Realigning Confidence with Temporal Saliency Information for Point-Level Weakly-Supervised Temporal Action Localization

Ziying Xia[1], Jian Cheng[1]✉, Siyu Liu[1], Yongxiang Hu[1], Shiguang Wang[1], Yijie Zhang[1], Liwan Dang[1,2]
[1]University of Electronic Science and Technology of China
[2]The Second Research Institute of Civil Aviation Administration of China
chengjian@uestc.edu.cn {zyxia, syliu, huyongx}@std.uestc.edu.cn
{xiaohu_wyyx, zhangyijieuestc, caacsri_dangwanli}@163.com

Abstract

Point-level weakly-supervised temporal action localization (P-TAL) aims to localize action instances in untrimmed videos through the use of single-point annotations in each instance. Existing methods predict the class activation sequences without any boundary information, and the unreliable sequences result in a significant misalignment between the quality of proposals and their corresponding confidence. In this paper, we surprisingly observe the most salient frame tend to appear in the central region of the each instance and is easily annotated by humans. Guided by the temporal saliency information, we present a novel proposal-level plug-in framework to relearn the aligned confidence of proposals generated by the base locators. The proposed approach consists of Center Score Learning (CSL) and Alignment-based Boundary Adaptation (ABA). In CSL, we design a novel center label generated by the point annotations for predicting aligned center scores. During inference, we first fuse the center scores with the predicted action probabilities to obtain the aligned confidence. ABA utilizes the both aligned confidence and IoU information to enhance localization completeness. Extensive experiments demonstrate the generalization and effectiveness of the proposed framework, showcasing state-of-the-art or competitive performances across three benchmarks. Our code is available at https://github.com/zyxia1009/CVPR2024-TSPNet.

1. Introduction

The recent spotlight on the utilization of single-point timestamp annotations in each action instance for precise temporal action localization (P-TAL) [33] within untrimmed videos has captured the attention of the research community. This approach holds considerable significance in real-world applications, including video retrieval [17, 44], security monitoring [36], and video grounding [25, 31, 50]. In comparison to fully-supervised temporal action localization (TAL) [23, 28–30, 41, 47] and weakly-supervised temporal action localization (W-TAL) [12, 14, 15, 22, 39, 40, 42, 55, 56, 58, 59], which rely solely on video-level annotations, P-TAL strikes a delicate balance between labor costs and localization precision. Consequently, the utilization of the additional point annotations is crucial for achieving accurate localization.

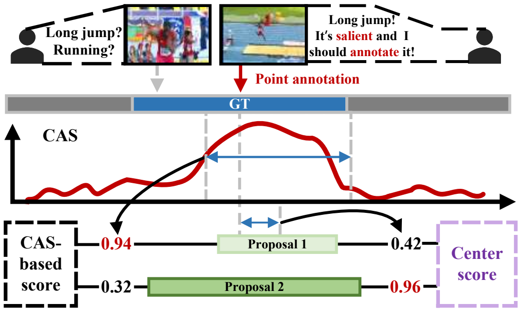

Figure 1. Existing methods calculate the misaligned confidence based on CAS information. Driven by human tendency during annotating, we propose to predict the aligned center score based on the offset between the proposal center and the annotation point.

Many existing P-TAL [11, 21, 26, 33, 52] methods follow a common framework known as multi-instance learning (MIL). This involves mapping the feature sequence extracted from the untrimmed video at the snippet level (spanning several frames) to a class activation sequence (CAS) for each action category. The CAS indicates the probability of the action occurrence. Subsequently, action proposals are derived from the CAS using a manually set multi-threshold strategy. These insightful techniques contribute to the generation of abundant high-quality proposals, facilitating effective localization for each action instance.

However, as depicted in Figure 1, the **alignment** between confidence and the most complete proposals is often deficient (*e.g.*, an average of 65.2% of the final proposals in LAC[21] have better alternatives). We argue that this mis-

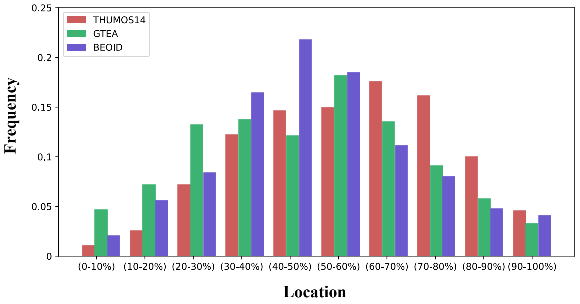

Figure 2. The frequency statistics show the distribution of annotation points appearing in different locations across three benchmarks. It is observed that the annotators consistently label the points in the central regions of the action instances.

alignment arises from the fact that the confidence score of each proposal is directly generated or computed by the unreliable CAS, such as the outer-inner-contrast score [43]. Moreover, the sensitive scores are less conducive to obtaining accurate boundaries. In essence, there are two critical challenges in enhancing localization performance: (1) How to align the quality of the proposal with the appropriate confidence correctly. (2) How to adjust the boundaries to improve the proposal completeness with aligned confidence.

To address these challenges, we observe that the most salient frame tends to appear in the central region of the each instance and is easily annotated by humans (Figure 2), we call the prior as "temporal saliency information". Such information reveals the approximate locations of instance centers and provides an evidence of alignment. we introduce a plug-and-play **T**emporal **S**aliency-Driven **P**roposal Learning framework (TSP-Net) for P-TAL. In general, we establish a direct proposal-level supervisory relationship between proposals and point annotations to realign confidence and enhance boundary precision.

Specifically, we first propose the Center Score Learning (CSL) to reduce the impact of confidence misalignment. CSL first generates the novel center labels based on the annotated points. The center labels guide the model to predict the aligned center scores. After that, the original points inevitably contain noise from personal preferences and deviate from the true instance center. To alleviate the annotation noise, we propose an updating strategy during the learning process. We mine pseudo-positive proposals and update the points continuously. After that the center labels are regenerated and utilized to facilitate the learning of more aligned center scores.

However, the proposal boundaries still rely on the CAS, and proposals with different durations may exhibit similar center scores. To tackle this, we learn actionness and action scores and integrate them with center scores to form aligned confidence. Subsequently, we introduce Alignment-based Boundary Adaptation (ABA) to incorporate temporal-related boundary and IoU information, thereby enhancing the completeness of confident proposals during inference.

Our contributions can be summarized as follows:

(1) We meticulously analyze the limitations of current P-TAL methods, specifically focusing on the misalignment of proposals and confidence. We observe that humans tend to annotate at the instance center region, and this prior is beneficial to improve the misalignment problem.

(2) For P-TAL, we introduce a plug-in framework aimed at aligning the proposal quality with confidence. We propose center score learning to dynamically predict the alignment from the saliency information. Additionally, alignment-based boundary adaptation is introduced to improve the completeness of confident proposals.

(3) We assess the framework on three benchmarks and compare it to existing P-TAL methods, we achieve improvements of at least 3.6%, 4.1%, and 6.3% in average mAP. Furthermore, we integrate TSP-Net into various P-TAL methods to demonstrate its generalization capabilities, and the results indicate an average improvement of 5.8%.

2. Related Work

Temporal action localization (TAL) requires accurate time coordinate annotations to locate action instances in untrimmed videos. [4, 23, 27–30, 35, 41, 47, 51]. Similar to the task of image object detection [13, 38], existing TAL methods can be categorized into one-stage and two-stage approaches based on the localization pipeline. The two-stage TAL methods [4, 27–30, 47] first densely generate potential instance proposals. After that, the independently trained proposal-based detectors will classify and refine the proposals. On the other hand, one-stage [41, 51] methods adopt an end-to-end pipeline to localize and recognize the actions. Although the impressive localization achieved by TAL technologies is captivating, the labor-intensive nature of the annotation process remains unaffordable due to the rapidly increasing number of videos.

Weakly-supervised temporal action localization (W-TAL) aims to localize action instances using only video-level annotations [12, 14, 15, 22, 32, 39, 40, 42, 43, 55, 56, 58, 59], offering an effective and cost-efficient approach that has garnered extensive discussion within the research community. Leveraging advancements in multi-instance learning theory, the CAS learned by MIL can explicitly localize instances with snippet-level information and subsequent manual post-processing. Approaches such as AUMN [32] construct the memory bank for action units to produce the discriminative CAS. To introduce more global information, some W-TAL methods [14, 39, 59] make the temporal reason at the proposal level. Such as P-MIL [39] extends MIL in the proposal level to keep the score consistent between training and testing. Furthermore, existing proposal

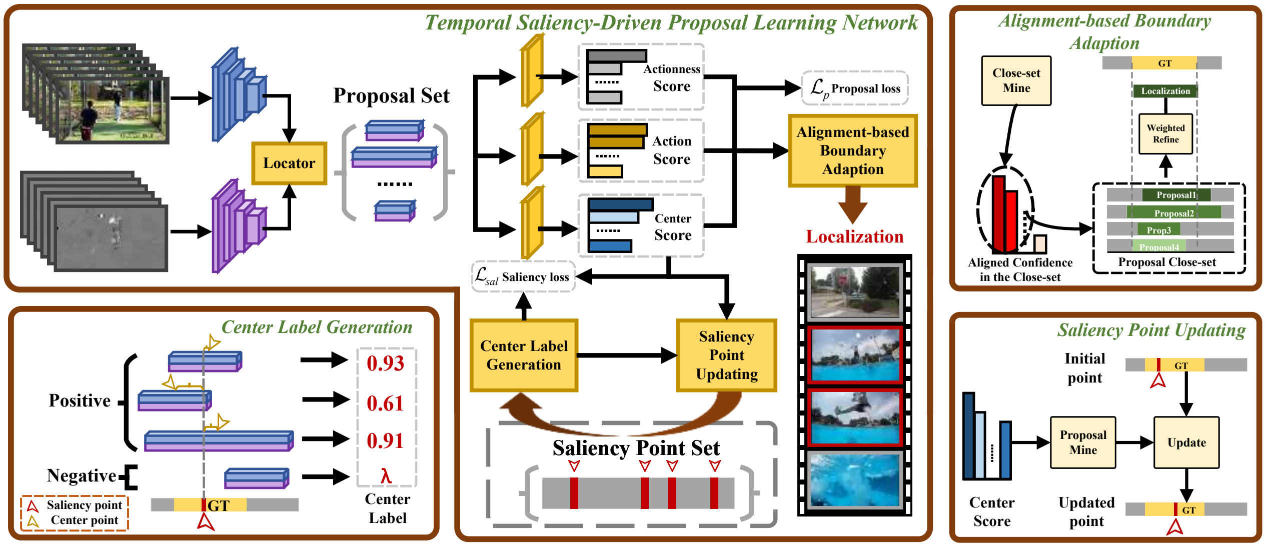

Figure 3. The proposed TSP-Net initially extracts proposal features based on the proposals generated by the CAS-based locator. These features are then mapped to actionness scores and action scores. A saliency branch consists of center label generation and saliency point updating is designed for learning aligned center scores. The mentioned scores are combined as aligned confidence and the proposed alignment-based boundary adaptation enhances the proposal completeness.

refining strategies [59] adjust boundaries and confidence with different CAS information, lacking a comprehensive analysis of completeness with proposals.

Point-level weakly-supervised temporal action localization. Utilizing only point annotations to realize action localization [6, 34, 49] is attractive because of the trade-off between the labor cost (at least 6 × more economical [21]) and performance. The majority of P-TAL methods [11, 19, 21, 24, 26, 33, 52] only utilize the point annotations for mining reliable pseudo points of the instance. For example, LAC [21] marks pseudo background points and searches the optimal sequence for completeness learning with the point annotations together. CRRC-Net [11] builds and updates the class-wise prototypes for more reliable classification. At the proposal-level, DCM [19] generates proposals with the points to estimate the locations individually. Yin *et al.* [53] introduce text information for evaluating action proposals. In contrast to prior works, we propose a proposal learning framework driven by temporal saliency information to ensure the alignment between the quality of proposals and the confidence scores.

3. Methodology

In this section, we first illustrate the formulation of the P-TAL (Sec. 3.1) and introduce the framework of the TSP-Net in Sec. 3.2. After that, we propose center score learning (Sec. 3.3) which consists of center label generation and saliency point updating to predict the aligned center scores between proposals and instances. After getting the aligned confidence, we propose alignment-based boundary adaption (Sec. 3.4) to correct the boundaries of high-quality proposals. At last, we state the training and inference process in Sec. 3.5.

3.1. Problem Formulation

For the untrimmed video v_i in training set $V = \{v_1, ...v_n\}$, we only annotate single timestamp for each action instance and get the point set $Y_i = \{(t_j, y_j)\}_{j=1}^{N_i}$, where N_i is the number of instances, and t_j, y_j denote the annotated timestamp and the category of the action, respectively. The target of the P-TAL [33] is to utilize only the Y as the supervise information to localize the instance $(\hat{s}_i, \hat{e}_i, \hat{y}_i, c_i)$ in the untrimmed video, where $\hat{s}_i, \hat{e}_i$ and c_i represent the start point, end point, and confidence score, respectively.

3.2. Overall Framework

The overall framework of the TSP-Net is shown in Figure 3. The snippet-level feature sequences extracted by the backbone are utilized to generate the coarse proposals. For each proposal, we extract proposal-level features to learn three different semantic scores. At last, we fuse the scores into the confidence score and refine the boundary to get the final localization.

Snippet-level feature extraction. We first partition each video into T snippets, each comprising 16 frames. After that, the pre-trained feature extractor [1, 3, 10] is applied to extract the snippet-level feature sequence. Following the methodology of earlier P-TAL approaches [11, 21, 26], we use I3D [3] pre-trained on Kinetics-400 [20] to extract features in both RGB and flow modalities. The features extracted from these modalities are concatenated to form the input feature sequence $\mathbf{X} \in \mathbb{R}^{T\times D}$, where D denotes the dimension of each snippet feature.

Proposal-level feature extraction. As referred, the existing P-TAL approaches [11, 21] effectively generate abundant proposals due to the locally responsive CAS and the multi-threshold generation strategy. In this work, we use $\mathbf{X}$ to reproduce the P-TAL method (we employ LAC [21] as the **baseline**) to generate the proposal set $P = \{(s_i, e_i, y_i, \mathbf{X}_i)\}_{i=1}^{M}$ without the misaligned confidence and post-processing operations, where $\mathbf{X}_i$ denotes the feature subsequence corresponding to the i-th proposal, $y_i \in \mathbb{R}^{C+1}$ is the one-hot category label based on the point set, C denotes the number of action categories, and the C+1-th category represents the background. Concretely, if none of the points in Y fall within the interval $[s_i, e_i]$, we label it as the background, and vice versa with the corresponding action category. To increase the discriminability of $\mathbf{X}_i$, we apply the surrounding contrastive feature extraction (SCFE) [39] to acquire the proposal-level features $\mathbf{X}_p \in \mathbb{R}^{M\times D_p}$.

$$\mathbf{X}_{p,i} = \text{SCFE}(\mathbf{X}_i). \tag{1}$$

Proposal learning. Effective proposal learning (PL) should synthesize different semantic information into confidence. Following [11, 21, 39], we design three independent branches and input $\mathbf{X}_p$ for learning the actionness scores, action scores, and center scores, respectively. Each branch consists of a $1D$ convolution layer with a kernel size of 3 and a ReLU layer. For the head of each branch, we use the $1D$ convolution layer with a kernel size of 1 to get $s_a \in \mathbb{R}^{M\times 1}$, $s_c \in \mathbb{R}^{M\times (C+1)}$, and $s_{cen} \in \mathbb{R}^{M\times 1}$. After that, we apply the sigmoid function to s_a and s_{cen}. Especially, s_a denotes whether the proposal includes an action instance. s_c indicates which action is occurring. The proposed center scores s_{cen} in Sec. 3.3 are learned based on human salient perception and aligned with the quality of the proposal. The scores are fused into aligned confidence $c_i \in \mathbb{R}^{M\times 1}$ for the i-th class. At last, based on c_i and IoU information, we selectively refine the proposal boundary s and e to realize more precise localization by the proposed alignment-based boundary adaptation (Sec. 3.4).

3.3. Center Score Learning

In this section, we introduce a novel center score learning approach to evaluate proposal quality and predict quality-aligned scores for proposals. The proposed learning paradigm comprises center label generation and saliency point updating. Center label generation establishes explicit supervision to align confidence, while saliency point updating continuously and dynamically mines more valid information for obtaining more accurate center labels.

Center label generation. There is a natural tendency for humans to concentrate on the most salient timestamp within an action instance rather than the ambiguous boundaries. This distinction is also a key factor contributing to the significant difference in labor costs between TAL and P-TAL. The inherent inclination indicates that point annotations are consistently situated in the central region of the action instance. Motivated by this observation, we introduce the concept of center label to tackle the aforementioned issue. Concretely, we first introduce saliency point (can be seen as a pseudo-center) for each instance to represent the most salient timestamp, and the point can serve as a guide to the center of the instance. In practice, we initialize the saliency point set P_{sal} with the point annotations Y. Subsequently, considering that the center of the most complete proposal, along with its corresponding ground truth instance, should be proximate to or aligned with the saliency point, we manually generate the hard center label for the i-th proposal as:

$$y_{cen,i} = \begin{cases} 1 - 2\left|\frac{t-s_i}{e_i-s_i} - 0.5\right| & ,\exists t \in P_{sal}, t \in [s_i, e_i] \\ 0 & ,\forall t \in P_{sal}, t \notin [s_i, e_i] \end{cases}. \tag{2}$$

The normalized center shift of bilateral symmetry between the proposal center and the saliency point indicates the center consistency. We use the center label as the learning objective and optimize the model to predict a aligned center score. Therefore, in the absence of boundary information, this score potentially reflects the value of the proposal and reduces the sensitivity of confidence to the CAS. After label generation, we incorporate the saliency loss $\mathcal{L}_{sal}$ to learn the center score as follows:

$$\mathcal{L}_{sal} = \frac{1}{M}\sum_{i=1}^{M}(y_{cen,i} - s_{cen,i})^2. \tag{3}$$

Although the hard center labels y_{cen} can explicitly supervise an aligned score, negative proposals lacking any saliency point may still encompass a portion of the instance. Forcing the model to predict zero scores can easily lead to performance deterioration. Thus, we propose the soft center label as:

$$y^{s}_{cen,i} = \begin{cases} 1 - 2\left|\frac{t-s_i}{e_i-s_i} - 0.5\right| & ,\exists t \in P_{sal}, t \in [s_i, e_i] \\ \lambda & ,\forall t \in P_{sal}, t \notin [s_i, e_i] \end{cases}, \tag{4}$$

where the λ is the soft-value set manually and related to distribution of proposals. The introduction of soft-term supervision serves to reduce model uncertainty arising from incomplete supervisory information.

Saliency point updating. Although the aligned center score can be regarded as critical evidence to quantify the proposal quality, the following limitations remain: (1) The initial saliency point from the annotator inevitably includes a great deal of personal preference. (2) The most salient point may be off the center region of the instance. We regard the limitations as the saliency noise. To alleviate the

Algorithm 1: Update saliency point set P_{sal}

```
Input: annotated points $Y = \{t_i\}_{i=1}^{N}$, proposal
       coordinates $C = \{(s_j, e_j)\}_{j=1}^{M}$, center
       scores $s_{cen}$, update threshold $\theta_{up}$
Output: updated saliency point set $P_{sal}$
for $i = 1$ to $N$ do
    $S_{up} \leftarrow \{(s_m, e_m), s_{cen,m}\}_{m=1}^{N_{up}}$ where
      $(s_{cen,m} > \theta_{up} \& t_{sal,i} \in [s_m, e_m])$
    $\Delta t_{sal,i} \leftarrow \sum_{S_{up}} (t_i - \frac{s_m+e_m}{2}) \frac{s_{cen,m}}{\sum_{m=1}^{N_{up}} s_{cen,m}}$
      $P_{sal,i} \leftarrow t_i + \Delta t_{sal,i}$
end
return $P_{sal}$
```

noise, we propose the saliency point updating strategy to mine the reliable center information of proposals and update the initial saliency points adaptively and continually. For brevity, we describe the strategy as Algorithm 1. For each saliency point, we first mine the proposal set S_{up} that contains the point and has s_{cen} greater than the update threshold θ_{up}. Secondly, the centers in the mined proposal coordinates are referred to as pseudo points, containing valuable center information. Following this, we update the points by the weighted sum of the difference between the saliency point and the pseudo points. Finally, we repeat the operation to update the entire saliency point set. To maintain effectiveness and stability, we update the saliency points every p_{up} iterations based on the initial annotated points Y. After the updating, we regenerate the center labels to learn better s_{cen}.

3.4. Alignment-based Boundary Adaption

After continuous updating and learning the center score for each proposal, we apply the softmax function to s_c, yielding p_c. Subsequently, we fuse the scores generated by TSP-Net to obtain the aligned confidence $c_i \in \mathbb{R}^{M \times 1}$ of the proposals for i-th category as:

$$c_i = s_a * p_{c,i} * s_{cen}. \tag{5}$$

We apply c to replace the original confidence scores generated by the base locator. While the relearned confidence scores exhibit high quality during inference, the subsequent post-processing operation can result in information loss [59]. Additionally, the complete reliance of the proposal scope on the original CAS may lead to inaccurate localization. Notably, proposals with distinct scopes but close centers may learn similar s_{cen}, resulting in comparable confidence scores.

To address those problems, based on the aligned confidence, we propose alignment-based boundary adaption to refine the most confident proposal by its close proposals, adaptively. Inspired by the thought of NMS and [59], for each category, we first decrease c_i to find the proposal $\hat{p} = \{s, e\}$ with the maximum confidence. Secondly, we calculate the IoUs between $\hat{p}$ and the proposal set P, and the proposals with IoU larger than the refine threshold θ_{re} are regarded as the close-set $\hat{P}_r = \{s_i, e_i\}_{i=1}^{M_r}$ for $\hat{p}$. Thirdly, we compute the updating weight $\mathbf{w} \in \mathbb{R}^{(M_r+1)\times 1}$ for boundary adaption. An intuitive approach is to calculate directly based on confidence, but this approach does not take into account the proposal correlation. Specifically, aligned confidence determines the importance of boundary information for $\hat{p}$, while the IoU reflects the information relevance. Thus, we utilize both aligned confidence and IoU information to calculate $\mathbf{w}$ as follows:

$$w_i = c_i * \overline{IoU}_i / (\sum_{j=0}^{M_r} c_j * \overline{IoU}_j), \tag{6}$$

where $\overline{IoU}$ denotes the normalized IoU in $\{\hat{p}, \hat{P}_r\}$. After that, we adapt the boundaries of $\hat{p}$ as:

$$\hat{b} = w_0 b + \sum_{i=1}^{M_r} w_i b_i, b \in \{s, e\}. \tag{7}$$

At last, we remove $\hat{p}$ and $\hat{P}_r$ from the proposal set and repeat the above refine operation until the set is empty. Notice that all proposals in close-sets are retained instead of discarded. Thus, the adaptation can further eliminate the degradation of proposal completeness by the original CAS. For the aligned proposals with various durations, a voting-like adaption approach can mitigate the impact of proposal generation at the extreme threshold on the final localization.

3.5. Training and Inference

Training. As mentioned in Sec. 3.2, TSP-Net needs to synthesize different semantic information. The objective function of TSP-Net is described as:

$$\mathcal{L} = \mathcal{L}_p + \alpha \mathcal{L}_{sal}, \tag{8}$$

where α represents the multi-task balance factor. $\mathcal{L}_p$ utilizes a cross-entropy loss to learn the class-agnostic probability and class-wise probability simultaneously in a proposal-level and is denoted as:

$$\mathcal{L}_p = \frac{1}{M} \sum_{i=1}^{M} CE(y_i, softmax(\bar{s}_{c,i})), \tag{9}$$

where $\bar{s}_c = s_a * s_c$. $\mathcal{L}_{sal}$ denotes the saliency loss in Eq. 3.

Inference. For generated proposals P in Sec. 3.2, we input it into TSP-Net and get the final aligned confidence as Eq. 5. After that, alignment-based boundary adaption (ABA) is utilized to refine the proposals as:

$$\hat{P} = \text{ABA}(P). \tag{10}$$

At last, the Soft-NMS [2] is applied to $\hat{P}$ to get the final localization.

Table 1. Comparison results with the state-of-the-art methods on THUMOS'14. The proposed method outperforms previous W-TAL and P-TAL methods by a large margin in terms of mAP and has a competitive performance with the fully supervised methods.

Supervision	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 (TAL)	G-TAD [48](CVPR'20)	-	-	54.5	47.6	40.2	30.8	23.4	-	39.3	-
	TCANet [37](CVPR'21)	-	-	60.6	53.2	44.6	36.8	26.7	-	44.4	-
	RTD-Net [45](ICCV'21)	-	-	68.3	62.3	51.9	38.8	23.7	-	49.0	-
	GCM [54](TPAMI'21)	72.5	70.9	66.5	60.8	51.9	-	-	64.5	-	-
	VSGN [57](ICCV'21)	-	-	66.7	60.4	52.4	41.0	30.4	-	50.2	-
Video-level (W-TAL)	DELU [5](ECCV'22)	71.5	66.2	56.5	47.7	40.5	27.4	15.3	56.5	37.4	46.4
	RSKP [16](CVPR'22)	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
	ASM-Loc [14](CVPR'22)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
	PivoTAL [40](CVPR'23)	74.1	69.6	61.7	52.1	42.8	30.6	16.7	60.1	40.8	49.6
	Zhou et al. [59](CVPR'23)	74.0	69.4	60.7	51.8	42.7	26.2	13.1	59.7	38.9	48.3
	P-MIL [39](CVPR'23)	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
Point-level (P-TAL)	SF-Net [33](ECCV'20)	68.3	62.3	52.8	42.2	30.5	20.6	12.0	51.2	31.6	41.2
	LAC [21](ICCV'21)	75.7	71.4	64.6	56.5	45.3	34.5	21.8	62.7	44.5	52.8
	DCM [19](ICCV'21)	70.2	63.5	55.6	44.7	32.3	22.0	12.3	53.3	33.4	42.9
	BackTAL [52](TPAMI'21)	-	-	54.4	45.5	36.3	26.2	14.8	-	35.4	-
	CRRC-Net [11](TIP'22)	77.8	73.5	67.1	57.9	46.6	33.7	19.8	64.6	45.1	53.8
	PCL [26](ESWA'23)	74.6	70.2	63.3	55.9	44.4	-	-	61.7	-	-
	Ours	**82.3**	**77.6**	**70.1**	**60.0**	**49.4**	**37.6**	**24.5**	**67.9**	**48.3**	**57.4**

Table 2. Comparison results with the state-of-the-art methods on GTEA and BEOID in terms of mAP at different IoUs. * represents the reproduced results.

Dataset	Method	mAP@IoU (%)				AVG
		0.1	0.3	0.5	0.7	(0.1:0.7)
GTEA	SF-Net [33]	58.0	37.9	19.3	11.9	31.0
	LAC [21]	63.9	55.7	33.9	20.8	43.5
	DCM [19]	59.7	38.3	21.9	18.1	33.7
	PCL [26]	65.2	56.8	34.3	**21.2**	44.9
	LAC* [21]	64.1	50.9	37.0	13.8	41.9
	Ours	**74.6**	**60.9**	**39.5**	16.6	**49.0**
BEOID	SF-Net [33]	62.9	40.6	16.7	3.5	30.9
	LAC [21]	76.9	61.4	42.7	25.1	51.8
	DCM [19]	63.2	46.8	20.9	5.8	34.9
	BackTAL [52]	60.1	40.9	21.2	11.0	32.5
	PCL [19]	78.7	63.3	44.1	**26.9**	53.3
	LAC* [21]	72.2	62.7	45.7	16.0	51.5
	Ours	**83.8**	**73.0**	**51.1**	23.8	**59.6**

Table 3. Results of incorporating TSP-Net into other P-TAL methods on THUMOS'14. * represents the reproduced results.

Method	mAP@IoU (%)				
	0.1	0.3	0.5	0.7	AVG
SF-Net* [33]	68.1	52.3	29.1	10.9	40.4
SF-Net+Ours	**75.5**	**57.9**	**33.1**	**12.3**	**44.8**
BackTAL* [52]	66.8	54.4	36.3	14.8	43.7
BackTAL+Ours	**77.1**	**63.5**	**43.4**	**19.7**	**51.7**
LAC* [21]	74.8	63.7	45.9	21.8	52.5
LAC+Ours	**82.3**	**70.1**	**49.4**	**24.5**	**57.4**

4. Experiments

4.1. Experimental Setup

Dataset. In this work, we use three widely used P-TAL benchmarks. **THUMOS'14** [18] is a widely used TAL dataset. It contains 413 untrimmed videos with 20 sports action categories. We use 200 validation videos as the training set and test on 213 testing videos. **BEOID** [8] includes 58 videos with 34 operation classes in 6 locations. We follow previous work to use 46 videos for training and 12 videos for testing. **GTEA** [9] includes 28 untrimmed videos recorded in the kitchen and contains 7 fine-grained actions. 21 and 7 videos are split for training and testing, respectively.

Evaluation Metrics. For a fair comparison, we follow the previous P-TAL works [11, 21, 33, 52] and use mean Average Precision (mAP) as the evaluation metric. Concretely, we evaluate mAP under different IoU thresholds and compute their average.

As the center score learning is to learn the center alignment score. We design and compute the ratio of center alignment r_{IoU} between the instances and final proposals under different IoUs as Eq. 11, where c_g^i denotes center of the i-th instance, c_p^* denotes the center of matched proposal with instance, l_g denotes instance length, and N refer to the number of instances. $\mathbb{I}_{match}(\cdot)$ is the indicator function that gets 1 when the instance is matched with the proposal.

$$r_{IoU} = \frac{1}{N}\sum_{i=0}^{N-1}(1 - \mathbb{I}_{match}(\left|c_p^* - c_g^i\right|/l_g^i)) \qquad (11)$$

Implementation Details. We first utilize Denseflow [46] to extract the flow fields. For snippet-level feature extraction, we follow previous work to use I3D [3] pre-trained on Kinetics-400 in an open toolkit [7] as the extractor. The dimension of the feature sequence is 2048, and the proposal level feature dimension is 1024. The soft-value λ is set to 0.4. For saliency point updating, the update threshold θ_{up} and update period p_{up} are set to 0.8 and 200, respectively. The refine threshold in alignment-based boundary adaption is set to 0.4. In the training stage, the learning rate and

Table 4. Ablation results for the proposed components in terms of average mAP on THUMOS'14.

Setup				AVG mAP@IoU(%)		
PL	CLG	SPU	ABA	0.1:0.5	0.3:0.7	0.1:0.7
				62.1	44.6	52.5
√				65.9	45.6	55.0
√			√	66.5	46.5	55.7
√	√			67.2	47.3	56.5
√	√	√		67.6	47.7	56.9
√	√	√	√	**67.9**	**48.3**	**57.4**

Table 5. Impact of hard center labels and soft labels with different λ on localization performance on THUMOS'14.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
CLG-hard	81.5	68.7	48.1	23.9	56.2
CLG-soft ($\lambda = 0.2$)	82.0	69.0	48.2	23.6	56.6
CLG-soft ($\lambda = 0.3$)	81.9	69.5	49.0	24.5	57.1
CLG-soft ($\lambda = 0.4$)	**82.3**	**70.1**	**49.4**	**24.5**	**57.4**
CLG-soft ($\lambda = 0.5$)	82.2	69.5	48.9	23.9	56.8

Table 6. Ablation study of different types of updating strategy on THUMOS'14.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
iterative	81.5	69.0	49.0	23.9	56.6
SPU	**82.3**	**70.1**	**49.4**	**24.5**	**57.4**

weight decay are set to $5e$-5 and 0.001, respectively. The balance factor α is set to 20, the batch size is 10, and the total number of training iterations is 2000.

4.2. Comparison with The State-of-the-arts

We first evaluate the effectiveness of the TSP-Net on THUMOS'14. The results are shown in Table 1 and indicate that the performance of the proposed method largely outperforms previous P-TAL and W-TAL methods in terms of mAP at all IoU thresholds. Compared with CRRC-Net, TSP-Net gets 3.3%, 3.2%, and 3.6% improvements on the metrics of average mAP (0.1:0.5), average mAP (0.3:0.7) and average mAP (0.1:0.7), respectively. We also compare with the fully temporal supervised methods, and TSP-Net can still achieve competitive performance. After that, we conduct experiments on GTEA and BEOID, and the proposed method surpasses the state-of-the-art P-TAL methods by a large margin in terms of average mAP (0.1:0.7) by 4.1% and 6.3%, respectively.

4.3. Generalization

The proposed method is exclusively focused on the operation of proposals and is agnostic to the structure of the base locator. Consequently, TSP-Net can be easily integrated into any P-TAL method. We employ three excellent P-TAL methods to assess the generalization of TSP-Net on THUMOS'14. The results in Table 3 demonstrate a substantial improvement in the performance of all methods when TSP-Net is introduced to learn the proposals generated by the official models. In particular, we observed that our approach exhibited less improvement on SF-Net [33] due to its omission of multi-threshold proposal generation, resulting in a significantly smaller number of proposals. This underscores the importance of having an ample number of proposals for the effective center score learning and boundary adaptation.

Table 7. Ratio of center alignment at different IoU(%).

Setup	0.1	0.3	0.5	0.7	AVG
Baseline	**65.6**	69.1	61.0	41.3	60.9
Ours	65.4	**70.7**	**64.2**	**49.0**	**63.6**

Table 8. Ablation of the information in proposals need to be used for refining on THUMOS'14.

Setup	mAP@IoU(%)				
	0.1	0.3	0.5	0.7	AVG
w/ conf. refine	78.1	67.0	46.4	23.1	54.3
w/o maintain	78.9	67.0	46.6	20.6	54.1
w/o IoU info.	82.2	69.7	**49.5**	24.3	57.2
ABA	**82.3**	**70.1**	49.4	**24.5**	**57.4**

Table 9. Ablation of various point annotations.

	Gaussian	Manual	Center	Start	End
AVG mAP	**57.4**	56.3	57.1	52.3	47.1

4.4. Ablation Study

In this section, we conduct extensive ablation studies on the THUMOS'14 to demonstrate the proposed mechanisms' effectiveness and deeply analyze their effect. In each study, we adopt the hyper-parameters set in implementation details if not explicitly stated.

Effectiveness of the components. Table 4 indicates the results of different combinations of components. 'PL' denotes the proposal learning without the saliency branch in Sec. 3.2, 'CLG' and 'SPU' represent center label generation for learning aligned scores and saliency point updating, respectively, and 'ABA' denotes the proposed adaption strategy in Sec. 3.4. Regarding average mAP (0.1:0.7), the baseline only achieves 52.5% without any component. After introducing proposal learning, the performance gets a 2.5% improvement. It indicates that using point annotations to retrain the proposals can get more information about the action and actionness. The aligned center score provides the evidence to match the most complete proposal with the ground truth and gets 56.5%. The combination of CLG and SPU can achieve a 0.4% improvement due to the utilization of more superior saliency points. In addition, whether center score learning is introduced or not, ABA is effective and achieves 55.7% and 57.4% in two conditions, respectively. In addition to the final metric, results in Table 7 indicate that the proposed framework benefits center alignment, especially when IoU is high.

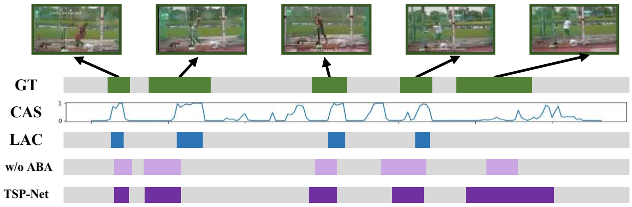

Figure 4. Visualization of the localization of the base LAC, CAS and the proposed method that applies ABA or not (IoU > 0.3). The aligned confidence from TSP-Net can suppress the low-quality proposals and increase the proposal completeness.

Impact of soft label in CLG. As mentioned in Sec.3.3, the hard center label may cause the degradation of the model. We conduct experiments on the hard-term center score learning and the soft-term with various soft values λ in Table 5. The hard-term label can only achieve 56.2% in terms of average mAP (0.1:0.7) as the tough supervision inevitably results in a wrong prediction for proposals containing partial instances. The soft-term supervision with a proper soft value can alleviate the issue, especially when $\lambda = 0.4$, TSP-Net can achieve the best performance.

How to update saliency points. A conventional method for updating manually constructed points or features involves iterative updating using previously updated information during the updating process [11]. In our approach, we update the points based on the initial points and compare different strategies in Table 6. Results indicate that iteration-based updating introduces cumulative noise, whereas the initial-based approach effectively captures superior saliency information.

How to adapt boundaries based on saliency information. Compared with the boundary refining in [59], ABA does not refine the learned confidence scores, retains the proposals in the close-set, and introduces IoU information for refinement. The results in Table 8 indicate that only the boundaries need to be refined when the confidence is aligned. The IoU information enables differentiation in the adaptation process, and the retained sub-optimal proposals are also valuable when action instances occur consecutively.

Influence of point annotations. We conduct an ablation about the influence of various point annotations in Table 9. The performance of manual annotation is worse as some points fall in the boundary regions, but our method still works well. When all the points are at boundaries, our method fails, which also proves the value of our motivation. In fact, ideal absolute center is suboptimal in our work, it breaks the commonality between features from the action recognition backbone, CAS-based proposals [56], and our motivation to always produce confident responses to salient/distinguishable frames. Gaussian-based points suppress the points in the boundary regions while maintaining a human-like annotation distribution and get the best result.

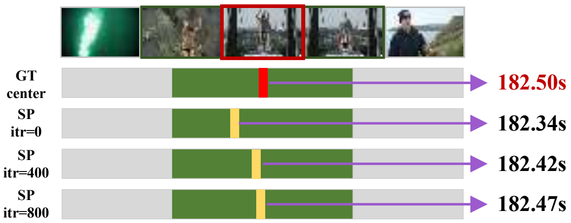

Figure 5. Illustration of the proposed saliency point updating strategy across various training iterations. As robust proposals are identified and utilized, the initial saliency points progressively converge toward their respective instance centers.

4.5. Qualitative Analyze

Qualitative comparison. In Figure 4, we qualitatively compare TSP-Net with the baseline method LAC [21] at the IoU larger than 0.3. The visualization indicates that the base proposals with high confidence always overlap less with the ground-truths as the confidence is seriously affected by CAS. To make matters worse, many high-quality proposals with high IoU are suppressed by post-processing operations because of misaligned scores (the 5th instance in Figure 4). After introducing center score learning, better-aligned confidence can increase proposal completeness and successfully locate hard instances. Moreover, The completeness of localization can be further improved when boundaries are adapted using information from neighboring proposals.

Saliency point and instance center. As stated in Sec. 3.3, the closer the saliency point is to the center region, the more valuable alignment information can be learned. We visualize the process of the updating strategy in Figure 5. The proposed SPU can continually adjust the saliency point to generate more valuable labels.

5. Conclusion

In this study, we scrutinize the misalignment challenges in P-TAL and introduce a plug-in proposal learning framework for realigning confidence with proposals. We first introduce the center score learning to get the aligned center scores, and the alignment-based boundary adaption is proposed to enhance the localization completeness during inference. Experimental evaluations validate the effectiveness, showcasing competitive or state-of-the-art performance. Moreover, the framework is generalized for any P-TAL method, revealing a novel relationship between point labels and proposal quality.

6. Acknowledgments

This work was supported in part by the Natural Science Foundation of Sichuan, China under Grant 2023NSFSC0484, in part by the NNSFC&CAAC under Grants U2233209 and U2133211, and in part by the National Natural Science Foundation of China under Grant 62071104.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 5

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 6

[4] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. Dcan: improving temporal action detection via dual context aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 248–257, 2022. 2

[5] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 6

[6] Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, and Cordelia Schmid. A flexible model for training action localization with varying levels of supervision. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[7] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 6

[8] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, page 3, 2014. 6

[9] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 6

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[11] Jie Fu, Junyu Gao, and Changsheng Xu. Compact representation and reliable classification learning for point-level weakly-supervised action localization. *IEEE Transactions on Image Processing*, 31:7363–7377, 2022. 1, 3, 4, 6, 8

[12] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022. 1, 2

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[14] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13925–13935, 2022. 1, 2, 6

[15] Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8002–8011, 2021. 1, 2

[16] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022. 6

[17] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6565–6574, 2023. 1

[18] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 6

[19] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13455–13464, 2021. 3, 6

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[21] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13648–13657, 2021. 1, 3, 4, 6, 8

[22] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1854–1862, 2021. 1, 2

[23] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2373–2383, 2023. 1, 2

[24] Sangjin Lee, Jaebin Lim, Jinyoung Moon, and Chanho Jung. An improved point-level supervision method for temporal action localization. *IEEE Access*, 2023. 3

[25] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023. 1

[26] Ping Li, Jiachen Cao, and Xingchao Ye. Prototype contrastive learning for point-supervised temporal action detection. *Expert Systems with Applications*, 213:118965, 2023. 1, 3, 6

[27] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11499–11506, 2020. 2

[28] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2021. 1

[29] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[30] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 1, 2

[31] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1683–1691, 2022. 1

[32] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9969–9979, 2021. 2

[33] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. Sf-net: Single-frame supervision for temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 420–437. Springer, 2020. 1, 3, 6, 7

[34] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 437–453. Springer, 2016. 3

[35] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Post-processing temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18837–18845, 2023. 2

[36] Prashant W Patil, Akshay Dudhane, Sachin Chaudhary, and Subrahmanyam Murala. Multi-frame based adversarial learning approach for video surveillance. *Pattern Recognition*, 122:108350, 2022. 1

[37] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 485–494, 2021. 6

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[39] Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2394–2404, 2023. 1, 2, 4, 6

[40] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 1, 2, 6

[41] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 1, 2

[42] Haichao Shi, Xiao-Yu Zhang, and Changsheng Li. Stochasticformer: Stochastic modeling for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 32:1379–1389, 2023. 1, 2

[43] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018. 2

[44] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029, 2022. 1

[45] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13526–13535, 2021. 6

[46] Shiguang* Wang, Zhizhong* Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. denseflow. https://github.com/open-mmlab/denseflow, 2020. 6

[47] Kun Xia, Le Wang, Sanping Zhou, Nanning Zheng, and Wei Tang. 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*, pages 13884–13893, 2022. 1, 2

[48] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165, 2020. 6

[49] Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. Point-supervised video temporal grounding. *IEEE Transactions on Multimedia*, 2022. 3

[50] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 1

[51] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020. 2

[52] Le Yang, Junwei Han, Tao Zhao, Tianwei Lin, Dingwen Zhang, and Jianxin Chen. Background-click supervision for temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9814–9829, 2021. 1, 3, 6

[53] Yuan Yin, Yifei Huang, Ryosuke Furuta, and Yoichi Sato. Proposal-based temporal action localization with point-level supervision. *arXiv preprint arXiv:2310.05511*, 2023. 3

[54] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6209–6223, 2021. 6

[55] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Nanning Zheng, David Doermann, Junsong Yuan, and Gang Hua. Adaptive two-stream consensus network for weakly-supervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4136–4151, 2022. 1, 2

[56] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 1, 2, 8

[57] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021. 6

[58] Tao Zhao, Junwei Han, Le Yang, and Dingwen Zhang. Equivalent classification mapping for weakly supervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3019–3031, 2022. 1, 2

[59] Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, and Hongsheng Li. Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23003–23012, 2023. 1, 2, 3, 5, 6, 8