Transferable Knowledge-Based Multi-Granularity Fusion Network for Weakly Supervised Temporal Action Detection

Haisheng Su, Xu Zhao, *Member, IEEE*, Tianwei Lin, Shuming Liu, and Zhilan Hu

***Abstract*—Despite remarkable progress, temporal action detection is still limited for real application due to the great amount of manual annotations. This issue motivates interest in addressing this task under weak supervision, namely, locating the action instances using only video-level class labels. Many current works on this task are mainly based on the Class Activation Sequence (CAS), which is generated by the video classification network to describe the probability of each snippet being in a specific action class of the video. However, the CAS generated by a simple classification network can only focus on local discriminative parts instead of locating the entire interval of target actions. In this paper, we present a novel framework to handle this issue. Specifically, we propose to utilize convolutional kernels with varied dilation rates to enlarge the receptive fields, which can transfer the discriminative information to the surrounding non-discriminative regions. Then, we design a cascaded module with the proposed Online Adversarial Erasing (OAE) mechanism to further mine more relevant regions of target actions by feeding the erased-feature maps of discovered regions back into the system. In addition, inspired by the transfer learning method, we adopt an additional module to transfer the knowledge from trimmed videos to untrimmed videos to promote the classification performance on untrimmed videos. Finally, we employ a boundary regression module embedded with Outer-Inner-Contrastive (OIC) loss to automatically predict the boundaries based on the enhanced CAS. Extensive experiments are conducted on two challenging datasets, THUMOS14 and ActivityNet-1.3, and the experimental results clearly demonstrate the superiority of our unified framework.**

***Index Terms*—Boundary regression, cascaded dilated classification block, class activation sequence, knowledge transfer, temporal action detection, weak supervision.**

Manuscript received August 18, 2019; revised March 10, 2020 and May 17, 2020; accepted May 18, 2020. Date of publication June 1, 2020; date of current version May 26, 2021. This work was supported in part by NSFC (61673269, 61273285) and in part by the project funding of the Institute of Medical Robotics at Shanghai Jiao Tong University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. M. Murshed. *(Corresponding author: Xu Zhao.)*

Haisheng Su, Tianwei Lin, and Shuming Liu are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: suhaisheng@sjtu.edu.cn; wzmsltw@sjtu.edu.cn; shumingliu@sjtu.edu.cn).

Xu Zhao is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaoxu@sjtu.edu.cn).

Zhilan Hu is with the Central Media Technology Institute of Huawei Co. Ltd., Shenzhen 518129, China (e-mail: huzhilan@huawei.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2999184

I. Introduction

With the increasing development of computer vision and great amount of media resources, intelligent video content analysis has attracted much attention from many researchers in recent years. Videos in realistic life are usually long and untrimmed and may contain multiple action instances with arbitrary durations. Thus, there is an important yet challenging task for video analysis: temporal action detection, which requires one to accurately classify the untrimmed videos into specific categories and precisely locate the temporal boundaries of action instances. Although substantial progress has been achieved in this task [1]–[8], it remains limited for industrial applications due to the huge amount of temporal annotations used for training such a deep-learning-based model in a fully supervised manner, which are labor intensive to annotate especially for a large-scale dataset. In contrast, weak labels such as video-level labels, are much easier to obtain, so many current studies attempt to handle this problem under weak supervision.

Analogous to Weakly Supervised Object Detection (WSOD) in images, Weakly Supervised Temporal Action Detection (WSTAD) can be considered a temporal version of WSOD for locating action instances using only video-level class labels. However, WSTAD is much more challenging than WSOD because of the larger video content variation and uncertain temporal length of action instances. A prevalent practice of WSTAD adopts the idea [9] to generate a 1-D Class Activation Sequence (CAS) to highlight the discriminative regions that contribute most to the video classification results, which was originally used for locating object in images. Nevertheless, a high-quality CAS for temporally locating the action boundaries should possess two properties: (1) the CAS can completely cover the temporal interval of target actions; (2) the CAS can densely locate the action instances with fewer missing detections.

To generate a high-quality CAS, many recent methods adopt the "*localization by classification*" framework with different improvement strategies. Singh *et al.* [10] present a data augmentation method by randomly hiding regions during the training phase to force the classification network to look for other discriminative areas. Wei *et al.* [11] employ the adversarial erasing method to discover more relevant regions by iteratively training the classifiers with discriminative region erasing of different degrees. In addition, Nguyen *et al.* [12] introduce the attention mechanism to distinguish between the action and the

background. However, these works on achieving a better-quality CAS may have some drawbacks: (1) without effective guidance, the random hiding strategy only makes sense in images, not in videos; (2) training of a classification network with iterative region erasing is somewhat impractical and inefficient; and (3) the CAS fails to densely highlight the action instances in the video, causing many missing detections.

To address these issues and effectively locate the action instances under weak supervision, we propose the Multi-Granularity Fusion Network (MGFN), which first adopts a cascaded dilated classification block to enhance the quality of the CAS; then, it employs a boundary regression module to directly predict the temporal boundaries of action instances based on the CAS. To generate a high-quality CAS, our cascaded dilated classification block utilizes two main sub-modules to achieve this goal. Specifically, the multi-dilated convolution module augments the simple classification network with multiple convolutional kernels of different dilation rates to transfer the discriminative information of initial seeds to the surrounding non-discriminative regions, which expands the visible areas of the classification network. Then, the cascaded classification module adapts two classifiers with identical architecture to further discover other potential action regions using an Online Adversarial Erasing (OAE) mechanism that do not appear in the initial localization sequence. With this mechanism, the feature maps of discriminative regions discovered in the first stage are dynamically erased from the video sequence, which is subsequently fed into the second-stage classifier for further mining. Unlike previous methods, our approach is more efficient and intuitive. To further improve the CAS quality, we incorporate the transfer learning idea and learn transferable knowledge between trimmed videos and untrimmed videos to promote the classification performance on untrimmed videos. Finally, instead of performing temporal action detection by directly thresholding the CAS, which may not be robust to noises in the CAS, we adopt a boundary regression module to predict the boundaries. For the segment-level supervision used for boundary regression, the Outer-Inner-Contrastive (OIC) loss is employed. The entire framework is optimized in an end-to-end fashion.

It is non-trivial to effectively and efficiently enhance the quality of the CAS since the CAS generated by the simple classification network can only focus on local discriminative parts, making it inferior and not qualified for the temporal action detection task. Hence, to achieve a good performance on this task, a high-quality CAS is a prerequisite. For the sake of end-to-end optimization, our OAE mechanism is time efficient and only needs to train a model for the entire region mining. Therefore, with the integrated training process, the augmented classification network based on transferable knowledge and the boundary regression module can collaborate with each other to achieve a better performance.

The main contributions of our work are four-fold:

1) To the best of our knowledge, we are the first to incorporate high-quality CAS generation and boundary regression into a unified framework for weakly supervised temporal action detection, which can coarsely locate the complete regions of actions in the videos and finely predict the temporal boundaries through regression.
2) We introduce two effective yet efficient modules to enhance the quality of the CAS, where the multi-dilated convolution module is used to expand the local discriminative regions, while the cascaded classification module is used to further locate other potential action instances.
3) A knowledge transfer module is introduced to learn transferable knowledge between untrimmed videos and trimmed videos to promote the classification performance on untrimmed videos.
4) Extensive experiments demonstrate that our approach achieves state-of-the-art performance on both the THUMOS14 and ActivityNet-1.3 datasets.

The below content is organized as follows. Section II summarizes the related work on action recognition, weakly supervised object detection and weakly supervised temporal action detection. Section III explains our proposed method in detail. Section IV presents the related experiments results. Section V concludes our work.

II. Related Work

A. Action Recognition

Action recognition is an essential branch of video content analysis, which aims to classify manually trimmed videos into specific categories and has been extensively explored in recent years [13]–[16]. Earlier methods such as improved Dense Trajectory (iDT) [17], [18], which mainly adopt the extracted hand-crafted features, with the trajectories including HOG, HOF and MBH, have won a leading place. With the rapid development of deep learning, tremendous progress [19]–[22] has been made in this field. Typically, a two-stream network [19], [20], [21] is utilized to learn both appearance feature and motion information through two branches, which are based on the RGB frame and stacked optical flow field, respectively. The C3D network [22] directly captures the spatial and temporal information from raw videos using a series of 3D convolutional kernels. These action recognition networks are usually adopted to extract visual features of long and untrimmed videos at the snippet level.

B. Weakly Supervised Object Detection

Weakly supervised object detection aims to perform object detection using only image-level labels. Bottom-up [23]–[25] and top-down [9]–[11], [26], [27] mechanisms are two main streams in current works. In the bottom-up approaches, candidate proposals are first extracted using selective search [28] or edge boxes [29], which are subsequently fed into the deep convolutional neural networks for classification, and the scores from all proposals are merged together to match the image-level labels.

In the top-down approaches, Zhou *et al.* [9] and Zhang *et al.* [26] first explore the relationship between the image classification results and the neural responses of image regions, and

they subsequently locate the areas of high activations as detection results. Meanwhile, Zhou *et al.* validate the localization ability of a Global Average Pooling (GAP) layer in image classification networks. Singh *et al.* [10] propose to improve the quality of Class Activation Mapping (CAM) by randomly hiding some patches in images during the training phase so that the classification network is forced to look for other discriminative regions. However, this data augmentation strategy without effective guidance is blind and inefficient. Wei *et al.* [11] introduce the Adversarial Erasing (AE) idea, which repeatedly trains the classification networks with the discriminative regions iteratively erased. However, this idea requires the networks to be trained several times, which is time consuming and impractical. In our work, we propose the Online Adversarial Erasing (OAE) method, which combines two classifiers of the same structure in a cascaded manner, and the video features of discriminative regions discovered in the first classifier are dynamically and automatically erased from the video sequence for the second classifier. In addition, Wei *et al.* [30] utilize dilated convolution to transfer the surrounding discriminative information to non-discriminative object regions in images, thus promoting the emergence of these regions in the localization maps to be more integral, but they still neglect some missing regions of interest that do not appear in the initial localization sequence. We expand this idea to the temporal field and equip the two-stage video classifiers with 1-D dilated convolutional kernels of varied dilation rates. Hence, more discriminative regions can be mined, and a high-quality dense localization map can be generated in an end-to-end framework.

C. Weakly Supervised Temporal Action Detection

As a counterpart of weakly supervised image object detection, the goal of weakly supervised temporal action detection is to locate action instances in untrimmed videos including temporal boundaries and action categories by relying only on video-level class labels. Similar to the idea in [23], Wang *et al.* [31] first generate temporal action proposals with the priors of an action shot and subsequently adopt a classification module and a selection module to perform action classification and important segment selection. However, the use of the softmax function across proposals prevents this approach from distinguishing multiple action instances in the same video. Singh *et al.* [10] also implement this idea on this task, but it makes little sense due to the complexity and varied lengths of videos. Sujoy *et al.* [32] propose a co-activity loss to train a weakly supervised network. All the localization parts of these methods are based on thresholding on the final Class Activation Sequence (CAS), which causes inaccurate boundaries of detections. Instead of applying a simple threshold on the Class Activation Sequence (CAS) to directly perform action localization, Shou *et al.* [33] propose the Outer-Inner-Contrastive (OIC) loss to provide segment-level supervision for training a boundary regressor via the anchor mechanism, which is intuitive and effective. However, AutoLoc directly adopts a pretrained video classifier to generate the CAS for boundary prediction, and the CAS quality is inferior without improvement. In our work, we incorporate a high-quality CAS generation and boundary prediction process into a unified network for collaborative optimization.

III. Our Approach

A. Problem Definition

We denote an untrimmed video as $\mathbf{X}_v = \{x_t\}_{t=1}^{l_v}$, where l_v is the number of frames, and x_t is the t-th frame in $\mathbf{X}_v$. Each video $\mathbf{X}_v$ is annotated with a set of temporal action instances $\mathbf{\Phi}_v = \{\phi_n = (t_n^s, t_n^e, \varphi_n)\}_{n=1}^{N_v}$, where N_v is the number of temporal action instances in $\mathbf{X}_v$, and t_n^s, t_n^e, and φ_n are the starting time, ending time and category of instance ϕ_n, respectively, with $\varphi_n \in \{1, \ldots, K\}$, where K is the number of action categories. During the training phase, only the video-level action label set $\mathbf{\Psi}_v = \{\varphi_n\}_{n=1}^{N_v}$ is given; during the test phase, $\mathbf{\Phi}_v$ must be predicted.

B. Video Feature Encoding

To apply our algorithm, first, the feature representations must be extracted to describe the visual content of the input video in our work. In this paper, the prevalent pretrained architectures, namely, UntrimmedNet [31], is employed for feature extraction since this type of architecture using multiple two-stream networks has shown great performance and its use has become a prevalent practice in action recognition and temporal action localization tasks. A two-stream network contains two branches: the spatial network operates on a single RGB frame to capture the appearance feature, and the temporal network handles the stacked optical flow field to capture motion information.

Given a video containing l_v frames, the video unit is used as the basic processing unit in our framework for computational efficiency. Hence, the video is divided into l_v/n_u consecutive units without overlap, where n_u is the frame number of a unit. Then, we compose a unit sequence $\mathbf{U} = \{\mathbf{u}_j\}_{j=1}^{l_u}$ from video $\mathbf{X}_v$, where l_u is the number of units. A video unit can be represented as $\mathbf{u}_j = \{x_t\}_{t=f_s}^{f_s+n_u}$, where f_s is the starting frame, and $f_s + n_u$ is the ending frame. Each unit is fed into the pretrained visual encoder to extract the representation. Concretely, the center RGB frame inside a unit is processed by the spatial network, and the stacked optical flow derived around the center frame is processed by the temporal network; then, we concatenate output scores of video feature encoders in the fc-action layer to form the feature vector $\mathbf{f}_{u_j} = \{\mathbf{f}_{S,u_j}, \mathbf{f}_{T,u_j}\}$, where $\mathbf{f}_{S,u_j}$ and $\mathbf{f}_{T,u_j}$ are the output scores of the spatial and temporal networks, respectively, with length G. Finally, the unit-level feature sequence $\mathbf{F} = \{\mathbf{f}_{u_j}\}_{j=1}^{l_u}$ is used as the input of our MGFN.

C. Multi-Granularity Fusion Network

We propose a novel architecture to effectively detect the action instances with entire temporal regions and accurate boundaries under weak supervision. As shown in Fig. 1, we design a cascaded dilated classification block to mine more relevant regions of target actions by implementing convolutional kernels of varied dilation rates and cascaded mechanism, which enhances the quality of the Class Activation Sequence (CAS). Instead of locating temporal action instances by directly applying a simple thresholding method, which is not robust to the noise of the CAS,

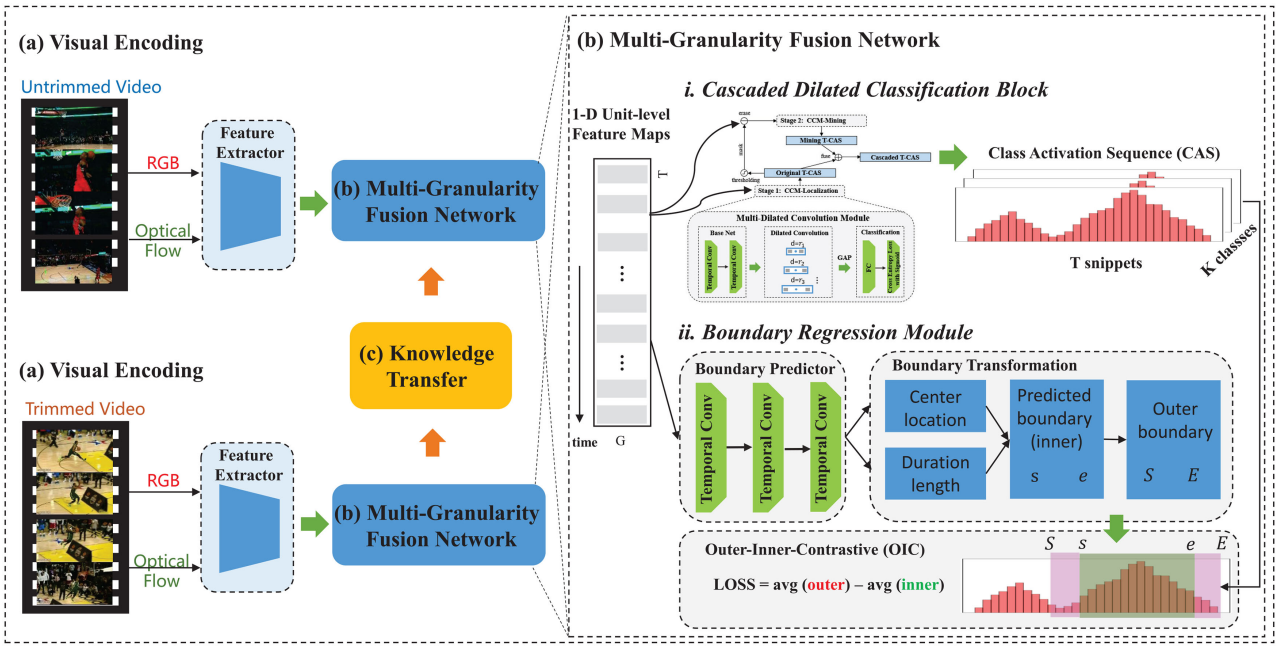


Fig. 1. The framework of our proposed method. (a) A two-stream network is used to encode video features at the snippet level for our algorithm to perform action recognition and temporal boundary prediction under weak supervision. (b) Architecture of the Multi-Granularity Fusion Network: (i) the *cascaded dilated classification block* handles the extracted visual features as input for video classification by adopting the multi-dilated convolution module and cascaded classification module to discover entire class-specific temporal regions; (ii) the *boundary regression module* handles the input features to directly predict the boundary, which is embedded with the Outer-Inner-Contrastive loss to optimize the boundary regressor based on the generated CAS. (c) Finally, we transfer knowledge from trimmed videos to untrimmed videos to promote the classification performance on untrimmed videos and further enhance the quality of the Class Activation Sequence (CAS).

we adopt a boundary regression module to automatically regress to the accurate boundaries using the Outer-Inner-Contrastive (OIC) loss. In addition, since the classification performance of a pretrained video classifier is bound to decrease on untrimmed videos due to the existence of background noise, we introduce the transfer learning mechanism and learn transferable knowledge between trimmed videos and untrimmed videos to promote the performance of the classification network.

Network Architecture: The architecture of our MGFN is illustrated in Fig. 1, which mainly contains two parts: the cascaded dilated classification block and boundary regressor. As shown in Fig. 2, the cascaded dilated classification block includes two sub-modules: the multi-dilated convolution module and cascaded classification module. The *multi-dilated convolution module* is designed to augment the simple classification network by gradually enlarging the receptive field of kernels with multi-dilated convolution branches, which can effectively incorporate the surrounding context and transfer the semantic information from discriminative regions to non-discriminative regions to expand the highlighted areas related to the actions. Then, the *cascaded classification module* is a two-stage model that combines the two multi-dilated convolution modules of the same structure with an online adversarial erasing method to discover more relevant regions of target actions and generate a CAS of high quality. Based on the enhanced CAS, we adopt a *boundary regression module* to directly predict the segment boundary via the anchor mechanism; then, we inflate the inner segment boundary to obtain the outer segment boundary and

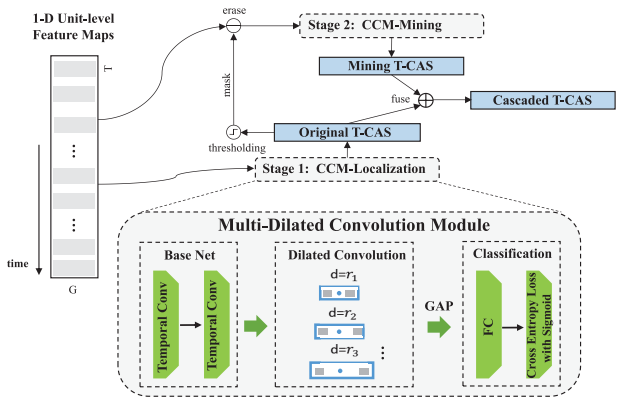


Fig. 2. Architecture of the cascaded dilated classification block. The extracted unit-level video features are concurrently fed into the two-stage classifiers for video classification. For each stage, a standard classification network equipped with multiple dilated convolutional branches of varied dilation rates is used to generate a dense localization sequence. Then, the two-stage classifiers are combined using an online adversarial erasing mechanism, where the video features of the discriminative regions highlighted in the first-stage localization sequence are dynamically erased from the video sequence for the second-stage classifier to prompt the second classifier to leverage other related regions of target actions. Finally, the localization sequences generated in the two stages are fused to obtain a better quality.

optimize the regressor with the OIC loss to provide segment-level supervision.

Multi-Dilated Convolution Module: The goal of this module is to augment the simple classification network with dilated

convolution. Since the Class Activation Sequence (CAS) generated by a simple classification network can only highlight discriminative regions of target actions, which is not qualified for the temporal action detection task, we introduce the dilated convolution, which is promising for incorporating the surrounding context. By enlarging the receptive field of convolutional kernels with varied dilation rates, the semantic information for supporting the classification result can be transferred from the initially discriminative regions to other surrounding regions, thus enhancing their discriminativeness. Fig. 4 illustrates how the dilation step enables the information to be transferred with the temporal dimension.

Specifically, first, we adopt two temporal convolutional layers, which serve as the base net to handle the temporal information of the input video feature sequence; then, multi-dilated convolutional branches with varied dilation rates (*i.e.,* $d = r_i, i = 1, \ldots, k$) are appended to the base net to discover action-relevant temporal regions perceived by varied receptive fields. After the Global Average Pooling (GAP) layer, the pooled representations are further passed through a fully connected layer for classification. Then, we optimize the augmented classification network with the sigmoid cross-entropy loss and produce the class-specific localization sequence for each branch. Finally, we obtain the dense Class Activation Sequence (CAS) by fusing the localization sequences from multiple branches.

Specifically, our dilated convolution module mainly includes two types of operation. 1) Standard convolutional kernels with dilation rate $d = 1$ are employed to generate the original localization sequence H_0, where discriminate regions are effectively highlighted despite some missing true positive regions. 2) Convolutional kernels with varied dilation rates are employed to expand the discriminative information to surrounding areas. However, we observe that when the receptive field of kernels is set too large, this will also introduce some true negative temporal regions. Hence, we choose small dilation rates (*i.e.,* $d =$ 2, 3, 5) in this paper. The final localization sequence H for temporal action region generation is subsequently fused with $\mathbf{H} = \mathbf{H}_0 + \frac{1}{n_d}\sum_{i=1}^{n_d} \mathbf{H}_i$, where n_d is the number of dilated convolution branches.

Cascaded Classification Module: This module aims to further mine more relevant regions of target actions. Although the dilated convolution can be used to expand the initially discriminative regions to be more integral, it fails to locate other regions of interest that do not appear in the initial localization sequence. To further promote the quality of the CAS, we adopt a cascaded mechanism with the Online Adversarial Erasing (OAE) step to enforce two classification network of the same architecture to locate different but complementary regions of target actions.

In the first stage, the dilated convolution module handles the video feature sequence as the input to generate the localization sequence $\mathbf{H}$. Then, we apply a threshold on $\mathbf{H}$ to generate a mask, which represents the discriminative regions detected by the first classifier. Next, we use this mask to erase the input video feature maps from the video sequence, which is subsequently fed into the second stage. Thus, the second classifier with the erased-input-feature maps is forced to discover other action-related regions to support the video-level class labels.

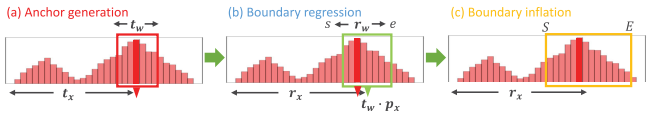


Fig. 3. Illustration of the boundary prediction procedure, which consists of three steps sequentially: (a) **anchor generation** to obtain the predefined boundary hypothesis; (b) **boundary regression** to obtain the predicted boundary of the action segment (denoted as the inner boundary); and (c) **boundary inflation** to obtain the outer boundary for OIC loss implementation.

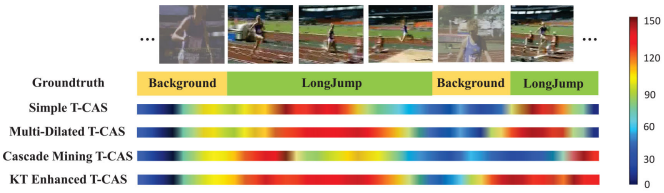


Fig. 4. Illustration of the comparison of ground-truth temporal intervals, simple T-CAS, multi-dilated T-CAS, cascade mining T-CAS and knowledge transfer enhanced T-CAS for the *LongJump* action class.

Concretely, the second classifier will generate a new initial seed and use dilated convolution to expand it. Finally, we integrate the two generated Temporal Class Activation Sequences (T-CASs), $\mathbf{H}$ and $\bar{\mathbf{H}}$ to form the cascaded localization sequence $\mathbf{H}_t^k(Cas) = \max\{\mathbf{H}_t^k, \bar{\mathbf{H}}_t^k\}$, where $\mathbf{H}_t^k(Cas)$ indicates the t-th element in the cascaded localization sequence of class k.

Boundary Regression Module: The goal of this module is to learn to directly predict the segment boundary based on the enhanced CAS obtained above. The multi-anchor mechanism has shown great effectiveness in the fully supervised temporal action detection task, which generates the detections by constantly regressing the predefined multi-scale anchors at each temporal position, including the center location and temporal length. However, without temporal annotations for the weakly supervised counterpart, it is vital to leverage other priors to provide segment-level supervision. Based on the idea in [33], we employ the Outer-Inner-Contrastive (OIC) loss to optimize the predictor. The boundary prediction procedure is illustrated in Fig. 3.

Specifically, given the encoded video features as the input, a boundary predictor first stacks three same temporal convolutional layers to handle the temporal information. Each temporal convolutional layer has the same configurations: 128 filters, kernel size 3, and stride 1 with ReLu activation. Then, another temporal convolutional layer with $2M$ filters, kernel size 3 and stride 1 is employed to predict the boundary regression values p_x and p_w for each position, where M is the number of anchor scales. For anchor generation, we denote t_x and t_w as the temporal position and temporal length of each anchor in the output feature maps, and each cell of output feature maps in the prediction layer is associated with multi-scale anchors. Then, for each anchor at location t_x, we use the output regression values to adjust the segment, where the center localization is $r_x = t_x + t_w \cdot p_x$, and the temporal length is $r_w = t_w \cdot exp(p_w)$; hence, the predicted inner boundary can be computed by $s = r_x - r_w/2$ and $e = r_x + r_w/2$. Then, to implement the OIC loss, we inflate

the inner boundary by a ratio γ to obtain the outer boundary $S = s - r_w \cdot \gamma$ and $E = e + r_w \cdot \gamma$.

The OIC loss is introduced to measure how likely the anchor covers the actions and subsequently discard the negative segments. Concretely, it can be denoted as the average activations of the outer red area minus the average activations of the inner green area among the enhanced CAS obtained previously.

D. Transfer Learning Mechanism

Transfer learning technology has been widely explored in modeling the shifts of data distributions across different domains [34], [35]. The performance of video encoders pretrained on trimmed videos is bound to decrease on untrimmed videos due to the great amount of background noise, which greatly affects the quality of the CAS. To promote the classification performance of untrimmed videos and make full use of the large-scale trimmed video datasets, we introduce the transfer learning mechanism to learn transferable knowledge between trimmed videos and untrimmed videos. Since the trimmed videos are precisely annotated with the action of interest, decisive clues in high layers ($i.e.$, the classification layer) can be utilized for action recognition even in untrimmed videos. As shown in Fig. 1, we take the trimmed branch as the source branch and the untrimmed branch as the target branch. Then, we mine informative knowledge from the trimmed videos to improve the performance on untrimmed videos via knowledge transfer.

For the trimmed branch, we use settings identical to those for the untrimmed one. Specifically, trimmed videos are also fed into the two-stream network for feature extraction; then, the visual feature sequence serves as the input of our framework. Unlike the untrimmed branch, since the trimmed videos are well segmented, the boundary regression module is no longer necessary. As a result, the trimmed branch is trained by minimizing the classification loss on the trimmed video dataset the same as for the untrimmed branch. After the trimmed branch converges, we extract the output features in the GAP layer of the classification stage as decisive knowledge to be fed into the knowledge transfer module. Then, we utilize the Maximum Mean Discrepancy (MMD) [36] to measure the distance of the output distribution of the two branches. Then, an instructive clue is leveraged from the trimmed branch for the untrimmed branch to improve the overall performance.

E. Integrated Training

The multi-dilated convolution module with the cascaded classification module is designed to improve the quality of the Class Activation Sequence (CAS). Meanwhile, the boundary regression module is proposed to perform boundary prediction on the CAS, which is more robust to noise. In addition, with the provided trimmed videos, the knowledge transfer module further boosts the classification performance on untrimmed videos by leveraging the decisive information from trimmed videos. To promote the quality of the CAS greatly and predict the accurate boundaries efficiently, we adopt a multi-task learning approach to jointly train these modules, where the loss functions of three parts are combined in an end-to-end framework. The training details of our algorithm are introduced in this section.

Training Data Construction: As described in Section III. B, for a given video $\mathbf{X}_v$, we form the unit sequence $\mathbf{U}$ and extract the corresponding feature sequence $\mathbf{F}$ with length l_u. Then, we try three sampling strategies to simplify the input feature sequence for computational cost reduction and long-range video modeling, including uniform sampling, sparse sampling [21] and shot-based sampling [37]. We adopt the sparse sampling during training, which achieves the optimal performance. We think that this is mainly because the sparse sampling method can improve the long-range modeling capability; meanwhile, it randomly selects a snippet from K segments each to form the feature sequence during the training phase, which can be regarded as a data augmentation strategy to increase the feature diversity. After sampling, we construct the training data of each pair of untrimmed and trimmed videos as $\mathbf{\Theta}_{untri}$ and $\mathbf{\Theta}_{tri}$ with the same action category, where $\mathbf{\Theta}(\mathbf{X}_v) = \{\mathbf{U}^{'}(X_v), \mathbf{F}^{'}(X_v), \mathbf{\Psi}_v\}$. Finally, each pair of $\mathbf{\Theta}_{untri}$ and $\mathbf{\Theta}_{tri}$ is separately fed into the two branches for our algorithm implementations.

Loss of the Cascaded Dilated Classification Block: Taking a video feature sequence as the input, the multi-dilated convolution module utilizes multiple convolutional kernels with varied dilation rates to perform video classification, while the cascaded classification module adopts the online adversarial erasing method to combine the same two classifiers with different input video features for entire region mining in a cascaded manner. The standard multi-label sigmoid cross-entropy loss function is computed on video-level action labels to concurrently train the two cascaded classifiers, which can be defined as:

$$L_{class} = \frac{1}{N_{train}} \sum_{v=1}^{N_{train}} -log(\mathbf{y}_v^{(\mathbf{\Psi}_v)}), \tag{1}$$

where $\mathbf{y}_v^{(\mathbf{\Psi}_v)} = \frac{1}{1+exp(-\mathbf{y}_{cls,v}^{(\mathbf{\Psi}_v)})}$, and $\mathbf{y}_{cls,v}^{(\mathbf{\Psi}_v)}$ is the predicted class score of action label $\mathbf{\Psi}_v$ for video v.

For the online adversarial erasing mechanism, we also test different erasing thresholds ξ_{era} from 0.6 to 0.9, and the evaluation results are shown in Section IV.

Loss of the Boundary Regression Module: As shown in the bottom-right of Fig. 1, we can compute the OIC loss for each anchor attached to the output feature maps. As described before, each predicted anchor consists of the inner boundary $[s, e]$, inflated outer boundary $[S, E]$ and action category k. We denote the class activation at position t in the CAS of action k as $a_k(t)$. Hence, the OIC loss of the prediction τ can be defined as the average activation $A_o(\tau)$ of the outer area ($i.e.$, $[S, s]$ and $[e, E]$) minus the average activation $A_i(\tau)$ of the inner area ($i.e.$, $[s, e]$):

$$\begin{aligned} L_{OIC} &= A_o(\tau) - A_i(\tau) \\ &= \frac{\int_S^s a_k(t)dt - \int_e^E a_k(t)dt}{(s - S + 1) - (E - e + 1)} - \frac{\int_s^e a_k(t)dt}{(e - s + 1)}. \end{aligned} \tag{2}$$

During training, we only consider the CAS of the ground-truth action category ψ_v. If the activation $a_k(t)$ at position t in the CAS

is less than 0.1, we discard all anchors attached to this temporal position. In addition, among the M anchors of each remaining temporal position, we only keep the one with the lowest OIC loss, which indicates the most likely scale to contain an action instance. Then, the kept anchors with OIC loss greater than -0.3 are further removed. Finally, we conduct Soft Non-Maximum Suppression (Soft-NMS) [38] on the remaining anchors with predefined overlap IoU threshold θ_{NMS}. The total loss of the boundary regression module is the summation of the OIC losses of all kept anchors. With the OIC loss, the optimization process will encourage the higher activation in the inner area but the lower activation in the outer area.

Loss of the Knowledge Transfer Module: The loss function of the knowledge transfer module can be defined as the squared Maximum Mean Discrepancy (MMD) loss [36]:

$$\begin{aligned} L_{KT} &= L_{FC} = MMD^2(\mathbf{T}, \mathbf{U}) \\ &= \frac{1}{N_T^2}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T} k(\mathbf{t}_i, \mathbf{t}_j) + \frac{1}{N_U^2}\sum_{i=1}^{N_U}\sum_{j=1}^{N_U} k(\mathbf{u}_i, \mathbf{u}_j) \\ &\quad - \frac{2}{N_T \cdot N_U}\sum_{i=1}^{N_T}\sum_{j=1}^{N_U} k(\mathbf{t}_i, \mathbf{u}_j), \end{aligned} \tag{3}$$

where N_T and N_U indicate the number of trimmed and untrimmed videos with the same class, respectively. $\mathbf{t}$ and $\mathbf{u}$ indicate the video-level representation before a Fully Connected (FC) layer of trimmed and untrimmed videos, respectively, in the classification network. k denotes the Gaussian kernel function.

Objective for Training: The training objective of the transferable-knowledge-based multi-granularity fusion network is to solve a multi-task optimization problem. The overall loss function is a weighted sum of the classification loss (L_{class}), boundary regression loss (L_{OIC}, knowledge transfer loss (L_{KT}) and L_2 loss for regularization:

$$L = L_{class} + \alpha \cdot L_{OIC} + \beta \cdot L_{KT} + \lambda \cdot ||\mathbf{\Xi}||_2^2, \tag{4}$$

where α, β and λ are the weight terms for balancing the loss functions of multiple tasks. L_{class} is the classification loss in the two-stage cascaded classification module. α is set to 10, β is set to 10, and λ is set to 0.0001 by empirical validation. $\mathbf{\Xi}$ is the unified model.

F. Inference During Prediction

During prediction, using the enhanced CAS and boundary predictor, we can obtain the detections with entire regions and accurate boundaries in three steps:

1) First, we derive the CAS of the augmented classification network based on the idea of [9] and subsequently combine the CASs of multiple dilated branches and two cascaded stages to obtain the enhanced CAS.
2) Then, we perform boundary prediction on the generated CAS and obtain the confidence score of each kept detection by fusing the OIC score, mean activation score and classification score.
3) Finally, we conduct the postprocessing step to suppress redundant detections based on their confidence scores.

Boundary Prediction: To implement boundary prediction on the CAS, first, we derive the CAS of the augmented classification network including the multi-dilated convolution module and cascaded classification module. During prediction, we adopt the uniform sampling method to sample the input video feature sequence for time saving and stable results and consider the CAS of all action categories to perform boundary regression and obtain the class-specific segment prediction results that satisfy the conditions as described in the training procedure.

Score Fusion: After the boundary prediction, we can obtain the prediction set $\mathbf{\Gamma} = \{\tau_n\}_{n=1}^{N_p}$, where N_p is the number of candidate detections. To obtain the confidence score of each prediction for retrieval, we adopt the score fusion of three parts for a reliable evaluation: mean activation score, Outer-Inner-Contrastive (OIC) score and classification score. Specifically, denoting prediction τ as $[t_{start}, t_{end}]$, we first calculate the mean activation score among the temporal range of the detection as p_{act}:

$$\begin{aligned} p_{act} &= \frac{1}{t_{end} - t_{start} + 1}\sum_{t=t_{start}}^{t_{end}} \mathbf{H}_t^k(Cas) = \frac{1}{t_{end} - t_{start} + 1} \\ &\sum_{t=t_{start}}^{t_{end}} \left(\max\left\{ \mathbf{H}_0 + \frac{1}{n_d}\sum_{i=1}^{n_d}\mathbf{H}_i, \bar{\mathbf{H}}_0 + \frac{1}{n_d}\sum_{i=1}^{n_d}\bar{\mathbf{H}}_i \right\} \right), \end{aligned} \tag{5}$$

and denote the OIC score p_{oic} of τ as 1 minus its OIC loss; then, we obtain the confidence score p_{conf} by fusing p_{act}, p_{oic} and p_{cls} with multiplication:

$$p_{conf} = p_{act} \cdot p_{oic} \cdot p_{class}. \tag{6}$$

Redundant Detection Suppression: Since our algorithm generates detections based on densely distributed anchors, the candidate predictions may overlap with each other to different degrees. To suppress redundant detections and improve the recall, we conduct the Soft-NMS [38] method, which suppresses redundant detections using a Gaussian decaying score function to re-rank the prediction set:

$$p'_{conf} = \begin{cases} p_{conf,n}, & iou\,(\tau_m, \tau_n) < \theta, \\ p_{conf,n} \cdot e^{-\frac{iou(\tau_m,\tau_n)^2}{\varepsilon}}, & iou\,(\tau_m, \tau_n) > \theta, \end{cases} \tag{7}$$

where τ_m is the prediction with the maximum score, ε is a parameter of the Gaussian function, and θ is the predefined threshold. After postprocessing, we obtain the final prediction set $\mathbf{\Gamma}' = \{\tau_n' = \{t_{start}^n, t_{end}^n, p_{conf,n}', k_n\}\}_{n=1}^{N_p'}$, where N_p' is the number of final predictions.

IV. Experiments

A. Datasets and Setup

Datasets: ActivityNet-1.3 [39] is a large-scale video dataset for action recognition and temporal action detection tasks used in the ActivityNet Challenge 2016, 2017 and 2018, which contains 19994 videos and is divided into training, validation and testing subsets at a ratio of 2:1:1, with 200 action classes annotated. Each video is annotated with 1.5 temporal action instances on

average. The **THUMOS14** [40] dataset uses the UCF-101 [41] dataset as the training set, which includes 13320 trimmed videos for the action recognition task, and contains 1010 untrimmed videos for validation and 1574 untrimmed videos for testing with video-level labels of 101 action classes, while only a subset of 200 videos in the validation set and 213 videos in the testing set are temporally annotated among 20 classes.

We train our model with the validation subset without using the temporal annotations and trimmed videos in the UCF-101 dataset of the same 20 classes among the 101 classes for knowledge transfer. For the ActivityNet-1.3 dataset, we observe that there are 30 classes that belong to the same classes among the two datasets. Therefore, we also adopt the corresponding trimmed videos for knowledge transfer.

In this section, we compare our method with state-of-the-art methods on both ActivityNet-1.3 and THUMOS14. Exploration studies are performed on THUMOS14.

Evaluation metrics: Following the conventions, we use the mean Average Precision (mAP) as the evaluation metric, where the Average Precision (AP) is separately calculated on each class. We report the mAP values at different Intersection over the Union (IoU) thresholds. On ActivityNet-1.3, mAPs with IoU thresholds of $\{0.5, 0.75, 0.95\}$ and average mAPs with IoU thresholds set at $\{0.5 : 0.05 : 0.95\}$ are used. On THUMOS14, mAPs with IoU thresholds of $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ are used.

Implementation Details: For visual feature encoding, we adopt the two-stream networks with the architecture described in [31], where the ResNet network [42] is the spatial network, and the BN-Inception network [43] is the temporal network. For the RGB stream, we perform a center crop of size 224×224; for the optical stream, we utilize the TV-L1 optical flow algorithm. During feature extraction, the inputs to the two-stream network are stacks of unit frames n_u sampled at 30 fps. n_u is set to 16 on both datasets. The two-stream network is implemented using Caffe [44].

In KT-MGFN, since the duration of videos on the ActivityNet-1.3 dataset is limited, we rescale the feature sequence of each input video to a fixed length window $l_w = 256$ by linear interpolation as in our previous work [6], [45]. We train our KT-MGFN with the multi-task objective using the Adam optimizer, and the implementations are based on PyTorch. On both datasets, the batch size is set to 16, and the learning rate is set to 0.001 for 30 epochs without optimization of the boundary regression module and then to 0.0001 for another 70 epochs with overall optimization. The erasing threshold ξ_{era} in the cascaded classification module is set to 0.8, and the dilation rates d in the multiple dilated convolution branches are 1, 2, 3 and 5. For IoU threshold θ_{NMS} in Soft-NMS, we set θ_{NMS} to 0.4 on ActivityNet-1.3 and 0.5 on THUMOS14 based on empirical validation. ε in the Gaussian function is set to 0.75. We choose anchors of snippet-level length 1, 2, 4, 8, 16, and 32 for the THUMOS14 dataset and 4, 8, 16, 32, 64, and 128 for the ActivityNet-1.3 dataset.

B. Comparison with State-of-the-Art Methods

Action Recognition: We first evaluate the video classification performance on untrimmed videos and compare the results with other state-of-the-art methods. Table I illustrates the evaluation results on the THUMOS14 dataset, which show that with our augmented classification network and knowledge transfer module, our method can achieve a better performance than the existing methods. Table II illustrates the classification results on the ActivityNet-1.3 dataset.

TABLE I
COMPARISON RESULTS OF THE CLASSIFICATION ACCURACY (%) ON THE THUMOS14 DATASET. NOTE THAT MGFN IS A SIMPLER VERSION OF KT-MGFN, WHICH EXCLUDES THE KNOWLEDGE TRANSFER MODULE

Supervision	Method	RGB	Optical Flow	Fusion
Strong	iDT+FV [18]	-	-	63.1
	TSN (3 seg) [21]	-	-	78.5
	Two-Stream	68.2	71.6	73
Weak	UntrimmedNets [31]	-	-	82.2
	MGFN	79.3	80.1	85.4
	KT-MGFN	**79.9**	**80.7**	**86.0**

TABLE II
COMPARISON RESULTS OF THE CLASSIFICATION ACCURACY (%) ON THE ACTIVITYNET-1.3 DATASET. NOTE THAT MGFN IS A SIMPLER VERSION OF KT-MGFN, WHICH EXCLUDES THE KNOWLEDGE TRANSFER MODULE

Method	RGB	Optical Flow	Fusion
Two-Stream [46]	71.1	73.5	79.2
MGFN	78.2	80.7	85.6
KT-MGFN	**79.6**	**82.3**	**88.1**

Action Detection: Then, we evaluate the action detection performance of our method under weak supervision and compare the results with other state-of-the-art approaches in both fully supervised and weakly supervised manners. Table III illustrates the results on the THUMOS14 dataset. We observe that our method significantly outperforms other existing weakly supervised methods and is even competitive with some fully supervised approaches. The detection performance of our MGFN without the knowledge transfer module demonstrates that our augmented classification network can generate a higher-quality CAS than AutoLoc [33], and the end-to-end training approach contributes to the improvement, which will be discussed in Section VI.C. Compared to the methods proposed in [10], the mAP of our method under $tIoU = 0.5$ is nearly two times higher, which confirms that our Online Adversarial Erasing (OAE) mechanism is more effective and intuitive. The comparison results for the ActivityNet-1.3 dataset in Table IV also show the superiority and generalizability of our method. For better comparison, we list the per-class Average Precision of several fully supervised methods under $tIoU = 0.5$, which show that our weakly supervised temporal action detector can also achieve competitive performance with fully supervised methods even with weak labels.

C. Exploration Study

Architecture of KT-MGFN: We first conduct ablation studies on the THUMOS14 dataset to investigate the contribution of each module introduced in this paper. As shown in Table V,

TABLE III
COMPARISON OF OUR METHOD WITH STATE-OF-THE-ART METHODS ON THE THUMOS14 DATASET FOR ACTION DETECTION, INCLUDING BOTH STRONG SUPERVISION AND WEAK SUPERVISION. UNTF AND I3DF ARE ABBREVIATIONS FOR THE UNTRIMMEDNET FEATURES AND I3D FEATURES, RESPECTIVELY

Supervision	Method	Feature	mAP@*tIoU* (α)						
			0.1	0.2	0.3	0.4	0.5	0.6	0.7
Strong	Oneata et al. [47]	-	36.6	33.6	27.0	20.8	14.4	-	-
	Richard et al. [48]	-	39.7	35.7	30.0	23.2	15.2	-	-
	Shou et al. [2]	-	47.7	43.5	36.3	28.7	19.0	10.3	5.3
	Yuan et al. [49]	-	51.0	45.2	36.5	27.8	17.8	-	-
	Lin et al. [4]	-	50.1	47.8	43.0	35.0	24.6	15.3	7.7
	Zhao et al. [3]	-	60.3	56.2	50.6	40.8	29.1	-	-
	Gao et al. [5]	-	60.1	56.7	50.1	41.3	**31.0**	19.1	9.9
Weak	Singh et al. [10]	-	36.4	27.8	19.5	12.7	6.8	-	-
	Wang et al. [31]	UNTF	44.4	37.7	28.2	21.1	13.7	-	-
	Nguyen et al. [12]	UNTF	45.3	38.8	31.1	23.5	16.2	9.8	5.1
	Su et al. [45]	UNTF	47.1	41.6	32.8	24.7	16.1	10.1	5.5
	Sujoy et al. [32]	UNTF	49.0	42.8	32.0	26.0	18.8	-	6.2
	Sujoy et al. [32]	I3DF	53.7	48.5	39.2	29.9	22.0	-	7.3
	Shou et al. [33]	UNTF	46.4	41.5	35.8	29.0	21.2	13.4	5.8
	KT-MGFN (Ours)	UNTF	**52.3**	**46.6**	**39.0**	**30.7**	**22.7**	**13.6**	**5.9**
	KT-MGFN (Ours)	I3DF	**56.5**	**51.3**	**42.2**	**33.8**	**25.6**	**17.0**	**7.5**

TABLE IV
COMPARISON RESULTS ON THE VALIDATION SET OF ACTIVITYNET-1.3 IN TERMS OF MAP@*TIOU* AND AVERAGE MAP

Supervision	Method	0.5	0.75	0.95	Average
Strong	Singh et al. [46]	34.5	-	-	-
	Heilbron et al. [50]	40.00	17.90	4.70	21.70
	Wang et al. [51]	42.28	3.76	0.05	14.85
	Shou et al. [52]	43.83	25.88	0.21	22.77
	Xiong et al. [53]	39.12	23.48	5.49	23.98
	Lin et al. [54]	48.99	32.91	7.87	32.26
Weak	Su et al. [45]	39.29	24.09	6.71	24.42
	Nguyen et al. [12]	29.3	16.9	2.6	-
	Liu et al. [55]	34.0	20.9	5.7	21.2
	KT-MGFN (Ours)	**39.89**	**24.56**	**7.30**	**24.75**

TABLE V
ABLATION STUDIES WITH RESPECT TO ARCHITECTURE CHOICES ON THUMOS14. ✓DENOTES THAT THE SETTING OF THE CORRESPONDING COLUMN IS EMPLOYED. OTHERWISE, THE SIMPLE CAS GENERATED WITHOUT THE MULTI-DILATED CONVOLUTION MODULE (MDCM), CASCADED CLASSIFICATION MODULE (CCM), KNOWLEDGE TRANSFER MODULE (KTM), AND BOUNDARY REGRESSION MODULE (BRM) IS ADOPTED

Model	MDCM	CCM	KTM	BRM	mAP (α = 0.5)
KT-MGFN					13.8
	✓				15.4
	✓	✓			17.3
	✓	✓	✓		18.2
	✓	✓	✓	✓	**22.7**

the original Class Activation Sequence (CAS) generated by the simple classification network is chosen as the baseline of our work. The multi-dilated convolution module and cascaded classification module augment the classification network to generate a high-quality CAS, which promotes the localization performance. The knowledge transfer module further boosts the performance with the considered informative information. Finally, using the boundary regression module, accurate boundaries can be obtained. These observations reveal that each module of our KT-MGFN is effective and indispensable.

Dilation Rate: The dilation rate d in the multi-dilated convolution module is used to augment the simple classification network by enlarging the receptive field. Previous work in [30] shows that convolutional kernels with varying dilation rates can transfer the discriminative knowledge of sparsely highlighted regions to other object regions. In this manner, the discriminativeness of other surrounding object-related regions that have not been discovered can be effectively enhanced. However, a convolutional kernel with a large dilation rate can also introduce irrelevant regions, so we only use smaller dilation rates in this paper (*i.e.,* $d = 2, 3, 5$). In addition, even with smaller dilation rates, some irrelevant regions can still be misrecognized, i.e., true negative regions. True negative regions usually show diversity under different dilations, while true positive shows consistency in different localization maps of multiple dilated convolution branches. Therefore, we employ a fusion strategy to remove noise by averaging multiple class activation sequences. The evaluation results of different dilations are illustrated in Fig. 5.

Erasing Threshold: The erasing threshold ξ_{era} mentioned in the online adversarial erasing step of the cascaded classification module is used to dynamically mask out the extracted features of discriminative regions discovered by the first-stage classifier, which forces the second-stage classifier to leverage other

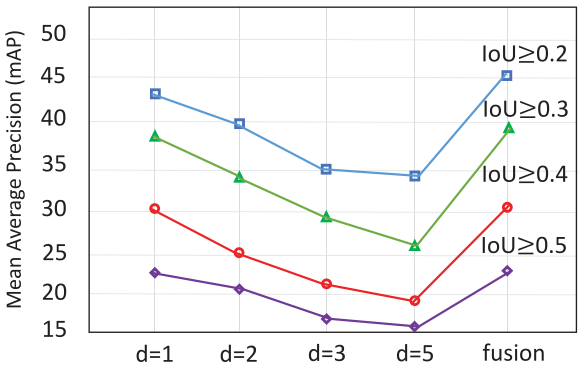


Fig. 5. Evaluation results of different dilation rates in the multi-dilated convolution module.

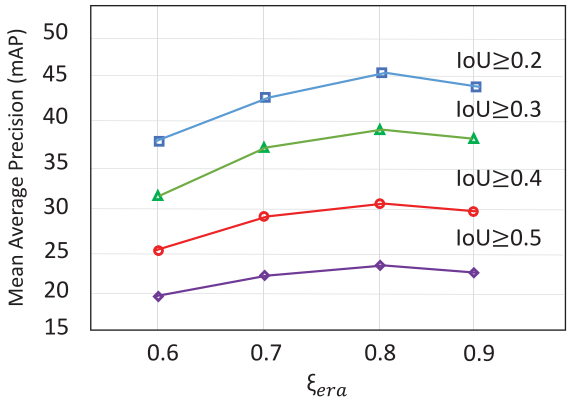


Fig. 6. Evaluation results of different erasing thresholds in the cascaded classification module.

TABLE VI
STUDY OF THE NUMBER OF CLASSIFIERS OF KT-MGFN USED TO GENERATE THE CAS ON THE THUMOS14 DATASET IN TERMS OF mAP@*tIoU*

Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7
KT-MGFN(one)	48.6	42.4	33.3	23.5	18.6	9.2	5.1
KT-MGFN(three)	49.2	43.5	36.1	26.8	19.9	11.1	5.4
KT-MGFN(two)	**52.3**	**46.6**	**39.0**	**30.7**	**22.7**	**13.6**	**5.9**

supportive regions for classification. Hence, the generated CAS can locate relatively complete regions of target actions. However, ξ_{era} is a hyperparameter, which can affect the results to varying degrees for different settings. Concretely, a larger ξ_{era} would fail to discover more useful information since fewer regions are effectively erased, while a smaller ξ_{era} would decrease the performance due to the introduced background information. In this paper, we test ξ_{era} from 0.6 to 0.9 and find that when $\xi_{era} = 0.8$, our method achieves the best performance over different $tIoU$. The evaluation results are shown in Fig. 6.

Cascaded Classifiers: In the cascaded classification module, we perform an ablation study on the number of classifiers. The experimental results are shown in Table VI. The comparison results suggest that there is no significant improvement in the localization performance of KT-MGFN with three classifiers. Therefore, adding the third stage is not necessary, and two stages

TABLE VII
PERFORMANCE COMPARISONS OF STAGE-WISE TRAINING AND END-TO-END TRAINING ON THE THUMOS14 DATASET

	Stage-wise	End-to-end
mAP (α = 0.5)	21.8	**22.7**

TABLE VIII
STUDY OF THE CONTRIBUTIONS OF EACH COMPONENT TO SCORE FUSION ON THE THUMOS14 DATASET. ✓ INDICATES THAT THE CORRESPONDING SCORE IS EMPLOYED FOR FUSION

p_{act}	✓			✓	✓		✓
p_{class}		✓		✓		✓	✓
p_{oic}			✓		✓	✓	✓
mAP (α = 0.5)	19.4	20.6	21.8	21.7	22.4	22.2	**22.7**

TABLE IX
PER-CLASS AVERAGE PRECISION (AP) AT THE IOU THRESHOLD OF 0.5 ON THE THUMOS14 DATASET (%). NOTE THAT ALL COMPARED METHODS ARE FULLY SUPERVISED

Action	[2]	[4]	[56]	[57]	Ours
Baseball Pitch	14.9	**29.3**	19.3	26.1	14.1
Basketball Dunk	20.1	9.2	38.5	**54.0**	12.3
Billiards	7.6	4.7	4.6	8.3	7.7
Clean and Jerk	24.8	35.6	**54.1**	27.9	21.0
Cliff Diving	27.5	46.1	**63.9**	49.2	27.6
Cricket Bowling	15.7	10.0	15.1	**30.6**	12.4
Cricket Shot	**13.8**	1.9	10.3	10.9	8.2
Diving	17.6	17.6	26.9	26.2	21.1
Frisbee Catch	15.3	6.6	**22.0**	20.1	7.9
Golf Swing	18.2	13.3	20.5	16.1	**25.2**
Hammer Throw	19.1	51.6	41.6	43.2	38.2
High Jump	20.0	21.6	22.0	30.9	22.1
Javelin Throw	18.2	42.7	**52.0**	47.0	42.3
Long Jump	34.8	71.3	**71.7**	57.4	70.5
Pole Vault	32.1	58.1	48.9	42.7	47.0
Shotput	12.1	21.8	16.0	19.4	11.2
Soccer Penalty	19.2	13.3	**26.4**	15.8	17.2
Tennis Swing	**19.3**	8.8	12.3	16.6	7.6
Throw Discus	24.4	24.8	7.4	29.2	**40.3**
Volleyball Spiking	4.6	3.9	10.8	5.6	10.1
mAP@0.5	19.0	24.6	**29.2**	28.9	22.7

are usually sufficient for locating the integral temporal action intervals.

Network Training Pattern: Stage-wise vs. End-to-end. KT-MGFN is designed to jointly optimize the augmented classification network and boundary regressor. It is also possible to separately train the augmented classification network and boundary regressor, in which they do not collaborate with each other. Such a training scheme is called stage-wise training. Table VII illustrates a comparison between these two approaches, from which we observe that the unified training approach can outperform stage-wise training with identical settings. This result clearly demonstrates the importance of jointly optimizing the augmented classification network and boundary regressor.

Score Fusion for Retrieval: We evaluate the score fusion with different combinations of classification score p_{class}, activation score p_{act} and OIC score p_{oic}. The evaluation results are shown in Table VIII, which suggest that using the classification score p_{class}, activation score p_{act} or OIC score p_{oic} alone

Fig. 7. Qualitative examples generated by KT-MGFN on the THUMOS14 dataset (top two rows) and ActivityNet-1.3 dataset (bottom row).

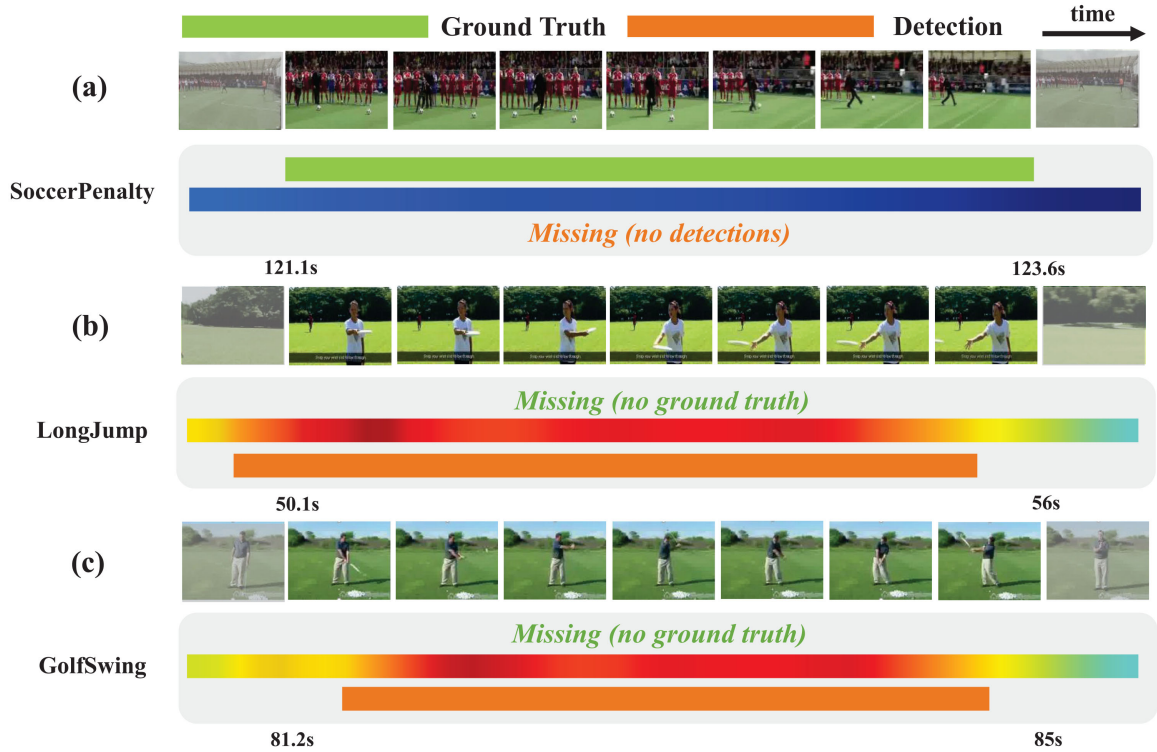


Fig. 8. Qualitative examples of low-quality detections generated by KT-MGFN on the THUMOS14 dataset.

is worse than combining two of them, and by fusing the three scores together, we can obtain the best performance.

D. Qualitative Analysis

As shown in Fig. 7, we visualize the localization performance of three examples predicted by our methods on the THUMOS14 and ActivityNet-1.3 datasets. In Fig. 7(a), the given video contains two ground truth action instances with different classes, and our method can generate the class-specific detections. In Fig. 7(b), a video of the *HammerThrow* action is densely annotated, where the durations of action instances greatly vary. However, our method can generate dense detections to cover the ground truths as complete as possible. In Fig. 7(c), the duration of the *LongJump* action instance almost occupies the entire video, but our method is sufficiently robust to discover the most discriminative regions even with some missing areas.

We also illustrate some low-quality prediction examples of our method on the THUMOS14 dataset in Fig. 8. In Fig. 8(a), there is a potential action instance with the *SoccerPenalty* class that our method failed to detect, and the reason may be ascribed to the similar appearance and little dynamic motions along the temporal dimension. In Fig. 8(b) and Fig. 8(c), our method generates detections of the *LongJump* and *GolfSwing* classes, while there is no corresponding ground truth in the video. A possible reason is that the annotator failed to segment them. From these examples, we conclude that our method can be well generalized to different scenarios and generate high-quality detections with high recall.

V. Conclusion

In this paper, we propose a unified network for weakly supervised temporal action detection. Our method can generate a high-quality Class Activation Sequence (CAS) by augmenting the simple classification network with the cascaded dilated convolution block, where the multi-dilated convolution module employs convolutional kernels with varying dilation rates for local discriminative region expansion, while the cascaded classification module adapts two cascaded classifiers for entire region mining. In addition, to improve the classification performance on untrimmed videos, informative knowledge has been transferred from trimmed videos to untrimmed videos with a knowledge transfer module. Finally, a boundary regression module is adopted to perform boundary prediction on the enhanced CAS. Experiments conducted on the THUMOS14 dataset and ActivityNet-1.3 dataset demonstrate the effectiveness of our approach.

References

[1] Z. Zhou, F. Shi, and W. Wu, "Learning spatial and temporal extents of human actions for action detection," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 512–525, Apr. 2015.

[2] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1049–1058.

[3] Y. Zhao *et al.*, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2914–2923.

[4] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. ACM Multimedia Conf*, 2017, pp. 988–996.

[5] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," in *Proc. Brit. Mach. Vision Conf.*, 2017, pp. 52.1–52.11.

[6] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 3–19.

[7] S. Hao, X. Wu, Z. Bing, Y. Wu, and Y. Jia, "Temporal action localization in untrimmed videos using action pattern trees," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 717–730, Mar. 2019.

[8] D. Guo, L. Wei, and X. Fang, "Fully convolutional network for multi-scale temporal action proposals," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3428–3438, Dec. 2018.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2921–2929.

[10] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 3544–3553.

[11] Y. Wei *et al.*, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, vol. 1, no. 2, 2017.

[12] P. Nguyen, T. Liu, G. Prasad, and B. Han, "Weakly supervised action localization by sparse temporal pooling network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6752–6761.

[13] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019.

[14] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 634–644, Mar. 2018.

[15] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, Mar. 2018.

[16] J. Hou, X. Wu, Y. Sun, and Y. Jia, "Content-attention representation by factorized action-scene network for action recognition," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1537–1547, Jun. 2018.

[17] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 3169–3176.

[18] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 3551–3558.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[20] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1933–1941.

[21] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2016, pp. 20–36.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4489–4497.

[23] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2846–2854.

[24] P. Bai, X. Tang, X. Wang, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4322–4328.

[25] Y. Tang, X. Wang, E. Dellandréa, and L. Chen, "Weakly supervised learning of deformable part-based models for object detection via region proposals," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 393–407, Feb. 2017.

[26] J. Zhang *et al.*, "Top-down neural attention by excitation backprop," in *Proc. Int. J. Comput. Vision*, Springer, 2018, vol. 126, no. 10, pp. 1084–1102.

[27] Q. Tao, H. Yang, and J. Cai, "Exploiting web images for weakly supervised object detection," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1135–1146, May 2019.

[28] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," in *Proc. Int. J. Comput. Vision*, Springer, 2013, vol. 104, no. 2, pp. 154–171.

[29] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 391–405.

[30] Y. Wei *et al.*, "Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7268–7277.

[31] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4325–4334.

[32] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-TALC: Weakly-supervised temporal activity localization and classification," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 563–579.

[33] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 154–171.

[34] C. Yan *et al.*, "Cross-modality bridging and knowledge transferring for image understanding," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2675—2685, Oct. 2019.

[35] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1050–1062, May 2017.

[36] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[37] H. Su, X. Zhao, T. Lin, and H. Fei, "Weakly supervised temporal action detection with shot-based temporal pooling network," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 426–436

[38] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5561–5569.

[39] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 961–970.

[40] Y. Jiang *et al.*, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. Comput. Vision-Eur. Conf. Comput. Vision Workshop*, 2014. [Online]. Available: http://crcv.ucf.edu/THUMOS14/

[41] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 448–456.

[44] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[45] H. Su, X. Zhao, and T. Lin, "Cascaded pyramid mining network for weakly supervised temporal action localization," in *Proc. Asian Conf. Comput. Vision*, Springer, 2018, pp. 558–574.

[46] G. Singh and F. Cuzzolin, "Untrimmed video classification for activity detection: submission to activitynet challenge," 2016, *arXiv:1607.01979*.

[47] D. Oneata, J. Verbeek, and C. Schmid, "The lear submission at thumos2014," *THUMOS Action Recognit. Challenge*, Tech. Rep., 2014.

[48] A. Richard and J. Gall, "Temporal action detection using a statistical language model," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3131–3140.

[49] Z.-H. Yuan, J. C. Stroud, T. Lu, and J. Deng, "Temporal action localization by structured maximal sums." in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, vol. 2, p. 7.

[50] F. C. Heilbron, W. Barrios, V. Escorcia, and B. Ghanem, "SCC: Semantic context cascade for efficient action detection." in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3175–3184.

[51] R. Wang and D. Tao, "Uts at activitynet 2016," *ActivityNet Large Scale Activity Recognit. Challenge*, vol. 1, p. 8, 2016.

[52] Z. Shou *et al.*, "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1417–1426.

[53] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," 2017, *arXiv:1703.02716*.

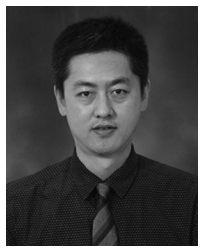
[54] T. Lin, X. Zhao, and Z. Shou, "Temporal convolution based action proposal: Submission to activitynet 2017," 2017, *arXiv:1707.06750*.

[55] D. Liu, J. Tingting, and W. Yizhou, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1298–1307

[56] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proc. Brit. Mach. Vision Conf.*, 2017, vol. 2, p. 7.

[57] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3d network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5783–5792.

Haisheng Su received the master's degree from the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 2020, advised by Prof. X. Zhao. He won the Temporal Action Localization task championship of ActivityNet challenge 2018 and was second runner-up in 2019. His research interests include machine learning, visual understanding and analysis of humans, especially action recognition, temporal action detection, and weakly supervised learning.



Xu Zhao (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2011. He was a visiting scholar with the Beckman Institute, UIUC from November 2007 to December 2008 and a Postdoctoral Research Fellow with Northeastern University from August 2012 to August 2013. In November 2013, he joined SJTU, where he is currently a Full Professor with the Department of Automation, School of Electronic Information and Electrical Engineering. He is also an Adjunct Professor and an Assistant Dean of the Institute of Medical Robotics, SJTU. His research interests lie in computer vision and machine learning, specifically in human-centric visual computing.

Tianwei Lin received the B.Eng. degree from the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2016, and the master's degree from Shanghai Jiao Tong University, Shanghai, China, in 2019, advised by Prof. X. Zhao. His research interests include computer vision, deep learning, action recognition, temporal action detection, and GAN.

Shuming Liu received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 2018. He is currently working toward the a master's degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, advised by Prof. X. Zhao. His research interests include temporal action detection and action recognition.

Zhilan Hu received the B.Sc. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2004 and 2009, respectively. She was a researcher with the Samsung Advanced Institute of Technology in China from 2009 to 2015. Since 2016, she has been a principle Researcher with the Central Media Technology Institute of Huawei Company, Ltd., Shenzhen, China. Her research interests include computer vision, pattern recognition, and machine learning.