Pattern Recognition 162 (2025) 111421

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Beyond boundaries: Hierarchical-contrast unsupervised temporal action localization with high-coupling feature learning

Yuanyuan Liu [a], Ning Zhou [a], Yuxuan Huang [a], Shuyang Liu [a], Leyuan Liu [b], Wujie Zhou [c], Chang Tang [a], Ke Wang [a,*]

[a] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, 430074, China
[b] School of National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, 430079, China
[c] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, 310018, China

ARTICLE INFO

Keywords:
Unsupervised temporal action localization
Coarse-to-fine
Video-level CL
Instance-level CL
Boundary-level CL

ABSTRACT

Current unsupervised temporal action localization (UTAL) methods mainly use clustering and localization with independent learning mechanisms. However, these individual mechanisms are low-coupled and struggle to finely localize action-background boundary information due to the lack of feature interactions in the clustering and localization process. To address this, we propose an end-to-end Hierarchical-Contrast UTAL (HC-UTAL) framework with high-coupling multi-task feature learning. HC-UTAL incorporates coarse-to-fine contrastive learning (CL) at three levels: *video level*, *instance level* and *boundary level*, thus obtaining adaptive interaction and robust performance. We first employ the *video-level CL* on video-level and cluster-level feature learning, generating video action pseudo-labels. Then, using the video action pseudo-labels, we further devise the *instance-level CL* on action-related feature learning for coarse localization and the *boundary-level CL* on ambiguous action-background boundary feature learning for finer localization, respectively. We conduct extensive experiments on THUMOS'14, ActivityNet v1.2, and ActivityNet v1.3 datasets. The results demonstrate that our method achieves state-of-the-art performance. The code and trained models are available at: https://github.com/bugcat9/HC-UTAL.

1. Introduction

Temporal action localization (TAL) in videos is a challenging task in multimedia analysis and mining [1,2]. It involves identifying and localizing action instances within untrimmed video data [3,4]. TAL has numerous applications in real-life multimedia intelligence analysis and management, such as video summarization, video highlight detection, video content retrieval, multimedia content understanding, etc. [4,5]. Recently, deep convolutional neural networks (DCNNs) [6] have achieved promising TAL results, but they require a large amount of annotated data for model training. Annotating TAL data is known to be a labor-intensive and time-consuming process. For example, it can take an experienced annotator approximately 3–5 min to annotate action instances in a 1-minute video [7]. Hence, enabling DCNN models to learn from unlabeled action videos, which are easier to collect, has become important for multimedia intelligent applications, such as resource-efficient video annotation [5], surveillance and security [4], etc.

Recently, unsupervised TAL (UTAL) on unlabeled action videos has garnered increased interest from researchers [8,9]. To achieve UTAL performance, existing methods typically employ a two-stage learning framework involving clustering and localization. Gong et al. [8] introduced the unsupervised Temporal Co-Attention Model (TCAM) that first incorporates a clustering algorithm to divide videos into different groups and then employ an attention mechanism to locate action instances. Liu et al. [9] proposed Action-positive Separation Learning (APSL) for UTAL. APSL introduced an CL loss for action feature separation learning, obtaining the salient action features to improve the localization performance. In addition, with the rise of large vision-language models or large vision models recently, the advantages of large models could be enhanced with an UTAL method. Despite achieving the progress, the UTAL task can be extremely challenging due to two primary obstacles below: (1) *Low-Coupling learning*. As shown in Fig. 1(a) , clustering and localization employ distinct learning mechanisms with minimal interaction, presenting a challenge for effectively mining semantic features related to video actions. (2) *Ambiguous action-background boundaries*. Due to the lack of effective supervision, UTAL methods often struggle to distinguish action boundaries from the

* Corresponding author.
E-mail address: wk2023@cug.edu.cn (K. Wang).

https://doi.org/10.1016/j.patcog.2025.111421
Received 6 July 2024; Received in revised form 30 November 2024; Accepted 26 January 2025
Available online 3 February 2025

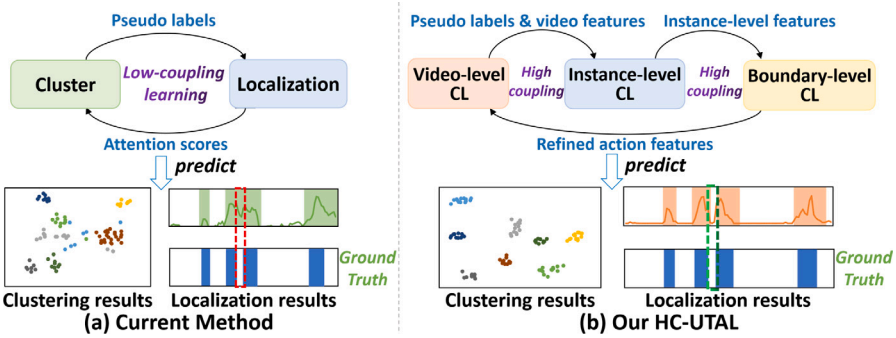

Fig. 1. The motivation of our method. (a) and (b) show the frameworks, action clustering and localization results for the current method and our method, respectively. Unlike the current low-coupling clustering and localization learning process in the two-stage approach, which does not consider feature interactions, our HC-UTAL is a unified CL learning process, which enables pseudo-labels and action-related features to be learnt interactively in a high-coupling manner at each CL level, resulting in more accurate UTAL results. It is worth noting that both action clustering and localization in (b) are more accurate than (a). For instance, the red dashed box in (a) indicates the incorrect localization result of TCAM, while the green dashed box in (b) indicates the correctly localized boundary result. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

background, as the minimal differences between them make precise localization challenging.

To tackle the above-mentioned limitations, leveraging the powerful self-supervised and unsupervised learning capabilities of contrastive learning (CL) [10,11], we introduce a novel three-layer coarse-to-fine contrastive learning (CL) framework for the first time. This framework, termed Hierarchical-Contrast UTAL (HC-UTAL), employs high-coupling feature interaction learning at the video, instance, and boundary levels to achieve effective UTAL performance. In this work, the high-coupling mechanism in HC-UTAL strengthens the connection between clustering and localization through hierarchical contrastive learning at the video, instance, and boundary levels. This approach facilitates both label transmission and feature sharing, promoting adaptive feature interaction and enabling more precise delineation of action-background boundaries. Fig. 1 presents an intuitive motivation and the results of HC-UTAL. Unlike existing two-stage UTAL methods that typically involve two distinct, loosely coupled learning stages for obtaining action pseudo-labels and semantic features, our HC-UTAL framework simultaneously and efficiently facilitates interactive learning between these processes through an elaborated high-coupling approach. Additionally, the integrated three-layer CL mechanism collaborates adaptively to extract multi-level semantic features from coarse to fine, effectively discovering ambiguous action-background boundary information and enhancing the robustness of UTAL performance.

More specifically, in video-level CL, we first employ bi-directional CL mechanisms: horizontal CL learns video-level features, while vertical CL clusters video groups to generate video action pseudo-labels. Then, with the generated pseudo-labels and video-level features, instance-level CL introduce a contrast loss to separate action-related and unrelated features, ensuring the former are tightly clustered while the latter are well-separated, enabling coarse action clip localization. While the video-level and instance-level CL enhance UTAL performance, ambiguous action-background boundaries (discussed in Fig. 4) can impact localization accuracy. To address this, we propose a boundary-level CL component, which extracts finer-grained boundary information. To ensure seamless collaboration across these learning processes, we implement an adaptive multi-task weighting mechanism for end-to-end interaction among video-level, instance-level, and boundary-level CLs, improving both robustness and overall UTAL performance.

In general, the major contributions of this paper are as follows:

(1) We introduce HC-UTAL, a novel coarse-to-fine contrastive learning framework that captures fine-grained multi-level semantic features through highly coupled multi-task learning. As the first hierarchical CL framework in UTAL, HC-UTAL achieves state-of-the-art performance in weakly supervised and unsupervised temporal action localization on THUMOS'14, ActivityNet v1.2, and ActivityNet v1.3 datasets.

(2) In video-level CL, we introduce bi-directional mechanisms: horizontal CL for video-level action learning and vertical CL for cluster-level feature learning. Together, these mechanisms generate action pseudo-labels for each untrimmed video.

(3) In instance-level CL, we use contrast loss to align action-related features with the same pseudo-labels while separating unrelated features, enhancing coarse action localization. At the boundary level, we introduce joint learning objectives for non-boundary and boundary feature learning to improve fine action and background boundary localization.

(4) We introduce an adaptive multi-task weighting mechanism that integrates video, instance, and boundary-level CL. This innovative approach enables strong interactions and mutual reinforcement, enhancing the precision of action category identification and proposal generation.

2. Related work

2.1. Temporal action localization

TAL aims to identify the temporal intervals of specific actions in untrimmed videos. Most existing works fall into two categories: fully and weakly supervised methods.

Fully supervised TAL methods mainly use frame-level annotations to generate and classify temporal action proposals [5]. For example, GTAN [12] uses Gaussian kernels to optimize the generation of proposals. Recently, graph neural networks have also been utilized to explore the relationships between action proposals with frame-level supervision.

Weakly supervised TAL methods require video-level annotations to locate the action instances in videos. For instance, UntrimmedNets addresses it with Multi-Instance Learning (MIL) approach. W-TALC [3] obtains action proposals by using metric learning to make similar action features closer to each other. FC-CRF [13] finds new foreground clips progressively via step-by-step erasion from a complete input video. HAM-Net [14] proposes mixed attention weights to localize complete action instances through multiple parallel and complementary branch learning. CoLA [15] utilizes snippet contrastive learning to improve localization results. FTCL [16] utilizes two complementary modules: Fine-grained Sequence Distance (FSD) contrasting and Longest Common Subsequence (LCS) contrasting. FSD examines the relationships between action and background proposals, while LCS identifies the longest common subsequences in videos. These modules synergistically enhance each other, improving action-background separation and addressing the classification-localization task gap. LSBF [17] introduces the first anchor-free TAL method, by employing novel boundary pooling to enhance proposal features, and several consistency constraints to

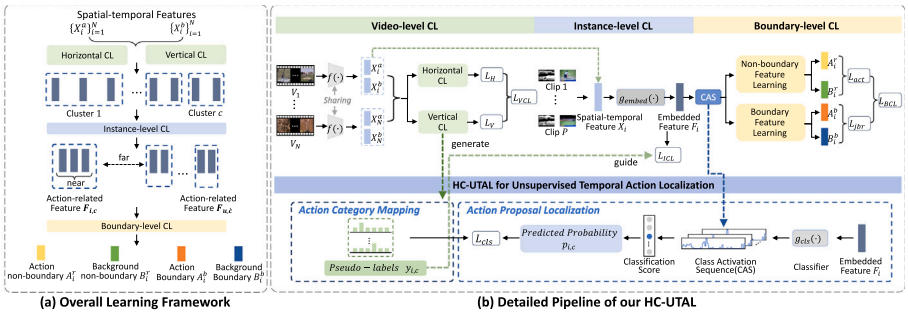

Fig. 2. The training pipeline of the HC-UTAL for unsupervised temporal action localization. Using the untrimmed video input, we apply three hierarchical CL, namely video-level CL, instance-level CL, and boundary-level CL, for obtaining the action categories and action localization proposals in a coarse-to-fine multi-task learning manner. Furthermore, these three hierarchical CL can be learned iteratively in an end-to-end and mutually reinforcing manner.

ensure accurate boundary detection for arbitrary proposals.

Compared with existing methods, HC-UTAL adopts a completely different learning mechanism to separate action and background boundary features through a hierarchical contrastive learning method without any annotation information. At the video level, traditional methods usually rely on the relationship between different clips in the same video to improve the positioning accuracy, ignoring the extraction of cluster-level features of different videos. This limitation hinders the model's ability to capture common action information between different videos. In addition, at the instance level, most past methods use attention mechanisms to distinguish foreground and background actions, thereby improving the accuracy of action localization. However, these methods often ignore the distinction between action-related features and irrelevant features, which may cause action-related features to be mixed with background information, making it difficult to accurately locate action instances. Overall, weakly supervised TAL relies heavily on video-level action labels. Annotating action categories for large amounts of untrimmed video is challenging and costly [7].

2.2. Unsupervised temporal action localization

An increasing number of recent research efforts focus on the task of UTAL as it does not rely on any video annotations [8]. For instance, TCAM [8] proposes the first unsupervised action localization method with a two-step "clustering + localization" iterative procedure. However, TCAM introduces noise from raw videos while using the overall video features for two individual clustering and localization procedures. To address this issue, APSL [9] introduces an iterative process of feature separation, clustering, and localization for the UTAL task. APSL improves the TCAM method by utilizing separation learning to extract important action features for clustering and localization.

Although the APSL is our previous work, we would like to clarify that there are significant differences between the HC-UTAL and APSL. Firstly, they do not share the same motivation. The motivation of APSL is to improve clustering and localization by separating salient action features through feature separation. while the motivation of HC-UTAL is to address the issues of low-coupling learning and boundary problems in UTAL. Secondly, the learning mechanisms of HC-UTAL and APSL are significant different. HC-UTAL employs a multi-level CL for video-level, instance-level and boundary-level feature learning in both action clustering and localization, while APSL uses two separated learning stages with the spectral clustering algorithm and contrastive learning for action localization. Thirdly, both APSL and TCAM require knowledge of the number of clusters during training, whereas our HC-UTAL treats it as a hyperparameter with enhanced adaptive learning capability.

2.3. Contrastive learning

Another recent advancement in vision and multimedia tasks is to use contrastive learning (CL) in self-supervised and unsupervised manners [10,11]. CL aims to map features of samples onto a unit hypersphere such that the feature distances of the positive sample pairs on the sphere are similar while the feature distances of the negative sample pairs are pushed apart. Popular CL-based methods, such as SimCLR [10] and MoCo [11], often use InfoNCE loss to learn a latent representation that is beneficial to downstream tasks. SimCLR [10] proposes a negative sample selection scheme by using the augmented views of other items in a minibatch during training. MoCo [11] uses a momentum-updated memory bank of old negative representations to remove the batch size restriction and enable the consistent use of negative samples. Furthermore, several works propose to deploy CL for video understanding tasks [18]. For example, Pace [19] uses video clips of the same action instance but with different visual rhythms to construct positive sample pairs for CL. SeCo [20] uses different frames from the same video to construct positive samples. Lastly, IIC [21] and CVRL [18] use different clips from the same video as positive samples. In summary, most approaches to CL-based video-related tasks implement CL by constructing sample pairs from different segments of the same video.

3. Methodology

The overall architecture of HC-UTAL for unsupervised temporal action localization (UTAL) is illustrated in Fig. 2, which consists of three hierarchical contrastive learning (CL) mechanisms: *the video-level CL*, *instance-level CL*, and *boundary-level CL*. Unlike current UTAL methods that mainly employ different learning schemes for action category clustering and action proposal localization, individually, our HC-UTAL is an end-to-end multi-level CL mechanism in a high-coupling multi-task manner, thereby enhancing multi-level semantic representations from coarse to fine, for improved performance in UTAL. In the following sections, we first provide details on each learning stage of the HC-UTAL approach, then describe how HC-UTAL is utilized for UTAL. Finally, we present the inference pipeline for obtaining action position proposals in untrimmed videos.

3.1. HC-UTAL pipeline

3.1.1. Overview

To better describe our method, we first formulate the HC-UTAL pipeline. In HC-UTAL, we first employ the *video-level CL* to generate video pseudo-labels for each unlabeled video by introducing horizontal and vertical CL for video-level and cluster-level feature learning, respectively. After obtaining the video pseudo-labels, we employ *the instance-level CL* to extract action-related features (same pseudo-label)

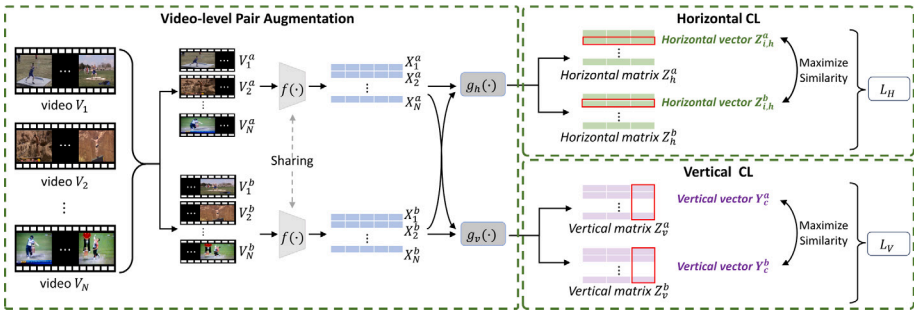

Fig. 3. The implementation pipeline of the video-level CL. Through properly video-level pair augmentation, the video-level CL employs two parallel CL mechanisms, namely vertical CL and horizontal CL, to generate video pseudo-labels.

and action-unrelated features (different pseudo-label) from the video-level features. This learning process encourages action-related features to be brought closer together while also pushing action-unrelated features apart, thus achieving coarse action localization. Furthermore, to address the challenge of capturing action-background boundary features in videos, we further devise *the boundary-level CL* on finer boundary semantic representation learning, thereby augmenting the performance of action proposal localization.

Formally, taken unlabeled videos as input, our HC-UTAL employs three-level distinct CL losses, *i.e.*, video-level CL loss denoted as L_{VCL}, instance-level CL loss denoted as L_{ICL}, and boundary-level CL loss denoted as L_{BCL}, to learn fine-grained multi-level semantic features from its input, respectively. Therefore, our overall pipeline of HC-UTAL can be described by:

$$L_{HC-UTAL} = \alpha L_{VCL} + \beta L_{ICL} + \gamma L_{BCL} \tag{1}$$

where α, β and γ are three dynamic weights to adaptively balance the coarse-to-fine learning objectives in the multi-task learning manner according to their contributions. To facilitate high-coupling interactions and mutual reinforcement among these three-level learning schemes, we employ the Dynamic Weight Average (DWA) to obtain these three parameters during training. We also show in the experiments that adding the dynamic weight learning improves performance (see Table 6), demonstrating the usefulness of adaptive cooperation of three-hierarchy CL schemes. By adaptively optimizing these three losses, HC-UTAL extracts more discriminative multi-level semantic features for robust UTAL performance in unlabeled videos. The details of each level CL in HC-UTAL is described below.

3.1.2. Video-level CL

To generate video pseudo-labels for guiding action proposal localization, the video-level CL aims to learn both video-level features and cluster-level features, where video-level features contain per video-specific action information and the cluster-level features encompass common action information. To this end, we devise two directional CL mechanisms, namely horizontal CL and vertical CL. The horizontal CL is utilized to learn video-level features, while the vertical CL is designed to learn common cluster-level features that can be used for the generation of action pseudo-labels. As the pipeline of the video-level CL shown in Fig. 3, we first construct video-level sample pairs and then execute the horizontal CL and vertical CL in parallel.

Video-level Pair Augmentation. Inspired by [22], we employ a video-level sample pair augmentation method, which consists of horizontal pair augmentation and vertical pair augmentation, as depicted in Fig. 3.

We denote an unlabeled video set in a training batch as $\mathcal{V} = \{V_i\}_{i=1}^N$, where i indexes the videos in $\mathcal{V}$ and N is the number of videos in $\mathcal{V}$. Each video V_i consists of RGB frames and the corresponding optic flow, and is divided into P non-overlapping small clips, represented as $V_i = \{v_{i,t}\}_{t=1}^P$, where t is the clip index in a video. To construct video-level data pairs, we split each video V_i in half into two parts, namely $V_i^a = \{v_{i,1}, ..., v_{i,P/2}\}$ and $V_i^b = \{v_{i,P/2+1}, \dots, v_{i,P}\}$. For each half part of a video, we utilize a shared pre-trained I3D (Inflated 3D) network as the backbone to extract its spatial–temporal features. The backbone network, denoted as $f(\cdot)$, can be any suitable architecture such as UntrimmedNet or I3D. It is important to note that the backbone $f(\cdot)$ is independent of the specific architecture and the discussions on additional choices for the backbone $f(\cdot)$ can be found in Section 1.2 of the supplementary material. As a result, the spatial–temporal features can be represented as $X_i^a = f(V_i^a)$ and $X_i^b = f(V_i^b)$, respectively, with a shape of $2d \times \frac{P}{2}$, where d represents the feature size of each clip.

With the spatial–temporal features X_i^a and X_i^b, we further employ two separate branches: $g_h(\cdot)$ and $g_v(\cdot)$. The $g_h(\cdot)$ is used to obtain the augmented horizontal feature pair, represented as $Z_{i,h}^a \in \mathbb{R}^{1\times C}$ and $Z_{i,h}^b \in \mathbb{R}^{1\times C}$. The $g_v(\cdot)$ extracts the augmented vertical feature pair as, $Z_{i,v}^a \in \mathbb{R}^{1\times C}$ and $Z_{i,v}^b \in \mathbb{R}^{1\times C}$. Here, $g_h(\cdot)$ and $g_v(\cdot)$ are two-layer nonlinear Multilayer Perceptron (MLP), and C is the number of clusters.

Horizontal CL. Building upon the horizontal pair augmentation, we can acquire two horizontal feature matrices in the training batch, represented as $\mathcal{Z}_h^a \in \mathbb{R}^{N\times C} = \{Z_{i,h}^a\}_{i=1}^N$ and $\mathcal{Z}_h^b \in \mathbb{R}^{N\times C} = \{Z_{i,h}^b\}_{i=1}^N$, where N is the video amount in the batch. Each row vector, *e.g.*, $Z_{i,h}^a$ or $Z_{i,h}^b$ in the two matrices, describes the video-level features for the video i. We consider horizontal feature vectors from the same video as positive pairs and horizontal feature vectors from various videos as negative pairs. The horizontal CL aims to maximize the similarities of positive pairs, while minimizing the similarities of negative pairs.

For a specific video i, the positive pair comes from the same video, represented as $Z_{i,h}^a$ and $Z_{i,h}^b$, and the remaining $2N-2$ horizontal feature vectors in the two horizontal feature matrices are negative pairs. Therefore, we define the horizontal CL loss as:

$$L_h = -\frac{1}{2N}\sum_{i=1}^{N}\Big(\frac{\exp(d(Z_{i,h}^a, Z_{i,h}^b)/\tau_h)}{\sum_{j=1,j\neq i}^{N}[\exp(d(Z_{i,h}^a, Z_{j,h}^a)/\tau_h) + \exp(d(Z_{i,h}^a, Z_{j,h}^b)/\tau_h)]} + \frac{\exp(d(Z_{i,h}^b, Z_{i,h}^a)/\tau_h)}{\sum_{j=1,j\neq i}^{N}[\exp(d(Z_{i,h}^b, Z_{j,h}^b)/\tau_h) + \exp(d(Z_{i,h}^b, Z_{j,h}^a)/\tau_h)]}\Big) \tag{2}$$

where τ_h is the horizontal temperature hyperparameter, and $d(\cdot)$ is cosine similarity. The loss encourages positive pairs to have higher similarities and negative pairs to have lower similarities, thus obtaining rich video-level information.

Vertical CL. In the vertical CL, the objective is to cluster videos into different action clusters for learning common cluster-level features. The vertical CL achieves this by maximizing the cluster similarity of positive pairs and minimizing the cluster similarity of negative pairs.

More specifically, using the $Z_{i,v}^a$ and $Z_{i,v}^b$ in a training batch, we first integrate them to form two vertical feature matrices, denoted as $\mathcal{Z}_v^a \in \mathbb{R}^{N\times C} = \{Z_{i,v}^a\}_{i=1}^N$ and $\mathcal{Z}_v^b \in \mathbb{R}^{N\times C} = \{Z_{i,v}^b\}_{i=1}^N$. In each feature matrix, the number of column corresponds to the predefined cluster count C. In other words, each column feature vector represents action cluster information that can be used to generate the pseudo-labels of action categories for each video. Based on this, we sample

the vertical feature vectors $Y_c^a \in \mathcal{Z}_v^a$ and $Y_c^b \in \mathcal{Z}_v^b$ with the same column c (cluster) as vertical positive pairs (see Fig. 3). Consequently, the remaining $2C-2$ column vectors in the training batch serve as vertical negative pairs. Both the positive and negative pairs are passed through the vertical CL to learn the cluster-level representations with different action categories. Overall, the vertical CL loss is written as:

$$L_v = -\frac{1}{2C}\sum_{c=1}^{C}\Big(\log\frac{\exp(d(Y_c^a,Y_c^b)/\tau_v)}{\sum_{j=1,j\neq c}^{C}[\exp(d(Y_c^a,Y_j^a)/\tau_v)+\exp(d(Y_c^a,Y_j^b)/\tau_v)]} + log\frac{\exp(d(Y_c^b,Y_c^a)/\tau_v)}{\sum_{j=1,j\neq c}^{C}[\exp(d(Y_c^b,Y_j^b)/\tau_v)+\exp(d(Y_c^b,Y_j^a)/\tau_v)]} - M(Y_c^a)-M(Y_c^b)\Big) \tag{3}$$

where τ_v is a temperature hyperparameter. Following [22], we employ $M(\cdot)$, which is the entropy of cluster assignment probabilities, to prevent the network from assigning all videos to a cluster.

Finally, the total objective L_{VCL} of video-level CL includes both the horizontal CL loss and vertical CL loss, as:

$$L_{VCL} = L_h + L_v. \tag{4}$$

By optimizing this objective, the video-level CL learns both video-level and cluster-level features, aiding in the generation of the pseudo-labels for videos.

3.1.3. Instance-level CL

After obtaining the video-level features and their corresponding pseudo-labels through the video-level CL, we employ the instance-level CL to discriminate between video-level features sharing the same pseudo-label, denoted as action-related features, and those with different action pseudo-labels, denoted as action-unrelated features. Through proper training with instance-level CL, the action-related features are encouraged to converge while simultaneously pushing away the action-unrelated features, aiding in achieving coarse action localization.

Formally, given the ith video's spatial–temporal features $X_i = \{X_i^a, X_i^b\}$ as input, we first use a temporal convolutional operator $g_{embed}(\cdot)$ with a ReLU activation function to embed X_i into the task-specific feature space. This is represented as $F_i = g_{embed}(X_i)$, where $F_i \in \mathbb{R}^{2d\times P}$ is the task-specific embedded features, and d and P represent the feature size of each clip and the number of the clips in a video, respectively. Using the F_i, the instance-level CL would extract the action-related feature and the action-unrelated features more properly. Next, to perform the instance-level contrastive training, we consider the action-related features (such as $F_{i,c}$ and $F_{j,c}$) with the same pseudo-label/cluster c as the positive pair, while the rest features $F_{u,c'}$ with different pseudo-labels/clusters c' as the negative pairs. Thereby, the instance-level CL loss L_{ICL} can be given by,

$$L_{ICL} = -\log\frac{\exp({F_{i,c}}^T \cdot F_{j,c}/\tau_{in})}{\exp({F_{i,c}}^T \cdot F_{j,c}/\tau_{in}) + \sum_u \exp({F_{i,c}}^T \cdot F_{u,c'}/\tau_{in})} \tag{5}$$

where τ_{in} is the temperature hyperparameter for the instance-level CL. By optimizing the instance-level CL loss, the instance-level CL encourages similar action features to have high similarity scores while reducing the similarity scores between action-related and action-unrelated features. While the process helps to achieve coarse action localization, it is still difficult to obtain fine-grained localization of action proposals, *e.g.*, a specific action category of a clip in a video.

3.1.4. Boundary-level CL

Since action and background boundary features in videos often contain very similar information, current CL schemes are difficult to capture and localize such ambiguous boundary changes efficiently. Therefore, to achieve finer action-background boundary localization, the HC-UTAL further introduces the boundary-level CL which focuses on refined action-background boundary feature learning within per video. To address this, we first define two distinct boundaries in a video: the action boundary and the background boundary. The former denotes the boundary feature that transitions from the action to the background, and the latter represents the feature that transitions from the background to the action. Then, we devise the boundary-level CL with two joint learning objectives, namely non-boundary feature learning and boundary feature learning, for effectively identifying these two types of features, leading to finer action proposal localization.

Non-boundary Feature Learning. The non-boundary feature learning aims to extract the action and background non-boundary features from per video. Action non-boundary features refer to easily distinguishable action features that are not at the boundary, while background non-boundary features refer to easily distinguishable background features that are not at the boundary. As shown in Fig. 2, given the embedded features F_i for the video i, a linear classifier $g_{cls}(\cdot)$ is employed to predict the Class Activation Sequence (CAS), denoted as $S_i = g_{cls}(F_i)$. The CAS $S_i \in \mathbb{R}^{P\times C}$ contains C sets of action scores, each representing an action category that occurs in all P clips of the ith video. Next, we calculate the total action score of the video by summing the S_i. To separate the action non-boundary features A_i^r from background non-boundary features B_i^r, we sort the total action score and select the clip corresponding to the top-K scores as the action non-boundary features A_i^r, and the bottom-K scores as the background non-boundary features B_i^r. To make the A_i^r and B_i^r as far apart as possible, we introduce a ranking loss as the non-boundary refinement loss L_{act}, which is given by:

$$L_{act} = \frac{1}{N}\sum_{i=1}^{N}(max(0, q-\|A_i^r\|)+\|B_i^r\|)^2 \tag{6}$$

where $\|\cdot\|$ denotes a norm function, and q is a predefined maximum feature magnitude. This learning process effectively separates action and background non-boundary features in a video feature space.

Boundary Feature Learning. With the obtained action and background non-boundary features, we further refine the hard-to-locate boundary information, aiming to obtain more precise action proposal positions. To achieve this, we devise a joint CL mechanism on the action and background boundary feature learning, respectively, thus obtaining a more robust and fine-grained boundary representation. In particular, the action and background boundary features are initially obtained using erosion and dilation operations [23]. We represent the action boundary features as A_i^b, and the background boundary features as B_i^b, respectively. For the action boundary features A_i^b, we select the corresponding action non-boundary features A_i^r as the positive pair, while the background non-boundary features B_i^r as the negative pair. Conversely, for the background boundary features B_i^b, we select the corresponding background non-boundary feature B_i^r as the positive pair, while the action non-boundary features A_i^r as the negative pair. To effectively distinguish the A_i^b and B_i^b, we employ two InfoNCE losses [11] for joint learning, so that we can differentiate between challenging action and background proposals. enenFor learning together, we define the boundary refinement loss L_{jbr} as:

$$L_{jbr} = -\sum_{i=1}^{N}\Big(\log\frac{\exp((\overline{A_i^b})^T \cdot \overline{A_i^r}/\tau_f)}{\exp((\overline{A_i^b})^T \cdot \overline{A_i^r}/\tau_f) + \sum_r \exp((\overline{A_i^b})^T \cdot B_i^r/\tau_f)} + \log\frac{\exp((\overline{B_i^b})^T \cdot \overline{B_i^r}/\tau_f)}{\exp((\overline{B_i^b})^T \cdot \overline{B_i^r}/\tau_f) + \sum_r \exp((\overline{B_i^b})^T \cdot A_i^r/\tau_f)}\Big) \tag{7}$$

where τ_f is a temperature hyperparameter in CL. To conform to the form of InfoNCE loss [11], we perform an averaging operation $\overline{(\cdot)}$ on A_i^b and B_i^b, respectively.

Overall, the total optimization objective L_{BCL} of boundary-level CL contains the non-boundary refinement loss L_{act} and boundary refinement loss L_{jbr} as:

$$L_{BCL} = L_{act} + L_{jbr} \tag{8}$$

Combining the two refinement objectives, the boundary-level CL achi eves the finer action and background boundary features, allowing more robust UTAL performance.

3.2. HC-UTAL for unsupervised temporal action localization

In this section, we provide the detailed implementation of our proposed HC-UTAL for UTAL.

3.2.1. Action category mapping

Using the video-level CL in HC-UTAL, we cluster the videos into C clusters, each corresponding to a video pseudo-label. A mapping process is applied to assign specific action categories to each video pseudo-label. Since video pseudo-labels may contain multiple action classes, the mapped action categories for a video are considered in a multi-label manner. To perform the mapping process, we follow the approach described in the previous work TCAM [8]. In this approach, we count the occurrence of action class labels within each cluster. It is worth to note that the number of action labels are only used for mapping and are not involved in training the model. Suppose that the most frequently occurring action class in cluster c appears t times, we retain the action classes for each video in this cluster *w.r.t* the occurrence times of that action classes $\geq \frac{t}{2}$. As a result, we obtain the final action categories of each video with a one-hot vector, represented as $y_{i,c}$, in a multi-label manner.

3.2.2. Action proposal localization

Since TAL involves both action category classification and proposal localization, we need to further classify the action proposals into the appropriate action categories, thus facilitating the final fine-grained localization. Therefore, with the mapped action categories $y_{i,c}$ and the finer CAS via the boundary-level CL, we perform action proposal localization with robustness and effectiveness below.

CAS Aggregation. Using the CAS $S_i = \{s_{i,c}\}_{c=1}^{C}$ obtained by the boundary-level CL, we aggregate the top-l scores of CAS for each action category $y_{i,c}$, and compute the average to obtain the per-action classification score $a_{i,c}$ for each video i. The CAS aggregation is performed as follows:

$$a_{i,c} = \frac{1}{l} max \sum_{l} s_{i,c}. \tag{9}$$

where $s_{i,c} \in \mathbb{R}^{1\times P}$ represents the corresponding action score for per video. Following [15], we empirically set $l = 5$. The per-action classification score $a_{i,c}$ represents the confidence of the action class prediction for the ith video.

Action Proposal Classification. To classify action proposals into the corresponding action categories, a multi-label cross-entropy loss L_{cls} is employed to optimize the classifier g_{cls} for multiple action class prediction. The loss is defined as:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} y_{i,c}\log(p_{i,c}) \tag{10}$$

where N is the number of videos, $y_{i,c}$ represents the mapped action labels, and $p_{i,c} = \frac{\exp(a_{i,c})}{\sum_{c=1}^{C}\exp(a_{i,c})}$ is the predicted probability of the action class for the ith video.

3.2.3. Overall learning objective for UTAL

In general, the total learning objective for UTAL is the summation of the above-mentioned HC-UTAL loss $L_{HC-UTAL}$ and the multi-label cross-entropy loss L_{cls}. Mathematically, the total loss L can be written as:

$$L = L_{cls} + L_{HC-UTAL}. \tag{11}$$

3.3. Inference pipeline

During inference, we first cluster and map the action classes, and then employ a threshold θ_{class} to select the predicted action class probabilities $p_{i,c}$ that exceed the threshold. With the selected action classes, another threshold θ_{act} is applied to the corresponding CAS values to generate a set of localization proposals. Each proposal is represented as (b_c, e_c, y_c), where b_c and e_c indicate the start and end times of the action class y_c, respectively. Finally, non-maximum suppression (NMS) is performed on all proposals to eliminate duplicate proposals. NMS helps generate the final set of localized action segments by selecting the most confident and non-overlapping proposals. Following these steps, the proposed framework enables the effective localization of multiple action proposals without any action annotations.

4. Experiments and analysis

4.1. Datasets

In this section, we provide comprehensive experimental results on two large video-based TAL datasets, including THUMOS'14 [24], ActivityNet v1.2 and ActivityNet v1.3 [25]. Both datasets consist of a large number of untrimmed videos. In other words, these videos contain action and non-action background clips.

THUMOS'14 [24]. THUMOS'14 includes 13,320 untrimmed videos. The video duration is highly variable, each video may contain multiple action instances. For a fair comparison, we followed the previous methods [26,27] and utilized 413 videos for unsupervised or weakly supervised temporal action localization. Among them, 200 videos are sourced from the validation set of the THUMOS'14 for training, while 213 videos are sourced from the test set of the THUMOS'14 for evaluation.

ActivityNet v1.2 and ActivityNet v1.3 [25]. ActivityNet v1.2 and ActivityNet v1.3 are two different versions of the popular large-scale action localization benchmark dataset. ActivityNet v1.2 contains a total of 9682 videos, including 4819 videos in the training set, 2383 videos in the validation set, and 2480 videos in the test set, covering 100 action categories. In contrast, ActivityNet v1.3 contains a total of 20,000 videos, including about 10,024 videos in the training set, about 4926 videos in the validation set, and 5044 videos in the test set, with the action categories increased to 200 categories. For ActivityNet v1.2, we use a training set of 4819 videos for training and evaluate on a test set of 2480 videos. For ActivityNet v1.3, we train the model on a training set of 10,024 videos and evaluate it on a validation set of 4926 videos.

4.2. Implementation details

4.2.1. Evaluation protocols

We evaluated our method with mean Average Precision (mAP) under several Intersections over Union (IoU) thresholds, *i.e.*, the standard evaluation metrics for temporal action localization. Both datasets used the benchmark code provided by ActivityNet for evaluation [25]. For THUMOS'14, the threshold range used is from 0.1 to 0.7, and the average mAP within this range is taken as the final metric. For ActivityNet, the threshold range used is from 0.5 to 0.95, and the average mAP within this range is taken as the final metric. In addition, we employed the normalized mutual information (NMI) score and adjusted rand index (ARI) to measure the clustering performance, which has been widely used in clustering tasks. In addition, the baseline in our work, is a typical two-stage framework using the traditional spectral clustering algorithm [28] for action clustering and the classification loss (L_{cls}) for action localization.

Table 1
The key training parameters involved in this work.

Parameters	Description of the parameters	Values
d	Tensor dimension of each clip	1024
τ_h	The hyperparameter in the horizontal CL in Eq. (2)	0.4
τ_v	The hyperparameter in the vertical CL in Eq. (3)	0.9
τ_{in}	The hyperparameter in the instance-level CL in Eq. (5)	0.07
q	Predefined maximum feature magnitude in Eq. (6)	150
τ_f	The hyperparameter in the boundary-level CL in Eq. (7)	0.07

4.2.2. Training details

The number of clips P in a video was set to 750, 50 for THUMOS'14 and ActivityNet, respectively. For training, we utilized the Adam optimizer with an initialized learning rate of 0.0001. For clarification, other key training parameters are shown in Table 1.

4.2.3. Inference details

For THUMOS'14 and ActivityNet, we set θ_{class} to 0.2 and 0.1 to determine which action classes will be localized. We used multiple thresholds for proposal generation. For THUMOS'14, we set θ_{act} to [0.325:0.375:0.025] and then performed non-maximum suppression (NMS) using a threshold of 0.55. For ActivityNet, we set θ_{act} to [0.0:0.105:0.015] and then performed non-maximum suppression (NMS) using a threshold of 0.7.

4.3. Comparisons with the state-of-the-arts

4.3.1. Evaluation on THUMOS'14

Table 2 summarizes the results of the THUMOS'14 dataset for the fully supervised, weakly supervised, and unsupervised TAL, respectively, when the IoU threshold varies between 0.1 and 0.7. HC-UTAL outperforms previous weakly supervised and unsupervised methods in almost all IoU metrics on the THUMOS14 dataset. In the unsupervised case, our method achieved favorable performance of 30.1% mAP@0.5 and 37.4% mAP@Avg. When compared with the SOTA methods APSL [9] and FEEL [34], HC-UTAL achieves an absolute improvement of 2.2% and 0.8% in average mAP and mAP at an IoU threshold of 0.5, respectively. It demonstrates the effectiveness of our coarse-to-fine method through a highly coupled learning process for unsupervised temporal action localization. In the weakly supervised case, where only weak annotations, *i.e.*, action category labels were provided, our method still achieved the best result of 43.6% mAP@Avg on THUMOS'14. Compared with FTCL [16] and P-MIL [31], our methods are also better than theirs, implying that the HC-UTAL is still valid for the weakly-supervised framework. This result is close to the fully supervised TAL performance. In addition, Table 5 reports the results of our HC-UTAL for action category clustering. We achieved good results on two widely-used clustering metrics, *i.e.*, the adjusted rand index (ARI) and normalized mutual information score (NMI), on THUMOS'14, obtaining 0.705 on ARI and 0.860 on NMI, whereas APSL only obtained 0.639 on ARI and 0.821 on NMI. It shows that the highly coupled learning process is also helpful for clustering.

4.3.2. Evaluation on ActivityNet v1.2 and ActivityNet v1.3

Table 3 displays the evaluation results on ActivityNet v1.2, comparing the performance of various action localization methods, including unsupervised, weakly supervised, and fully supervised approaches. Obviously, our method also achieved state-of-the-art performance on the ActivityNet1.2 datasets.

In the unsupervised case, where no annotations are available for the videos, our method achieved a noteworthy 28.3% mAP@Avg, surpassing the performance of some other unsupervised methods. Compared with the SOTA methods APSL [9], FEEL [34], and UGCT [35], our method achieved absolute gains of 0.7%, 3.8%, and 5.6% in terms of average mAP, respectively. This result indicates that our method can successfully localize actions in videos even without any explicit supervision. In the weakly supervised case, our method outperformed the state-of-the-art methods. Compared with the SOTA methods APSL [9] and CASE [36], we achieved absolute gains of 1.0% and 1.3% in terms of average mAP, respectively. This improvement signifies the effectiveness of our approach in leveraging limited supervision to localize actions in videos accurately.

As shown in Table 4, we conducted additional evaluation experiments on the ActivityNet v1.3 dataset to further demonstrate the effectiveness of our proposed method in both unsupervised and weakly supervised settings. Under the weakly supervised scenario, our method achieves a notable performance of 28.9% mAP@Avg, surpassing other SOTA methods. This result highlights our approach's ability to accurately localize actions in videos with limited supervision. In the unsupervised setting, our method also exhibited a significant advantage, with an absolute gain of 2.3% mAP@Avg compared to TSCN [33]. In summary, our method consistently delivers strong performance in action localization, whether under limited supervision or in fully unsupervised conditions.

Additionally, as shown in Table 5, we evaluated the clustering performance of our method on ActivityNet v1.2. The ARI obtained was 0.692, indicating a high degree of agreement between the predicted clusters and the ground truth. Moreover, the NMI yielded a value of 0.894, which signifies a moderate level of agreement between the predicted clusters and the ground truth.

4.4. Ablation studies

4.4.1. Effect of hierarchical contrastive learning

Table 6 presents the results of ablation studies conducted on the THUMOS'14 dataset in the unsupervised case, aiming to analyze the contribution of each hierarchical contrast loss. Introducing the video-level CL loss (L_{VCL}) to replace the traditional spectral clustering algorithm substantially improved the performance by 3.8% in mAP@0.5. This improvement can be attributed to the fact that video-level CL enables pseudo-labels and action-related features to be learnt interactively in a high-coupling manner, resulting in enhanced action localization. As shown in Table 6, the integration of the loss L_{ICL} improved the performance by 3.1%, and further addition of the loss L_{BCL} resulted in an increase of 8.1%. Finally, using all the losses (L_{cls}, L_{VCL}, L_{ICL}, and L_{BCL}) in combination with dynamic weighting (DWA) for training the action localization model yielded the best result of 30.1% in mAP@0.5. This demonstrates the effectiveness of incorporating all losses and employing dynamic weighting to optimize the model's performance.

4.4.2. Effect of different clustering methods

The subsection presents the results of various clustering methods applied to the THUMOS'14 dataset. The performance metrics, such as adjusted rand index (ARI) and normalized mutual information (NMI), are reported in Table 7. It can be seen that our video-level CL-based clustering method with integrated high-coupling learning performs much better than other clustering methods. Video-level CL achieved 0.70 on ARI, an improvement of 0.31 over spectral clustering. Moreover, video-level CL achieved 0.86 on NMI, which is 0.17 higher than spectral clustering. This suggests that the best clustering performance can be achieved by using video-level CL, thanks to more effective feature interactions during the integrated hierarchical CL process. Furthermore, from the localization results, it can be observed that video-level CL achieved the best performance with mAP@Avg of 37.4%, indicating that video-level CL produces more accurate pseudo-labels for achieving better localization results.

In addition, we have expanded our analysis of the impact of various data augmentation in horizontal-level and vertical-level CL, respectively. The results are shown in Table 8. At the horizontal level, when only horizontal pair augmentation is used, the ARI is 0.10 and the NMI is 0.37; meanwhile, in vertical-level, when only vertical pair

Table 2
Comparison of state-of-the-art methods on the THUMOS'14 dataset in fully supervised, weakly supervised, and unsupervised learning settings, respectively. We denote the fully supervised, weakly supervised, and unsupervised as FS, WS, and US, respectively. The best results are in bold. *indicates the results obtained by reproducing the code.

Supervision	Method	mAP@t-IoU (%)							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
FS	S-CNN(2016) [6]	47.7	43.5	36.3	28.7	19	–	–	–
	SSN(2017) [5]	66	59.4	51.9	41	29.8	–	–	–
	GTAN(2019) [12]	69.1	63.7	57.8	47.2	38.8	–	–	–
WS	W-TALC(2018) [3]	55.2	49.6	40.1	31.1	22.8	–	7.6	–
	CMCS(2019) [26]	57.4	50.8	41.2	32.1	23.1	15	7	32.4
	DGAM(2020) [27]	60	54.2	46.8	38.2	28.8	19.8	11.4	37
	TCAM(2020) [8]	–	–	46.9	38.9	30.1	19.8	10.4	–
	HAM-Net(2021) [14]	65.9	59.6	52.2	43.1	32.6	21.9	12.5	41.1
	SAPS(2021) [29]	61.7	56	50	38.8	28.8	18.7	9.5	–
	IONet(2021) [30]	58.0	50.3	41.1	32.3	24.0	15.1	7.2	–
	CoLA(2021) [15]	66.2	59.5	51.5	41.9	32.2	22	**13.1**	40.9
	EGANet(2022) [2]	64.5	58.4	50.0	41.4	31.5	21.0	10.7	–
	FTCL(2022) [16]	69.6	63.4	**55.2**	**45.2**	35.6	23.7	12.2	43.6
	P-MIL(2023)*[31]	65.3	61.3	53.7	45.0	36.5	24.2	12.8	42.8
	A-TSCN(2023) [32]	65.3	59.0	52.1	42.5	33.6	23.4	12.7	32.9
	CAM(2023) [1]	64.7	57.6	49.2	38.0	31.0	22.9	12.1	39.3
	APSL(2023) [9]	69.1	62.4	53.7	43.6	33.6	23.8	12.8	42.7
	Ours	**69.7**	**63.5**	54.8	44.9	**35.6**	**24.4**	12.5	**43.6**
US	TCAM(2020) [8]	–	–	39.6	32.9	25	16.7	8.9	–
	TSCN(2020) [33]	57.1	51.6	43.9	35.3	26.0	15.7	6.0	33.7
	FEEL(2023) [34]	–	–	–	–	29.3	22.6	**11.5**	–
	APSL(2023) [9]	57.7	52.4	44.1	35.9	27.9	18.5	10	35.2
	Ours	**60.8**	**55.1**	**47.2**	**38.3**	**30.1**	20.2	10.2	**37.4**

Table 3
Comparison of state-of-the-art methods on the ActivityNet v1.2 dataset in fully supervised, weakly supervised, and unsupervised learning settings, respectively. We denote fully supervised, weakly supervised, and unsupervised as FS, WS, and US, respectively. The best results are in bold. * indicates the results obtained by reproducing the code.

Supervision	Method	mAP@t-IoU (%)			
		0.5	0.75	0.95	Avg
FS	SSN(2017) [5]	41.3	27	6.1	26.6
WS	DGAM(2020) [27]	41	23.5	5.3	24.4
	SAPS(2021) [29]	38.7	23.2	5.7	–
	CoLA(2021) [15]	42.7	25.7	5.8	26.1
	EGANet(2022) [2]	41.3	24.7	5.5	25.4
	CAM(2023) [1]	42.9	25.5	9.3	25.9
	UGCT(2023) [35]	43.1	26.6	6.1	26.9
	P-MIL(2023)*[31]	44.2	26.1	5.3	26.5
	A-TSCN(2023) [32]	39.6	25.1	5.8	25.6
	ASE(2023) [36]	43.8	27.2	6.7	27.9
	APSL(2023) [9]	44.3	28.5	6.2	28.2
	SRHN(2024) [37]	44.3	26.7	5.3	26.8
	Ours	**46.3**	**29.4**	**6.4**	**29.2**
US	TCAM(2020) [8]	35.2	21.4	3.1	21.1
	TSCN(2020) [33]	22.3	13.6	2.1	13.6
	UGCT(2023) [35]	37.4	23.8	4.9	22.7
	FEEL(2023) [34]	38.0	25.6	3.4	24.5
	APSL(2023) [9]	43.7	28.1	5.8	27.6
	Ours	**44.1**	**29.1**	**6.3**	**28.3**

augmentation is used, the ARI is 0.28 and the NMI is 0.60;. We conducted experiments on the THUMOS'14 dataset to evaluate the effects of horizontal CL and vertical CL on model performance, demonstrating how various augmentation strategies can improve overall effectiveness.

4.5. Visualization and qualitative results

4.5.1. Visualization on the hierarchical CL

To explore the role of contrastive learning at each level, we visualized the obtained CAS score of the ThrowDiscus action category using different CL levels in HC-UTAL, as shown in Fig. 4. From the figure, the blue rectangles represent the ground truth of the ThrowDiscus action category in the video. The lines with different colors represent the CAS scores obtained by the baseline and different CL levels, respectively. The figure illustrates that as the hierarchical CL components are added, the action score gradually decreases in the background region outside the ground truth. This implies that adding hierarchical CL refines the boundary regions and leads to more accurate localization of the ThrowDiscus action. In summary, the visualization results demonstrate the effectiveness of incorporating hierarchical and highly-coupling contrastive learning in improving the localization and refinement of action boundaries, particularly in the difficult-to-locate boundaries of background regions.

4.5.2. Visualization of clustering results

Fig. 5 shows the results of clustering visualization using the traditional spectral clustering method and our proposed video-level CL

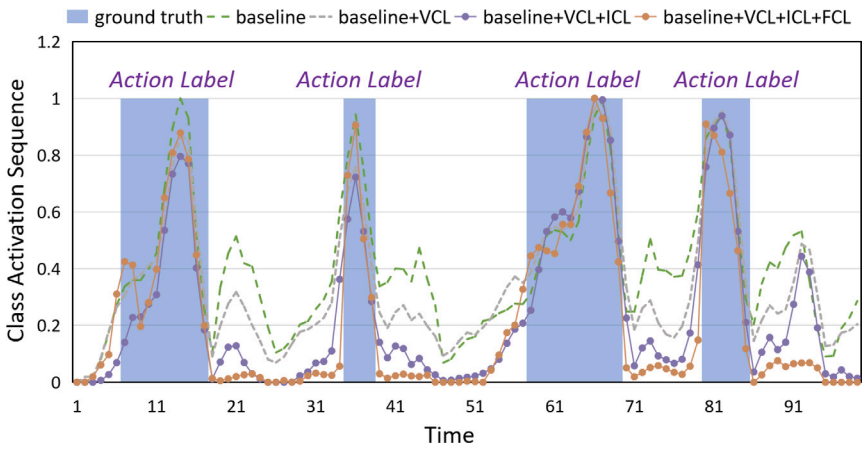

Fig. 4. Visualization on the CAS scores of the action category *ThrowDiscus* via CL of different levels. Obviously, adding hierarchical contrastive learning refines the boundary regions, leading to more accurate localization in the category *ThrowDiscus*.

Table 4
Comparison of state-of-the-art methods on the ActivityNet v1.3 dataset in fully supervised, weakly supervised, and unsupervised learning settings, respectively. We denote fully supervised, weakly supervised, and unsupervised as FS, WS, and US, respectively. The best results are in bold.

Supervision	Method	mAP@t-IoU (%)			
		0.5	0.75	0.95	Avg
FS	SSN(2017) [5]	43.26	28.7	5.6	28.28
WS	IONet(2021) [30]	34	20.3	5.0	22.2
	EGANet(2022) [2]	35.4	22.5	4.5	22.4
	DCC(2022) [38]	38.8	24.2	5.7	24.3
	UGCT(2023) [35]	41.2	24.4	5.9	25.5
	P-MIL*(2023) [31]	41.8	25.4	5.2	25.5
	A-TSCN(2023) [32]	37.9	23.1	5.6	23.6
	CASE(2023) [36]	43.2	26.2	**6.7**	26.8
	SRHN(2024) [37]	41.7	26.1	6.1	26.2
	SPCC-Net(2024) [39]	41.0	25.0	6.4	25.4
	ISSF(2024) [40]	39.4	25.8	6.4	25.8
	Ours	**45.8**	**29.8**	5.9	**28.9**
US	TSCN(2020) [33]	25.9	16.2	3.8	16.3
	Ours	**29.4**	**19.0**	**4.3**	**18.6**

Table 5
Comparison of clustering results on the THUMOS'14 dataset and the ACTIVITYNET V1.2 dataset, respectively. The best results are in bold.

Dataset	Method	ARI	NMI
THUMOS'14	APSL(2023) [9]	0.639	0.821
	Ours	**0.705**	**0.86**
ACTIVITYNET V1.2	APSL(2023) [9]	0.574	0.795
	Ours	**0.692**	**0.894**

Table 6
Ablation study of different contrast losses on the THUMOS'14 dataset in the unsupervised case. The best results are in bold.

Baseline	L_{VCL}	L_{ICL}	L_{BCL}	Dynamic weighting	mAP@0.5 (%)
✓					13.8
✓	✓				17.6
✓	✓	✓			20.7
✓	✓	✓	✓		28.8
✓	✓	✓	✓	✓	**30.1**

method, respectively. We performed the clustering for generating action pseudo-labels on the THUMOS'14 dataset and used the t-SNE feature maps for visualization. Fig. 5(a) shows the feature map for the spectral clustering, and (b) shows the feature map for the video-level CL. It can be observed that the clusters obtained by video-level CL are more compact than the spectral clustering, with smaller cluster class

Table 7
Comparison of different clustering methods on the THUMOS'14 dataset.

Method	ARI	NMI	mAP@Avg (%)
Kmeans	0.04	0.39	28.6
Agglomerative	0.18	0.58	31.7
Spectral clustering	0.39	0.69	33.5
Separated learning with Video-level CL	0.63	0.82	34.4
Integrated learning with Video-level CL	**0.70**	**0.86**	**37.4**

Table 8
The impact of horizontal CL and vertical CL in video-level CL on the clustering performance of the model.

Horizontal CL	Vertical CL	ARI	NMI
✓		0.10	0.37
	✓	0.28	0.60
✓	✓	**0.70**	**0.86**

distances for each cluster, indicating the advantages of video-level CL in clustering.

4.5.3. Visualization of localization results

Fig. 6(a) and (b) show the localization results of single actions on the THUMOS'14 dataset under unsupervised conditions by different methods. Blue represents the true action region, green represents the localization results of the baseline method without hierarchical contrastive learning, gray shows the localization results after adding video-level CL, purple shows the results after adding both video-level and instance-level CL, and orange shows the effects after adding video-level, instance-level, and boundary-level CL. As shown in Fig. 6(a) and (b), with the introduction of three-level CL, the localization boundary of single actions gradually becomes clearer, highlighting the key role of hierarchical CL in boundary localization. To further demonstrate the importance of hierarchical CL in multi-action scenarios, we perform boundary localization on instances of multiple action categories. Fig. 6(c) shows that with the addition of three-level CL, the boundaries of multi-action localization also become clearer, indicating that hierarchical CL also plays an important role in multi-action boundary localization.

In Fig. 7, we visualized the extracted task-specific embedded features F_i with different settings by using the t-SNE on the THUMOS'14 dataset, where blue represents background and orange represents action. As shown in Fig. 7(a), one can see that the task-specific embedded features F_i extracted by the baseline contain lots of confusing information between action and background features. As shown in Fig. 7(b), after adding video-level CL to the baseline, there are still many confusing pieces of information existing between action and background.

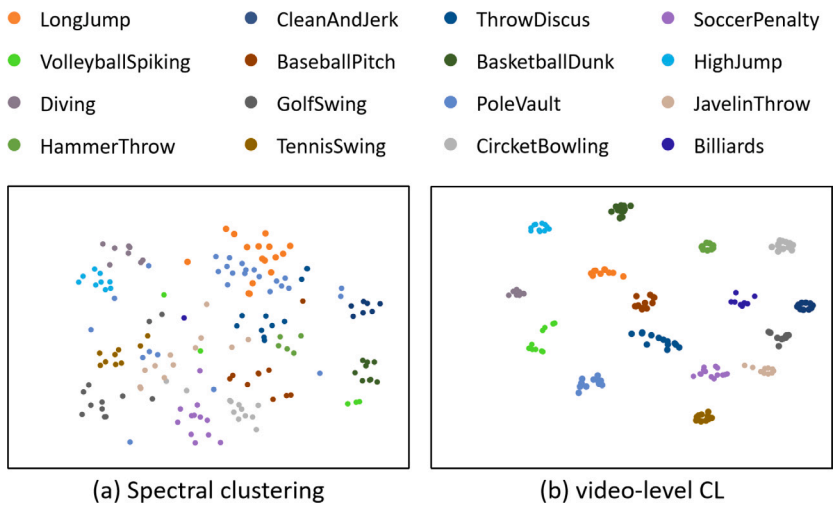

Fig. 5. Visualization for video-level feature clustering on the THUMOS'14 dataset. (a) The feature map learned by the spectral clustering, and (b) the feature map learned by our video-level CL. Obviously, the clusters obtained by video-level CL are more compact than spectral clustering, with smaller cluster class distances for each cluster.

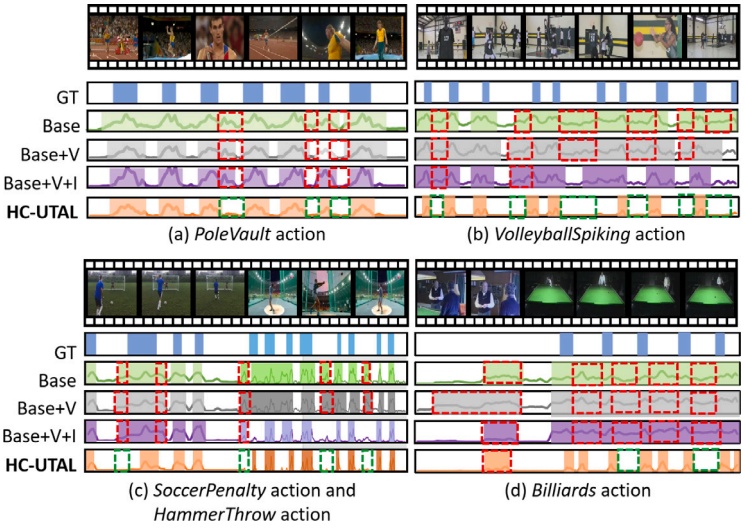

Fig. 6. Visualization results on THUMOS'14 for videos with a single action category or multiple action analogies in the unsupervised case. Adding hierarchical CL components gradually improves the clarity and accuracy of action instance localization. The localization results become clearer and more refined, especially in locating the boundaries of actions. The red dotted boxes represent incorrect localization results, while the green dotted boxes represent the corresponding correct localization results of our HC-UTAL. Among them, (a) represents the *PoleVault* action, (b) represents the *VolleyballSpiking* action, (c) represents the *SoccerPenalty* and *HammerThrow* action, and (d) represents the failure cases *Billiards* action. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

One possible analysis for this is that video-level CL may contribute more to clustering improvement, but it does not provide significant assistance in distinguishing between backgrounds and actions with clarity. In comparison to Fig. 7(b), 7(c) demonstrates that after adding instance-level CL, there is some separation between background and action. However, there are still some confusions present between action and background due to the challenge boundary information. Furthermore, as shown in Fig. 7(d), the addition of boundary-level CL results in a clear separation between action and background in F_i due to fully considering boundary learning. This separation allows for better identification and recognition of action and background.

4.5.4. Visualization of failure cases

As shown in Fig. 6(d), despite the effectiveness of our method, we found that there are still some rooms in which our HC-UTAL could be improved. Firstly, some extremely minor visual changes remain challenging for current methods to capture, which has an impact on the methods' performance. Additionally, due to the limited information provided by unsupervised learning, we need to pre-define the number of action categories in our method, which affects the performance of real-world UTAL applications.

5. Conclusion

In this work, we propose a novel and end-to-end hierarchical contrastive learning framework, called HC-UTAL, for robust UTAL. Specifically, HC-UTAL contains three-hierarchy contrastive learning (CL), *i.e.*, *video-level CL, instance-level CL, and boundary-level CL*, for effectively mining multi-level, high-coupling semantic features related to actions in untrimmed videos. First, we introduce video-level CL for video-level feature clustering to generate video pseudo-labels. Then, with the generated pseudo-labels, we introduce instance-level CL to learn action-related and action-unrelated features, making the action features with the same categories closer together. Next, we further design boundary-level CL on more fine-grained action-background boundary semantic representation learning, thus improving the finer action localization performance. By training the coarse-to-fine CL for high-coupling feature

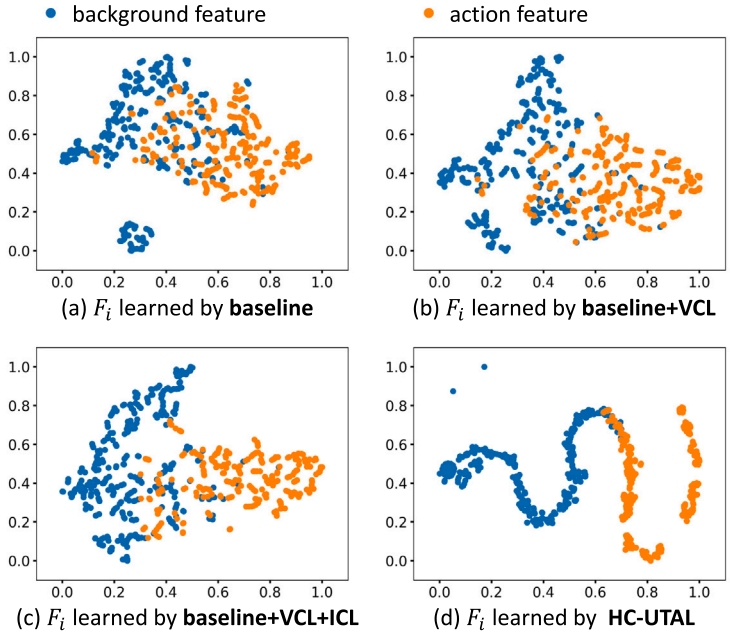

Fig. 7. Comparison of the distributions of action, background with different settings using the t-SNE visualization, where blue represents background and orange represents action. (a) the F_i learned by the baseline, (b) the F_i learned by baseline and video-level CL, (c) the F_i learned by baseline, video-level CL and Instance-CL, (d) the F_i learned by HC-UTAL. Obviously, owing to considering feature interaction in three-level CL, our HC-UTAL achieves a more discriminatory distribution of features, even the easily confused action and background boundary features, thus obtaining finer action-background boundary localization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

learning and interaction via adaptation weighting, HC-UTAL can obtain more fine-grained action semantic representations, thus improving the performance of both action proposal classification and localization. Extensive experiments conducted on the THUMOS'14, ActivityNet v1.2, and ActivityNet v1.3 datasets demonstrate that our method outperforms the baseline method by a significant margin and achieves state-of-the-art performance. In future work, I will introduce physic-informed prior knowledge as prompts into our framework to explore more comprehensive action information for UTAL task in unconstrained environment.

CRediT authorship contribution statement

Yuanyuan Liu: Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Conceptualization. **Ning Zhou:** Writing – review & editing, Writing – original draft, Methodology. **Yuxuan Huang:** Visualization, Validation, Investigation. **Shuyang Liu:** Visualization, Validation. **Leyuan Liu:** Supervision, Formal analysis. **Wujie Zhou:** Supervision, Formal analysis. **Chang Tang:** Supervision, Conceptualization. **Ke Wang:** Validation, Supervision, Investigation.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this manuscript and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province, China grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2025.111421.

Data availability

The code link was shared in the footnote on the home page of the article.

References

[1] C. Wang, J. Wang, P. Liu, Complementary adversarial mechanisms for weakly-supervised temporal action localization, Pattern Recognit. 139 (2023) 109426.
[2] Y. Cheng, Y. Sun, H. Fan, T. Zhuo, J.-H. Lim, M. Kankanhalli, Entropy guided attention network for weakly-supervised action localization, Pattern Recognit. 129 (2022) 108718.
[3] S. Paul, S. Roy, A.K. Roy-Chowdhury, W-talc: Weakly-supervised temporal activity localization and classification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 563–579.
[4] Y. Ge, X. Qin, D. Yang, M. Jagersand, Deep snippet selective network for weakly supervised temporal action localization, Pattern Recognit. 110 (2021) 107686.
[5] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, D. Lin, Temporal action detection with structured segment networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 2914–2923.

[6] Z. Shou, D. Wang, S.-F. Chang, Temporal action localization in untrimmed videos via multi-stage cnns, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1049–1058.
[7] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, Z. Shou, SF-Net: Single-frame supervision for temporal action localization, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 420–437.
[8] G. Gong, X. Wang, Y. Mu, Q. Tian, Learning temporal co-attention models for unsupervised video action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9819–9828.
[9] Y. Liu, N. Zhou, F. Zhang, W. Wang, Y. Wang, K. Liu, Z. Liu, APSL: Action-positive separation learning for unsupervised temporal action localization, Inform. Sci. 630 (2023) 206–221.
[10] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[11] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[12] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, T. Mei, Gaussian temporal awareness networks for action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 344–353.
[13] J.-X. Zhong, N. Li, W. Kong, T. Zhang, T.H. Li, G. Li, Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 35–44.
[14] A. Islam, C. Long, R. Radke, A hybrid attention mechanism for weakly-supervised temporal action localization, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (2) 2021, pp. 1637–1645.
[15] C. Zhang, M. Cao, D. Yang, J. Chen, Y. Zou, Cola: Weakly-supervised temporal action localization with snippet contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16010–16019.
[16] J. Gao, M. Chen, C. 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, 2022, pp. 19999–20009.
[17] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Learning salient boundary feature for anchor-free temporal action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3320–3329.
[18] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, Y. Cui, Spatiotemporal contrastive video representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6964–6974.
[19] J. Wang, J. Jiao, Y.-H. Liu, Self-supervised video representation learning by pace prediction, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 504–521.
[20] T. Yao, Y. Zhang, Z. Qiu, Y. Pan, T. Mei, Seco: Exploring sequence supervision for unsupervised representation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 10656–10664.
[21] L. Tao, X. Wang, T. Yamasaki, Self-supervised video representation learning using inter-intra contrastive framework, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2193–2201.
[22] Y. Li, P. Hu, Z. Liu, D. Peng, J.T. Zhou, X. Peng, Contrastive clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 8547–8555.
[23] J.Y. Gil, R. Kimmel, Efficient dilation, erosion, opening, and closing algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1606–1617.
[24] H. Idrees, A.R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, M. Shah, The thumos challenge on action recognition for videos “in the wild”, Comput. Vis. Image Underst. 155 (2017) 1–23.
[25] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.
[26] D. Liu, T. Jiang, Y. Wang, Completeness modeling and context separation for weakly supervised temporal action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1298–1307.
[27] B. Shi, Q. Dai, Y. Mu, J. Wang, Weakly-supervised action localization by generative attention modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1009–1019.
[28] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (2007) 395–416.
[29] X.-Y. Zhang, Y. Zhang, H. Shi, J. Dong, SAPS: Self-attentive pathway search for weakly-supervised action localization with background-action augmentation, Comput. Vis. Image Underst. 210 (2021) 103256.
[30] X.-Y. Zhang, H. Shi, C. Li, P. Li, Z. Li, P. Ren, Weakly-supervised action localization via embedding-modeling iterative optimization, Pattern Recognit. 113 (2021) 107831.
[31] H. Ren, W. Yang, T. Zhang, Y. 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, 2023, pp. 2394–2404.
[32] Y. Zhai, L. Wang, W. Tang, Q. Zhang, N. Zheng, D. Doermann, J. Yuan, G. Hua, Adaptive two-stream consensus network for weakly-supervised temporal action localization, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 4136–4151.
[33] Y. Zhai, L. Wang, W. Tang, Q. Zhang, J. Yuan, G. Hua, Two-stream consensus network for weakly-supervised temporal action localization, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 37–54.
[34] H. Tang, H. Jiang, M. Xu, Y. Hu, J. Zhu, L. Nie, Unsupervised temporal action localization via self-paced incremental learning, 2023, arXiv preprint arXiv:2312.07384.
[35] W. Yang, T. Zhang, Y. Zhang, F. Wu, Uncertainty guided collaborative training for weakly supervised and unsupervised temporal action localization, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 5252–5267.
[36] Q. Liu, Z. Wang, S. Rong, J. Li, Y. Zhang, Revisiting foreground and background separation in weakly-supervised temporal action localization: A clustering-based approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10433–10443.
[37] Y. Zhao, H. Zhang, Z. Gao, W. Guan, M. Wang, S. Chen, A snippets relation and hard-snippets mask network for weakly-supervised temporal action localization, IEEE Trans. Circuits Syst. Video Technol. (2024).
[38] J. Li, T. Yang, W. Ji, J. Wang, L. Cheng, Exploring denoised cross-video contrast for weakly-supervised temporal action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19914–19924.
[39] Y. Shao, F. Zhang, C. Xu, Snippet-to-prototype contrastive consensus network for weakly supervised temporal action localization, IEEE Trans. Multimed. (2024).
[40] W. Yun, M. Qi, C. Wang, H. Ma, Weakly-supervised temporal action localization by inferring salient snippet-feature, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 6908–6916.

Yuanyuan Liu received the PhD degree from Central China Normal University. She is currently an associate professor at the China University of Geosciences (Wuhan) and a visiting scholar in Nanyang Technological University, Singapore. Her research interests include deep learning on emotion recognition and understanding.

Ning Zhou received the B.S. degree in Software Engineering from China University of Geosciences, Wuhan, China, in 2021. He is currently working toward the Master degree at China University of Geosciences in Wuhan, China. His research interests include video temporal action localization.

Yuxuan Huang received the B.S. degree in Computer Science and Technology from Southwest Minzu University in 2022; she is currently working toward the Master degree at China University of Geosciences in Wuhan, China. Her research interests include computer vision and affective computing.

Shuyang Liu received the B.S. degree in Network Engineering from China University of Geosciences, Wuhan, China, in 2021; he is currently working toward the Master degree at China University of Geosciences in Wuhan, China. His research interests include computer vision and affective computing.

Leyuan Liu received the Ph.D degree from Huazhong University of Science and Technology. He is currently an associate professor at the Central China Normal University, China. His main research interests include computer vision, pattern recognition, computer graphics, and human–computer interaction.

Wujie Zhou is an Associate Professor at Zhejiang University of Science and Technology and a Visiting Scholar at Nanyang Technological University, Singapore. He has authored over 70 articles in AI and multimedia signal processing, including 30+ in IEEE journals, 7 ESI hot papers, and 14 highly cited papers.

Chang Tang received the Ph.D degree from Tianjin University and is now a full professor at China University of Geosciences. He has published 80+ peer-reviewed papers in top venues like IEEE T-PAMI, ICCV, and CVPR, and won the best paper award at the 5th Asian Conference on Artificial Intelligence Technology.

Ke Wang received the master’s degree from Taiyuan University of Science and Technology in 2023. He is currently pursuing a Ph.D degree at China University of Geosciences (Wuhan). His research interests include computer vision and affective computing.