Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Enhancing action discrimination via category-specific frame clustering for weakly supervised temporal action localization*

Huifen XIA[1,3], Yongzhao ZHAN[‡1,2], Honglin LIU[1], Xiaopeng REN[1]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*

[2]*Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agricultural Applications, Zhenjiang 212013, China*

[3]*Changzhou Vocational Institute of Mechatronic Technology, Wujin, Changzhou 213164, China*

E-mail: 106338096@qq.com; yzzhan@ujs.edu.cn; 534110389@qq.com; 1815772718@qq.com

Received Jan. 13, 2023; Revision accepted July 6, 2023; Crosschecked

Abstract: Temporal action localization (TAL) is a task of detecting the start and end times of action instances and classifying them in an untrimmed video. As the number of action categories per video increases, existing weakly supervised temporal action localization (W-TAL) methods with only video-level labels cannot provide sufficient supervision. Single-frame supervision has attracted the interest of researchers. Existing paradigms model single-frame annotations from the perspective of video snippet sequences, neglect action discrimination of annotated frames, and do not pay sufficient attention to their correlations in the same category. Considering a category, the annotated frames exhibit distinctive appearance characteristics or clear action patterns. Thus, a novel method to enhance action discrimination via category-specific frame clustering for W-TAL is proposed. Specifically, the K-means clustering algorithm is employed to aggregate the annotated discriminative frames of the same category, which are regarded as exemplars to exhibit the characteristics of the action category. Then, the class activation scores are obtained by calculating the similarities between a frame and exemplars of various categories. Category-specific representation modelling can provide complimentary guidance to snippet sequence modelling in the mainline. As a result, a convex combination fusion mechanism is presented for annotated frames and snippet sequences to enhance the consistency properties of action discrimination, which can generate a robust class activation sequence for precise action classification and localization. Due to the supplementary guidance of action discriminative enhancement for video snippet sequences, our method outperforms existing single-frame annotation-based methods. Experiments conducted on three datasets THUMOS14, GTEA and BEOID show that our method achieves high localization performance compared with state-of-the-art methods.

unedited

Key words: Weakly supervised; Temporal action localization; Single-frame annotation; Category specific; Action discrimination
Manual; Word template
https://doi.org/10.1631/FITEE.2300024 **CLC number:**

1 Introduction

Temporal action localization (TAL) is one of the main research areas in computer vision and multimedia fields. It has a broad range of potential applications in anomaly detection (Zhou et al., 2021), video surveillance (Sultani et al., 2018) and so on. TAL is a vital and challenging task of locating the start and end timestamps of the action instances and classifying them in untrimmed videos. In the past several years, full supervision- or weak supervision-based TAL methods have been proposed. Specifically, fully supervised temporal action localization (F-TAL) methods require a large amount of labelled data, *i.e.*, frame-level temporal annotations. Although these methods have achieved promising progress, labelling the data is time-consuming, labor-intensive and error-prone, which limits the scalability in real-world scenarios. To overcome these limitations, weakly supervised temporal action localization

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61672268)
ORCID: Huifen XIA, https://orcid.org/0000-0003-3875-4919; Yongzhao ZHAN, https://orcid.org/0000-0001-7475-2895
© Zhejiang University Press 2023

(W-TAL) methods have become increasingly popular. Concretely, different kinds of information are used for supervision, such as the number of action instances (Narayan et al., 2019), video-level category labels (Wang et al., 2017; Nguyen et al., 2018; Paul et al., 2018), and recent single-frame annotation of each action instance (Ma et al., 2020). Since video-level labels need the weakest clue, indicating only whether an action occurs in an untrimmed video, it becomes the mainstream method. These methods have achieved some progress. However, due to the lack of precise action instance annotations, actions are difficult to distinguish from complex backgrounds. Additionally, action instances with the same category often appear repeatedly, or action instances of different categories occur alternately in a video. Therefore, supervision based on video-level labels cannot offer sufficient supervision, resulting in inferior performance.

To solve this issue, the method of single-frame annotation in W-TAL has been proposed in SF-Net (Ma et al., 2020). Specifically, only a single frame timestamp corresponding to its action category is annotated for each action instance during the training. Compared with video-level supervision, this supervision provides very useful positioning information with negligible extra labelling cost. However, current methods (Ju et al., 2021; Lee et al., 2021) only start from the video snippet sequence where the single-frame annotation is located, and they do not take the action discrimination of annotated frames and their correlations into account in the video. We believe that the annotated frames have discriminative characteristic representations of this action category. Hence, considering the modelling of annotated frames of the same category and their correlations are important issues for single-frame annotation temporal action localization.

The discriminative action expressions of annotated frames are still insufficient, which affects the improvement of action localization performance. To cope with this issue, a novel method to enhance action discrimination via category-specific frame clustering is proposed for W-TAL. By observing, we find that the annotated frames of the same category are the discriminative action frames, which have distinctive appearance characteristics or clear action patterns. Therefore, we employ the K-means clustering algorithm to aggregate the annotated frames of the same category, which are regarded as exemplars to exhibit the category-specific feature representation. By calculating the pairwise similarity between a frame and exemplars of various categories, we can obtain the class activation scores of all annotated frames. Therefore, the action discrimination of the annotated frames is fully exploited, which provides complementary guidance for video snippet sequence modelling. In addition, a convex combination fusion mechanism is presented between the annotated frames and video snippet sequences to ensure the consistency of action discrimination. After fusing, we obtain a more robust class activation sequence, which is used for precise temporal action classification and localization.

In addition, due to the lack of detailed frame-level annotations, W-TAL methods adopt the framework of localization-by-classification. For the classification task, the model only needs to detect the most discriminative action snippets in the video that contribute to the classification task. For the localization task, the model needs to detect the complete snippets of action instances. Therefore, the task of W-TAL is a trade-off between classification and localization, which has different feature preferences. For predicting complete action instances, GCRNet (Liao et al., 2021) proposes a global context relation network to model long-term related action snippets by introducing the self-attention mechanism. MDA (Chen et al., 2022) introduces the temporal relation block (TRB) to capture the relationship among long-term information. Different from them, we directly employ the VideoMAE model (Tong et al., 2022) pretrained on the Kinetics400 dataset to extract the features from the original video snippets and then obtain the global correlation along the temporal dimension. We do not need complex computation of the self-attention mechanism and obtain better global information. Finally, we fuse the local feature and global feature to generate a fruitful feature representation for action localization. In short, our contributions are summarized as follows:

(1) A novel method to enhance action discrimination via category-specific frame clustering is proposed for W-TAL. Specifically, the K-means clustering algorithm is employed to aggregate the annotated distinctive frames of the same category, and then

unedited

category-specific feature representation is obtained. It can provide complementary guidance to enhance action discrimination for video snippet sequence modelling.

(2) A convex combination fusion mechanism is presented, which ensures the consistency of action discrimination between the annotated frames and video snippet sequence. Then, we can obtain a more robust class activation sequence for precise action classification and localization.

(3) Experimental results conducted on three classic datasets, THUMOS14 (Jiang et al., 2014), GTEA (Lei et al., 2018) and BEOID (Damen et al., 2014), show that our localization performance outperforms state-of-the-art methods.

2 Related work

2.1 Action recognition

Action recognition, sometimes called action classification, is a subtask of temporal action localization. Recently, many novel action recognition networks (Zhu et al., 2022; Lei and Todorovic, 2018) have achieved significant performance. For example, TCLC (Zhu et al., 2022) was proposed to explore temporal correlations among neighboring frames and exploit cross-layer multiscale features for action recognition. These networks are helpful for extracting visual sequence features from untrimmed videos in temporal action localization tasks.

2.2 Fully supervised temporal action localization

Fully supervised temporal action localization (F-TAL) approaches rely on precise temporal action annotations, *i.e.*, frame-level labels during training. Due to the fine-grained annotations, they have made impressive progress. In existing works, there are two types: proposal-based paradigms (Chao et al., 2018; Zeng et al., 2019) and frame-based paradigms (Lin et al., 2017a; Long et al., 2019). Specifically, the proposed-based paradigm is a two-stage framework, which is inspired by the success of region-based convolution neural networks in object detection. They first generate action proposals, then classify them and conduct temporal boundary regression. On the other hand, the frame-based paradigm is a one-stage framework that directly predicts frame-level action category and location by some postprocessing techniques. Both paradigms require high annotation costs, and they are time consuming, labor intensive and error prone. Hence, these methods are not suitable for real-world scenarios. Weakly supervised temporal action localization (W-TAL) has attracted researchers' attention.

2.3 Weakly supervised temporal action localization

Compared with full supervision, weakly supervised temporal action localization (W-TAL) only requires coarse-grained annotations, which only require low labelling costs during training. Among the existing methods, there are different kinds of supervision, such as movie scripts (Bojanowski et al., 2013), web videos (Gan et al., 2016), temporal action orders (Huang et al., 2016; Bojanowski et al., 2014), and video-level category labels (Wang et al., 2017; Nguyen et al., 2018; Yang et al., 2021a). Due to the cost of labelling, W-TAL methods based on video-level category labels have been the mainstream. Existing video-level supervision methods mainly adopt snippet sequence-based methods for temporal action localization, which are mainly divided into two paradigms: a multi-instance learning (MIL)-based paradigm (Paul et al., 2018; Shou et al., 2018) and an attention-based paradigm (Nguyen et al., 2018; Nguyen et al., 2019; Ge et al., 2021). Specifically, the MIL-based paradigm treats the entire video and its video snippet sequences as a bag and instances, respectively. The snippet-level classifier is learned to generate a snippet-level class activation sequence. Then, the top-k mechanism is adopted to select the action snippets in the video. The attention-based paradigm estimates snippet-level action probabilities from the raw video directly and then selects snippets with high activation scores as action snippets. Both paradigms select the most discriminative action snippets in the video, which makes the action instances incomplete. Therefore, some researchers have studied the erasing mechanism on this basis, such as Hide and seek (Kumar Singh et al., 2017) and ASSD (Zhang et al., 2019). In short, the localization performance has been improved, but there is still a large gap compared with F-TAL methods. In addition, as the number of different action categories increases, video-level annotations cannot offer enough information for supervision.

unedited

Recently, on the basis of the balance between labelling cost and localization performance, single-frame annotation-based methods have been explored, which refer to the usage of a single frame of each action instance. This supervision provides more abundant information with an affordable labelling cost. Facing complex videos with repeated actions of the same category or alternately occurring actions of multiple categories, single-frame supervision brings more action location information. SF-Net (Ma et al., 2020) and LACP (Lee et al., 2021) adopt a pseudo-label mining strategy to obtain a better CAS for temporal action localization. Subsequently, DC (Ju et al., 2021) adopted a two-stage framework to bridge F-TAL to W-TAL. However, existing methods have not fully considered the action image information of the annotated single frame. Therefore, in our method, we propose a dual-branch architecture in parallel form to make full use of annotated single-frame image information as a supplement for snippet sequence modelling in the video.

3 Proposed method

3.1 Framework of our proposed method

3.1.1 Motivation and overview

Existing W-TAL methods of single-frame annotation only use video snippet sequences to model action or background and ignore the full utilization of the action discrimination of annotated frames, which makes the class activation sequence for temporal action localization insufficiently robust. This may be the main cause of inaccurate and imprecise detection of action instances. Therefore, a novel method to enhance action discrimination via category-specific frame clustering is introduced for temporal action localization. The framework of our method is shown in Fig. 1, which consists of the mainline, discriminative representation for action via frame clustering and the convex combination fusion mechanism. Specifically, the lower part of Fig. 1 exhibits the discrimination enhancement for the action of an annotated single frame, which is the first proposed for temporal action localization. It can provide complementary guidance information to the video snippet sequence branch in the mainline. In the right part of Fig. 1, a convex combination fusion mechanism is presented to ensure the consistency of action discrimination and make the temporal action localization algorithm more robust.

Unedited

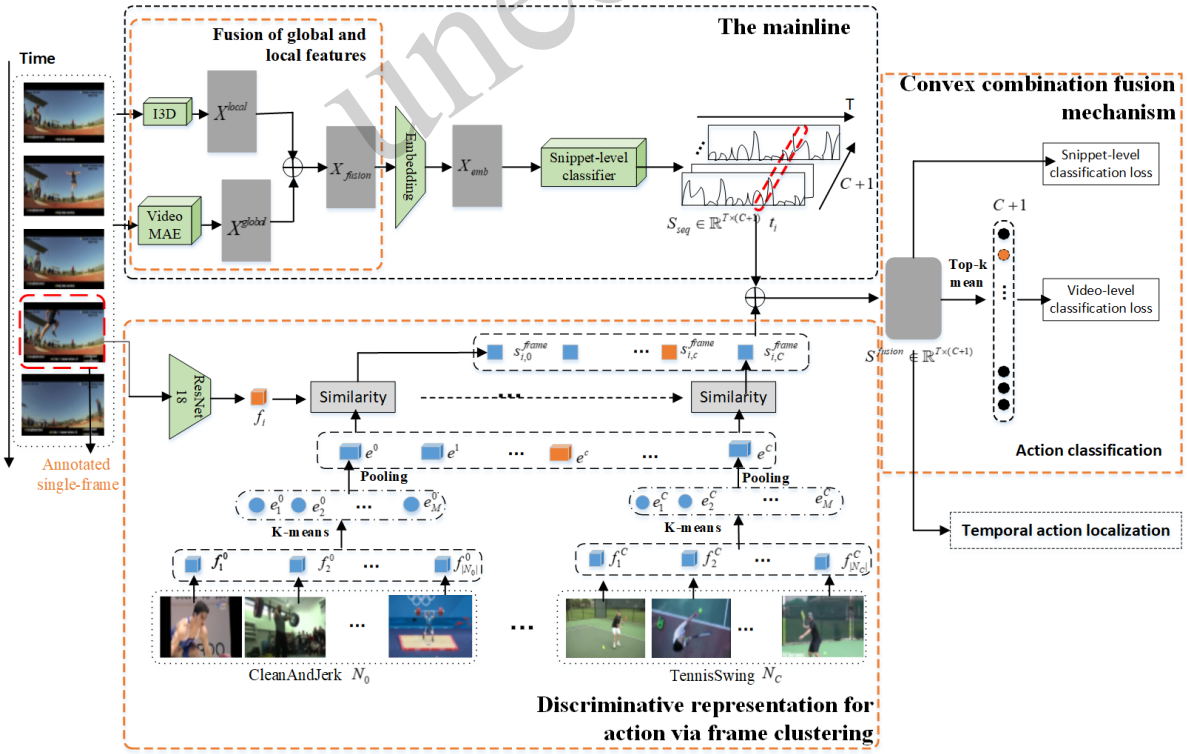


Fig. 1 The overall framework of our proposed method, which consists of the mainline, discriminative representation for action via frame clustering, and convex combination fusion mechanism of annotated frames and video snippet sequences.

Finally, a robust class activation sequence is generated for precise action classification and localization.

3.1.2 Problem Description

Following SF-Net (Ma et al., 2020), we set our problem of single frame-level supervision for W-TAL. Given an untrimmed input video v, a single timestamp and its action category for each action instance are provided, *i.e.*, $P_{act}=\{(t_i,y_{t_i})\}_{i=1}^{N_{act}}$, in which the *i*-th action instance is annotated at the t_i-th video snippet with the action label y_{t_i}. N_{act} is the number of action instances of video v. $y_{t_i}\in\mathbb{R}^C$ is a binary vector, *i.e.*, $y_{t_i,c}=1$, which represents the *i*-th action instance belonging to the *c*-th action category; otherwise, $y_{t_i,c}=0$, in which $c=0,1,\cdots,C-1$, for a total of *C* action categories. During testing, our goal is to predict a set of action instances with the form of $\{s,e,c,q\}$, which represents the start time, the end time, the action category and its confidence score for the predicted action instance *j*, respectively. Notably, each video may contain multiple action instances and multiple action categories. Therefore, the video-level label can be obtained by aggregating the annotations of the annotated frame along the temporal dimension, *i.e.*, $y_{vid,c}=\mathrm{I}[\sum_{i=1}^{N_{act}}y_{t_i,c}>0]$, where $\mathrm{I}[\cdot]$ is the indicator function.

3.2 The mainline

3.2.1 Local and global feature extraction and fusion

Following previous works (Paul et al., 2018; Lee et al., 2020), we divide it into nonoverlapping 16-frame snippets for a given input video. Then, we adopt the two-stream I3D network (Carreira et al., 2017), which is pretrained on the Kinetics400 dataset (Kay et al., 2017), to extract the snippet-level feature representations. Thus, we obtain two *D*-dimensional RGB features and flow features, *i.e.*, $X_{RGB}\in\mathbb{R}^{T\times D}$ and $X_{flow}\in\mathbb{R}^{T\times D}$, in which *T* is the number of video snippet sequences. Note that we call them the I3D features X_{I3D} collectively.

Due to the limitation of the receptive field in a 3D network, convolutions are designed to capture short-range information without long-range dependencies beyond the receptive field. However, as mentioned above, we know that the existing weak supervision methods adopt the pipeline of localization by classification. The I3D features focus on the local snippet-level features of the video, which are only effective for the classification task. The localization task requires the model to detect complete action instances. This indicates that the localization task has different feature preferences than the classification task. Therefore, we consider that X_{I3D} is not sufficient for W-TAL, resulting in incomplete action instances and limited localization performance improvement. We think that modelling long-term temporal snippet dependencies is also important in the task of action localization.

Yang et al. (Yang et al., 2021) proposed a novel MKR (multiple knowledge representation) framework, which learns information from different abstraction levels and different perspectives. Therefore, MKR has more explainable and generalizable feature representations. Inspired by MKR, the global information and local information are utilized together to obtain more discriminative features for action classification and localization in our method. Specifically, we use the transformer-based structure to model the global dependencies in the video to enhance the relationships between long-term temporal snippets and local common information. Different from GCRNet (Liao et al., 2021) and MDA (Chen et al., 2022), which directly perform a self-attention mechanism on I3D features, we employ the VideoMAE model (Tong et al., 2022), which is also pretrained on the Kinetics400 dataset (Kay et al., 2017), to extract a global feature representation, *i.e.*, $X_{global}\in\mathbb{R}^{T\times D}$. The inputs of the VideoMAE model are also nonoverlapping 16-frame video snippets, which are the same as the input of the I3D model. Then, we obtain the global correlation between a snippet sequence and the rest of the video along the temporal dimension. Finally, we fuse X_{I3D} and X_{global} to obtain novel fruitful feature representations X_{fusion} with both global and local feature information. The expression is shown as follows:

$$X_{fusion}=\alpha_1 X_{I3D}+(1-\alpha_1)X_{global} \tag{1}$$

unedited

in which α_1 is a hyperparameter. Note that we do not need complex computation of the self-attention mechanism on I3D features. Instead, through a simple fusion mechanism, the fruitful feature has the abilities of local and global feature discriminations, which is friendly to both classification and localization tasks.

The snippet-level video features X_{fusion} are fed into a 1-D temporal convolutional layer followed by the ReLU activation layer to generate the task-specific embedding feature X_{emb}. The formula is expressed as follows:

$$X_{emb} = \max(W_{emb} * X_{fusion} + b_{emb}, 0) \quad (2)$$

in which W_{emb}, b_{emb} are the weight and bias parameters. The dimension of X_{emb} is set to the same as the feature X_{fusion}.

3.2.2 Snippet-level class activation sequence

We feed the embedding features into the snippet-level classifier, which contains a 1-D convolutional layer to predict the snippet-level class activation scores S_{seq}. The formula is expressed as follows:

$$S_{seq} = conv(X_{emb}, W_{seq}) \quad (3)$$

in which $S_{seq} \in \square^{T\times(C+1)}$, C is the number of action classes, and the additional class is an auxiliary background class. After the convolutional layer, we set the softmax activation layer over the class dimension, *i.e.*, $\hat{s}_{t,c}^{seq} = soft\max(s_{t,c}^{seq})$, in which $c = 0,1,\cdots,C$, $t = 1,2,\cdots,T$, and $\hat{s}_{t,c}^{seq}$ is the class activation score of the *t*-th video snippet for the *c*-th action category of video v.

3.3 Discriminative enhancement representation for action

Existing single-frame annotation W-TAL methods only use video snippet sequences to model action or background and ignore the action discrimination of annotated frames, which makes the class activation sequence for action localization insufficiently robust. We argue that action instances of the same category should have the same appearance characteristics or motion patterns, which can be used to represent the action category. As far as we observe, all the annotated action frames are distinctive in action instances. Therefore, a novel method to enhance action discrimination via category-specific frame clustering is proposed. Specifically, we divide the annotated frames of the same category into *M* typical characteristics by the K-means clustering algorithm, in which *M* is a hyperparameter. Then, we take the average pooling on the *M* typical characteristics, which are used to represent the feature representation of each category. On this basis, the temporal action localization task is formulated as comparing a frame with an exemplar of each category. By calculating the similarities, we obtain the class activation scores. As a result, modelling action discrimination enhancement can provide complementary guidance information to the video snippet sequence of the mainline and make the temporal action localization algorithm robust to noise within the local video snippets.

3.3.1 Discriminative representation for action via category-specific frame clustering

First, all annotated frames in the same category *c* are grouped into a set N_c, in which $c = 0,1,2,\cdots,C$. Similar to the video snippet sequence, there are also *C* action classes and an additional background class. During the training phase, the annotated action frames are given, which are denoted as $P_{act} = \{(t_i, y_{t_i})\}_{i=1}^{N_{act}}$, where $act = 0,1,\cdots,C-1$. We generate discriminative background frames based on the class activation sequence $\hat{s}_{t,c}^{seq}$ in the mainline of the upper part of Fig. 1. Since there is an annotated frame for every action instance, there must be at least one background snippet between two adjacent annotated frames to separate them. Concretely, we select the snippets whose background scores $\hat{s}_{t,C}^{seq}$ are larger than the threshold β. If there is no snippet that meets this condition, we choose the snippet with the highest background score between the adjacent action frames. Finally, we put these background snippets together to obtain the pseudobackground set $P_{bkg} = \{(t_i, y_{t_i} = 1)\}_{i=1}^{N_{bkg}}$, in which $bkg = C$. Since the untrimmed original video has redundant temporal information, we regard the intermediate frame in the pseudobackground snippet as the sampled background frame.

unedited

We adopt ResNet18 to extract the features of the annotated action frames and sampled background frames f_i, $i=1,2,\cdots,|N_c|$, and $c=0,1,\cdots,C$, which represent the *i*-th annotated frame of video v. As shown in the lower part of Fig. 1, we cluster the annotated frames of class *c* into M typical characteristics via the K-means clustering algorithm, *i.e.*, $\varepsilon_c=[e_1^c,e_2^c,\cdots,e_M^c]$. By average pooling of *M* typical characteristics, we obtain the category feature representation e^c, which is regarded as the exemplar for class *c*. Therefore, we obtain the category-specific feature representation. This formula is shown as follows:

$$e^c = avg(e_1^c, e_2^c, \cdots, e_M^c) \tag{4}$$

in which $avg(\cdot)$ represents the average value of M characteristics, $c=0,1,\cdots,C$.

To classify the annotated frames correctly, we calculate the similarity scores between the features of each annotated frame and all exemplars of different categories to obtain the class activation scores. This formula is shown as follows.

$$s_{i,c}^{frame} = \cos(f_i, e^c) = \frac{f_i^T \cdot e^c}{\| f_i \|_2^2 \cdot \| e^c \|_2^2} \tag{5}$$

where $c=0,1,\cdots,C$, $i=1,2,\cdots,N_c$, and $s_{i,c}^{frame} \in \mathbb{R}^{C+1}$ represent the activation score of the *i*-th annotated frame for the *c*-th category. Note that the activations of the unannotated frames of the video are set to 0. Therefore, we obtain the activation scores of video v from the perspective of the annotated frame. The formula is calculated as follows:

$$s_{t,c}^{frame} = \begin{cases} s_{i,c}^{frame}, \ if \ t=t_i \\ 0, \ otherwise \end{cases} \tag{6}$$

in which $t_i \in P_{act} \cup P_{bkg}$, which indicates the annotated frame in the video. Then, we use the softmax normalization operation on the class activation scores of each annotated frame of video v along the class dimension, *i.e.*, $\hat{s}_{t,c}^{frame} = soft\max(s_{t,c}^{frame})$.

3.3.2 Convex combination fusion mechanism

Since the snippet-level class activation sequences from the mainline are not robust enough for precise action localization, we need to make full use of the action discrimination of annotated frames. To make the distinctive frames participate in the training to cooperate with snippet sequence modelling, we propose a convex combination fusion mechanism. It can enhance the consistency properties of action discrimination between annotated frames and video snippet sequences. After fusing, we obtain a more robust CAS $S^{fusion} \in \mathbb{R}^{T\times(C+1)}$ for precise action classification and localization. The formula is shown as follows:

$$s_{t,c}^{fusion} = \alpha_2 \hat{s}_{t,c}^{seq} + (1-\alpha_2)\hat{s}_{t,c}^{frame} \tag{7}$$

in which $\alpha_2 \in [0,1]$ is a hyperparameter to control the balance between them. $s_{t,c}^{fusion}$ represents the final class activation scores of the *t*-th video snippet for category c, and $t=1,2,\cdots,T$, $c=0,1,\cdots,C$.

To produce video-level class activation scores, we aggregate the new snippet-level activation scores by using top-k mean pooling along the temporal dimension. Formally, the video-level action score is calculated as follows.

$$s_{vid,c} = \frac{1}{k} \max_{\substack{|S'|=k \\ S'\subset S^{fusion}[:,c]}} \sum_{\forall a\in S'} a \tag{8}$$

in which $k=\left\lfloor \frac{T}{8} \right\rfloor$, S' represents the subset of *k* snippets in $S^{fusion}[:,c]$, which may contain actions. Similarly, we use softmax regularization along the category axis, *i.e.*, $\hat{s}_{vid,c} = soft\max(s_{vid,c})$. Therefore, the video-level binary cross-entropy loss is calculated as follows:

$$L_{vid} = -\sum_{c=0}^{C} [y_{vid,c} \log \hat{s}_{vid,c} + (1-y_{vid,c}) \log(1-\hat{s}_{vid,c})] \tag{9}$$

in which $y_{vid,c}$ is the video-level label, which is calculated by aggregating frame annotations of video v.

We use the snippet-level classification loss, which consists of the annotated action snippets L_{act} and pseudo background snippets L_{bkg}. Here, we employ the focal loss (Lin et al., 2017b) to train the snippet-level classification loss because it has been observed that the number of action instances of each video is different, and the annotated action frames of each action category are unbalanced. Therefore, the

unedited

classification loss for action snippets, where annotated frames are located, is calculated as follows:

$$L_{act} = -\frac{1}{N_{act}} \sum_{\forall (t_i, y_{t_i}) \in P_{act}} \{[\sum_{c=0}^{C} y_{t_i,c} (1-\hat{s}_{t_i,c}^{fusion})^{\gamma} \log \hat{s}_{t_i,c}^{fusion} + \sum_{c=0}^{C} (1-y_{t_i,c})(\hat{s}_{t_i,c}^{fusion})^{\gamma} \log(1-\hat{s}_{t_i,c}^{fusion})] + (\hat{s}_{t_i,C}^{fusion})^{\gamma} \log(1-\hat{s}_{t_i,C}^{fusion})\} \tag{10}$$

in which N_{act} indicates the number of annotated action frames in the video, and γ is the focusing parameter. To separate action snippets from complicated backgrounds, we need to select some pseudo-background snippets to complement the annotated action frames. Similar to the annotated action frames, we also use the focal loss function as the classification loss for the pseudobackground snippets. The formula is expressed as follows:

$$L_{bkg} = -\frac{1}{N_{bkg}} \sum_{\forall t_i \in P_{bkg}} [\sum_{c=0}^{C} (\hat{s}_{t_i,c}^{fusion})^{\gamma} \log(1-\hat{s}_{t_i,c}^{fusion}) + (1-\hat{s}_{t_i,C}^{fusion})^{\gamma} \log \hat{s}_{t_i,C}^{fusion}] \tag{11}$$

in which N_{bkg} indicates the number of pseudo background snippets in the video, and γ is the focusing parameter, which is the same as in formula (10).

3.4 Action completeness modelling

Since the annotated action frames are only a small part of the action instance, it is not enough to have the classification loss at the snippet-level and video-level for complete action instances. Following LACP (Lee et al., 2021), we employ Outer-Inner-Contrast (OIC) to calculate the completeness score $score(\pi_c)$ of the candidate action instances for searching the optimal action instances. Therefore, the optimal action sequence for class *c* is expressed as follows:

$$\pi_c^* = \arg\max score(\pi_c) = \{(t_{n,c}^{start}, t_{n,c}^{end}, z_{n,c})_{n=1}^{M_c}\}_{c=0}^{C} \tag{12}$$

in which $t_{n,c}^{start}$ and $t_{n,c}^{end}$ represent the start and end times of the *n*-th action instance for category *c*, respectively. M_c is the number of the optimal action instance for category *c*. $z_{n,c} \in \{0,1\}$ is a binary vector. $z_{n,c} = 1$ indicates that the *n*-th action instance belongs to the *c*-th category; otherwise, $z_{n,c} = 0$. Afterwards, we employ score separation loss and feature contrast loss to guide the completeness modelling. We know that the features of different action instances for the same category should be closer than any background instances of video v. Therefore, the formulas are expressed as follows:

$$L_{score} = \frac{1}{\sum_{c=0}^{C} y_{vid,c}} \sum_{c=0}^{C} y_{vid,c} (1-score(\pi_c^*))^2 \tag{13}$$

$$L_{feat} = \frac{1}{\sum_{c=0}^{C} \mathrm{I}[\sum_{n=1}^{M_c} z_{n,c} > 1]} \sum_{c=0}^{C} \mathrm{I}[\sum_{n=1}^{M_c} z_{n,c} > 1] \cdot l_{feat}^{c} \tag{14}$$

in which

$$l_{feat}^{c} = -\frac{1}{\sum_{n=1}^{M_c} z_{n,c}} \sum_{n=1}^{M_c} z_{n,c} \log \frac{\sum_{\forall n' \neq n} z_{n',c} \exp(X_{n,c} \cdot X_{n',c} / \tau)}{\sum_{\forall n'' \neq n} \exp(X_{n,c} \cdot X_{n'',c} / \tau)},$$

n' is another action instance, which is different from the n-th action instance of video v. n'' is a background sequence of the same video. τ is a hyperparameter, and $\mathrm{I}[\cdot]$ represents the indicator function.

3.5 Joint training and inference

Finally, the overall loss function of our model is as follows:

$$L_{total} = L_{vid} + \lambda_0 L_{act} + \lambda_1 L_{bkg} + \lambda_2 L_{score} + \lambda_3 L_{feat} \tag{15}$$

in which λ_0, λ_1, λ_2 and λ_3 are hyperparameters used to balance the loss functions.

After training our model, we use it to localize the action instance of test videos. During testing, we select the intermediate frame in each snippet of the test videos as the single frame in the static branch for extracting single frame-level features. Then, the fused snippet-level class activation sequence S^{fusion} is used for inference. Specifically, we leave action categories with classification scores larger than the threshold

unedited

θ_{cls}. For the remaining categories, we threshold on the snippet-level activation scores with the threshold θ_{act}. Afterward, the consecutive snippets are merged into the candidate proposals. A multithreshold approach is employed for θ_{act} to enrich action instances, and nonmaximum suppression (NMS) is performed to remove the high overlap proposals. Similarly, we use the OIC to calculate the confidence score of each proposal.

4 Experiments

4.1 Datasets and evaluation metrics

We conduct our experiments on three popular datasets: THUMOS14 (Jiang et al., 2014), BEOID (Damen et al., 2014), and GTEA (Lei et al., 2018). For fairness, during training, we use the single-frame annotation of each action instance labelled in SF-Net (Ma et al., 2020). THUMOS14 contains 200 validation videos for training and 213 testing videos for testing, with a total of 20 action categories. Each video consists of an average of approximately 15 action instances, and there are 3,007 single-frame annotations available in the training videos. Additionally, the lengths of video and action instance vary widely, so it is challenging. BEOID has a total of 58 videos in 34 action categories. Following Moltisanti et al. (Moltisanti et al., 2019), we split the videos with a proportion of 80-20% randomly for training and testing, and we obtain 594 single-frame-level annotations. GTEA covers 28 untrimmed videos in 7 categories of daily activities in a kitchen, including 21 videos for training and 7 videos for testing. Each video contains 17.5 action instances on average.

We evaluate the mean average precision (mAP) under different temporal intersection over union (t-IoU) thresholds for action localization following the standard protocol. The action proposal is regarded as correct when the predicted action class is correct and its t-IoU with the ground truth (GT) is larger than the preset thresholds. The localization performances at small t-IoU thresholds indicate the ability to detect actions, and those at high t-IoU thresholds indicate the completeness of the predicted action instances.

4.2 Implementation details

During the stage of feature extraction and fusion, we employ the I3D network (Carreira et al., 2017) and VideoMAE model (Tong et al., 2022), which are both pretrained on Kinetics400 (Kay et al., 2017), to extract the video local and global features, respectively. Each video is split into snippets with 16 nonoverlapping frames. The TV-L1 algorithm (Wedel et al., 2009) is used to obtain the optical flow maps. All the local RGB and optical flow features and global features are 1024 dimensions. Following LACP (Lee et al., 2021), we use the original number of snippets as T without sampling. We optimize our model by Adam (Kingma et al., 2014) with a learning rate of 10^{-4} and a batch size of 16. We set the video-level classification threshold θ_{cls} to 0.5 and the snippet-level localization threshold θ_{act} from 0 to 0.25 with a step size of 0.05. We determined hyperparameters by grid search: $\alpha_1 = 0.9$ in Eq. (1), $\alpha_2 = 0.9$ in Eq. (7), $\gamma = 2$ in Eq. (10) and (11), $\tau = 0.1$ in Eq. (14), and $\beta = 0.95$ in Section 3.3.1. $\lambda_0 = 0.5$ and $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in Eq. (15). NMS is performed with a threshold of 0.6.

4.3 Comparison with state-of-the-art methods

In Table 1, we compare our method with some state-of-the-art F-TAL and W-TAL methods on the THUMOS14 dataset. We show the mAP (%) under t-IoU thresholds from 0.1 to 0.7 and their average mAP (%). We divide the literature into three categories according to the supervision methods, *i.e.*, fully supervised methods with frame-level labels, weakly supervised methods with video-level labels and weakly supervised methods with single frame-level labels. It is worth noting that F-TAL methods with frame-level labels require more expensive annotation costs than W-TAL methods, while in W-TAL methods, the methods with a single frame-level label have annotation costs comparable to those of video-level methods.

In the comparisons, we can see that our method far outperforms RSKP (Huang et al., 202), which is a state-of-the-art video-level W-TAL method, and the average mAP value increases from 45.1% to 53.7%. This is because of single frame-level label supervision, which ensures that all action instances in the video can be detected, reducing the false positives (FP) and

unedited

false negatives (FN). Whether at the low or high t-IoU threshold, the localization performance is greatly improved. This indicates that both the ability to recognize actions and the completeness of action instances of our proposed model have been greatly improved. When compared with SF-Net (Ma et al., 2020), which is the first to propose single-frame annotation information for temporal action localization, the mAP @t-IoU=0.7 has been improved from 9.6% to 22.1%, nearly 2.5 times. This illustrates that the action instances detected by our method are more complete. Likewise, compared with LACP (Lee et al., 2021), which is a state-of-the-art single frame-level method, our performances are also improved, especially at high thresholds of 0.5 and 0.6, and the performances are improved by 1.5% and 1.3%, respectively. This verifies the effectiveness of our method for completeness learning, which further confirms that modelling the image information of annotated single frames is complementary to the modelling of video snippet sequences, and it is beneficial to temporal action localization. Furthermore, our method even performs on par with F-TAL methods at lower t-IoU thresholds with much lower annotation costs. However, our method still lags behind state-of-the-art methods because of the lack of frame-level annotation information.

Table 1 Comparisons with state-of-the-art methods of mAP (%) under different kinds of supervision at t-IoU thresholds from 0.1 to 0.7 and their average mAP (%) on the THUMOS14 dataset.

Super.	Method	mAP@t-IoU							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
Full (frame-level)	SSN (Zhao et al., 2017)	66.0	59.4	51.9	41.0	29.8	19.6	10.7	39.8
	TAL-Net (Chao et al., 2018)	-	-	53.2	48.5	42.8	33.8	20.8	-
	P-GCN (Zeng et al., 2019)	69.5	67.8	63.6	57.8	49.1	-	-	-
	G-TAD (Xu et al., 2020)	-	-	54.5	47.6	40.2	30.8	23.4	-
	AFSD (Lin et al., 2021)	-	-	67.3	62.4	55.5	43.7	31.1	-
Weak (video-level)	STPN (Nguyen et al., 2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	27.0
	W-TALC (Paul et al., 2018)	55.2	49.6	40.1	31.1	22.8	14.8	7.6	31.6
	CMCS (Liu et al., 2019)	57.4	50.8	41.2	32.1	23.1	15.0	7.0	32.4
	WSBM (Nguyen et al., 2019)	60.4	56.0	46.6	37.5	26.8	17.6	9.0	36.3
	Bas-Net (Lee et al., 2020)	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	DGAM (Shi et al., 2020)	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
	TSCN (Zhai et al., 2020)	63.4	57.6	47.8	37.7	28.7	19.4	10.2	37.8
	DSSN (Ge et al., 2021)	57.8	49.7	41.0	32.3	22.4	-	-	-
	UGCT (Yang et al., 2021a)	69.2	62.9	55.5	46.5	35.9	23.8	11.4	43.6
	FTCL (Gao et al., 2022)	69.6	63.4	55.2	45.2	35.6	23.7	12.2	43.6
	MDA (Chen et al., 2022)	69.7	63.1	55.2	46.6	35.6	25.0	14.4	44.2
	MMSD (Huang et al., 2022b)	69.7	64.3	54.6	45.0	36.4	23.0	12.3	43.6
	RSKP (Huang et al., 2022a)	71.3	65.3	55.8	47.5	38.2	25.4	12.5	**45.1**
Weak (single-frame-level)	SF-Net (Ma et al., 2020)	71.0	63.4	53.4	40.7	29.3	18.4	9.6	40.8
	DC (Ju et al., 2021)	72.8	64.9	58.1	46.4	34.5	21.8	11.9	44.3
	BackTAL (Yang et al., 2021b)	-	-	54.4	45.5	36.3	26.2	14.8	-
	LACP (Lee et al., 2021)	75.7	71.4	64.6	56.5	45.3	34.5	21.8	52.8
	Ours	76.8	72.3	65.1	57.2	46.8	35.8	22.1	**53.7**

unedited

Similarly, we compare our method with state-of-the-art single frame-level methods on the GTEA and BEOID datasets in Table 2. We show the

mAP (%) under t-IoU thresholds of 0.1, 0.3, 0.5, and 0.7 and their average mAP (%) from 0.1 to 0.7 at a step of 0.1. From both datasets, we can see that our method outperforms LACP (Lee et al., 2021), which is a state-of-the-art TAL method under the supervision of the single frame level. It is worth noting that the localization performance improves more at high t-IoU thresholds of 0.5 and 0.7. This indicates that our method can learn more complete action instances. As a result, modelling action discrimination of annotated single-frame images plays a good supplementary role for snippet sequence modelling in temporal action localization tasks, which confirms the effectiveness of our method.

Table 2 Comparisons with state-of-the-art methods of mAP (%) at different t-IoU thresholds on GTEA and BEOID datasets.

Dataset	Method	mAP@t-IoU				
		0.1	0.3	0.5	0.7	Avg
GTEA	SF-Net (Ma et al., 2020)	58.0	37.9	19.3	11.9	31.0
	DC (Ju et al., 2021)	59.7	38.3	21.9	18.1	33.7
	LACP (Lee et al., 2021)	63.9	55.7	33.9	20.8	43.5
	Ours	64.3	56.0	34.5	21.4	**43.9**
BEOID	SF-Net (Ma et al., 2020)	62.9	40.6	16.7	3.5	30.9
	DC (Ju et al., 2021)	63.2	46.8	20.9	5.8	34.9
	BackTAL (Yang et al., 2021b)	60.1	40.9	21.2	11.0	32.5
	LACP (Lee et al., 2021)	76.9	61.4	42.7	25.1	51.8
	Ours	77.2	61.9	43.1	25.8	**52.2**

unedited

From Tables 1 and 2, we can see that our method is superior to state-of-the-art methods at most t-IoU thresholds. In particular, at some high t-IoU thresholds, the localization performance boosts more. The main reason is that modelling action discrimination via category-specific frame clustering provides supplementary guidance for video snippet sequence modelling. Additionally, to ensure the consistency of action discrimination between annotated frames and video snippet sequences, we present a convex combination fusion mechanism. This allows us to generate a more robust class activation sequence for accurate and precise action localization.

4.4 Ablation study

We conduct extensive ablation studies to verify the effectiveness of our proposed two-branch network. Following the methods of DC (Ju et al., 2021) and LACP (Lee et al., 2021), we perform all ablation experiments on the THUMOS14 dataset.

4.4.1 Effectiveness of snippet-level fusion features

In the baseline of the upper part of Fig. 1, we adopt a feature fusion, which consists of I3D features and VideoMAE features. To verify the effectiveness of our fusion features, we conduct experimental comparisons of α_1 in Fig. 2. The abscissa is the value of α_1, *i.e.*, α_1=1, 0.9, 0.8 and 0.7, and the ordinate is the value of mAPs. We use mAPs at t-IoU thresholds of 0.1, 0.3, 0.5, and 0.7 for the performance metric.

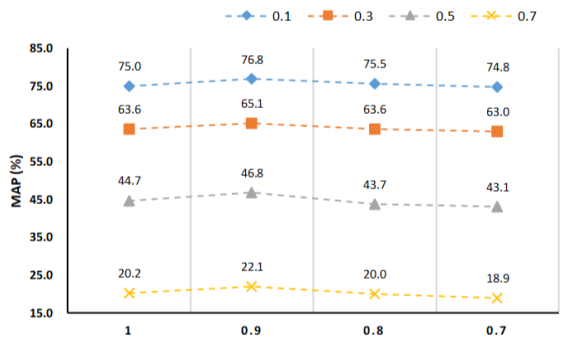


Fig. 2 The localization performance for mAP (%) with different fusion features on the THUMOS 14 dataset.

From Fig. 2, we can see that when α_1=1, only the I3D feature is used, and the mAPs @0.1, 0.3, 0.5 and 0.7 are 75.0%, 63.6%, 44.7% and 20.2%, respectively. When α_1=0.9, all localization performances are boosted. In particular, when the t-IoU

threshold is 0.5, the performance is improved by 2.1%. However, as α_1 gradually increases, the localization performance begins to decrease. This shows that when α_1=0.9, it is the best choice. Therefore, in our model, we set the hyperparameter α_1=0.9.

4.4.2 Comparisons of different M typical characteristics

In the action discriminative representation branch of the lower part of Fig. 1, the annotated frames in the same category are divided into M typical characteristics by the K-means clustering algorithm. We analyse the effect of clustering into different M classes on localization performance in Table 3. We divide each category into 5, 3 and 1 typical characteristics. Then, the typical characteristics are averaged to obtain the action features for each category. We use mAPs at t-IoU thresholds under 0.1, 0.3, 0.5, 0.7 and the average from 0.1 to 0.7 with a step of 0.1.

Table 3 The performance comparison of different M typical characteristics on the THUMOS 14 dataset.

Clustering	mAP @t-IoU				Avg
	0.1	0.3	0.5	0.7	
M=5	76.3	64.8	46.3	21.7	53.3
M=3	76.8	65.0	45.6	20.6	53.0
M=1	**76.8**	**65.1**	**46.8**	**22.1**	**53.7**

From the results of Table 3, we can see that when $M=1$, the localization performance at all t-IoU thresholds reaches optimality. This is reasonable, indicating that each category has a unique action feature representation. Therefore, in our model, we set the hyperparameter $M=1$.

4.4.3 Effectiveness of discriminative representation for action

In our model, we propose a novel method to enhance action discrimination via category-specific frame clustering. To verify the effectiveness of the annotated single-frame discriminative representation, we perform several experimental comparisons of α_2 in Fig. 3. The abscissa is the value of α_2, *i.e.*, α_2=1, 0.9, 0.8, 0.7, and the ordinate is the value of mAPs.

Similarly, we use mAPs at t-IoU thresholds of 0.1, 0.3, 0.5, and 0.7 for the performance metric. Note that when α_2=1, that is, only the video snippet sequence is modelled to generate CAS for temporal action localization without any action discrimination enhancement. When α_2=0.9, 0.8 and 0.7, they represent different proportions of action discriminative enhancement via frame clustering. From the results of Fig. 3, compared with α_2=1, we can see that when α_2=0.9, the localization performance at all t-IoU thresholds has improved. This shows that annotated single-frame modelling can provide complementary guidance information to video snippet sequence modelling, which further verifies that our method to enhance action discrimination via category-specific frame clustering is effective. However, as the value of α_2 decreases, we find that the localization performance decreases. Therefore, when α_2=0.9, the performance reaches optimality. Therefore, we set the hyperparameter α_2=0.9 in our model.

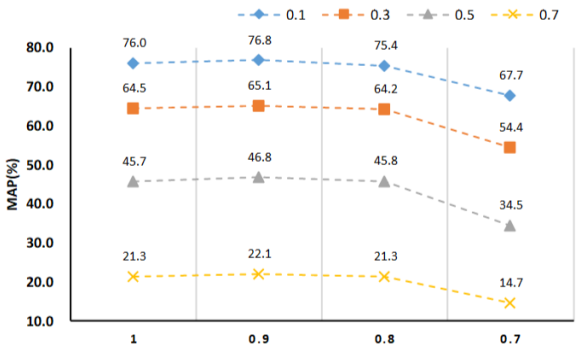


Fig. 3 The impact of annotated single-frame image information on localization performance on the THUMOS 14 dataset.

4.4.4 Complexity analysis for the clustering algorithm in the model

In our method, on the basis of the mainline, annotated single-frame image information is introduced as guidance to enhance action discrimination, and we adopt the K-means clustering algorithm to obtain category-specific action features. The discriminative representations of the action branch and the mainline branch are parallel. Therefore, the number of parameters of our model is unchanged when compared with LACP.

In the discriminative representation for the ac-

unedited

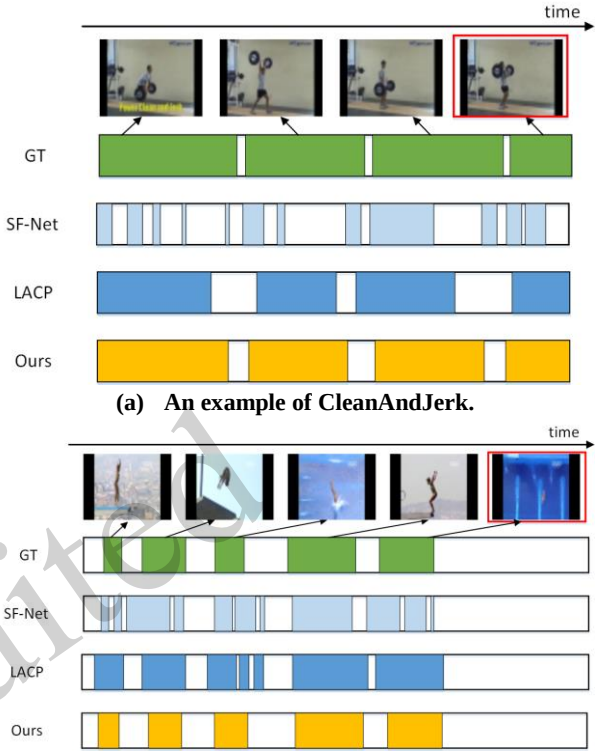
tion of the lower branch of Fig. 1, the K-means clustering algorithm is adopted, and the time complexity of discriminative representation for the action branch is $O(C*N_c*D*k)$, in which C represents the number of action categories, N_c represents the number of annotated frames for the c-th action category, and $C*N_c$ represents the number of annotated single-frame images. D represents the feature dimension of the annotated single-frame image, and k represents the number of iterations. Usually, the number of action categories C, the feature dimension D, and the number of iterations k are constants. Therefore, the time complexity for the clustering algorithm in our model is linearly related to N_c.

The convergence of our method requires a negligible extra time compared to LACP because we only have a fusion mechanism with single-frame image information. However, the accuracy of our detection performance is improved at all t-IoU thresholds.

4.5 Qualitative comparison

We conduct qualitative analysis on the THUMOS 14 dataset to obtain an intuitive understanding of our dual-branch network. We compare our method with SF-Net (Ma et al., 2020) and LACP (Lee et al., 2021), which are the first and last methods of single-frame annotations in Fig. 4. We provide two action examples with different action categories, namely, (a) *CleanAndJerk* and (b) *Diving*. The horizontal axes denote the timestamps of videos. The red boxes indicate that the action frames are not detected by SF-Net but are detected by our method and LACP. This shows that simply mining pseudo action frames and pseudo background frames is prone to generate false positives (FPs) and false negatives (FNs).

From Fig 4(a) and (b), we can see that SF-Net (Ma et al., 2020) produces many fragmentary action examples, while the action examples detected by our method and LACP (Lee et al., 2021) are relatively complete. This is because LACP employs two loss functions to model action completeness. Additionally, the predictions produced by our method are much closer to GT. This is because our proposed method uses action discrimination enhancement via category-specific frame clustering, and we can find that the detected action instances are more complete and accurate. This indicates that high-quality single-frame annotation information assists in generating discriminative class activation sequences, resulting in accurate action instances. This further verifies the feasibility of our method.



(a) An example of CleanAndJerk.

(b) An example of Diving.

Fig. 4 Qualitative comparisons with SF-Net (Ma et al., 2020) and LACP (Lee et al., 2021) on the THUMOS 14 dataset.

unedited

5 Conclusion

A novel method to enhance action discrimination via category-specific frame clustering is proposed for W-TAL. Specifically, we make full use of annotated frames in the video to enhance the action discriminative representation. Since all the annotated frames in the video are discriminative, we cluster the representative annotated frames from the same category to obtain category-specific representations. The single frame-level class activation score is generated by calculating the similarities between the frame and various categories. Then, a convex combination fusion mechanism between annotated frames and video snippet sequences is presented to ensure the consistency of action discrimination for generating a robust class activation sequence. Experiments conducted on three popular datasets validate that sin-

gle-frame image modelling can provide complementary guidance information to the video snippet sequence, and our method outperforms state-of-the-art methods. Our method can be effectively applied to untrimmed/trimmed videos, which have the same action categories and similar scenes. When the scene changes greatly and the action categories are different, the model needs to be retrained before it is applied. In the future, we will carry out research on cross-domain action localization to improve the generalizability of our method, which may have more real-world applications.

Contributors

Huifen XIA and Yongzhao ZHAN designed the research. Honglin LIU gave some theoretical guidance. Xiaopeng REN trained the model and processed the data. Huifen XIA drafted the paper. Yongzhao ZHAN revised the paper.

Compliance with ethics guidelines

Huifen XIA, Yongzhao ZHAN, Honglin LIU and Xiaopeng REN declare that they have no conflict of interest.

References

Bojanowski P, Bach F, Laptev I, et al., 2013. Finding actors and actions in movies. IEEE international conference on computer vision, p.2280-2287. https://doi.org/10.1109/ICCV.2013.283

Bojanowski P, Lajugie R, Bach F, et al., 2014. Weakly supervised action labeling in videos under ordering constraints. European Conference on Computer Vision. Springer, Cham, p.628-643. https://doi.org/10.1007/978-3-319-10602-1_41

Carreira J, Zisserman A, 2017. Quo vadis, action recognition? a new model and the kinetics dataset. IEEE Conference on Computer Vision and Pattern Recognition, p.6299-6308. https://doi.org/10.1109/CVPR.2017.502

Chao YW, Vijayanarasimhan S, Seybold B, et al., 2018. Rethinking the faster r-cnn architecture for temporal action localization. IEEE conference on computer vision and pattern recognition, p.1130-1139. https://doi.org/10.1109/CVPR.2018.00124

Chen ZY, Liu H, Zhang LL, et al., 2022. Multi-Dimensional Attention with Similarity Constraint for Weakly-Supervised Temporal Action Localization. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2022.3174344

Damen D, Leelasawassuk T, Haines O, et al., 2014. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. BMVC. volume2, p.3. https://dx.doi.org/10.5244/C.28.30

Gan C, Sun C, Duan LX, et al., 2016. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. European Conference on Computer Vision. Springer, Cham, p.849-866. https:/doi.org/10.1007/978-3-319-46487-9_52

Gao JY, Chen MY, Xu CS, 2022. Fine-grained Temporal Contrastive Learning for Weakly-supervised Temporal Action Localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.19999-20009. https://doi.org/10.1109/CVPR52688.2022.01937

Ge YX, Qin XL, Yang D, et al., 2021. Deep snippet selective network for weakly supervised temporal action localization. Pattern Recognition,110: p.107686. https://doi.org/10.1016/j.patcog.2020.107686

Huang DA, Li FF, Niebles JC, 2016. Connectionist temporal modeling for weakly supervised action labeling. European Conference on Computer Vision. Springer, Cham, p.137-153. https://doi.org/10.1007/978-3-319-46493-0_9

Huang LJ, Wang L, Li HS, 2022a. Weakly Supervised Temporal Action Localization via Representative Snippet Knowledge Propagation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.3272-3281. https://doi.org/10.1109/CVPR52688.2022.00327

Huang LJ, Wang L, Li HS, 2022b. Multi-modality self-distillation for weakly supervised temporal action localization. IEEE Transactions on Image Processing, 31: p.1504-1519. https://doi.org/10.1109/TIP.2021.3137649

Jiang YG, Liu J, Roshan Zamir A, et al., 2014. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14

Ju C, Zhao PS, Chen SH, et al., 2021. Divide and conquer for single-frame temporal action localization. IEEE/CVF International Conference on Computer Vision. P.13455-13464. https://doi.org/10.1109/ICCV48922.2021.01320

Kay W, Carreira J, Simonyan K, et al., 2017. The kinetics human action video dataset. https://arxiv.org/abs/ 1705.06950

Kingma D P, Ba J, 2014. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980

Kumar Singh K, Jae Lee Y, 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. IEEE International Conference on Computer Vision, p.3524-3533. https://doi.org/10.1109/ICCV.2017.381

Lee P, Byun H, 2021. Learning action completeness from points for weakly-supervised temporal action localization. IEEE/CVF International Conference on Computer Vision, p.13648-13657. https://doi.org/10.1109/ICCV48922.2021.01339

unedited

Lee P, Uh Y, Byun H, 2020. Background suppression network for weakly-supervised temporal action localization. Proceedings of the AAAI conference on artificial intelligence., 34(07): p.11320-11327. https://doi.org/10.1609/aaai.v34i07.6793

Lei P, Todorovic S, 2018. Temporal deformable residual networks for action segmentation in videos. IEEE conference on computer vision and pattern recognition. p.6742-6751. https://doi.org/10.1109/CVPR.2018.00705

Liao YG, Qiu CZ, Zhang ZY, et al., 2021. GCRNet: Global Context Relation Network for Weakly-Supervised Temporal Action Localization: Identify the target actions in a long untrimmed video and find the corresponding action start point and end point. The 5th International Conference on Video and Image Processing, p.184-190. https://doi.org/10.1145/3511176.3511204

Lin CM, Xu CM, Luo DH, et al., 2021. Learning salient boundary feature for anchor-free temporal action localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.3320-3329. https://doi.org/10.1109/CVPR46437.2021.00333

Lin TW, Zhao X, Shou Z, 2017. Single shot temporal action detection. Proceedings of the 25th ACM international conference on Multimedia, p.988-996. https://doi.org/10.1145/3123266.3123343

Lin TY, Goyal P, Girshick R, et al., 2017. Focal loss for dense object detection. IEEE international conference on computer vision, p.2980-2988. https://doi.org/10.1109/TPAMI.2018.2858826

Liu DC, Jiang TT, Wang YZ, 2019. Completeness modeling and context separation for weakly supervised temporal action localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.1298-1307. https://doi.org/10.1109/CVPR.2019.00139

Long FC, Yao T, Qiu ZF, et al., 2019. Gaussian temporal awareness networks for action localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 344-353.
https://doi.org/10.1109/CVPR.2019.00043

Ma F, Zhu LC, Yang Y, et al., 2020. Sf-net: Single-frame supervision for temporal action localization. European conference on computer vision. Springer, Cham, p.420-437. https://doi.org/10.1007/978-3-030-58548-8_25

Moltisanti D, Fidler S, Damen D, 2019. Action recognition from single timestamp supervision in untrimmed videos. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.9915-9924. https://doi.org/10.1109/CVPR.2019.01015

Narayan S, Cholakkal H, Khan F S, et al, 2019. 3c-net: Category count and center loss for weakly-supervised action localization. IEEE/CVF International Conference on Computer Vision, p. 8679-8687. https://doi.org/10.1109/ICCV.2019.00877

Nguyen P, Liu T, Prasad G, et al., 2018. Weakly supervised action localization by sparse temporal pooling network. IEEE Conference on Computer Vision and Pattern Recognition, p. 6752-6761. https://doi.org/10.1109/CVPR.2018.00706

Nguyen P X, Ramanan D, Fowlkes C C, 2019. Weakly-supervised action localization with background modeling. IEEE/CVF International Conference on Computer Vision, p.5502-5511. https://doi.org/10.1109/ICCV.2019.00560

Paul S, Roy S, Roy-Chowdhury A K, 2018. W-talc: Weakly-supervised temporal activity localization and classification. Proceedings of the European Conference on Computer Vision, p. 563-579. https://doi.org/10.1007/978-3-030-01225-0_35

Shi BF, Dai Q, Mu YD, et al., 2020. Weakly-supervised action localization by generative attention modeling. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.1009-1019. https://doi.org/10.1109/CVPR42600.2020.00109

Shou Z, Gao H, Zhang L, et al., 2018. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. Proceedings of the European Conference on Computer Vision, p.154-171. https://doi.org/10.1007/978-3-030-01270-0_10

Sultani W, Chen C, Shah M, 2018. Real-world anomaly detection in surveillance videos. IEEE conference on computer vision and pattern recognition, p. 6479-6488. https://doi.org/10.1109/CVPR.2018.00678

Tong Z, Song YB, Wang J, et al., 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. https://arxiv.org/abs/2203.12602

Wang, LM, Xiong, YJ, Lin, DH, et al, 2017. Untrimmednets for weakly supervised action recognition and detection. IEEE conference on Computer Vision and Pattern Recognition, p.4325-4334. https://doi.org/10.1109/CVPR.2017.678

Wedel A, Pock T, Zach C, et al., 2009. An improved algorithm for tv-l 1 optical flow. Statistical and geometrical approaches to visual motion analysis. Springer, Berlin, Heidelberg, p.23-45. https://doi.org/10.1007/978-3-642-03061-1_2

Xu MM, Zhao C, Rojas D S, et al., 2020. G-tad: Sub-graph localization for temporal action detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.10156-10165. https://doi.org/10.1109/CVPR42600.2020.01017

Yang L, Han JW, Zhao T, et al., 2021b. Background-click supervision for temporal action localization. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2021.3132058

Yang WF, Zhang TZ, Yu XY, et al., 2021a. Uncertainty guided collaborative training for weakly supervised temporal action detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.53-63.

unedited

https://doi.org/10.1109/CVPR46437.2021.00012

Yang Y, Zhuang Y, Pan Y, 2021. Multiple knowledge representation for big data artificial intelligence: Framework, applications, and case studies. Frontiers of Information Technology & Electronic Engineering, 22(12):1551-1558. https://doi.org/10.1631/fitee.2100463

Zeng RH, Huang WB, Tan MK, et al., 2019. Graph convolutional networks for temporal action localization. IEEE/CVF International Conference on Computer Vision, p.7094-7103. https://doi.org/10.1109/ICCV.2019.00719

Zhai YH, Wang L, Tang W, et al., 2020. Two-stream consensus network for weakly-supervised temporal action localization. European conference on computer vision. Springer, Cham, p.37-54. https://doi.org/10.1007/978-3-030-58539-6_3

Zhang CW, Xu YL, Cheng ZZ, et al., 2019. Adversarial seeded sequence growing for weakly-supervised temporal action localization. Proceedings of the 27th ACM international conference on multimedia, p.738-746. https://doi.org/10.1145/3343031.3351044

Zhao Y, Xiong YJ, Wang LM, et al., 2017. Temporal action detection with structured segment networks. IEEE International Conference on Computer Vision., p.2914-2923. https://doi.org/10.1109/ICCV.2017.317

Zhou H, Zhan YZ, Mao QR, 2021. Video anomaly detection based on space-time fusion graph network learning. J. Comput. Res. Dev. 58(1), 48-59. https://crad.ict.ac.cn/EN/10.7544

Zhu LC, Fan HH, Luo YW, et al., 2022. Temporal Cross-Layer Correlation Mining for Action Recognition. IEEE Transactions on Multimedia, p. 668-676. http://10.1109/tmm.2021.3057503

unedited