

# Supplementary Material

## AEI: Actors-Environment Interaction with Adaptive Attention for Temporal Action Proposals Generation

Khoa Vo<sup>1</sup>  
khoavoho@uark.edu

Hyekang Joo<sup>\*2</sup>  
hkjoo@cs.umd.edu

Kashu Yamazaki<sup>\*1</sup>  
kyamazak@uark.edu

Sang Truong<sup>1</sup>  
sangt@uark.edu

Kris Kitani<sup>3</sup>  
kkitani@cs.cmu.edu

Minh-Triet Tran<sup>4,5</sup>  
tmtriet@hcmus.edu.vn

Ngan Le<sup>1</sup>  
thile@uark.edu

<sup>1</sup> AICV Lab, University of Arkansas  
Fayetteville, AR, USA

<sup>2</sup> University of Maryland  
College Park, MD, USA

<sup>3</sup> Carnegie Mellon University  
Pittsburgh, PA, USA

<sup>4</sup> University of Science, VNU-HCM  
Ho Chi Minh City, Vietnam

<sup>5</sup> Vietnam National University  
Ho Chi Minh City, Vietnam

In this supplementary, we provide more a detailed description of our proposed Adaptive Attention Mechanism (ATT). Afterwards, network configurations of both CNN-based (AEI-B) and GCN-based (AEI-G) architectures are reported. Additionally, we conduct additional experiments to express the effectiveness of our proposed actors spectator and our proposed adaptive attention. Finally, we show qualitative results to illustrate the performance comparison between our proposed AEI-B and AEI-G with other SOTA methods on both temporal action proposal generation (TAPG) and temporal action detection (TAD) tasks. <sup>1</sup>

## 1 Adaptive Attention Mechanism

To begin, the environment feature  $f^e$  and actor features  $F^a$  are embedded into the same dimensional space by a multi-layer perceptron (MLP) parameterized by  $\theta$   $MLP_\theta(\cdot)$ :

$$\hat{f}^e = MLP_{\theta_e}(f^e) \quad (1)$$

$$\hat{F}^a = \{\hat{f}^a\}_{i=1}^{N_B} \text{ where } \hat{f}^a = MLP_{\theta_a}(f_i^a) \quad (2)$$

Then, both  $\hat{f}^e$  and  $\hat{F}^a$  are combined by element-wise addition (i.e.,  $\oplus$ ) to form a collaborative feature  $\hat{F}^c$ :  $F^c = \{f_i^c\}_{i=1}^{N_B}$ , where  $f_i^c = \hat{f}_i^a \oplus \hat{f}_i^e$ . Afterwards, we compute the L2-norm of each collaborative feature  $F^c$ . It is proven that features with the greatest L2-norm values carry meaningful information and better contribute to later modules [14], i.e.,  $A_c = \{a_i^c\}_{i=1}^{N_B}$ , where  $a_i^c = \|f_i^c\|_2$ .

Next, we re-scale all L2-norm values by softmax function to be summed up to 1.0, because L2-norm values are ranged arbitrarily:

$$A_c = \{\hat{a}_i^c\}_{i=1}^{N_B} \text{ where } \hat{a}_i^c = \frac{e^{a_i^c}}{\sum_{i=1}^{N_B} e^{a_i^c}} \quad (3)$$

We obtain main actor feature vectors:

$$\tilde{F}^a = \{f_i^a | \hat{a}_i^c > \tau\} \text{ where } \tau = \frac{1}{|\hat{A}_c|} \quad (4)$$

After that, we fuse a set of main actors feature vectors  $\tilde{F}^a$  into a single feature vector  $f^a$  by leveraging self-attention model proposed in [14].  $\tilde{F}^a$  is fed into three MLPs to build three intermediate features  $Q$ ,  $K$ , and  $V$ . Each of these has a specific role to form the relations between actors and re-weight each of them based on their relevance in the relations. Each  $q_i \in Q$ , corresponding to  $f_i^a$ , is used to generate an attention mask. Each value  $v_i \in V$  is re-weighted by this attention mask. The whole process is defined as

Table 1: The detailed architecture of CNN-based BMM.  $F$  is the input feature dimensions.  $T$  and  $D$  are the temporal length of the video and maximum duration of proposals in terms of number of snippets. The obtained outputs are  $O_T$  and  $O_P$ , which are corresponding to boundary-predictions and proposal actionness scores.

Layers	Input	Output
1DConv. $256 \times 3/1$ , ReLU	$I : F \times T$	$O_1 : 256 \times T$
1DConv. $128 \times 3/1$ , ReLU	$O_1 : 256 \times T$	$O_2 : 128 \times T$
1DConv. $256 \times 3/1$ , ReLU	$O_2 : 128 \times T$	$O_3 : 256 \times T$
1DConv. $2 \times 3/1$ , Sigmoid	$O_3 : 256 \times T$	$O_T : 2 \times T$
Matching layer	$O_2 : 128 \times T$	$O_5 : 128 \times 32 \times D \times T$
3DConv. $512 \times 32 \times 1 \times 1/(32,0,0)$ , ReLU	$O_5 : 128 \times 32 \times D \times T$	$O_6 : 512 \times 1 \times D \times T$
squeeze	$O_6 : 512 \times 1 \times D \times T$	$O_7 : 512 \times D \times T$
2DConv. $128 \times 1 \times 1/(0,0)$ , ReLU	$O_7 : 512 \times D \times T$	$O_8 : 128 \times D \times T$
2DConv. $128 \times 3 \times 3/(1,1)$ , ReLU	$O_8 : 128 \times D \times T$	$O_9 : 128 \times D \times T$
2DConv. $2 \times 1 \times 1/(0,0)$ , Sigmoid	$O_9 : 128 \times D \times T$	$O_P : 2 \times D \times T$

Table 2: The detailed architecture of GCN-based BMM.  $F$  is the input feature dimensions.  $T$  and  $D$  are the temporal length of the video and maximum duration of proposals in terms of number of snippets. The obtained outputs are  $O_T$  and  $O_P$ , which are corresponding to boundary-predictions and proposal actionness scores.

Layers	Input	Output
1DConv. $256 \times 3/1$ , ReLU	$I : F \times T$	$O_1 : 256 \times T$
G-Conv.	$O_1 : 256 \times T$	$O_2 : 256 \times T$
G-Conv.	$O_2 : 256 \times T$	$O_3 : 256 \times T$
1DConv. $2 \times 1/1$ , Sigmoid	$O_3 : 256 \times T$	$O_T : 2 \times T$
G-Conv.	$O_2 : 256 \times T$	$O_5 : 256 \times T$
Matching layer	$O_5 : 256 \times T$	$O_6 : 8192 \times D \times T$
2DConv. $512 \times 1 \times 1/(1,1)$ , ReLU	$O_6 : 8192 \times D \times T$	$O_7 : 512 \times D \times T$
2DConv. $128 \times 1 \times 1/(1,1)$ , ReLU	$O_7 : 512 \times D \times T$	$O_8 : 128 \times D \times T$
2DConv. $128 \times 3 \times 3/(1,1)$ , ReLU	$O_8 : 128 \times D \times T$	$O_9 : 128 \times D \times T$
2DConv. $128 \times 3 \times 3/(1,1)$ , ReLU	$O_9 : 128 \times D \times T$	$O_{10} : 128 \times D \times T$
2DConv. $2 \times 1 \times 1/(1,1)$ , Sigmoid	$O_{10} : 128 \times D \times T$	$O_P : 2 \times D \times T$

Table 3: G-Conv. layer

Layers	Input	Output
1DConv. $128 \times 1/1$ , ReLU	$I: 256 \times T$	$O_1: 128 \times T$
1DConv. $128 \times 3/1$ , ReLU	$O_1: 128 \times T$	$O_2: 128 \times T$
1DConv. $256 \times 1/1$	$O_2: 128 \times T$	$O_3: 256 \times T$
kNN	$I: 256 \times T$	$I': 512 \times T \times k$
2DConv. $128 \times 1 \times 1/(1,1)$ , ReLU	$I': 512 \times T \times k$	$O'_1: 128 \times T \times k$
2DConv. $128 \times 1 \times 1/(1,1)$ , ReLU	$O'_1: 128 \times T \times k$	$O'_2: 128 \times T \times k$
2DConv. $256 \times 1 \times 1/(1,1)$	$O'_2: 128 \times T \times k$	$O'_3: 256 \times T \times k$
Maxpool	$O'_3: 256 \times T \times k$	$O''_3: 256 \times T$
ReLU	$I + O_3 + O''_3$	$O_4: 256 \times T$

$\mathcal{A}(q_i, K, V) = \text{softmax}(\frac{q_i \cdot K^T}{\sqrt{d_K}})V$ , where  $d_K$  is the number of dimensions of features in  $K$ , following [14]. Finally, we obtain the actors visual feature  $f^a$  as:  $f^a = \frac{1}{|Q|} \sum_i^{|Q|} \mathcal{A}(q_i, K, V)$ .

## 2 Boundary Matching Module Configuration

In our proposed AEI framework, we examine the boundary matching module (BMM) with both CNN architecture and GCN architecture. The network configuration of CNN-based BMM is described as in Table 1, whereas that of GCN-based BMM is described as in Table 2 and the G-Conv. layer is defined in Table 3.

## 3 Additional Experiments

In this section, we first examine our proposed AEI-B and AEI-G on various features in Subsec. 3.1. We then investigate our proposed framework in the cases of with and without using the proposed actors spectator in Subsec 3.2. To prove the effectiveness of our proposed adaptive attention mechanism, we have conducted the comparison between adaptive attention and other attention mechanisms in Subsec. 3.3.

### 3.1 Various Features

The performance of our proposed CNN-based (AEI-B) and GCN-based (AEI-G) architectures on different features (i.e. C3D [14], 2Stream [15] and Slowfast [16]) is reported in Table 4 on ActivityNet-1.3 dataset [14]. The first part of the table is corresponding to the SOTA methods, whereas the second part of the table presents our performance on various types of features. As demonstrated, our proposed AEI outperforms other SOTA methods regardless of the boundary matching network backbones (either CNN-based or GCN-based) and regardless of feature network backbones (either C3D [14], 2Stream [15] or Slowfast [16]).

Table 4: Performance of AEI-B and AEI-G with **various features** i.e. C3D [10], 2Stream [15] and Slowfast [16] on TAPG and ActivityNet-13 [17] dataset

	Feature	AR@100	AUC(val)	AUC(test)
TCN [10]	2Stream	-	59.58	61.56
MSRA [15]	P3D	-	63.12	64.18
SSTAD [10]	C3D	73.01	64.40	64.80
CTAP [9]	2Stream	73.17	65.72	-
BSN [15]	2Stream	74.16	66.17	66.26
SRG [9]	2Stream	74.65	66.06	-
MGG [16]	I3D	74.54	66.43	66.47
BMN [16]	2Stream	75.01	67.10	67.19
DBG [9]	2Stream	76.65	68.23	68.57
BSN++ [16]	2Stream	76.52	68.26	-
TSI++ [16]	2Stream	76.31	68.35	68.85
MR[16].	I3D	75.27	66.51	-
AEI-B	C3D	77.25	69.43	69.94
	2Stream	76.64	68.48	69.21
	Slowfast	76.73	68.94	69.32
AEI-G	C3D	77.24	69.47	70.09
	2Stream	76.66	68.41	69.10
	Slowfast	77.00	69.18	69.66

Table 5: Effectiveness of our proposed **Actor** in **TAPG** on ActivityNet-1.3 [17]

		AR@100	AUC(val)	AUC(test)
AEI-B	W/O Actor	77.07	67.55	68.87
	<b>With Actor</b>	77.25	69.43	69.94
AEI-G	W/O Actor	76.33	68.68	68.74
	<b>With Actor</b>	77.24	69.47	70.09

## 3.2 Actors Spectator

Actors spectator is one of the main components in our proposed cognitive-based visual representation (CVR). In addition to the ablation study in the main manuscript, we conduct further experiments in the cases of with and without using our proposed actors spectator as follows:

- Case 1: Without actors spectator, CVR is mainly based on environment representation. This case is equivalent to the scenario where only environment is taken into consideration.
- Case 2: With actors spectator, CVR is computed based on both actor and environment, and then on the interaction between actor and environment. This case is our proposed framework.

The efficiency of actors spectator on both TAPG and TAD is examined and given in Tables 5, 6, 7, 8 corresponding to ActivityNet-1.3 dataset [17] and THUMOS-14 dataset [8]. As demonstrated, the proposed actors spectator helps to boost the performance of AEI by a large margin for both TAPG and TAD.



Table 6: Effectiveness of our proposed **Actor** in **TAPG** on THUMOS-14 [8]

		@50	@100	@200	@500	@1000
AEI-B	W/O Actor	37.68	46.48	53.89	61.22	65.45
	<b>With Actor</b>	44.97	50.13	57.34	64.43	67.78
AEI-G	W/O Actor	38.94	47.80	54.93	61.92	65.96
	<b>With Actor</b>	45.31	51.12	58.19	64.58	67.96

Table 7: Effectiveness of our proposed our proposed **Actor** in **TAD** on ActivityNet-1.3 [4]

		0.5	0.75	0.95	Average
AEI-B	W/O Actor	50.2	32.7	9.5	32.7
	<b>With Actor</b>	52.3	34.5	9.7	34.7
AEI-G	W/O Actor	50.3	32.7	9.5	32.8
	<b>With Actor</b>	52.4	34.5	9.6	34.7

Table 8: Effectiveness of **Actor** in **TAD** on THUMOS-14 [8]

		0.7	0.6	0.5	0.4	0.3
AEI-B	W/O Actor	22.2	36.6	50.5	59.5	66.8
	<b>With Actor</b>	22.3	37.9	52.0	60.4	67.6
AEI-G	W/O Actor	22.4	36.1	50.3	59.8	66.5
	<b>With Actor</b>	22.4	37.8	52.1	60.6	67.3

Table 9: Effectiveness of our proposed **Adaptive Attention** in **TAPG** on ActivityNet-1.3 [4]

		AR@100	AUC(val)	AUC(test)
AEI-B	Hard [14]	76.93	69.06	69.20
	Soft [17]	76.72	69.16	69.26
	<b>Adaptive</b>	77.25	69.43	69.94
AEI-G	Hard [14]	77.21	68.97	69.41
	Soft [17]	77.25	69.36	69.69
	<b>Adaptive</b>	77.24	69.47	70.09

### 3.3 Adaptive Attention

Adaptive attention is one of the main components in our proposed cognitive-based visual representation (CVR) that is designed to model the relationship between actor(s) and environment. In this section, we investigate the strength of our proposed adaptive attention by comparing it to hard-attention [14] and soft-attention [17] mechanisms. Tables 9, 10, 11 provide the performances of both AEI-B and AEI-G using three different attention mechanisms. While soft-attention mechanism [17] and hard-attention mechanism [14] yield equivalent performance, our adaptive attention mechanism achieves the SOTA performance on both TAGP and TAD.

## 4 Qualitative Results

In this section, we visualize some examples to illustrate the performance comparison between our proposed AEI-B and AEI-G with SOTA methods (e.g., BMN [14], DBG [9]) on

Table 10: Effectiveness of our proposed **Adaptive Attention** in **TAPG** on THUMOS-14 [8]

		@50	@100	@200	@500	@1000
AEI-B	Hard [14]	42.10	50.00	56.57	63.48	67.49
	Soft [14]	42.18	49.65	56.66	63.39	66.77
	<b>Adaptive</b>	44.97	50.13	57.34	64.43	67.78
AEI-G	Hard [14]	43.26	49.92	56.93	63.75	67.72
	Soft [14]	40.26	49.67	56.34	63.41	67.32
	<b>Adaptive</b>	45.31	51.12	58.19	64.58	67.96

Table 11: Effectiveness of our proposed **Adaptive Attention** in **TAD** on THUMOS-14 [8]

		0.7	0.6	0.5	0.4	0.3
AEI-B	Hard [14]	22.4	36.7	51.2	60.1	67.2
	Soft [14]	22.1	37.5	51.4	60.1	66.6
	<b>Adaptive</b>	22.3	37.9	52.0	60.4	67.6
AEI-G	Hard [14]	22.1	37.0	51.2	60.1	67.2
	Soft [14]	21.7	37.5	51.4	60.2	66.7
	<b>Adaptive</b>	22.4	37.8	52.1	60.6	67.3

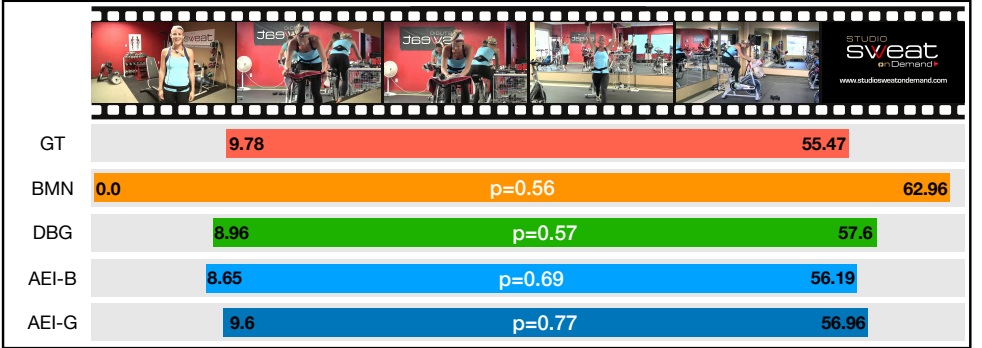


Figure 1: Qualitative comparison between our proposed method (i.e. AEI-B and AEI-G) with other SOTA methods (i.e. BMN [14], DBG [9]). The human being main actor occupies a large area in spatial domain.

TAPG as given in Figs. 1, 2, 3, 4. The performance of each method is shown in a tuple of starting time, ending time, and confident score. The performance comparison is made under following cases:

- **Human-being main actor:** Figs. 1 shows an example where the main actor occupies a big area of the spatial domain. In this case, spatial feature mainly presents the main actor who commits the action. In other words, the spatial feature carries the action information; thus, other SOTA methods (i.e. BMN [14], DBG [9]) obtains good performance while our AEI slightly improves. Figs. 2 shows an example where the main actor occupies a small area of the spatial domain. In this case, spatial feature mainly presents the environment that does not carry the action information. Thus, other SOTA methods are unable to perform well. On the other hand, our AEI with actors spectator and adaptive mechanism is able to focus on the main actor to capture action informa-

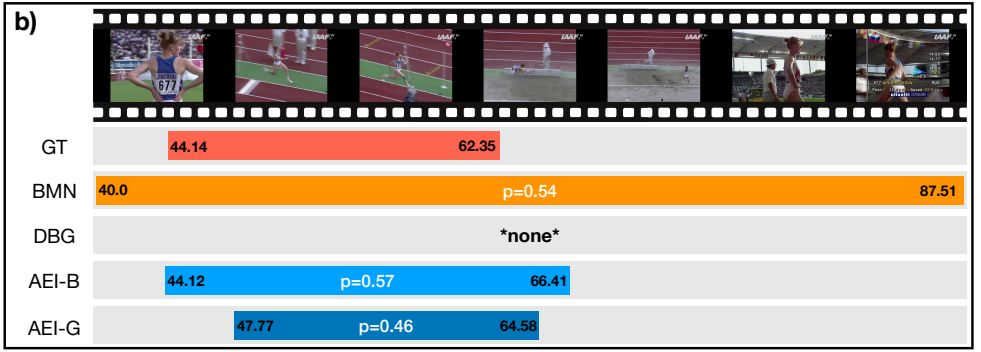


Figure 2: Qualitative comparison between our proposed method (i.e. AEI-B and AEI-G) with other SOTA methods (i.e. BMN [10], DBG [11]). The human-being main actor occupies a small area in spatial domain.

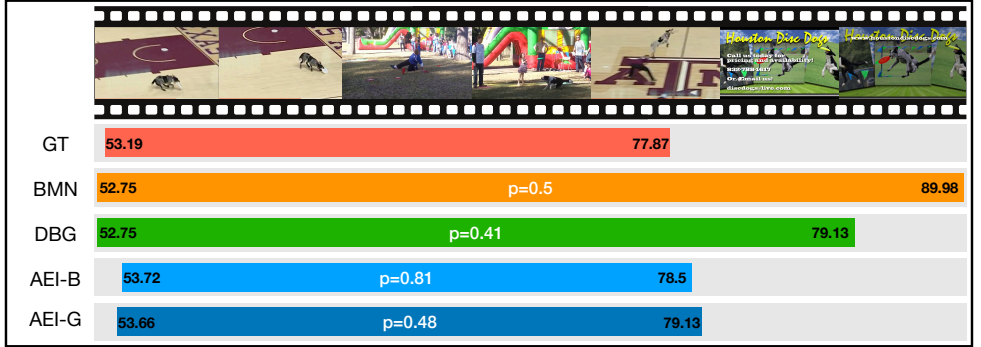


Figure 3: Qualitative comparison between our proposed method (i.e. AEI-B and AEI-G) with other SOTA methods (i.e. BMN [10], DBG [11]). The nonhuman-being main actor commits the action.

tion. Compared to other methods, our AEI outperforms by a large margin.

- Nonhuman-being main actors: (Fig. 3) shows an example where an action is committed by a nonhuman-being main actor (a dog), while (Fig. 4) shows an example where the action is committed by both human-being and nonhuman-being main actors (a dog). Our proposed cognitive-based visual representation (CVR), which is able to extract features from both actor and environment as well as model the relationship between actor and environment, aims to capture action environment feature in the case that nonhuman-being main actor exists. Thus, our proposed AEI (both AEI-B and AEI-G) obtains good performance in this case.

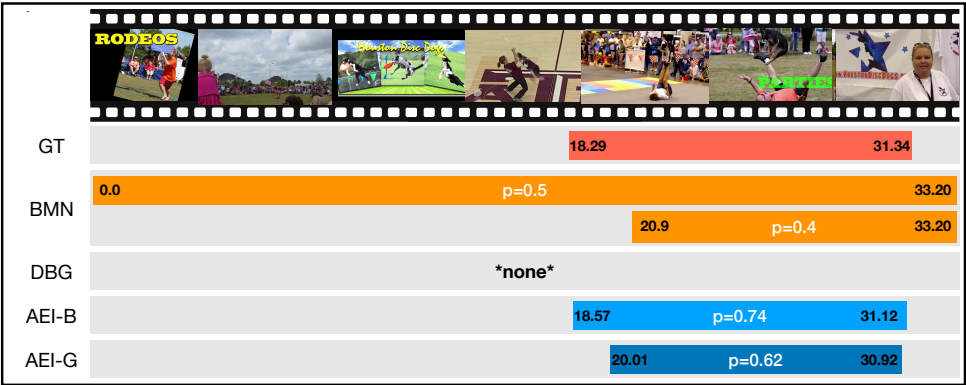


Figure 4: Qualitative comparison between our proposed method (i.e. AEI-B and AEI-G) with other SOTA methods (i.e. BMN [10], DBG [11]). Both human and nonhuman-being main actors commit the action.

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 *BMVC*, 2017.

[2] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, Oct 2017.

[3] H. Eun, S. Lee, J. Moon, J. Park, C. Jung, and C. Kim. Srg: Snippet relatedness-based temporal action proposal generator. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2019.

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

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210. IEEE, 2019.

[6] Jiyang Gao, Kan Chen, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, September 2018.

[7] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59.

[8] Y.-G. 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. <http://crcv.ucf.edu/THUMOS14/>, 2014.

- [9] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. *AAAI*, pages 11499–11506, Apr. 2020.
- [10] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, September 2018.
- [11] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, October 2019.
- [12] Shuming Liu, Xu Zhao, Haisheng Su, and Zhilan Hu. Tsi: Temporal scale invariant network for action proposal generation. In *ACCV*, November 2020.
- [13] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granularity generator for temporal action proposal. In *CVPR*, June 2019.
- [14] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *ECCV*, pages 3–20, 2018.
- [15] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS’14*, page 568–576, Cambridge, MA, USA, 2014. MIT Press.
- [16] Haisheng Su, Weihao Gan, Wei Wu, Junjie Yan, and Yu Qiao. BSN++: complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. In *ACCV*, 2020.
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30. Curran Associates, Inc., 2017.
- [18] T. Yao, Y. Li, Z. Qiu, F. Long, Y. Pan, D. Li, and T. Mei. Msr asia msm at activitynet challenge 2017: Trimmed action recognition, temporal action proposals and densecaptioning events in videos. In *CVPR Workshops*, 2017.
- [19] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *European Conference on Computer Vision*, pages 539–555. Springer, 2020.