Proposal Score Realignment Guided by Semantic Completeness for Weakly Supervised Temporal Action Localization

Maodong Li[1], Zhihao Wang[1], Jingxiong Wang[1], Jian Wang[1](✉), and Bing Li[1](✉)

School of Computer Science, Wuhan University, Wuhan, China
{limaodong2023,jianwang,bingli}@whu.edu.cn

Abstract. Due to the absence of temporal annotations, recent weakly supervised temporal action localization (WTAL) methods often adopt pseudo-label learning to boost localization. A segment-level WTAL baseline first generates proposals and constructs pseudo labels, which then supervise a fully supervised localization head. However, the construction of pseudo labels depends on the proposals' original confidence scores. Accumulated noise in proposal generation induces a mismatch between these confidence scores and proposal quality, thereby limiting the achievable quality of the pseudo labels. To address this, we propose SCLR, a proposal-level score realignment framework driven by Semantic Completeness Learning (SCL). SCL proceeds in three steps. First, for each action class, we construct a core semantic center that focuses on local high-response cues and a global teacher center that aggregates class-level contextual commonalities. Second, guided by the two centers, we self-supervise the decomposition of the core semantic center into generic and specific components, then reconstruct the class center by a weighted sum. Third, we derive soft labels from proposal–to–center similarity to supervise a per-proposal semantic-completeness score, which is then used for score realignment. Extensive experiments on two public datasets demonstrate that SCLR, when integrated as a plug-and-play framework, consistently reduces score misalignment across diverse baselines and improves localization accuracy. Further, leveraging re-aligned proposals as pseudo labels to train the localization head yields state-of-the-art performance.

Keywords: Weakly-supervised temporal action localization · Pseudo-label Learning · Proposal Score Realignment

1 Introduction

Weakly supervised temporal action localization (WTAL) aims to recognize and temporally localize actions in untrimmed videos using only video-level labels. Due to its low annotation cost, WTAL has garnered increasing attention in recent years and exhibits strong potential for practical applications [10, 38], including human–computer interaction and intelligent surveillance.

However, the lack of temporal annotations still leaves WTAL trailing behind fully supervised temporal action localization (FTAL), which relies on dense temporal labels and fully supervised localization heads. To narrow this gap, recent approaches [26, 33, 46, 47] learn from pseudo-temporal labels. A segment-level WTAL baseline first generates proposals and constructs instance-level pseudo labels. These labels then supervise a localization head following fully supervised designs to localize action instances. This line of work primarily devises strategies to exploit baseline priors better, in particular the original proposal confidence, to improve pseudo-label quality. A critical question, however, is whether these confidence scores are aligned with proposal quality. Since pseudo-labels are typically constructed and selected using original proposal confidence, a confidence-quality mismatch can compromise the reliability of the resulting pseudo-labels.

Specifically, most segment-level WTAL methods follow a localization-by-classification paradigm [6, 39, 50]. A representative instantiation is segment-level multiple-instance learning (S-MIL) pipeline [39]. As shown in Fig. 1, during training, the network predicts a class activation sequence (CAS) and a foreground-attention score for each segment (a short segment comprising multiple frames). In inference, a hand-crafted post-processing pipeline takes those two scores as input, aggregates segments into proposals, and computes proposal-level scores to yield the final instance predictions. In this pipeline, the training objective biases the model toward the most discriminative segments while neglecting ambiguous ones. This accelerates convergence but yields many misclassified segments. Moreover, the scoring objectives used in training and inference are inconsistent. Segment-level scoring errors accumulate when proposal scores are computed via hand-crafted rules, leading to situations where a high proposal score does not imply high proposal quality. Consequently, potentially high-quality proposals can be suppressed by standard post-processing. Therefore, it is beneficial to realign proposal scores after proposal generation. This step not only directly improves localization for WTAL baselines, but also provides a more reliable prior for pseudo-label construction, which in turn yields higher-quality pseudo labels and thus stronger supervision for training the localization head.

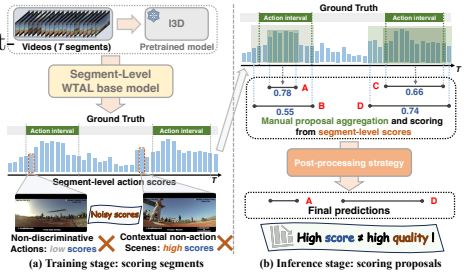

Fig. 1: Illustration of score–quality misalignment in segment-level WTAL. During training, segment-level scoring can assign low scores to non-discriminative action segments and high scores to contextual non-action segments. During inference, hand-crafted post-processing aggregates these noisy segment scores into proposal-level scores, so high proposal scores may not indicate high proposal quality.

Following this motivation, we start from the video-level classification signal, the only reliable supervision in WTAL, and introduce SCLR for proposal

score realignment. Through Semantic Completeness Learning (SCL), SCLR establishes explicit supervision between proposal features and class semantics, thereby reducing proposal-quality assessment to proposal–center similarity in semantic space. Here, semantic completeness refers to proposal–center alignment in semantic space rather than temporal completeness (full extent), since temporal boundaries are under-determined under video-level supervision. Concretely, SCL comprises three steps. First, for each ground-truth (GT) action class present in a video, we construct a core semantic center that emphasizes discriminative semantics from high-response local regions and a global teacher center that provides a complementary global context view by aggregating proposals over the whole video. Second, leveraging this local–global pair, we derive a class-agnostic video-level representation that captures video-specific factors shared across the video (e.g., characteristic camera motion). Guided by the two centers and the video representation, we self-supervise a decomposition of the core semantic center into a generic component (*what action is it?*) and a specific component (*how is this action instantiated in the current video?*). A weighted sum of these components reconstructs the class center. Finally, we learn a semantic-completeness score by comparing each proposal with the reconstructed class semantic center, which refines proposal quality assessment and aligns scores with quality.

Overall, our contributions are threefold: 1) We systematically characterize score–quality misalignment in recent instance-level pseudo-label WTAL pipelines and show that it can become a practical bottleneck for pseudo-label construction and selection. 2) We propose SCLR, a plug-and-play proposal-level score realignment framework driven by Semantic Completeness Learning (SCL). SCL constructs class centers and learns proposal–center similarity to reassess proposal quality in semantic space, producing more reliable proposal scores. 3) Extensive experiments on THUMOS14 and ActivityNet1.3 verify that SCLR consistently improves localization accuracy and reduces score–quality misalignment across diverse baselines. Moreover, training the localization head with re-aligned proposals as pseudo labels yields state-of-the-art performance.

2 Related Work

Fully supervised temporal action localization. FTAL methods are commonly grouped into single-stage [9,22,25,28,43] and two-stage detectors [4,5,20, 31,42]. Single-stage designs are compact, unifying proposal generation and classification, whereas two-stage pipelines first generate proposals and then classify them, allowing greater design flexibility [38]. Recent progress has been largely driven by the ActionFormer family [35, 45], which builds pyramidal temporal features and introduces Transformers into FTAL, yielding substantial accuracy gains. However, the reliance on dense temporal annotations remains the primary bottleneck to scaling FTAL in practice. Within FTAL, the work most related to ours is RefactorNet [40]. It uses temporal annotations to factorize raw features into action and co-occurrence components and then reconstructs them to enhance the input, which improves WTAL performance. Unlike RefactorNet,

SCLR requires no additional supervision. Instead of enhancing input features, it reassesses proposal quality via semantic-completeness learning.

Weakly supervised temporal action localization. WTAL pipelines are commonly divided into pre-classification [7,13,15,32,39,50] and post-classification [12, 27, 30, 34, 36, 48]. The former classifies each segment with temporal convolutions and then obtains a video-level prediction via Top-K pooling, etc. They are efficient but limited by receptive fields and prone to local dominance [38]. The latter estimates segment-wise attention and aggregates weighted features into a video-level representation. Although they leverage context, they may suppress fine-grained cues and thus suffer from the action-context confusion challenge [38]. Overall, these methods have pushed the field forward. Yet under weak supervision, the lack of frame-level labels breaks the segment–category correspondence, leaving a clear performance gap between WTAL and FTAL. To bridge this gap, recent studies have turned to pseudo-label learning with localization heads [26, 33, 46, 47]. However, these works mainly build proposal-level pseudo labels from baseline priors (e.g., proposal confidence and snippet cues) to train an FTAL-style localization head. In contrast, we target score–quality alignment at the proposal level. By recalibrating scores to reflect proposal quality, we improve the reliability of those priors and complement this line of work.

Among existing works, P-MIL [32] and SEAL [21] are most closely related to ours in that they explicitly introduce a proposal-level branch for completeness assessment. Nevertheless, their completeness estimation follows an attention-driven paradigm: high-attention proposals are first treated as pseudo labels, and the completeness of each candidate proposal is then measured by its overlap with this pseudo-label set. Their goal is to class-agnostically filter relatively low-quality proposals. Moreover, using only attention scores can introduce cross-class supervision in multi-class videos and overemphasize short high-foreground proposals, leading to model confusion. In contrast, we construct multiple semantically complete class centers and reassess proposal completeness from the perspective of class semantics using a unified criterion, thereby learning more reliable proposal scores. Our objective is not to filter out "incomplete" proposals, but to explicitly evaluate every proposal.

3 Method

Problem Definition. Let $\mathcal{V} = \{v_i\}_{i=1}^{N}$ be a set of N untrimmed videos, and let $\mathcal{Y} = \left\{ y_i \in \{0,1\}^{C+1} \right\}_{i=1}^{N}$ be the corresponding video-level labels, where C is the number of action classes and the additional class denotes background. With only $\mathcal{Y}$ available for training supervision, WTAL aims to predict, for a test video v, a set of temporal action instances $\left\{(s_j, e_j, c_j, q_j)\right\}_{j=1}^{M}$, where s_j, e_j, c_j, and q_j denote the start time, end time, action class, and confidence score, respectively.

Overview. As shown in Fig. 2, our pipeline is organized around the SCLR framework. First, we generate proposals using public baselines (Sec. 3.1). Next, taking the proposals as input, SCLR realigns their scores (Sec. 3.2). Finally, using the re-aligned proposals, we perform localization in two settings (Sec. 3.3).

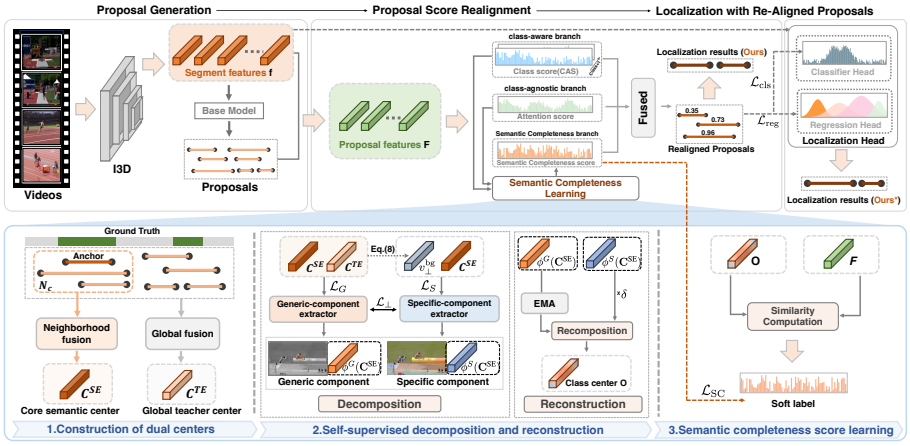

Fig. 2: Overview of the proposed pipeline. Our score realignment framework extends the pseudo-label learning paradigm and comprises three stages: (1) A baseline generates candidate proposals; (2) Given these proposals, SCLR employs Semantic Completeness Learning (SCL) to reassess proposal quality from semantic evidence and assign re-aligned scores; (3) The re-aligned proposals are used for **direct localization (Ours)** and to construct pseudo labels that **supervise a localization head (Ours*)**.

3.1 Proposal Generation and Framework Design

Proposal Generation. Following [6, 10, 33, 50], we uniformly partition each video into T non-overlapping segments. For each video, we extract RGB features $f^{\text{RGB}} \in \mathbb{R}^{T\times D}$ and optical-flow features $f^{\text{Flow}} \in \mathbb{R}^{T\times D}$ as inputs to the baseline model, yielding training candidates $P^{\text{train}} = \{(s_i, e_i)\}_{i=1}^{M_1}$ and test candidates $P^{\text{test}} = \{(s_i, e_i)\}_{i=1}^{M_2}$. During this step, we omit final post-processing (e.g., NMS) and discard baseline priors such as predicted class labels, keeping only the start–end timestamps of each proposal.

Framework Design. Our SCLR follows the design of S-MIL. Given the proposal set and the concatenated segment-level features $\mathbf{f} = [f^{\text{RGB}}, f^{\text{Flow}}]$, we obtain proposal-level features $\mathbf{F} \in \mathbb{R}^{M\times d}$ using a surrounding-contrast feature extractor [32] plus a projection layer for dimensionality reduction. We then introduce three independent branches to predict: proposal-level CAS $\mathbf{S}^{\text{base}} \in \mathbb{R}^{M\times(C+1)}$, foreground attention scores $\mathbf{A} \in \mathbb{R}^{M\times 1}$, and semantic-completeness scores $\mathbf{S}^{\text{sc}} \in \mathbb{R}^{M\times 1}$. Following S-MIL, under the assumption that background persists in videos and can be suppressed by $\mathbf{A}$, the background-suppressed CAS for each proposal is $\mathbf{S}^{\text{sup}} = \mathbf{S}^{\text{base}} \odot \mathbf{A}$ ($\odot$ denotes element-wise multiplication).

3.2 Proposal Score Realignment via Semantic Completeness Learning

Motivation. Under weak supervision, baseline-generated proposals are often over-complete or incomplete. Moreover, proposals of the same class may corre-

spond to different GT instances, so evaluating proposal quality solely by their temporal overlap with a few high-confidence proposals is unreliable.

We observe that proposals of the same class tend to share discriminative action semantics, and semantic similarity is often more robust than temporal overlap to duration variations. This motivates us to assess proposal quality from *semantic completeness* rather than IoU. However, using only discriminative cues may lead to trivial solutions, e.g., overemphasizing short but highly activated fragments.

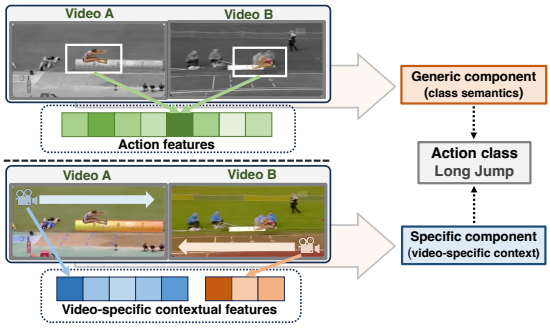

Fig. 3: Generic vs. specific components in long-jump videos. Top: class-stable action semantics shared across videos. Bottom: video-conditioned contextual cues that vary across videos.

As shown in Fig. 3, a semantically complete action instance is not only determined by class-discriminative cues but also by video-conditioned contextual cues (e.g., co-occurring context and camera motion). Accordingly, we reconstruct a semantically complete class center by decomposing per-class proposal features under self-supervision into generic and specific components, and recombining them:

- **Generic component (class-stable action semantics)**: shared across videos of the same class, capturing transferable and class-discriminative cues;
- **Specific component (video-conditioned contextual semantics)**: unique to the current video, capturing contextual cues (e.g., appearance, scene, and camera motion) that affect how an action appears in the current video.

The reconstructed class center serves as a stable semantic reference: proposal–center similarity yields semantic-completeness scores, which are then used to *realign* proposal confidence so that scores reflect proposal quality under the same criterion in both training and inference.

Dual-center construction. Under weak supervision, MIL-based baselines tend to assign high confidence to proposals with discriminative action evidence. However, even the highest-confidence candidate, denoted as the anchor P^*, may have an unreliable temporal extent, being biased toward short discriminative fragments or over-extended context. We thus aggregate P^* with its *predicted* same-class neighborhood to form a local core semantic center $\mathbf{C}^{\mathrm{SE}}$ via boundary-aware local aggregation, which strengthens local action cues. However, a purely local reference may entangle class semantics with video-specific contextual cues. We therefore introduce a complementary global view by aggregating proposals over the whole video to build a global teacher center $\mathbf{C}^{\mathrm{TE}}$. Together, the local–global pair provides a cross-view reference for decomposition: semantics

consistent across $\mathbf{C}^{\mathrm{SE}}$ and $\mathbf{C}^{\mathrm{TE}}$ form the class-stable generic component, while video-conditioned contextual cues are retained in the specific component.

For video v, let $\mathcal{I}$ denote the index set of all proposals and $\mathcal{N}_c \subseteq \mathcal{I}$ those predicted as class c. A naive anchor for class c is the proposal with the largest class-specific background-suppressed score $\mathbf{S}^{\mathrm{sup}}_{i,c}$, where $\mathbf{S}^{\mathrm{sup}} = \mathbf{S}^{\mathrm{base}} \odot \mathbf{A}$. However, $\mathbf{S}^{\mathrm{base}}$ and $\mathbf{A}$ are heterogeneous and their magnitudes can be mismatched or outlier-prone under weak supervision, making magnitude-based selection brittle. We therefore use a rank-fusion principle: we select the anchor that is jointly top-ranked in class evidence and actionness, which is invariant to monotone transformations and robust to magnitude noise. To implement this in a simple, scale-free manner, we adopt Hazen mid-rank normalization on any per-proposal score sequence $\mathbf{z} = (\mathbf{z}_i)_{i\in\mathcal{S}}$ over an index set $\mathcal{S}$:

$$\mathrm{Q}_{\mathcal{S}}[\mathbf{z}]_i \triangleq \frac{|\mathcal{S}| - \mathrm{rank}_{\mathcal{S}}(\mathbf{z}_i) + 0.5}{|\mathcal{S}|} \in (0,1), \tag{1}$$

where $\mathrm{rank}_{\mathcal{S}}(\mathbf{z}_i)$ denotes the descending rank of $\mathbf{z}_i$ within $\mathcal{S}$ (ties averaged). Applying Eq. (1) to $\mathcal{N}_c$, we define the joint score and select the anchor P^* as:

$$J_{i,c} = \mathrm{Q}_{\mathcal{N}_c}[\mathbf{S}^{\mathrm{base}}_c]_i \, \mathrm{Q}_{\mathcal{N}_c}[\mathbf{A}]_i, \ i \in \mathcal{N}_c, \qquad i^* = \arg\max_{i\in\mathcal{N}_c} J_{i,c}, \ P^* = P_{i^*}. \tag{2}$$

Based on P^*, we form its local neighborhood $\mathcal{N}^*_c = \{i \in \mathcal{N}_c \mid \mathrm{IoU}(P_i, P^*) \geq \theta\}$ for local fusion. To aggregate this neighborhood into a local center, the weight should account not only for class evidence and actionness but also for *relative temporal geometry*, since proposals with similar scores can still differ greatly in temporal coverage. We first use $\mathrm{IoU}(P_i, P^*)$ as a coarse overlap cue. However, IoU is necessary but insufficient: neighbors with similar $\mathrm{IoU}(P_i, P^*)$ can still have very different center displacement or duration, e.g., being shifted to context or temporally truncated. Therefore, we introduce a center–scale affinity k^{cen} to complement IoU by accounting for center shift and temporal scale, implemented as the Bhattacharyya coefficient [8] between Gaussian surrogates of two intervals. For an interval with start/end (s, e), let $\mu = (s+e)/2$, $\ell = e-s$, and $\sigma = \ell/\sqrt{12}$. With (μ^*, σ^*) for P^* and (μ_i, σ_i) for P_i, we define:

$$k^{\mathrm{cen}}(i,*) = \sqrt{\frac{2\,\sigma_i\,\sigma^*}{\sigma_i^2 + (\sigma^*)^2}} \exp\!\left(-\frac{(\mu_i - \mu^*)^2}{4(\sigma_i^2 + (\sigma^*)^2)}\right) \in (0,1], \tag{3}$$

which equals 1 iff $\mu_i = \mu^*$ and $\sigma_i = \sigma^*$. We then compute the fusion weights $w^{\mathrm{SE}}_{i,c}$ and obtain the core semantic center $\mathbf{C}^{\mathrm{SE}}_c$ as:

$$w^{\mathrm{SE}}_{i,c} = \mathrm{norm}_{\mathcal{N}^*_c}\Big(\mathbf{S}^{\mathrm{sup}}_{i,c} \cdot \mathrm{IoU}(P_i, P^*) \cdot k^{\mathrm{cen}}(i,*)\Big), \quad \mathbf{C}^{\mathrm{SE}}_c = \mathrm{norm}_2\Big(\sum_{i\in\mathcal{N}^*_c} w^{\mathrm{SE}}_{i,c}\,\mathbf{F}_i\Big), \tag{4}$$

where $\mathrm{norm}_{\mathcal{N}^*_c}(\cdot)$ denotes set-wise normalization over $\mathcal{N}^*_c$ and $\mathrm{norm}_2(\cdot)$ denotes ℓ_2 normalization.

For $\mathbf{C}^{\mathrm{TE}}_c$, unlike $\mathbf{C}^{\mathrm{SE}}_c$ built from $\mathcal{N}^*_c$, we aggregate features from all proposals indexed by $\mathcal{I}$. At this global scope, raw class-score magnitudes $\mathbf{S}^{\mathrm{base}}_c$ are

more susceptible to outliers and calibration drift; we therefore use the scale-free rank ordering $\mathrm{Q}_{\mathcal{I}}\big[\mathbf{S}_c^{\mathrm{base}}\big]$ as the class-conditioned prior. To complement the local evidence view and avoid over-concentration around the anchor, we introduce a non-local temporal gate that increases with the temporal distance from P^*:

$$k_{i\leftarrow P^*}^{\mathrm{time}} = 1 - \exp\!\left(-\frac{(\mu_i - \mu^*)^2}{2(\ell^*)^2}\right), \tag{5}$$

which vanishes at the anchor and asymptotically approaches 1 as $|\mu_i - \mu^*|$ grows, encouraging $\mathbf{C}_c^{\mathrm{TE}}$ to incorporate broader class-conditioned action evidence.

With these terms, we compute the weights and obtain $\mathbf{C}_c^{\mathrm{TE}}$:

$$w_{i,c}^{\mathrm{TE}} = \mathrm{norm}_{\mathcal{I}}\Big(\mathrm{Q}_{\mathcal{I}}[\mathbf{S}_c^{\mathrm{base}}]_i \cdot \mathbf{A}_i \cdot k_{i\leftarrow P^*}^{\mathrm{time}}\Big), \quad \mathbf{C}_c^{\mathrm{TE}} = \mathrm{norm}_2\Big(\sum_{i\in\mathcal{I}} w_{i,c}^{\mathrm{TE}}\,\mathbf{F}_i\Big). \tag{6}$$

Together, $\mathbf{C}^{\mathrm{SE}}$ and $\mathbf{C}^{\mathrm{TE}}$ provide complementary local and global semantic references, omitting the class index c for brevity.

Self-supervised decomposition and reconstruction. Given the local–global centers $\mathbf{C}^{\mathrm{SE}}$ and $\mathbf{C}^{\mathrm{TE}}$, we use two independent extractors ϕ^G and ϕ^S to separate (i) a *class-stable* generic component and (ii) a *video-conditioned* specific component, reducing factor entanglement under weak supervision. The extractor ϕ^G is trained to capture invariant semantics shared by both centers:

$$\mathcal{L}_G = 1 - \cos\big(\phi^G(\mathbf{C}^{\mathrm{SE}}),\, \phi^G(\mathbf{C}^{\mathrm{TE}})\big), \tag{7}$$

which aligns the generic components extracted from the two centers, yielding a shared class-stable representation.

Similarly, separating the specific component requires a video-conditioned reference for contrast. To prevent this reference from carrying class-stable action semantics, we first use the background weight $(1 - \mathbf{A}_i)$ to aggregate proposal features into a background/context descriptor, and then project out the generic subspace spanned by $\phi^G(\mathbf{C}^{\mathrm{SE}})$ and $\phi^G(\mathbf{C}^{\mathrm{TE}})$:

$$\begin{aligned} \mathbf{v}^{\mathrm{bg}} &= \frac{\sum_{i\in\mathcal{I}}(1-\mathbf{A}_i)\mathbf{F}_i}{\sum_{i\in\mathcal{I}}(1-\mathbf{A}_i)}, \quad U = [\phi^G(\mathbf{C}^{\mathrm{SE}}),\, \phi^G(\mathbf{C}^{\mathrm{TE}})], \\ \mathbf{v}_{\perp}^{\mathrm{bg}} &= \mathrm{norm}_2\Big(\big(\mathbf{I}_d - U(U^\top U + \varepsilon\mathbf{I}_2)^{-1}U^\top\big)\mathbf{v}^{\mathrm{bg}}\Big), \end{aligned} \tag{8}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Here, $\mathbf{v}^{\mathrm{bg}}$ is a raw video-specific background/context descriptor, while $\mathbf{v}_{\perp}^{\mathrm{bg}}$ is its residual after removing the generic action-semantic subspace, thereby reducing class-semantic leakage into the specific component. Using $\mathbf{v}_{\perp}^{\mathrm{bg}}$ as the specific reference, we align the specific component extracted from $\mathbf{C}^{\mathrm{SE}}$ with it:

$$\mathcal{L}_S = 1 - \cos\big(\phi^S(\mathbf{C}^{\mathrm{SE}}),\, \phi^S(\mathbf{v}_{\perp}^{\mathrm{bg}})\big). \tag{9}$$

To further disentangle the two components, we impose a decorrelation constraint between their representations extracted from the same $\mathbf{C}^{\mathrm{SE}}$:

$$\mathcal{L}_{\perp} = \Big(\cos\big(\phi^G(\mathbf{C}^{\mathrm{SE}}),\, \phi^S(\mathbf{C}^{\mathrm{SE}})\big)\Big)^2. \tag{10}$$

The total decomposition loss $\mathcal{L}_{\mathrm{DEC}}$ is:

$$\mathcal{L}_{\mathrm{DEC}} = \mathcal{L}_G + \mathcal{L}_S + \lambda_\perp \mathcal{L}_\perp, \tag{11}$$

where $\lambda_\perp$ balances the decorrelation term. Next, we reconstruct the class center by combining the generic component $\phi^G(\mathbf{C}^{\mathrm{SE}})$ and the specific component $\phi^S(\mathbf{C}^{\mathrm{SE}})$. By design, the generic component is updated across videos through a class-wise EMA for stable class-level semantics, whereas the specific component remains video-specific. Accordingly, the reconstructed center for class c is:

$$\mathbf{O}_c = \mathrm{norm}_2\Big(\mathrm{EMA}\left(\phi^G(\mathbf{C}_c^{\mathrm{SE}})\right) + \delta \cdot \phi^S(\mathbf{C}_c^{\mathrm{SE}})\Big). \tag{12}$$

Here, EMA$(\cdot)$ denotes the class-wise cross-video exponential moving average (momentum 0.9), and δ balances the specific term (set to 0.2).

Learning semantic-completeness scores. Building on $\mathbf{O}$, we measure a proposal's semantic completeness by its similarity to the corresponding center and use this as supervision for the semantic-completeness branch. However, two challenges remain in practice. (1) Under weak supervision, the base model inevitably generates proposals that do not belong to any GT class, so supervising them with their maximum similarity to class centers in the current video can introduce misleading gradients. (2) At early training or in very short videos, positives for some GT classes may be too sparse to define $\mathbf{C}_c^{\mathrm{SE}}$, leaving this branch without a learning target. To address this, we construct a conservative soft label for all proposals. We use similarity only when a proposal is predicted as a GT class and the corresponding class center is available. Otherwise, we use a constant γ.

Let $\hat{c}_i$ denote the predicted class of the i-th proposal in the current training video, and let $\mathcal{C}_v = \{c \mid y_c = 1 \wedge |\mathcal{N}_c| > 0\}$ denote the set of GT classes with available centers. The soft label and the loss are:

$$y_i^{\mathrm{sc}} = \begin{cases} \psi\big(\cos(\mathbf{O}_{\hat{c}_i}, \mathbf{F}_i)\big), & \hat{c}_i \in \mathcal{C}_v, \\ \gamma, & \text{otherwise.} \end{cases} \tag{13}$$

$$\mathcal{L}_{\mathrm{SC}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \varphi\big(\mathbf{S}_i^{\mathrm{sc}},\, y_i^{\mathrm{sc}}\big). \tag{14}$$

Here $\psi(x) = (1+x)/2$ and $\varphi(\cdot,\cdot)$ denotes the MSE loss. Both ϕ^G and ϕ^S preserve feature dimensionality ($\mathbb{R}^d \to \mathbb{R}^d$), making $\cos(\mathbf{O}_{\hat{c}_i}, \mathbf{F}_i)$ well-defined when $\hat{c}_i \in \mathcal{C}_v$.

3.3 Action Localization with Re-aligned Proposals

Model training. During training, the score-realignment objective of SCLR is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MIL}} + \mathcal{L}_{\mathrm{DEC}} + \lambda_{\mathrm{SC}} \mathcal{L}_{\mathrm{SC}}, \tag{15}$$

where $\mathcal{L}_{\mathrm{MIL}}$ is the multi-instance learning loss [6, 10, 32, 33] and λ_{SC} balances $\mathcal{L}_{\mathrm{SC}}$. Optimizing $\mathcal{L}$ jointly updates all three branches and learns a re-aligned score for each proposal, defined as:

$$\mathbf{S}_i^{\mathrm{re}} = \mathbf{S}_i^{\mathrm{sup}} \odot \mathbf{S}_i^{\mathrm{sc}}, \tag{16}$$

which makes training and inference use the same proposal-level score. Using the re-aligned scores, we construct instance-level pseudo labels with soft non-maximum suppression (Soft-NMS) [1] and Gaussian fusion [26, 41, 50]. Based on these pseudo labels, we further train a localization head. The localization head is built following ActionFormer [45] and TriDet [35], and its loss function is:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}, \tag{17}$$

where $\mathcal{L}_{\text{cls}}$ is focal loss [24] and $\mathcal{L}_{\text{reg}}$ is DIoU loss [49].

Model inference. We evaluate two protocols that directly address the questions raised in the introduction.

(i) SCLR only. Given the test proposal set P^{test}, SCLR produces re-aligned proposals $\hat{P}^{\text{test}}$. Applying Soft-NMS to $\hat{P}^{\text{test}}$ yields the final localization results. This setting removes the localization head and isolates the contribution of score realignment to baseline localization.

(ii) SCLR with localization head. The localization head is trained on pseudo labels constructed from the re-aligned training proposals $\hat{P}^{\text{train}}$. At test time, inference is performed with the head alone. This setting evaluates whether score realignment improves pseudo-label supervision and thereby enhances the head's localization performance.

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate on two public benchmarks. **THUMOS14** [14] contains 200 validation and 213 test videos over 20 classes. We train on the validation split and evaluate on the test split. **ActivityNet1.3** [2] comprises 10,024 training and 4,926 validation videos across 200 classes. We train on the training split and evaluate on the validation split.

Evaluation Metrics. Following the standard evaluation protocol, we report performance results using mean Average Precision **(mAP)** at various temporal Intersection over Union (t-IoU) thresholds. THUMOS14 uses [0.1:0.1:0.7] and ActivityNet1.3 uses [0.5:0.05:0.95]. In addition, we report **ECE@0.7** [11, 29] and **Kendall-τ** [17]. For ECE@0.7, a proposal is labeled positive if its maximum t-IoU to a same-class ground-truth instance is ≥ 0.7. With $B = 15$ equal-width score bins, ECE measures the average gap between accuracy and confidence (*lower is better*). Kendall-τ quantifies the per-video rank correlation between proposal scores and proposal quality (maximum t-IoU) and is averaged across videos (*higher is better*).

Implementation details. We use I3D [3] pre-trained on Kinetics [16] to extract both RGB and optical-flow features. For each modality, the feature dimension is D=1024, and it is reduced to d=512 by a dimensionality-reduction layer. The extractors ϕ^G and ϕ^S each consist of a single fully connected layer followed by a LayerNorm layer. To evaluate generalization, we adopt WTAL methods DELU [6], DDG-Net [37], and AFPS [23], as well as the P-WTAL method

Table 1: Comparison on THUMOS14. * means further trains a localization head using the re-aligned proposals as pseudo labels.

Sup	Method	Pub	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	RefactorNet [40]	CVPR'22	-	-	70.7	65.4	58.6	47.0	32.1	-	54.8	-
	ActionFormer [45]	ECCV'22	-	-	82.1	77.8	71.0	59.4	43.9	-	66.8	-
	TriDet [35]	CVPR'23	87.1	86.2	83.6	80.3	73.0	62.2	46.8	82.0	69.2	74.2
Point	LAC [19]	ICCV'21	75.7	71.4	64.6	56.5	45.3	34.5	21.8	62.7	44.5	52.8
Weak	DELU [6]	ECCV'22	71.5	66.2	56.5	47.7	40.5	27.2	15.3	56.5	37.4	46.4
	P-MIL [32]	CVPR'23	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
	PivoTAL [33]	CVPR'23	74.1	69.6	61.7	52.1	42.8	30.6	16.7	60.1	40.8	49.6
	DDG-Net [37]	ICCV'23	72.5	67.7	58.2	49.0	41.4	27.6	14.8	57.8	38.2	47.3
	ISSF [44]	AAAI'24	72.4	66.9	58.4	49.7	41.8	25.5	12.8	57.8	37.6	46.8
	AFPS [23]	NN'24	73.5	68.8	60.8	51.3	41.0	27.5	16.5	59.1	39.4	48.5
	NoCo [47]	AAAI'25	75.2	70.7	63.0	54.1	44.0	31.7	17.7	61.4	42.1	50.9
	PseudoFormer [26]	CVPR'25	76.8	72.6	65.4	55.6	44.1	31.3	17.1	62.9	42.7	51.9
	MLLM4WTAL [46]	CVPR'25	74.3	69.8	61.8	52.3	43.0	30.8	16.6	60.2	40.9	49.8
	Ours		78.0	73.0	62.5	52.4	42.7	28.2	16.0	61.7	40.4	50.4
	Ours*		**80.1**	**74.6**	**66.1**	**56.7**	**45.6**	**31.2**	**18.4**	**64.6**	**43.6**	**53.2**

Note. Controlled comparisons with recent proposal-level methods under identical proposal sources are provided in **Supplementary Tab. S2**.

LAC [19], as base models. LAC is a point-level weakly supervised method, i.e., it yields higher-quality proposals that are densely distributed around point labels. Prior work [41] reports that, on average, 65.2% of its proposals can be better replaced, which further stress-tests our fine-grained proposal scoring. We train with Adam [18]. For SCLR, the learning rates are 0.00005 for THUMOS14 and 0.000015 for ActivityNet1.3, with 200 training epochs and a batch size of 10. For the localization head, the learning rates are 0.0001 and 0.001, trained for 30 and 16 epochs, respectively. The neighborhood threshold θ is set to 0.7, γ in the soft labels to 0.1, and the balancing coefficients $\lambda_{\perp}$ and λ_{SC} to 0.05 and 10, respectively. Detailed hyperparameter analysis can be found in the **supplementary materials**.

4.2 Comparison with State-of-the-art Methods

In this section, using DDG-Net as the baseline, we report two sets of localization results: **Ours**—applying SCLR alone, and **Ours***—further adding a localization head trained on the re-aligned proposals. This setting gives a first answer to the two questions raised in the Introduction: (1) Can SCLR improve the localization of baseline proposals? (2) Do re-aligned proposal scores provide better supervision for pseudo-label learning?

Results on THUMOS14. As shown in Tab. 1, we compare against both weakly and fully supervised methods. Using SCLR alone improves the baseline by 3.1% for average mAP at (0.1:0.7), validating the framework and indicating that score–quality misalignment had suppressed the baseline's potential.

Moreover, with the localization head (Ours*), our method surpasses all competitors, including recent SOTA pseudo-label learning methods such as PseudoFormer and NoCo, achieving 5.9% over the baseline. This shows that once proposal scores are better aligned with true quality, even a simple pseudo-label construction yields substantial gains—precisely the motivation behind SCLR.

Table 2: Comparison on ActivityNet1.3. * means further trains a localization head using re-aligned proposals as pseudo labels.

Sup	Methods	Pub	mAP@IoU(%)			AVG
			0.5	0.75	0.95	(0.5:0.95)
Full	ActionFormer [45]	ECCV'22	54.7	37.8	8.4	36.6
	TriDet [35]	CVPR'23	54.7	38.0	8.4	36.8
Point	LAC [19]	ICCV'21	40.4	24.6	5.7	25.1
Weak	P-MIL [32]	CVPR'23	41.8	25.4	5.2	25.5
	PivoTAL [33]	CVPR'23	45.1	28.2	5.0	28.1
	DDG-Net [37]	ICCV'23	40.6	26.0	6.0	25.8
	ISSF [44]	AAAI'24	39.4	25.8	6.4	25.8
	AFPS [23]	NN'24	43.9	27.1	6.3	27.3
	PseudoFormer [26]	CVPR'25	46.9	28.6	6.5	29.0
	Ours		43.2	27.4	6.0	27.5
	Ours*		**49.2**	**29.7**	**7.1**	**30.5**

Results on ActivityNet1.3. Tab. 2 reports the comparisons on ActivityNet1.3. Consistent with THUMOS14, applying SCLR alone markedly boosts the baseline localization performance. With the additional localization head, our approach surpasses the current SOTA PseudoFormer by about 1.5% in avg mAP@[0.5:0.95]. These results indicate strong cross-dataset consistency of our framework.

4.3 Ablation Study

Effect of individual losses. Tab. 3 reports the ablation over all loss terms. Starting from $\mathcal{L}_{\text{MIL}}$ (47.0 AVG mAP), adding $\mathcal{L}_{\text{SC}}$ yields a clear gain (48.6), because Eq. 13–14 makes $\mathcal{L}_{\text{SC}}$ *class-conditional*: it enforces proposal–center similarity only for proposals predicted as the current class (when a valid center exists), while assigning the rest a conservative constant target γ, thereby shaping proposal scores even without the explicit decomposition regularizers. More importantly, adding $\mathcal{L}_{\text{G}}$ on top of $\mathcal{L}_{\text{MIL}}$+$\mathcal{L}_{\text{sc}}$ further improves performance to 49.5, whereas adding $\mathcal{L}_{\text{S}}$ alone brings no improvement (48.6, Tab. 3), providing direct evidence that the cross-center contrast in Eq. (7) (enabled by the dual centers) is the key driver. We also observe a non-trivial interaction between the two objectives: using $\mathcal{L}_{\text{G}}$+$\mathcal{L}_{\text{S}}$ without $\mathcal{L}_{\perp}$ slightly underperforms $\mathcal{L}_{\text{G}}$ alone (49.4 vs. 49.5), indicating optimization interference between the generic and specific objectives when $\mathcal{L}_{\perp}$ is absent. Adding $\mathcal{L}_{\perp}$ explicitly decorrelates the two components, mitigates such interference, and achieves the best result.

Table 3: Ablation study of loss functions on THUMOS14.

Exp	$\mathcal{L}_{\text{MIL}}$	$\mathcal{L}_{\text{SC}}$	$\mathcal{L}_{\text{G}}$	$\mathcal{L}_{\text{S}}$	$\mathcal{L}_{\perp}$	mAP@0.5	AVG mAP
1	✓					38.7	47.0
2	✓	✓				40.9	48.6
3	✓	✓	✓			41.6	49.5
4	✓	✓		✓		40.4	48.6
5	✓	✓	✓	✓		42.0	49.4
6	✓	✓	✓	✓	✓	**42.7**	**50.4**

Effect of score realignment on pseudo-label learning. With a localization head, our method attains SOTA accuracy on both datasets. As shown in Tab. 4,

we ablate the head's training pipeline to assess how score realignment affects pseudo-label learning. Compared with training the localization head on baseline proposals, incorporating SCLR yields consistent gains and further boosts the Gaussian fusion strategy, indicating that more reliable proposal scores lead to higher-quality pseudo labels and, in turn, more effective supervision for the localization head.

Table 4: Ablation of pseudo-label learning (localization head).

Method	mAP@IoU(%)			AVG
	0.3	0.5	0.7	(0.1:0.7)
Baseline	58.2	41.4	14.8	47.3
+ SCLR (Ours)	62.5	42.7	16.0	50.4
+ Loc head	62.1	41.0	15.6	49.1
+ Gaussian + Loc head	62.3	42.9	17.6	50.5
+ SCLR + Gaussian + Loc head (Ours*)	**66.1**	**45.6**	**18.4**	**53.2**

Class-center construction. We ablate the specific component $\phi^S(\mathbf{C}^{\mathrm{SE}})$, the balancing coefficient δ, and the EMA on the generic component. Tab. 5 shows that removing any of them degrades performance and the full model is best, validating the need for both cross-video stabilization (EMA) and a balanced video-specific injection.

Anchor selection strategies. The anchor determines the construction of $\mathbf{C}^{\mathrm{SE}}$. Tab. 6 compares Top-$\mathbf{A}$, Top-$\mathbf{S}^{\mathrm{base}}$/Top-$\mathbf{S}^{\mathrm{sup}}$, and Soft-argmax($J$), a continuous variant of our joint scheme J. From the results, the discrete joint anchor argmax(J) achieves the best performance. This advantage arises because Top-$*$ heuristics are sensitive to scale variations and noise, whereas Soft-argmax(J) smooths the score landscape and weakens local maxima. In contrast, selecting argmax(J) preserves the sharp joint peak induced by $\mathbf{A}$ and $\mathbf{S}^{\mathrm{base}}$, thereby improving separability.

Table 5: Ablation of components in the class center reconstruction.

Method	mAP@IoU(%)			AVG
	0.3	0.5	0.7	(0.1:0.7)
w/o $\phi^S(\mathbf{C}^{\mathrm{SE}})$	60.8	40.6	13.3	48.7
w/o EMA($\cdot$)	61.9	42.2	14.4	49.7
w/o δ	62.3	41.6	14.5	49.7
Ours	**62.5**	**42.7**	**16.0**	**50.4**

Table 6: Comparison of different anchor-selection strategies.

Method	mAP@IoU(%)			AVG
	0.3	0.5	0.7	(0.1:0.7)
Top-$\mathbf{A}$	60.1	40.0	13.0	47.9
Top-$\mathbf{S}^{\mathrm{base}}$	61.1	40.5	13.8	48.7
Top-$\mathbf{S}^{\mathrm{sup}}$	60.9	40.1	13.2	48.3
Soft-argmax(J)	61.8	40.9	13.5	49.0
argmax(J) (Ours)	**62.5**	**42.7**	**16.0**	**50.4**

Weighting design of $\mathbf{C}^{\mathrm{SE}}$ and $\mathbf{C}^{\mathrm{TE}}$. Both aggregation weights incorporate class evidence $\mathbf{S}^{\mathrm{base}}$ and class-agnostic actionness $\mathbf{A}$, but we treat class evidence

differently due to the local–global aggregation scopes. For the local center $\mathbf{C}^{\mathrm{SE}}$, we use $\mathbf{S}^{\mathrm{sup}}{=}\mathbf{S}^{\mathrm{base}}\cdot\mathbf{A}$ to emphasize high-confidence neighbors (sharper anchor) and add k^{cen} to encode *shift+scale* compatibility beyond IoU. For the global center $\mathbf{C}^{\mathrm{TE}}$, we replace raw score magnitudes with rank-based $Q_{\mathcal{I}}[\mathbf{S}^{\mathrm{base}}]$ to suppress outliers and use k^{time} to avoid over-peaked aggregation around the anchor by collecting contextual proposals away from it. Tab. 7 validates these choices: removing any cue or replacing the center-specific designs (e.g., $\mathbf{S}^{\mathrm{sup}}\rightarrow\mathbf{S}^{\mathrm{base}}$ in w^{SE}, $Q_{\mathcal{I}}[\mathbf{S}^{\mathrm{base}}]\rightarrow\mathbf{S}^{\mathrm{base}}$ in w^{TE}) consistently degrades performance.

Table 7: Ablation of the fusion weights for the dual centers.

Method	mAP@IoU(%)			AVG
	0.3	0.5	0.7	(0.1:0.7)
w/o k^{cen} in w^{SE}	61.4	41.6	13.8	49.2
w/o k^{time} in w^{TE}	61.9	41.9	13.9	49.5
w/o k^{cen} and k^{time}	61.5	41.7	13.8	49.2
w/o IoU in w^{SE}	61.6	42.0	14.2	49.4
w/o $\mathbf{A}$ in w^{TE}	61.2	41.0	13.7	49.0
Replace $\mathbf{S}^{\mathrm{sup}}$ with $\mathbf{S}^{\mathrm{base}}$ in w^{SE}	61.4	40.9	14.1	49.0
Replace $\mathrm{Q}_{\mathcal{I}}[\mathbf{S}^{\mathrm{base}}]$ with $\mathbf{S}^{\mathrm{base}}$ in w^{TE}	61.7	41.8	14.4	49.5

Generalization Experiments of Our Framework. A central question of this paper is whether SCLR, as a plug-and-play module, can mitigate score misalignment and thereby improve localization across WTAL baselines. We integrate SCLR into four representative backbones and observe consistent localization gains, as shown in Tab. 8. More importantly, SCLR delivers substantial score alignment: ECE@0.7(%) is sharply reduced across all models, and the average Kendall-τ (%) between proposal scores and proposal quality rises by about 7.2 points. These alignment gains directly translate into stronger localization and more effective pseudo-label training with the localization head.

Table 8: Generalization experiments of SCLR on THUMOS14.

Sup	Methods	mAP@IoU(%)			AVG	ECE@0.7	Kendall-τ
		0.3	0.5	0.7	(0.1:0.7)		
Weak	DELU	56.5	40.5	15.3	46.4	63.9	28.3
	+Ours	**60.3**	**41.2**	**15.6**	**49.2**	**5.5**	**38.5**
	DDG-Net	58.2	41.4	14.8	47.3	66.4	27.1
	+Ours	**62.5**	**42.7**	**16.0**	**50.4**	**6.2**	**38.5**
	AFPS	60.8	41.0	16.5	48.5	29.6	43.5
	+Ours	**64.8**	**45.7**	**17.3**	**52.2**	**4.8**	**44.7**
Point	LAC	64.6	45.3	21.8	52.8	37.8	41.0
	+Ours	**68.9**	**51.0**	**21.9**	**56.4**	**28.8**	**47.2**

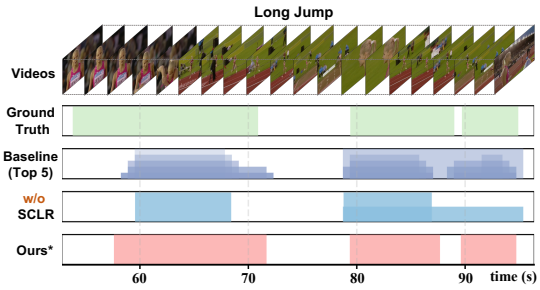

Fig. 4: Visual comparison. For the baseline, we visualize the top-5 proposals with the highest overlap with the ground truth. Compared with pseudo labels generated without SCLR, incorporating SCLR yields more reliable pseudo labels that better cover complete action extents.

Qualitative comparison. As shown in Fig. 4, we compare pseudo labels generated with and without SCLR. Score realignment consistently improves the reliability of pseudo-label construction and yields more accurate pseudo labels. More qualitative visualizations, including failure cases, are provided in the **Supplementary Material**.

5 Conclusion

In this work, we analyze and quantify score–quality misalignment in WTAL and present SCLR, a plug-and-play proposal-level framework. Guided by SCL, SCLR aligns proposal confidence with proposal quality from a semantic perspective. Extensive experiments show that SCLR consistently reduces misalignment and improves localization across diverse baselines. By explicitly improving proposal-score priors before pseudo-label construction, SCLR strengthens the instance-level pseudo-labeling paradigm and enables a localization head that reaches state-of-the-art accuracy. We hope this provides a simple and general recipe for future WTAL research.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62572362) and the Shenzhen Science and Technology Program (Grant No. CJGJZD20230724091659002).

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)

2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1130–1139 (2018)
5. Chen, G., Zhang, C., Zou, Y.: Afnet: Temporal locality-aware network with dual structure for accurate and fast action detection. IEEE Transactions on Multimedia **23**, 2672–2682 (2020)
6. Chen, M., Gao, J., Yang, S., Xu, C.: Dual-evidential learning for weakly-supervised temporal action localization. In: European conference on computer vision. pp. 192–208. Springer (2022)
7. Cheng, Y., Sun, Y., Fan, H., Zhuo, T., Lim, J.H., Kankanhalli, M.: Entropy guided attention network for weakly-supervised action localization. Pattern Recognition **129**, 108718 (2022)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, 2nd edn. (2000)
9. Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R.: Turn tap: Temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE international conference on computer vision. pp. 3628–3636 (2017)
10. Gao, J., Chen, M., Xu, C.: Vectorized evidential learning for weakly-supervised temporal action localization. IEEE transactions on pattern analysis and machine intelligence **45**(12), 15949–15963 (2023)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
12. Huang, L., Huang, Y., Ouyang, W., Wang, L.: Modeling sub-actions for weakly supervised temporal action localization. IEEE Transactions on Image Processing **30**, 5154–5167 (2021)
13. Huang, L., Wang, L., Li, H.: Multi-modality self-distillation for weakly supervised temporal action localization. IEEE Transactions on Image Processing **31**, 1504–1519 (2022)
14. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017)
15. Islam, A., Long, C., Radke, R.: A hybrid attention mechanism for weakly-supervised temporal action localization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1637–1645 (2021)
16. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
17. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (1938)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

19. Lee, P., Byun, H.: Learning action completeness from points for weakly-supervised temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13648–13657 (2021)
20. Li, J., Liu, X., Zong, Z., Zhao, W., Zhang, M., Song, J.: Graph attention based proposal 3d convnets for action detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4626–4633 (2020)
21. Li, M., Zheng, C., Wang, J., Li, B.: Similar modality enhancement and action consistency learning for weakly supervised temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4842–4850 (2025)
22. Li, X., Lin, T., Liu, X., Zuo, W., Li, C., Long, X., He, D., Li, F., Wen, S., Gan, C.: Deep concept-wise temporal convolutional networks for action localization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4004–4012 (2020)
23. Li, Z., Wang, Z., Liu, Q.: Weakly supervised temporal action localization with actionness-guided false positive suppression. Neural Networks **175**, 106307 (2024)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
25. Liu, Y., Ma, L., Zhang, Y., Liu, W., Chang, S.F.: Multi-granularity generator for temporal action proposal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3604–3613 (2019)
26. Liu, Z., Liu, Y.: Bridge the gap: From weak to full supervision for temporal action localization with pseudoformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8711–8720 (2025)
27. Liu, Z., Wang, L., Zhang, Q., Tang, W., Yuan, J., Zheng, N., Hua, G.: Acsnet: Action-context separation network for weakly supervised temporal action localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2233–2241 (2021)
28. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 344–353 (2019)
29. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29, pp. 2901–2907 (2015)
30. Nguyen, P.X., Ramanan, D., Fowlkes, C.C.: Weakly-supervised action localization with background modeling. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5502–5511 (2019)
31. Qing, Z., Su, H., Gan, W., Wang, D., Wu, W., Wang, X., Qiao, Y., Yan, J., Gao, C., Sang, N.: Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 485–494 (2021)
32. Ren, H., Yang, W., Zhang, T., Zhang, Y.: Proposal-based multiple instance learning for weakly-supervised temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2394–2404 (2023)
33. Rizve, M.N., Mittal, G., Yu, Y., Hall, M., Sajeev, S., Shah, M., Chen, M.: Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22992–23002 (2023)

34. Shi, B., Dai, Q., Mu, Y., Wang, J.: Weakly-supervised action localization by generative attention modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1009–1019 (2020)
35. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18857–18866 (2023)
36. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE international conference on computer vision. pp. 3524–3533 (2017)
37. Tang, X., Fan, J., Luo, C., Zhang, Z., Zhang, M., Yang, Z.: Ddg-net: Discriminability-driven graph network for weakly-supervised temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6622–6632 (2023)
38. Wang, B., Zhao, Y., Yang, L., Long, T., Li, X.: Temporal action localization in the deep learning era: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2171–2190 (2023)
39. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
40. Xia, K., Wang, L., Zhou, S., Zheng, N., Tang, W.: 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. pp. 13884–13893 (2022)
41. Xia, Z., Cheng, J., Liu, S., Hu, Y., Wang, S., Zhang, Y., Dang, L.: 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. pp. 18440–18450 (2024)
42. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision. pp. 5783–5792 (2017)
43. Yang, H., Wu, W., Wang, L., Jin, S., Xia, B., Yao, H., Huang, H.: Temporal action proposal generation with background constraint. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3054–3062 (2022)
44. Yun, W., Qi, M., Wang, C., Ma, H.: Weakly-supervised temporal action localization by inferring salient snippet-feature. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6908–6916 (2024)
45. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)
46. Zhang, Q., Fang, J., Yuan, R., Tang, X., Qi, Y., Zhang, K., Yuan, C.: Weakly supervised temporal action localization via dual-prior collaborative learning guided by multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24139–24148 (2025)
47. Zhang, Q., Qi, Y., Tang, X., Yuan, R., Lin, X., Zhang, K., Yuan, C.: Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10085–10093 (2025)
48. Zhang, X.Y., Shi, H., Li, C., Li, P., Li, Z., Ren, P.: Weakly-supervised action localization via embedding-modeling iterative optimization. Pattern Recognition **113**, 107831 (2021)

49. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12993–13000 (2020)
50. Zhou, J., Huang, L., Wang, L., Liu, S., Li, H.: 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. pp. 23003–23012 (2023)