*, 13925–13935.

Hong, F.-T.; Feng, J.-C.; Xu, D.; Shan, Y.; and Zheng, W.-S. 2021. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM international conference on multimedia*, 1591–1599.

Huang, L.; Wang, L.; and Li, H. 2021. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8002–8011.

Huang, L.; Wang, L.; and Li, H. 2022. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3272–3281.

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

Islam, A.; Long, C.; and Radke, R. 2021. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1637–1645.

Kang, H.; Kim, H.; An, J.; Cho, M.; and Kim, S. J. 2023. Soft-landing strategy for alleviating the task discrepancy problem in temporal action localization tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6514–6523.

Lee, P.; and Byun, H. 2021. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13648–13657.

Lee, P.; Uh, Y.; and Byun, H. 2020. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11320–11327.

Lee, P.; Wang, J.; Lu, Y.; and Byun, H. 2021. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1854–1862.

Li, G.; Cheng, D.; Ding, X.; Wang, N.; Wang, X.; and Gao, X. 2023. Boosting weakly-supervised temporal action localization with text information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10648–10657.

Li, X.; and Zhao, Z. 2021. Pixel level semantic understanding: From classification to regression. *SCIENTIA SINICA Informationis*, 51: 521–564.

Li, Z.; Wang, Z.; and Liu, Q. 2023. Actionness inconsistency-guided contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1513–1521.

Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3320–3329.

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3889–3898.

Lin, T.; Zhao, X.; and Shou, Z. 2017. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 988–996.

Liu, Y.; Chen, J.; Chen, Z.; Deng, B.; Huang, J.; and Zhang, H. 2021. The blessings of unlabeled background in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6176–6185.

Luo, W.; Zhang, T.; Yang, W.; Liu, J.; Mei, T.; Wu, F.; and Zhang, Y. 2021. Action unit memory network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9969–9979.

Luo, Z.; Guillory, D.; Shi, B.; Ke, W.; Wan, F.; Darrell, T.; and Xu, H. 2020a. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 729–745. Springer.

Luo, Z.; Guillory, D.; Shi, B.; Ke, W.; Wan, F.; Darrell, T.; and Xu, H. 2020b. Weakly-supervised action localization with expectation-maximization multi-instance learning. In

Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16, 729–745. Springer.

Ma, J.; Gorti, S. K.; Volkovs, M.; and Yu, G. 2021. Weakly supervised action selection learning in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7587–7596.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6752–6761.

Ren, H.; Yang, W.; Zhang, T.; and Zhang, Y. 2023. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2394–2404.

Rizve, M. N.; Mittal, G.; Yu, Y.; Hall, M.; Sajeev, S.; Shah, M.; and Chen, M. 2023. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22992–23002.

Shi, B.; Dai, Q.; Mu, Y.; and Wang, J. 2020. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1009–1019.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1049–1058.

Tejero-de Pablos, A.; Nakashima, Y.; Sato, T.; Yokoya, N.; Linna, M.; and Rahtu, E. 2018. Summarization of user-generated sports video by using deep action recognition features. *IEEE Transactions on Multimedia*, 20(8): 2000–2011.

Vignolo, A.; Noceti, N.; Rea, F.; Sciutti, A.; Odone, F.; and Sandini, G. 2017. Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, 4: 14.

Wang, B.; Zhao, Y.; Yang, L.; Long, T.; and Li, X. 2023. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, L.; Xiong, Y.; Lin, D.; and Van Gool, L. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 4325–4334.

Wang, Y.; Li, Y.; and Wang, H. 2023. Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18878–18887.

Xia, K.; Wang, L.; Zhou, S.; Zheng, N.; and Tang, W. 2022. Learning to refactor action and co-occurrence features for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13884–13893.

Xia, Z.; Cheng, J.; Liu, S.; Hu, Y.; Wang, S.; Zhang, Y.; and Dang, L. 2024. Realigning Confidence with Temporal Saliency Information for Point-Level Weakly-Supervised Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18440–18450.

Xu, M.; Perez Rua, J. M.; Zhu, X.; Ghanem, B.; and Martinez, B. 2021. Low-fidelity video encoder optimization for temporal action localization. *Advances in Neural Information Processing Systems*, 34: 9923–9935.

Xu, Y.; Zhang, C.; Cheng, Z.; Xie, J.; Niu, Y.; Pu, S.; and Wu, F. 2019. Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9070–9078.

Yang, H.; Wu, W.; Wang, L.; Jin, S.; Xia, B.; Yao, H.; and Huang, H. 2022. Temporal action proposal generation with background constraint. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3054–3062.

Yang, L.; Peng, H.; Zhang, D.; Fu, J.; and Han, J. 2020. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29: 8535–8548.

Yang, W.; Zhang, T.; Yu, X.; Qi, T.; Zhang, Y.; and Wu, F. 2021. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 53–63.

Yun, K.; Kwon, Y.; Oh, S.; Moon, J.; and Park, J. 2019. Vision-based garbage dumping action detection for real-world surveillance platform. *Etri Journal*, 41(4): 494–505.

Yun, W.; Qi, M.; Wang, C.; and Ma, H. 2024. Weakly-Supervised Temporal Action Localization by Inferring Salient Snippet-Feature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6908–6916.

Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; and Gan, C. 2021. Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6209–6223.

Zhai, Y.; Wang, L.; Tang, W.; Zhang, Q.; Yuan, J.; and Hua, G. 2020. 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*, 37–54. Springer.

Zhang, C.; Cao, M.; Yang, D.; Chen, J.; and Zou, Y. 2021. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16010–16019.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhao, Y.; Zhang, H.; Gao, Z.; Guan, W.; Nie, J.; Liu, A.; Wang, M.; and Chen, S. 2022. A temporal-aware relation and attention network for temporal action localization. *IEEE Transactions on Image Processing*, 31: 4746–4760.