

Graph Convolutional Networks for Temporal Action Localization

Runhao Zeng¹ Wenbing Huang^{2,5} Mingkui Tan^{1,4†} Yu Rong²
Peilin Zhao² Junzhou Huang² Chuang Gan³

¹School of Software Engineering, South China University of Technology, China

²Tencent AI Lab ³MIT-IBM Watson AI Lab ⁴Peng Cheng Laboratory, Shenzhen

⁵Department of Computer Science and Technology, Tsinghua University, State Key Lab. of Intelligent Technology and Systems, Tsinghua National Lab. for Information Science and Technology (TNList)

Frunhaozeng.cs, ganchuang1990G@gmail.com, hwenbing@126.com, mingkui tan@scut.edu.cn

Abstract

Most state-of-the-art action localization systems process each action proposal individually, without explicitly exploiting their relations during learning. However, the relations between proposals actually play an important role in action localization, since a meaningful action always consists of multiple proposals in a video. In this paper, we propose to exploit the proposal-proposal relations using Graph Convolutional Networks (GCNs). First, we construct an action proposal graph, where each proposal is represented as a node and their relations between two proposals as an edge. Here, we use two types of relations, one for capturing the context information for each proposal and the other one for characterizing the correlations between distinct actions. Then we apply the GCNs over the graph to model the relations among different proposals and learn powerful representations for the action classification and localization. Experimental results show that our approach significantly outperforms the state-of-the-art on THUMOS14 (49.1% versus 42.8%). Moreover, augmentation experiments on ActivityNet also verify the efficacy of modeling action proposal relationships.

1. Introduction

Understanding human actions in videos has been becoming a prominent research topic in computer vision, owing to its various applications in security surveillance, human behavior analysis and many other areas [10, 35, 38, 12, 13, 14, 15, 16, 42]. Despite the fruitful progress in this vein, there are still some challenging tasks demanding further explo-

indicates equal contributions. This work was done when Runhao Zeng was served as a research intern in Tencent AI Lab under the supervision of Wenbing Huang.

[†]Corresponding author

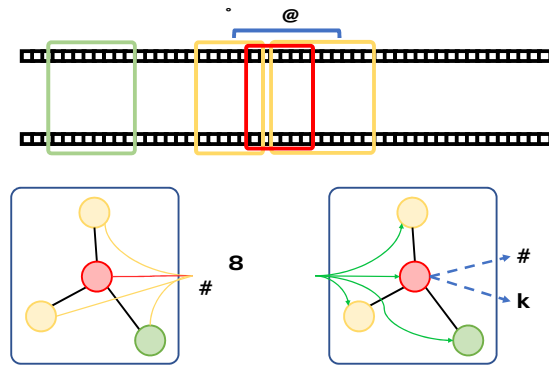


Figure 1. Schematic depiction of our approach. We apply graph convolutional networks to model the proposal-proposal interactions and boost the temporal action localization performance.

ration — *temporal action localization* is such an example. To deal with real videos that are untrimmed and usually contain the background of irrelevant activities, temporal action localization requires the machine to not only classify the actions of interest but also localize the start and end time of every action instance. Consider a sport video as illustrated in Figure 1, the detector should find out the frames where the action event is happening and identify the category of the event.

Temporal Action localization has attracted increasing attention in the last several years [6, 18, 26, 33, 34]. Inspired by the success of object detection, most current action detection methods resort to the two-stage pipeline: they first generate a set of 1D temporal proposals and then perform classification and temporal boundary regression on each proposal individually. However, processing each proposal separately in the prediction stage will inevitably neglect the semantic relations between proposals.

We contend that exploiting the proposal-proposal relations in the video domain provides more cues to facilitate the recognition of each proposal instance. To illustrate this,

we revisit the example in Figure 1, where we have generated four proposals. On the one hand, the proposals p_1 , p_2 and p_3 overlapping with each other describe different parts of the same action instance (*i.e.*, the start period, main body and end period). Conventional methods perform prediction on p_1 by using its feature alone, which we think is insufficient to deliver complete knowledge for the detection. If we additionally take the features of p_2 and p_3 into account, we will obtain more contextual information around p_1 , which is advantageous especially for the temporal boundary regression of p_1 . On the other hand, p_4 describes the background (*i.e.*, the sport field), and its content is also helpful in identifying the action label of p_1 , since what is happening on the sport field is likely to be sport action (*e.g.* “discus throwing”) but not the one happens elsewhere (*e.g.* “kissing”). In other words, the classification of p_1 can be partly guided by the content of p_4 even they are temporally disjointed.

To model the proposal-proposal interactions, one may employ the self-attention mechanism [39] — as what has been conducted previously in language translation [39] and object detection [22] — to capture the pair-wise similarity between proposals. A self-attention module can affect an individual proposal by aggregating information from all other proposals with the automatically learned aggregation weights. However, this method is computationally expensive as querying all proposal pairs has a quadratic complexity of the proposal number (note that each video could contain more than thousands of proposals). On the contrary, Graph Convolutional Networks (GCNs), which generalize convolutions from grid-like data (*e.g.* images) to non-grid structures (*e.g.* social networks), have received increasing interests in the machine learning domain [25, 47]. GCNs can affect each node by aggregating information from the adjacent nodes, and thus are very suitable for leveraging the relations between proposals. More importantly, unlike the self-attention strategy, applying GCNs enables us to aggregate information from only the local neighbourhoods for each proposal, and thus can help decrease the computational complexity remarkably.

In this paper, we regard the proposals as nodes of a specific graph and take advantage of GCNs for modeling the proposal relations. Motivated by the discussions above, we construct the graph by investigating two kinds of edges between proposals, including the *contextual edges* to incorporate the contextual information for each proposal instance (*e.g.*, detecting p_1 by accessing p_2 and p_3 in Figure 1) and the *surrounding edges* to query knowledge from nearby but distinct proposals (*e.g.*, querying p_4 for p_1 in Figure 1).

We then perform graph convolutions on the constructed graph. Although the information is aggregated from local neighbors in each layer, message passing between distant nodes is still possible if the depth of GCNs increases. Besides, we conduct two different GCNs to perform classifica-

tion and regression separately, which is demonstrated to be effective by our experiments. Moreover, to avoid the overwhelming computation cost, we further devise a sampling strategy to train the GCNs efficiently while still preserving desired detection performance. We evaluate our proposed method on two popular benchmarks for temporal action detection, *i.e.*, THUMOS14 [24] and ActivityNet1.3 [4].

To sum up, our contributions are as follow:

- To the best of our knowledge, we are the first to exploit the proposal-proposal relations for temporal action localization in videos.
- To model the interactions between proposals, we construct a graph of proposals by establishing the edges based on our valuable observations and then apply GCNs to do message aggregation among proposals.
- We have verified the effectiveness of our proposed method on two benchmarks. On THUMOS14 especially, our method obtains the mAP of 49.1% when $\text{tIoU} = 0.5$, which significantly outperforms the state-of-the-art, *i.e.* 42.8% by [6]. Augmentation experiments on ActivityNet also verify the efficacy of modeling action proposal relationships.

2. Related work

Temporal action localization. Recently, great progress has been achieved in deep learning [5, 9, 19, 53], which facilitates the development of temporal action localization. Approaches on this task can be grouped into three categories: (1) methods performing frame or segment-level classification where the smoothing and merging steps are required to obtain the temporal boundaries [33, 28, 51]; (2) approaches employing a two-stage framework involving proposal generation, classification and boundary refinement [34, 46, 52]; (3) methods developing end-to-end architectures integrating the proposal generation and classification [48, 1, 26].

Our work is built upon the second category where the action proposals are first generated and then used to perform classification and boundary regression. Following this paradigm, Shou *et al.* [34] propose to generate proposals from sliding windows and classify them. Xu *et al.* [46] exploit the 3D ConvNet and propose a framework inspired by Faster R-CNN [30]. The above methods neglect the context information of proposals, and hence some attempts have been developed to incorporate the context to enhance the proposal feature [8, 17, 18, 52, 6]. They show encouraging improvements by extracting features on the extended receptive field (*i.e.*, boundary) of the proposal. Despite their success, they all process each proposal individually. In contrast, our method has considered the proposal-proposal interactions and leveraged the relations between proposals.

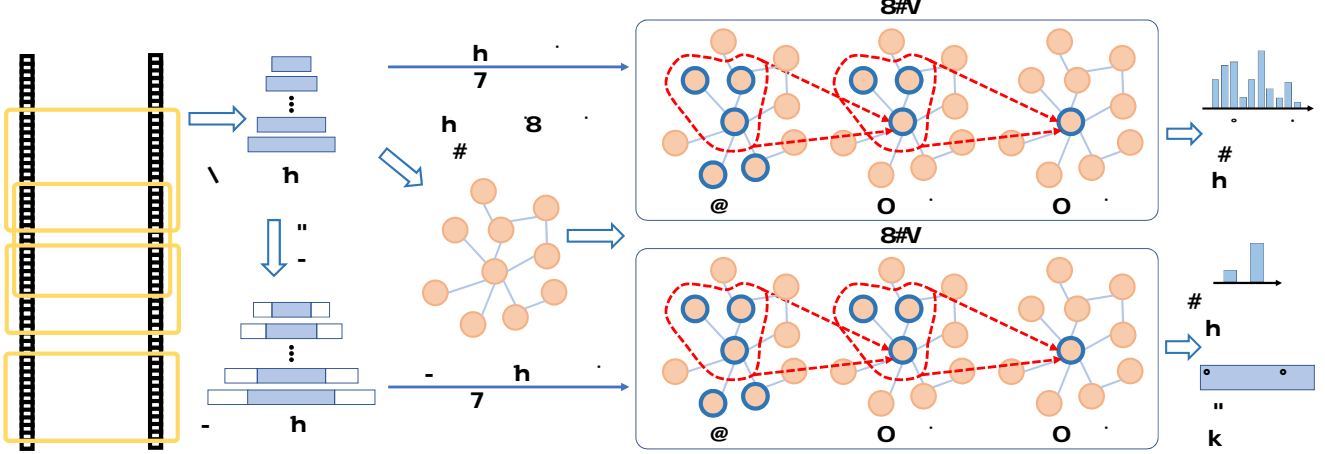


Figure 2. Schematic of our P-GCN model. Given a set of proposals from the input untrimmed video, we instantiate the nodes in the graph by each proposal. Then, edges are established among nodes to model the relations between proposals. We employ two separate GCNs on the same constructed graph with different input features (*i.e.*, the original feature and the extended feature). Finally, P-GCN model outputs the predicted action category, completeness and boundary regression results for all proposals simultaneously.

Graph Convolutional Networks. Kipf *et al.* [25] propose the Graph Convolutional Networks (GCNs) to define convolutions on the non-grid structures [37]. Thanks to its effectiveness, GCNs have been successfully applied to several research areas in computer vision, such as skeleton-based action recognition [47], person re-identification [32], and video classification [45]. For real-world applications, the graph can be large and directly using GCNs is inefficient. Therefore, several attempts are posed for efficient training by virtue of the sampling strategy, such as the node-wise method SAGE [20], layer-wise model FastGCN [7] and its layer-dependent variant AS-GCN [23]. In this paper, considering the flexibility and implementability, we adopt SAGE method as the sampling strategy in our framework.

3. Our Approach

3.1. Notation and Preliminaries

We denote an untrimmed video as $V = \{I_t\}_{t=1}^T$, where I_t denotes the frame at the time slot t with height H and width W . Within each video V , let $P = \{p_i \mid p_i = (x_i, (t_{i,s}, t_{i,e}))\}_{i=1}^N$ be the action proposals of interest, with $t_{i,s}$ and $t_{i,e}$ being the start and end time of a proposal, respectively. In addition, given proposal p_i , let $x_i \in \mathbb{R}^d$ be the feature vector extracted by certain feature extractor (*e.g.*, the I3D network [5]) from frames between $I_{t_{i,s}}$ and $I_{t_{i,e}}$.

Let $G(V, E)$ be a graph of N nodes with nodes $v_i \in V$ and edge $e_{ij} = (v_i, v_j) \in E$. Furthermore, let $A \in \mathbb{R}^{N \times N}$ be the adjacency matrix associated with G . In this paper, we seek to exploit graphs $G(P, E)$ on action proposals in P to better model the proposal-proposal interactions in videos. Here, each action proposal is treated as a node and the edges

in E are used to represent the relations between proposals.

3.2. General Scheme of Our Approach

In this paper, we use a proposal graph $G(P, E)$ to present the relations between proposals and then apply GCN on the graph to exploit the relations and learn powerful representations for proposals. The intuition behind applying GCN is that when performing graph convolution, each node aggregates information from its neighborhoods. In this way, the feature of each proposal is enhanced by other proposals, which helps boost the detection performance eventually.

Without loss of generality, we assume the action proposals have been obtained beforehand by some methods (*e.g.*, the TAG method in [52]). In this paper, given an input video V , we seek to predict the action category \hat{y}_i and temporal position $(\hat{t}_{i,s}, \hat{t}_{i,e})$ for each proposal p_i by exploiting proposal relations. Formally, we compute

$$\{(\hat{y}_i, (\hat{t}_{i,s}, \hat{t}_{i,e}))\}_{i=1}^N = F(GCN(\{x_i\}_{i=1}^N, G(P, E))), \quad (1)$$

where F denotes any mapping functions to be learned. To exploit GCN for action localization, our paradigm takes both the proposal graph and the proposal features as input and perform graph convolution on the graph to leverage proposal relations. The enhanced proposal features (*i.e.*, the outputs of GCN) are then used to jointly predict the category label and temporal bounding box. The schematic of our approach is shown in Figure 2. For simplicity, we denote our model as **P-GCN** henceforth.

In the following sections, we aim to answer two questions: (1) how to construct a graph to represent the relations between proposals; (2) how to use GCN to learn representations of proposals based on the graph and facilitate the action localization.

3.3. Proposal Graph Construction

For the graph $G(P, E)$ of each video, the nodes are instantiated as the action proposals, while the edges E between proposals are demanded to be characterized specifically to better model the proposal relations.

One way to construct edges is linking all proposals with each other, which yet will bring in overwhelming computations for going through all proposal pairs. It also incurs redundant or noisy information for action localization, as some unrelated proposals should not be connected. In this paper, we devise a smarter approach by exploiting the temporal relevance/distance between proposals instead. Specifically, we introduce two types of edges, the contextual edges and surrounding edges, respectively.

Contextual Edges. We establish an edge between proposal p_i and p_j if $r(p_i, p_j) > \text{ctx}$, where ctx is a certain threshold. Here, $r(p_i, p_j)$ represents the relevance between proposals and is defined by the tIoU metric, *i.e.*,

$$r(p_i, p_j) = \text{tIoU}(p_i, p_j) = \frac{I(p_i, p_j)}{U(p_i, p_j)}, \quad (2)$$

where $I(p_i, p_j)$ and $U(p_i, p_j)$ compute the temporal intersection and union of the two proposals, respectively. If we focus on the proposal p_i , establishing the edges by computing $r(p_i, p_j) > \text{ctx}$ will select its neighbourhoods as those have high overlaps with it. Obviously, the non-overlapping portions of the highly-overlapping neighbourhoods are able to provide rich contextual information for p_i . As already demonstrated in [8, 6], exploring the contextual information is of great help in refining the detection boundary and increasing the detection accuracy eventually. Here, by our contextual edges, all overlapping proposals automatically share the contextual information with each other, and these information are further processed by the graph convolution.

Surrounding Edges. The contextual edges connect the overlapping proposals that usually correspond to the same action instance. Actually, distinct but nearby actions (including the background items) could also be correlated, and the message passing among them will facilitate the detection of each other. For example in Figure 1, the background proposal p_4 will provide a guidance on identifying the action class of proposal p_1 (*e.g.*, more likely to be sport action). To handle such kind of correlations, we first utilize $r(p_i, p_j) = 0$ to query the distinct proposals, and then compute the following distance

$$d(p_i, p_j) = \frac{|c_i - c_j|}{U(p_i, p_j)}, \quad (3)$$

to add the edges between nearby proposals if $d(p_i, p_j) < \text{sur}$, where sur is a certain threshold. In Eq. (3), c_i (or c_j) represents the center coordinate of p_i (or p_j). As a complement of the contextual edges, the surrounding edges en-

able the message to pass across distinct action instances and thereby provides more temporal cues for the detection.

3.4. Graph Convolution for Action Localization

Given the constructed graph, we apply the GCN to do action localization. We build K -layer graph convolutions in our implementation. Specifically for the k -th layer ($1 \leq k \leq K$), the graph convolution is implemented by

$$\mathbf{X}^{(k)} = \mathbf{A} \mathbf{X}^{(k-1)} \mathbf{W}^{(k)}. \quad (4)$$

Here, \mathbf{A} is the adjacency matrix; $\mathbf{W}^{(k)} \in \mathbb{R}^{d_k \times d_k}$ is the parameter matrix to be learned; $\mathbf{X}^{(k)} \in \mathbb{R}^{N \times d_k}$ are the hidden features for all proposals at layer k ; $\mathbf{X}^{(0)} \in \mathbb{R}^{N \times d}$ are the input features.

We apply an activation function (*i.e.*, ReLU) after each convolution layer before the features are forwarded to the next layer. In addition, our experiments find it more effective by further concatenating the hidden features with the input features in the last layer, namely,

$$\mathbf{X}^{(K)} = \mathbf{X}^{(K)} \parallel \mathbf{X}^{(0)}, \quad (5)$$

where \parallel denotes the concatenation operation.

Joining the previous work [52], we find that it is beneficial to predict the action label and temporal boundary separately by virtue of two GCNs—one conducted on the original proposal features \mathbf{x}_i and the other one on the extended proposal features \mathbf{x}_i . The first GCN is formulated as

$$\{\hat{y}_i\}_{i=1}^N = \text{softmax}(\text{FC}_1(\text{GCN}_1(\{\mathbf{x}_i\}_{i=1}^N, G(P, E)))), \quad (6)$$

where we apply a Fully-Connected (FC) layer with softmax operation on top of GCN_1 to predict the action label \hat{y}_i . The second GCN can be formulated as

$$\{(\hat{t}_{i,s}, \hat{t}_{i,e})\}_{i=1}^N = \text{FC}_2(\text{GCN}_2(\{\mathbf{x}_i\}_{i=1}^N, G(P, E))), \quad (7)$$

$$\{\hat{c}_i\}_{i=1}^N = \text{FC}_3(\text{GCN}_2(\{\mathbf{x}_i\}_{i=1}^N, G(P, E))), \quad (8)$$

where the graph structure $G(P, E)$ is the same as that in Eq. (6) but the input proposal feature is different. The extended feature \mathbf{x}_i is attained by first extending the temporal boundary of p_i with $\frac{1}{2}$ of its length on both the left and right sides and then extracting the feature within the extended boundary. Here, we adopt two FC layers on top of GCN_2 , one for predicting the boundary $(\hat{t}_{i,s}, \hat{t}_{i,e})$ and the other one for predicting the completeness label \hat{c}_i , which indicates whether the proposal is complete or not. It has been demonstrated by [52] that, incomplete proposals that have low tIoU with the ground-truths could have high classification score, and thus it will make mistakes when using the classification score alone to rank the proposal for the mAP test; further applying the completeness score enables us to avoid this issue.

Adjacency Matrix. In Eq. (4), we need to compute the adjacency matrix \mathbf{A} . Here, we design the adjacency matrix by assigning specific weights to edges. For example, we can apply the cosine similarity to estimate the weights of edge e_{ij} by

$$A_{ij} = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \cdot \|\mathbf{x}_j\|_2}. \quad (9)$$

In the above computation, we compute A_{ij} relying on the feature vector \mathbf{x}_i . We can also map the feature vectors into an embedding space using a learnable linear mapping function as in [44] before the cosine computation. We leave the discussion in our experiments.

3.5. Efficient Training by Sampling

Typical proposal generation methods usually produce thousands of proposals for each video. Applying the aforementioned graph convolution (Eq. (4)) on all proposals demands hefty computation and memory footprints. To accelerate the training of GCNs, several approaches [7, 23, 20] have been proposed based on neighbourhood sampling. Here, we adopt the SAGE method [20] in our method for its flexibility.

The SAGE method uniformly samples the fixed-size neighbourhoods of each node layer-by-layer in a top-down passway. In other words, the nodes of the $(k-1)$ -th layer are formulated as the sampled neighbourhoods of the nodes in the k -th layer. After all nodes of all layers are sampled, SAGE performs the information aggregation in a bottom-up fashion. Here we specify the aggregation function to be a sampling form of Eq. (4), namely,

$$\mathbf{x}_i^{(k)} = \frac{1}{N_s} \sum_{j=1}^{N_s} A_{ij} \mathbf{x}_j^{(k-1)} + \mathbf{x}_i^{(k-1)} \mathbf{W}^{(k)}, \quad (10)$$

where node j is sampled from the neighbourhoods of node i , *i.e.*, $j \in N(i)$; N_s is the sampling size and is much less than the total number N . The summation in Eq. (10) is further normalized by N_s , which empirically makes the training more stable. Besides, we also enforce the self addition of its feature for node i in Eq. (10). We do not perform any sampling when testing. For better readability, Algorithm 1 depicts the algorithmic Flow of our method.

4. Experiments

4.1. Datasets

THUMOS14 [24] is a standard benchmark for action localization. Its training set known as the UCF-101 dataset consists of 13320 videos. The validation, testing and background set contain 1010, 1574 and 2500 untrimmed videos, respectively. Performing action localization on this dataset

Algorithm 1 The training process of P-GCN model.

Input: Proposal set $\mathbf{P} = \{\mathbf{p}_i \mid \mathbf{p}_i = (\mathbf{x}_i, (t_{i,s}, t_{i,e}))\}_{i=1}^N$; original proposal features $\{\mathbf{x}_i^{(0)}\}_{i=1}^N$; extended proposal features $\{\mathbf{x}_i^{(0)}\}_{i=1}^N$; graph depth K ; sampling size N_s

Parameter: Weight matrices $\mathbf{W}^{(k)}$, $k \in \{1, \dots, K\}$

- 1: instantiate the nodes by the proposals \mathbf{p}_i , $\mathbf{p}_i \in \mathbf{P}$
- 2: establish edges between nodes
- 3: obtain a proposal graph $G(\mathbf{P}, E)$
- 4: calculate adjacent matrix using Eq. (9)
- 5: **while** not converges **do**
- 6: **for** $k = 1 \dots K$ **do**
- 7: **for** $\mathbf{p} \in \mathbf{P}$ **do**
- 8: sample N_s neighborhoods of \mathbf{p}
- 9: aggregate information using Eq. (10)
- 10: **end for**
- 11: **end for**
- 12: predict action categories $\{\hat{y}_i\}_{i=1}^N$ using Eq. (6)
- 13: perform boundary regression using Eq. (7)
- 14: predict completeness $\{\hat{c}_i\}_{i=1}^N$ using Eq. (8)
- 15: **end while**

Output: Trained P-GCN model

is challenging since each video has more than 15 action instances and its 71% frames are occupied by background items. Following the common setting in [24], we apply 200 videos in the validation set for training and conduct evaluation on the 213 annotated videos from the testing set.

ActivityNet [4] is another popular benchmark for action localization on untrimmed videos. We evaluate our method on ActivityNet v1.3, which contains around 10K training videos and 5K validation videos corresponded to 200 different activities. Each video has an average of 1.65 action instances. Following the standard practice, we train our method on the training videos and test it on the validation videos. In our experiments, we contrast our method with the state-of-the-art methods on both THUMOS14 and ActivityNet v1.3, and perform ablation studies on THUMOS14.

4.2. Implementation details

Evaluation Metrics. We use mean Average Precision (mAP) as the evaluation metric. A proposal is considered to be correct if its temporal IoU with the ground-truth instance is larger than a certain threshold and the predicted category is the same as this ground-truth instance. On THUMOS14, the tIoU thresholds are chosen from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$; on ActivityNet v1.3, the IoU thresholds are from $\{0.5, 0.75, 0.95\}$, and we also report the average mAP of the IoU thresholds between 0.5 and 0.95 with the step of 0.05.

Features and Proposals. Our model is implemented under the two-stream strategy [35]: RGB frames and optical flow

fields. We first uniformly divide each input video into 64-frame segments. We then use a two-stream Inflated 3D ConvNet (I3D) model pre-trained on Kinetics [5] to extract the segment features. In detail, the I3D model takes as input the RGB/optical-flow segment and outputs a 1024-dimensional feature vector for each segment. Upon the I3D features, we further apply max pooling across segments to obtain one 1024-dimensional feature vector for each proposal that is obtained by the BSN method [27]. Note that we do not fine-tune the parameters of the I3D model in our training phase. Besides the I3D features and BSN proposals, our ablation studies in § 5 also explore other types of features (e.g. 2-D features [27]) and proposals (e.g. TAG proposals [52]).

Proposal Graph Construction. We construct the proposal graph by fixing the values of ctx as 0.7 and sur as 1 for both streams. More discussions on choosing the values of ctx and sur could be found in the supplementary material. We adopt 2-layer GCN since we observed no clear improvement with more than 2 layers but the model complexity is increased. For more efficiency, we choose $N_s = 4$ in Eq. (10) for neighbourhood sampling unless otherwise specified.

Training. The initial learning rate is 0.001 for the RGB stream and 0.01 for the Flow stream. During training, the learning rates will be divided by 10 every 15 epochs. The dropout ratio is 0.8. The classification \hat{y}_i and completeness \hat{c}_i are trained with the cross-entropy loss and the hinge loss, respectively. The regression term $(\hat{t}_{i,s}, \hat{t}_{i,e})$ is trained with the smooth L_1 loss. More training details can be found in the supplementary material.

Testing. We do not perform neighbourhood sampling (i.e. Eq. (10)) for testing. The predictions of the RGB and Flow streams are fused using a ratio of 2:3. We multiply the classification score with the completeness score as the final score for calculating mAP. We then use Non-Maximum Suppression (NMS) to obtain the final predicted temporal proposals for each action class separately. We use 600 and 100 proposals per video for computing mAPs on THUMOS14 and ActivityNet v1.3, respectively.

4.3. Comparison with state-of-the-art results

THUMOS14. Our P-GCN model is compared with the state-of-the-art methods in Table 1. The P-GCN model reaches the highest mAP over all thresholds, implying that our method can recognize and localize actions much more accurately than any other method. Particularly, our P-GCN model outperforms the previously best method (i.e. TAL-Net [6]) by 6.3% absolute improvement and the second-best result [27] by more than 12.2%, when $\text{tIoU} = 0.5$.

ActivityNet v1.3. Table 2 reports the action localization results of various methods. Regarding the average mAP, P-GCN outperforms SSN [52], CDC [33], and TAL-Net [6] by 3.01%, 3.19%, and 6.77%, respectively. We observe that

Table 1. Action localization results on THUMOS14, measured by mAP (%) at different tIoU thresholds

tIoU	0.1	0.2	0.3	0.4	0.5
Oneata <i>et al.</i> [29]	36.6	33.6	27.0	20.8	14.4
Wang <i>et al.</i> [40]	18.2	17.0	14.0	11.7	8.3
Caba <i>et al.</i> [3]	-	-	-	-	13.5
Richard <i>et al.</i> [31]	39.7	35.7	30.0	23.2	15.2
Shou <i>et al.</i> [34]	47.7	43.5	36.3	28.7	19.0
Yeung <i>et al.</i> [48]	48.9	44.0	36.0	26.4	17.1
Yuan <i>et al.</i> [49]	51.4	42.6	33.6	26.1	18.8
Escorcia <i>et al.</i> [11]	-	-	-	-	13.9
Buch <i>et al.</i> [2]	-	-	37.8	-	23.0
Shou <i>et al.</i> [33]	-	-	40.1	29.4	23.3
Yuan <i>et al.</i> [50]	51.0	45.2	36.5	27.8	17.8
Buch <i>et al.</i> [1]	-	-	45.7	-	29.2
Gao <i>et al.</i> [18]	60.1	56.7	50.1	41.3	31.0
Hou <i>et al.</i> [21]	51.3	-	43.7	-	22.0
Dai <i>et al.</i> [8]	-	-	-	33.3	25.6
Gao <i>et al.</i> [17]	54.0	50.9	44.1	34.9	25.6
Xu <i>et al.</i> [46]	54.5	51.5	44.8	35.6	28.9
Zhao <i>et al.</i> [52]	66.0	59.4	51.9	41.0	29.8
Lin <i>et al.</i> [27]	-	-	53.5	45.0	36.9
Chao <i>et al.</i> [6]	59.8	57.1	53.2	48.5	42.8
P-GCN	69.5	67.8	63.6	57.8	49.1

Table 2. Action localization results on ActivityNet v1.3 (val), measured by mAP (%) at different tIoU thresholds and the average mAP of IoU thresholds from 0.5 to 0.95. (*) indicates the method that uses the external video labels from UntrimmedNet [41].

tIoU	0.5	0.75	0.95	Average
Singh <i>et al.</i> [36]	34.47	-	-	-
Wang <i>et al.</i> [43]	43.65	-	-	-
Shou <i>et al.</i> [33]	45.30	26.00	0.20	23.80
Dai <i>et al.</i> [8]	36.44	21.15	3.90	-
Xu <i>et al.</i> [46]	26.80	-	-	-
Zhao <i>et al.</i> [52]	39.12	23.48	5.49	23.98
Chao <i>et al.</i> [6]	38.23	18.30	1.30	20.22
P-GCN	42.90	28.14	2.47	26.99
Lin <i>et al.</i> [27] *	46.45	29.96	8.02	30.03
P-GCN*	48.26	33.16	3.27	31.11

the method by Lin *et al.* [27] (called LIN below) performs promisingly on this dataset. Note that LIN is originally designed for generating class-agnostic proposals, and thus relies on external video-level action labels (from UntrimmedNet [41]) for action localization. In contrast, our method is self-contained and is able to perform action localization without any external label. Actually, P-GCN can still be modified to take external labels into account. To achieve this, we assign the top-2 video-level classes predicted by UntrimmedNet to all the proposals in that video. We provide more details about how to involve external labels in P-GCN in the supplementary material. As summarized in

Figure 3. Action localization results on THUMOS14 with different backbones, measured by mAP@tIoU=0.5.

Table 2, our enhanced version P-GCN* consistently outperforms LIN, hence demonstrating the effectiveness of our method under the same setting.

5. Ablation Studies

In this section, we will perform complete and in-depth ablation studies to evaluate the impact of each component of our model. More details about the structures of baseline methods (such as MLP and MP) can be found in the supplementary material.

5.1. How do the proposal-proposal relations help?

As illustrated in § 3.4, we apply two GCNs for action classification and boundary regression separately. Here, we implement the baseline with a 2-layer MultiLayer-Perceptron (MLP). The MLP baseline shares the same structure as GCN except that we remove the adjacent matrix \mathbf{A} in Eq. (4). To be specific, for the k -th layer, the propagation in Eq. (4) becomes $\mathbf{X}^k = \mathbf{X}^{k-1}\mathbf{W}^k$, where \mathbf{W}^k are the trainable parameters. Without using \mathbf{A} , MLP processes each proposal feature independently. By comparing the performance of MLP with GCN, we can justify the importance of message passing along proposals. To do so, we replace each GCN with an MLP and have the following variants of our model including: (1) $\text{MLP}_1 + \text{GCN}_2$ where GCN_1 is replaced; (2) $\text{GCN}_1 + \text{MLP}_2$ where GCN_2 is replaced; and (3) $\text{MLP}_1 + \text{MLP}_2$ where both GCNs are replaced. Table 3 reads that all these variants decrease the performance of our model, thus verifying the effectiveness of GCNs for both action classification and boundary regression. Overall, our model P-GCN significantly outperforms the MLP protocol (*i.e.* $\text{MLP}_1 + \text{MLP}_2$), validating the importance of considering proposal-proposal relations in temporal action localization.

Table 3. Comparison between our P-GCN model and MLP on THUMOS14, measured by mAP (%).

mAP@tIoU=0.5	RGB	Gain	Flow	Gain
$\text{MLP}_1 + \text{MLP}_2$	34.75	-	43.68	-
$\text{MLP}_1 + \text{GCN}_2$	35.94	1.19	44.59	0.91
$\text{GCN}_1 + \text{MLP}_2$	35.82	1.07	45.26	1.58
P-GCN ($\text{GCN}_1 + \text{GCN}_2$)	37.27	2.52	46.53	2.85

Table 4. Comparison between our P-GCN model and mean-pooling (MP) on THUMOS14, measured by mAP (%).

mAP@tIoU=0.5	RGB	Gain	Flow	Gain
$\text{MP}_1 + \text{MP}_2$	35.32	-	43.97	-
$\text{MP}_1 + \text{GCN}_2$	36.50	1.18	45.78	1.81
$\text{GCN}_1 + \text{MP}_2$	36.22	0.90	44.42	0.45
P-GCN ($\text{GCN}_1 + \text{GCN}_2$)	37.27	1.95	46.53	2.56

Table 5. Comparison of different types of edge functions on THUMOS14, measured by mAP (%).

mAP@tIoU=0.5	RGB	Flow
MLP	34.75	43.68
P-GCN(cos-sim)	35.55	44.83
P-GCN(cos-sim, self-add)	37.27	46.53
P-GCN(embed-cos-sim, self-add)	36.81	46.89

5.2. How does the graph convolution help?

Besides graph convolutions, performing mean pooling among proposal features is another way to enable information dissemination between proposals. We thus conduct another baseline by first adopting MLP on the proposal features and then conducting mean pooling on the output of MLP over adjacent proposals. The adjacent connections are formulated by using the same graph as GCN. We term this baseline as MP below. Similar to the setting in § 5.1, we have three variants of our model including: (1) $\text{MP}_1 + \text{MP}_2$; (2) $\text{MP}_1 + \text{GCN}_2$; and (3) $\text{GCN}_1 + \text{MP}_2$. We report the results in Table 4. Our P-GCN outperforms all MP variants, demonstrating the superiority of graph convolution over mean pooling on capturing between-proposal connections. The protocol $\text{MP}_1 + \text{MP}_2$ in Table 4 performs better than $\text{MLP}_1 + \text{MLP}_2$ in Table 3, which again reveals the benefit of modeling the proposal-proposal relations, even we pursue it using the naive mean pooling.

5.3. Influences of different backbones

Our framework is general and compatible with different backbones (*i.e.*, proposals and features). Beside the backbones applied above, we further perform experiments on TAG proposals [52] and 2D features [27]. We try different combinations: (1) BSN+I3D; (2) BSN+2D; (3) TAG+I3D; (4) TAG+2D, and report the results of MLP and P-GCN in Figure 3. In comparison with MLP, our P-GCN leads to significant and consistent improvements in all types of features

Table 6. Comparison of two types of edge on THUMOS14, measured by mAP (%).

mAP@tIoU=0.5	RGB	Gain	Flow	Gain
w/ both edges (P-GCN)	37.27	-	46.53	-
w/o surrounding edges	35.84	-1.43	45.89	-0.64
w/o contextual edges	36.81	-0.46	45.57	-0.96
w/o both edges (MLP)	34.75	-2.52	43.68	-2.85

Table 7. Comparison of different sampling size and training time for each iteration on THUMOS14, measured by mAP@tIoU=0.5.

N_s	1	2	3	4	5	10
RGB	36.0	36.92	35.68	37.27	36.11	36.37
Flow	46.15	45.06	45.13	46.53	46.28	46.14
Time(s)	0.10	0.23	0.33	0.41	0.48	1.72

and proposals. These results conclude that, our method is generally effective and is not limited to the specific feature or proposal type.

5.4. The weights of edge and self-addition

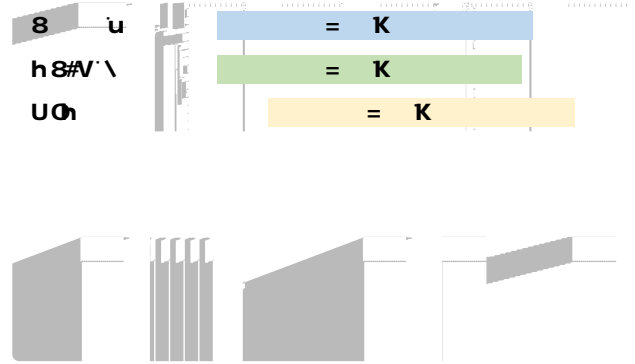
We have defined the weights of edges in Eq. (9), where the cosine similarity (cos-sim) is applied. This similarity can be further extended by first embedding the features before the cosine computation. We call the embedded version as embed-cos-sim, and compare it with cos-sim in Table 5. No obvious improvement is attained by replacing cos-sim with embed-cos-sim (the mAP difference between them is less than 0.4%). Eq. (10) has considered the self-addition of the node feature. We also investigate the importance of this term in Table 5. It suggests that the self-addition leads to at least 1.7% absolute improvements on both RGB and Flow streams.

5.5. Is it necessary to consider two types of edges?

To evaluate the necessity of formulating two types of edges, we perform experiments on two variants of our P-GCN, each of which considers only one type of edge in the graph construction stage. As expected, the result in Table 6 drops remarkably when either kind of edge is removed. Another crucial point is that our P-GCN still boosts MLP when only the surrounding edges are remained. The rationale behind this could be that, actions in the same video are correlated and exploiting the surrounding relation will enable more accurate action classification.

5.6. The efficiency of our sampling strategy

We train P-GCN efficiently based on the neighbourhood sampling in Eq. (10). Here, we are interested in how the sampling size N_s affects the final performance. Table 7 reports the testing mAPs corresponded to different N_s varying from 1 to 5 (and also 10). The training time per iteration is also added in Table 7. We observe that when $N_s = 4$ the model achieves higher mAP than the full model (*i.e.*,



References

- [1] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Proceedings of the British Machine Vision Conference*, 2017. **2, 6**
- [2] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6373–6382. IEEE, 2017. **6**
- [3] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1914–1923, 2016. **6**
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. **2, 5**
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. **2, 3, 6**
- [6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. **1, 2, 4, 6**
- [7] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: fast learning with graph convolutional networks via importance sampling. *International Conference on Learning Representations*, 2018. **3, 5**
- [8] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. **2, 4, 6**
- [9] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018. **2**
- [10] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3059–3069, 2018. **1**
- [11] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Proceedings of the European Conference on Computer Vision*, pages 768–784, 2016. **6**
- [12] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. **1**
- [13] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, pages 5589–5597, 2018. **1**
- [14] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015. **1**
- [15] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yuet-ing Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016. **1**
- [16] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*, pages 923–932, 2016. **1**
- [17] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3628–3636, 2017. **2, 6**
- [18] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *BMVC*, 2017. **1, 2, 6**
- [19] Yong Guo, Qi Chen, Jian Chen, Qingyao Wu, Qinfeng Shi, and Mingkui Tan. Auto-embedding generative adversarial networks for high resolution image synthesis. *TMM*, 2019. **2**
- [20] Will Hamilton, Zitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. **3, 5**
- [21] Rui Hou, Rahul Sukthankar, and Mubarak Shah. Real-time temporal action localization in untrimmed videos by sub-action discovery. In *BMVC*, 2017. **6**
- [22] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. **2**
- [23] Wenbing Huang, Tong Zhang, Yu Rong, and Junzhou Huang. Adaptive sampling towards fast graph representation learning. In *Advances in Neural Information Processing Systems*, pages 4558–4567, 2018. **3, 5**
- [24] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. **2, 5**
- [25] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. **2, 3**
- [26] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 988–996, 2017. **1, 2**
- [27] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *The European Conference on Computer Vision*, September 2018. **6, 7**
- [28] Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *1st NIPS Workshop on Large Scale Computer Vision Systems*, 2016. **2**

- [29] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. The lear submission at thumos 2014. 2014. **6**
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. **2**
- [31] Alexander Richard and Juergen Gall. Temporal action detection using a statistical language model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3131–3140, 2016. **6**
- [32] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *The European Conference on Computer Vision*, September 2018. **3**
- [33] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. **1, 2, 6**
- [34] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016. **1, 2, 6**
- [35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. **1, 5**
- [36] Gurkirt Singh and Fabio Cuzzolin. Untrimmed video classification for activity detection: submission to activitynet challenge. *ActivityNet Large Scale Activity Recognition Challenge*, 2016. **6**
- [37] Mingkui Tan, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Junbin Gao, Fuyuan Hu, and Zhen Zhang. Learning graph structure for multi-label image classification via clique generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4100–4109, 2015. **3**
- [38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. **1**
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. **2**
- [40] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014. **6**
- [41] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. **6**
- [42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 20–36. Springer, 2016. **1**
- [43] Ruxin Wang and Dacheng Tao. Uts at activitynet 2016. *ActivityNet Large Scale Activity Recognition Challenge*, 2016. **6**
- [44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. **5**
- [45] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *The European Conference on Computer Vision*, September 2018. **3**
- [46] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. **2, 6**
- [47] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018. **2, 3**
- [48] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016. **2, 6**
- [49] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kasim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016. **6**
- [50] Zehuan Yuan, Jonathan C. Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. **6**
- [51] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 2019. **2**
- [52] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. **2, 3, 4, 6, 7**
- [53] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware channel pruning for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 875–886, 2018. **2**