ACTION COHERENCE NETWORK FOR WEAKLY SUPERVISED TEMPORAL ACTION LOCALIZATION

Yuanhao Zhai[1] *Le Wang*[1⋆] *Ziyi Liu*[1] *Qilin Zhang*[2] *Gang Hua*[3] *Nanning Zheng*[1]

[1]Xi'an Jiaotong University [2]HERE Technologies [3]Wormpex AI Research

ABSTRACT

Most prominent temporal action localization methods are of the fully-supervised type, which rely heavily on frame-level labels, which could be prohibitively expensive to annotate. Thanks to recent developments on the Weakly-supervised Temporal Action Localization (W-TAL), this alternative paradigm requires only video-level labels in training, alleviating such annotation efforts. Specifically, we present Action Coherence Network (ACN) for W-TAL, which features a new coherence loss that better supervises action boundary learning and facilitate proposal regression. In addition, a purpose-built fusion module is proposed for localization inference based on features extracted by two streams of convolutional neural network. Overall, the proposed ACN achieves state-of-the-art W-TAL performance on two challenging datasets (THUMOS14 and ActivityNet1.2, particularly ACN attains mAP of 24.2% on THUMOS14 under IoU threshold 0.5), which is approaching some recent fully-supervised TAL methods.

***Index Terms*—** weakly-supervised, temporal action localization, coherence loss

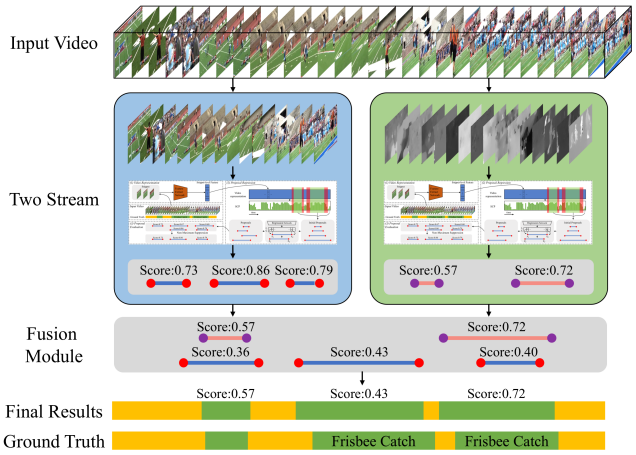

Fig. 1. Overview of the proposed ACN inference. With each untrimmed video, the RGB and optical flow modalities are independently processed by two streams and produce respective action proposals, which are ultimately reconciled by a purpose-built fusion module for final localization.

1. INTRODUCTION

Temporal action localization is an important learning problem for high-level video understanding tasks, such as event detection, video summarization, and visual question answering. Thanks to the advances of deep learning, multiple breakthroughs have been made on temporal action localization [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Most of these prominent methods require full supervision during training, *i.e.*, training videos need precise annotations of all the start and end temporal locations of all action instances. However, large-scale labeling is expensive and time-consuming, and could be inconsistent due to ambiguous transitional actions.

In contrast, video-level label could potentially be automatically obtained with textual search terms on video sharing websites, without incurring significant financial cost or labeling delays. With only the video-level labels in training data, the Weakly-supervised Temporal Action Localization (W-TAL) paradigm proposed by Sun et al. [13] offers an appealing alternative and multiple efforts have been made in W-TAL. Hide-and-Seek [14] randomly hides parts of the input to guide the network to learn the most relevant parts. UntrimmedNet [15] is an end-to-end framework, which uses classification results to detect important snippets. AutoLoc [16] proposes the Outer-Inner-Contrastive (OIC) loss to detect action boundaries and uses it to regress proposals. W-TALC [17] achieves the state-of-the-art results, which divides the network into two separate sub-networks and introduces co-activity similarity loss.

Despite these recent efforts, two major challenges still persist. (1) Most current W-TAL methods only exploit the classification scores and attention weights to determine action instance boundaries without explicit constrains on frame appearance changes, which could lead to sub-optimal performance. To address this challenge, we proposed the new coherence loss that accounts for both appearance coherence and snippet-level classification activation. (2) Inspired by two-stream CNNs [18], we speculate that the characteristics of RGB and optical flow modalities are largely ignored in the de-facto standard practice of direct concatenation-based stream-fusion. Intuitively, RGB stream is sensitive to scene transi-

⋆Corresponding author, *lewang@xjtu.edu.cn*

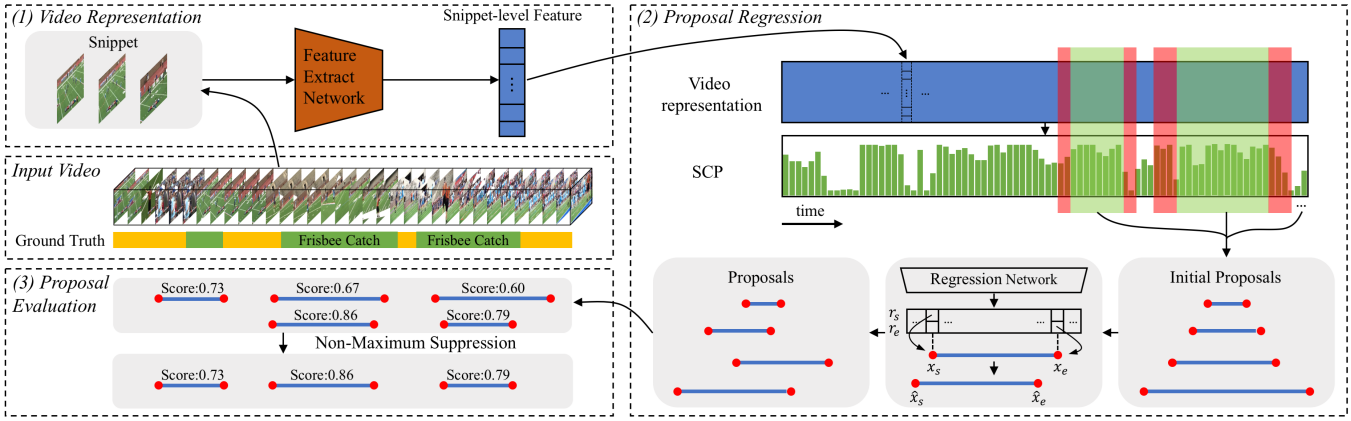

Fig. 2. RGB stream of the proposed ACN. (1) UntrimmedNet is used to encode the snippet-level feature. (2) We slide multiple regression networks along the feature sequence and Snippet-level Classification Prediction (SCP), with each network generating and regressing proposals independently. (3) Non-Maximum Suppression (NMS) eliminatess redundant proposals.

tions, but tends to neglect slight movements. Flow stream is more sensitive to slight movements but may introduce lots of noises during scene transitions or camera movements. In this paper, a purpose-built fusion module is proposed to account for the different modality characteristics and produces empirically better temporal localization inferences. The overview of our proposed Action Coherence Network (ACN) is shown in Figure 1.

2. COHERENCE LOSS

An ideal action instance is assumed to have distinctive temporal boundaries, which are estimated in Autoloc [16] with the OIC loss. Formally, given the Snippet-level Classification Prediction (SCP) $\mathbf{S} \in \mathbb{R}^{C\times T}$ of a video with T snippets and C action categories, the OIC loss for a proposal $[x_s, x_e]$ of the action class $k \in 1, \cdots, C$ is

$$L_{OIC} = \frac{\int_{X_s}^{X_e} S(k,u)du - \int_{x_s}^{x_e} S(k,u)du}{(X_e - X_s + 1) - (x_e - x_s + 1)} - \frac{\int_{x_s}^{x_e} S(k,u)du}{x_e - x_s + 1}, \tag{1}$$

where X_s, X_e are the inflated boundaries and $S(k,n)$ denotes the matrix element corresponding to SCP for the n-th snippet of class k.

Therefore, the OIC loss focuses exclusively on high activation for snippet-level classification without explicit terms relevant to the action instance content. Intuitively, we argue that a well-defined loss function should explicitly promote "clear" appearance distinctions between its preceding and succeeding snippets in both the RGB and Flow representation. Following such intuition, the coherence term in the coherence loss is formulated as an arithmetic average of the cosine similarities between the action area and its "start area" and "end area". Given a proposal $[x_s, x_e]$ and inflated boundary $[X_s, X_e]$, its start area feature $\mathbf{R}_s$, end area feature $\mathbf{R}_e$ and action area feature $\mathbf{R}_a$ are defined respectively as

$$\mathbf{R}_s = \frac{\int_{X_s}^{x_s} \mathbf{F}(u)du}{x_s - X_s + 1}, \mathbf{R}_e = \frac{\int_{x_e}^{X_e} \mathbf{F}(u)du}{X_e - x_e + 1}, \mathbf{R}_a = \frac{\int_{x_s}^{x_e} \mathbf{F}(u)du}{x_e - x_s + 1}, \tag{2}$$

and its corresponding coherence term L_c is computed as

$$L_c = \frac{1}{2}\left(\frac{\langle \mathbf{R}_a, \mathbf{R}_s\rangle}{\langle \mathbf{R}_a, \mathbf{R}_a\rangle^{\frac{1}{2}} \langle \mathbf{R}_s, \mathbf{R}_s\rangle^{\frac{1}{2}}} + \frac{\langle \mathbf{R}_a, \mathbf{R}_e\rangle}{\langle \mathbf{R}_a, \mathbf{R}_a\rangle^{\frac{1}{2}} \langle \mathbf{R}_e, \mathbf{R}_e\rangle^{\frac{1}{2}}}\right). \tag{3}$$

where $\mathbf{F}(n)$ denotes the n-th snippet feature representation, $\langle , \rangle$ denotes inner product.

With the new coherence term L_c accounting for the appearance distinctions, the proposed coherence loss L is introduced as a hybrid of appearance coherence and snippet-level classification activation,

$$L = \alpha L_{OIC} + (1-\alpha)(L_c - 1), \tag{4}$$

where $\alpha \in (0,1)$ is a trade-off constant empirically set to 0.6 and fixed thereafter.

3. ACTION COHERENCE NETWORK

3.1. Video Representation

We first divide each video into non-overlapping fixed length (15 frames) snippets and extract snippet-level feature with UntrimmedNet [15] (soft selection method). The Inception network with Batch Normalization is used as network backbone for each stream. Features are extracted as the 1024-dimensional tensor at the global average pooling layer. Given a video with T snippets, we denote the RGB feature and optical flow feature as $\mathbf{F}_r, \mathbf{F}_f \in \mathbb{R}^{1024\times T}$ respectively. In addition, the RGB and optical flow attention weight $\mathbf{A}_r, \mathbf{A}_f \in \mathbb{R}^T$ and SCP $\mathbf{S}_r, \mathbf{S}_f \in \mathbb{R}^{C\times T}$ are also obtained by UntrimmedNet, respectively.

3.2. Proposal Regression

For conciseness, we illustrate only the RGB stream of ACN in Figure 2 as an example of proposal regression[1]. Inspired by the Faster R-CNN [19], we feed video representation $\mathbf{R}_r$ into multiple regression networks, each consisting of 3 temporal convolutional layers and assigned with a fixed anchor size P.The first two convolutional layers have 256 dilated filters with kernel size 3 temporally and stride 1. The receptive field of all regression networks are identical to their anchor size P to ensure sufficient (but not excessively redundant) context information. Specifically, the dilation of the first two layers is $\frac{P}{4}$, leading to a receptive field of $(\frac{P}{4}+\frac{P}{4}+2)\times 2+1\approx P$. The last convolutional layer has 2 filters with kernel size 1 and stride 1. In addition, zero padding is used for the first two layers to ensure the matching temporal location outputs.

Given anchor size P, we initialize action proposals at all possible temporal snippet locations as $\{(x_{s,i},x_{e,i})\}_{i=1}^{T-P}$, such that $x_{e,i}-x_{s,i}=P$. Subsequently, these initial proposals are refined by the corresponding regression network (details in Section 4.2), with which generating respective temporal regression results $\{r_{s,i}\}_{i=1}^{T}$ and $\{r_{e,i}\}_{i=1}^{T}$, so that the estimated proposal boundaries $(\hat{x}_{s,i},\hat{x}_{e,i})$ are:

$$\hat{x}_{s,i}=x_{s,i}+P\cdot\left(\text{sigmoid}(r_{s,x_{s,i}})-\frac{1}{2}\right), \tag{5}$$

$$\hat{x}_{e,i}=x_{e,i}+P\cdot\left(\text{sigmoid}(r_{e,x_{e,i}})-\frac{1}{2}\right). \tag{6}$$

In this manner, each boundary is able to regress to any temporal location within its receptive field.

3.3. Proposal Evaluation

We inflate the temporal boundaries of each proposal $[x_s,x_e]$ to $[X_s,X_e]$, where $X_s=x_s-\frac{P}{4}$ and $X_e=x_e+\frac{P}{4}$ to account for the context information. We define the confidence score of a proposal as the negation of its coherence loss, namely $-L$.

During training, the algorithm traverses every action category with classification score over the predefined threshold of 0.1 to check the SCP values. If the SCP of a temporal snippet position is lower than a threshold (set to 0.1), all proposals containing this position are discarded. Subsequently, we only keep one proposal per snippet location which achieves the highest score among all proposals covering this location of different anchor sizes P and discard all others. The overall score is the arithmetic average of all remaining proposal scores.

During testing, Non-Maximum Suppression (NMS) is performed with overlap Intersection-of-Union (IoU) threshold 0.4, which is empirically determined via cross-validation.

[1]The Flow stream shares similar settings.

3.4. Proposal Fusion

With the obtained proposals from RGB and Flow streams, a fusion module is proposed for ACN to select and reconcile such proposals. Empirically, the Flow stream typically provides more accurate proposals thanks to its sensitivity to even subtle motions, which statistically corresponds well with the start and end areas of action instances. Based on this observation, we use Flow stream as the primary source and RGB stream as the auxiliary one. Let $\{p_{f,j}\}_{j=1}^{N_f}$ and $\{p_{r,j}\}_{j=1}^{N_r}$ denote proposals from the Flow and RGB steam, respectively, with N_f and N_r denoting the number of kept proposals in Flow and RGB stream, respectively.

The fusion module first retains all $p_{f,j}$, $j=1,\cdots,N_f$ and simultaneously discounts all RGB proposal confidence score by a factor[2] of 2. Subsequently, for each $p_{r,j}$, we calculate its overlap IoUs with all $p_{f,j}$, $j=1,\cdots,N_f$ and obtain a retention score $I(p_{r,j})$ by max-pooling,

$$I(p_{r,j})=\max\left(\text{IoU}(p_{r,j},p_{f,1}),\cdots,\text{IoU}(p_{r,j},p_{f,N_f})\right). \tag{7}$$

The reconciled proposals are the union of all $p_{f,j}$, $j=1,\cdots,N_f$ and the set of $p_{r,j}$ with $I(p_{r,j})<0.4$.

4. EXPERIMENT

4.1. Dataset and Evaluation

THUMOS14 [20] dataset contains 200 untrimmed videos in validation set and 213 untrimmed videos in test set with 20 classes for the temporal action localization task. We use the validation set for training and the test set for testing.

ActivityNet v1.2 [21] covers 100 action classes, with training set containing 4819 untrimmed videos and the validation set containing 2383 untrimmed videos, which are used in our training and testing, respectively. This dataset is processed according to the settings in UntrimmedNet [15] and Autoloc [16] for fair comparison.

We use mean Average Precision (mAP) at different levels of IoU thresholds to measure the performance of all localization results.

4.2. Implementation Details

ACN is implemented on PyTorch [22]. We train each stream 3 epochs with the stochastic gradient descent optimizer, an initial learning rate of 0.001, and a decay factor of 10 per epoch. The mini-batch size is set to 4. To alleviate the background noise, attention thresholding is employed during testing, all snippets with attention weight lower than a threshold (empirically fixed at 5 for Flow stream and 7 for RGB stream) are discarded. Using grid search in cross-validation, we set α to 0.6 in Equation (4) for both streams and both datasets. We

[2]To alleviating its overfitting tendency, the factor 2 is empirically determined via cross validation.

Table 1. Comparison with state-of-the-art weakly-supervised methods on ActivityNet v1.2 validation set.

Method	mAP@IoU										Avg
	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	
UntrimmedNet [15]	7.4	6.1	5.2	4.5	3.9	3.2	2.5	1.8	1.2	0.7	3.6
Autoloc [16]	27.3	24.9	22.5	19.9	17.5	15.1	13.0	10.0	6.8	3.3	16.0
Ours-ACN	**30.4**	**27.2**	**24.3**	**20.5**	**18.0**	**15.4**	**13.2**	**10.3**	**7.5**	**3.7**	**17.0**

Table 2. Comparison with state-of-the-art weakly-supervised methods on THUMOS14 test set.

Method	mAP@IoU				
	0.3	0.4	0.5	0.6	0.7
Hide-and-Seek [23]	19.5	12.7	6.8	-	-
UntrimmedNet [15]	28.2	21.1	13.7	8.3	4.2
STPN [24]	35.5	25.8	16.9	9.9	4.3
AutoLoc [16]	35.8	29.0	21.2	13.4	5.8
W-TALC+UNTF [17]	32.0	26.0	18.8	-	6.2
Ours-ACN	**35.9**	**30.7**	**24.2**	**15.7**	**7.4**

choose anchor sizes P of the snippet length 1, 2, 4, 8, 16, 32 for THUMOS14 and 16, 32, 64, 128, 256, 512 for ActivityNet.

4.3. Comparison with State-of-the-arts

Performance results[3] on ActivityNet v1.2 validation set are presented in Table 1, where the proposed ACN outperforms all existing state-of-the-art W-TAL methods. Particularly, UntrimmedNet [15] did not report their temporal localization performance on ActivityNet 1.2 in their paper, but they publicly released their trained model and source codes, based on which we re-evaluate and obtain the UntrimmedNet [15] performance in Table 1.

The mAP comparison on THUMOS14 test set is summarized in Table 2, where ACN outperforms all competing methods at all IoU thresholds.

4.4. Sensitivity and Ablation Study

Sensitivity and ablation analysis is carried out with localization performance evaluated with different mAP@IoU thresholds and with variations of ablated coherence loss.

Sensitivity Analysis on α. We test different α in Equation (4) during both training and evaluating phases. The results are measured with mAP@IoU 0.5 and are summarized in Table 3, which justifies our empirical choice of $\alpha = 0.6$.

Ablation Study on Proposal Scoring Method. As presented in Section 3.3, the negation of Coherent loss L is also used to score proposals. We included two additional variants of ablated scoring methods (for both training and evaluating phases), the coherence term $-L_c$ only and OIC term $-L_{OIC}$ only. The temporal localization performance on the RGB stream-only and Flow stream-only is summarized in Table 4 and Table 5, respectively. The performance advantage of $-L$ as scoring method verifies that both terms are indispensable and both are jointly contributing to the scoring.

[3]For fair comparison, the results of W-TALC+I3DF are not listed here, because they used a more complicated backbone.

Table 3. Sensitivity analysis: α in Equation (4) on THUMOS14.

Modality	mAP@IoU 0.5 at α values				
	0.4	0.5	0.6	0.7	0.8
RGB	8.8	9.1	**9.7**	9.6	9.4
optical flow	21.5	22.2	**22.8**	22.7	22.6

Table 4. RGB stream-only localization performance with different scoring methods on THUMOS14 test set.

Method	mAP@IoU				
	0.3	0.4	0.5	0.6	0.7
$-L_c$	11.3	7.8	4.2	1.9	0.7
$-L_{OIC}$	19.6	14.4	9.1	4.5	1.4
$-L$	**20.9**	**14.8**	**9.7**	**4.5**	**1.9**

Table 5. Flow stream-only localization performance with different scoring methods on THUMOS14 test set.

Method	mAP@IoU				
	0.3	0.4	0.5	0.6	0.7
$-L_c$	13.7	10.0	6.8	3.9	1.7
$-L_{OIC}$	33.4	28.1	22.1	14.4	6.9
$-L$	**35.2**	**29.7**	**22.8**	**15.0**	**7.2**

5. CONCLUSION

In this paper, we have proposed the ACN for weakly-supervised temporal action localization with a new coherence loss and a purpose-built fusion module reconciling both optical flow and RGB-based action proposals. Experiments on two datasets have verified the performance advantage, with additional sensitivity and ablation analysis demonstrating some design intuitions.

6. ACKNOWLEDGEMENT

This work was supported partly by National Key R&D Program of China Grant 2017YFA0700800, NSFC Grants 61629301 and 61773312, China Postdoctoral Science Foundation Grant 2019M653642, and Young Elite Scientists Sponsorship Program by CAST Grant 2018QNRC001.

7. REFERENCES

[1] Zheng Shou, Dongang Wang, and Shih-Fu Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016, pp. 1049–1058.

[2] Zhao Yue, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin, "Temporal action detection with structured segment networks," in *ICCV*, 2017.

[3] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang, "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *CVPR*, 2017, pp. 1417–1426.

[4] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: region convolutional 3d network for temporal activity detection," in *ICCV*, 2017, pp. 5794–5803.

[5] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou, "Exploring temporal preservation networks for precise temporal action localization," in *AAAI*, 2018.

[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 *CVPR*, 2018, pp. 1130–1139.

[7] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *ECCV*, 2018.

[8] Haisheng Su, Xu Zhao, and Tianwei Lin, "Cascaded pyramid mining network for weakly supervised temporal action localization," *arXiv preprint arXiv:1810.11794*, 2018.

[9] Yunlu Xu, Chengwei Zhang, Zhanzhan Cheng, Jianwen Xie, Yi Niu, Shiliang Pu, and Fei Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," *arXiv preprint arXiv:1811.07460*, 2018.

[10] Zijian Kang, Le Wang, Ziyi Liu, Qilin Zhang, and Nanning Zheng, "Extracting action sensitive features to facilitate weakly-supervised action localization," in *AIAI*, 2019.

[11] Zhanning Gao, Le Wang, Qilin Zhang, Zhenxing Niu, Nanning Zheng, and Gang Hua, "Video imprint segmentation for temporal action detection in untrimmed videos," in *AAAI*, 2019.

[12] Ziyi Liu, Le Wang, Gang Hua, Qilin Zhang, Zhenxing Niu, Ying Wu, and Nanning Zheng, "Joint video object discovery and segmentation by coupled dynamic markov networks," *TIP*, pp. 5840–5853, 2018.

[13] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *ACM MM*, 2015, pp. 371–380.

[14] Krishna Kumar Singh and Yong Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*, 2017, pp. 3524–3533.

[15] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *CVPR*, 2017, pp. 6402–6411.

[16] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *ECCV*, 2018, pp. 162–179.

[17] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *ECCV*, 2018, pp. 588–607.

[18] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[20] YG Jiang, Jingen Liu, A Roshan Zamir, G Toderici, I Laptev, Mubarak Shah, and Rahul Sukthankar, "Thumos challenge: Action recognition with a large number of classes," 2014.

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015, pp. 961–970.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[23] Krishna Kumar Singh and Yong Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*, 2017, pp. 3524–3533.

[24] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han, "Weakly supervised action localization by sparse temporal pooling network," in *CVPR*, 2018, pp. 6752–6761.