Supplemental Material for Contextual Proposal Network for Action Localization

He-Yen Hsieh, Ding-Jie Chen and Tyng-Luh Liu
Institute of Information Science, Academia Sinica, Taiwan
heyen@iis.sinica.edu.tw, djchen.tw@gmail.com, liutyng@iis.sinica.edu.tw

In this supplementary material, we first discuss the RNN-based boundary scoring mechanism of our CPN and show more visualization results to support its false-positive reduction ability. Then, we provide the action detection results, including more recent state-of-the-art results. Finally, we detail the way of label assignment in model training.

0.1. Discussion of our RNN-based boundary scoring

This part discusses the benefits of our boundary scoring mechanism. Table 1 summarizes the ablation study to assess our model's various configurations. Please notice that we assume that each model configuration in Table 1 equips with our complete feature enhancing mechanism. [1] Table 1 has six model configurations, and only the first configuration, *i.e.*, baseline-BS, is implemented without using RNNs and directly predicts the probabilities of p^s and p^e akin to BSN yet with the enhanced feature **F**. In contrast, the other rows are RNN-based boundary predictions using the hidden states of RNNs.

0.1.1 What are the advantages of CPN

Capturing temporal context via bidirectional RNNs is useful. Comparing row 1 and row 2 shows that exploiting the temporal context of actionness via the bi-directional RNNs is beneficial, which improves AUC by 1.38%. The recent Gao's model [4] also supports this configuration.

Formulating the background context brings notable performance gain. One special property of our CPN is leveraging both the actionness and background via bi-directional RNNs to co-estimate the action boundaries. The design is derived from the observations that *i)* the features for describing the long-time actionness/background are more consistent along the temporal dimension than the short-time; *ii)* a snippet could be an action-starting boundary if its previous snippet is a background snippet, and its subsequent snippet is an action instance. On the other hand, a snippet could be an action ending boundary if its previous snippet is an action instance, and its subsequent snippet is a background. Comparing row 2 and row 3 shows the performance gain of 1.2% AUC compared with merely using the actionness formulation.

Formulating the snippet-level probabilities p^{se} and p^c is helpful. Comparing row 3 to row 4 and row 5 shows that p^{se} and p^c's snippet-level probabilities improve AUC by 0.67% and 0.63%, respectively. Simultaneously using two probabilities further improves AUC by 1.26% in comparison with the configuration in row 3.

Our CPN obviously reduces the false-positive boundary predictions. Our CPN is able to significantly reduce the false positives of action boundary estimations, as shown in Figure 1. The reduction is derived from the boundary co-estimation by leveraging the actionness and background via bi-directional RNNs.

0.2. More experimental results

In this part, Figure 2 shows some examples of retrieving the high-quality proposals of our CPN on the ActivityNet-1.3 dataset. We also provide more recent results of the state-of-the-art temporal action detectors on the THUMOS-14 testing split in Table 2.

0.3. Label assignment

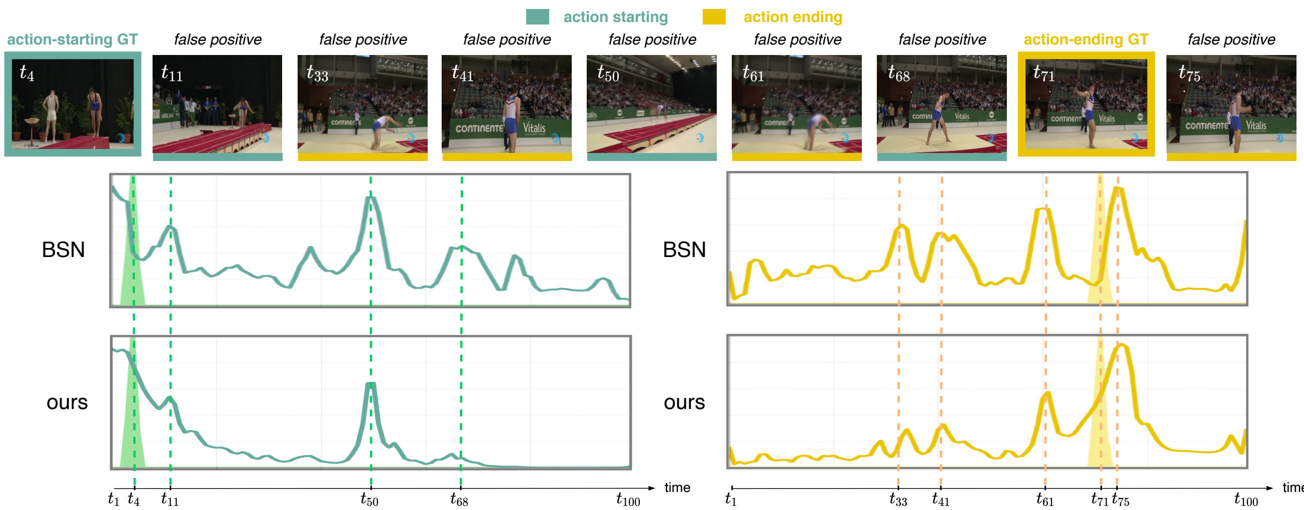
While training our CPN model, we separately define *i)* the ground-truth snippet-level probabilities: actionness, background, action starting, and action ending; *ii)* the ground-truth proposal-level probabilities: confidence and boundary-relation.

Given a video sequence $\mathcal{X}$ of l_v frames comprising N_g action instances, the n-th ground-truth action instance $(\hat{t}^n_s, \hat{t}^n_e)$ means an action occurring from the $\hat{t}^n_s$-th frame to the $\hat{t}^n_e$-th frame. We define three temporal intervals for the

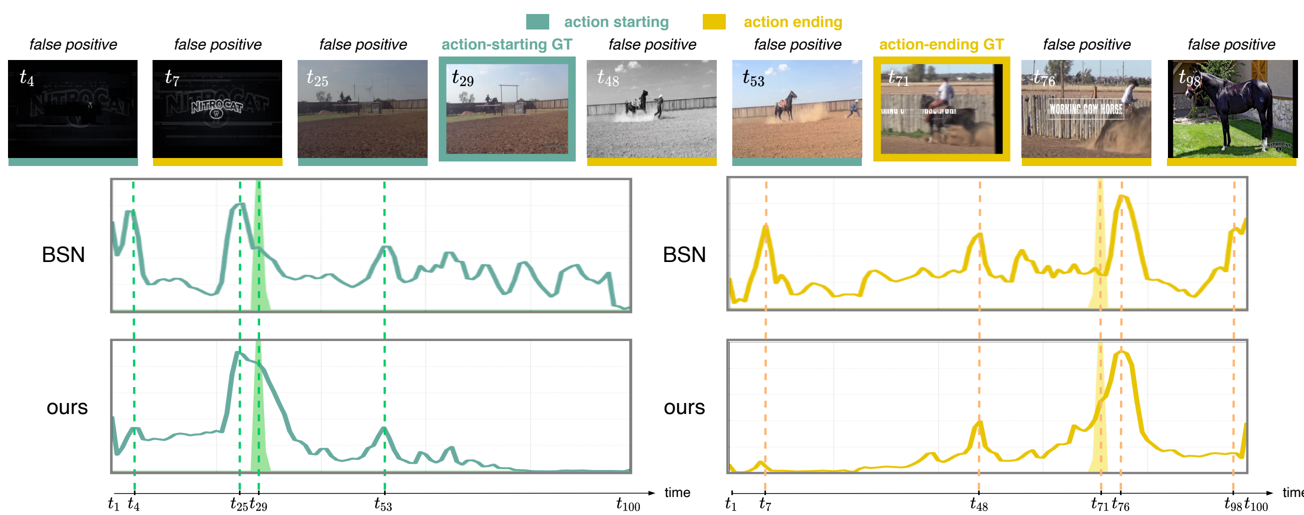
[1] Table 1 is the revised version of Table 3 in the main paper, with brief explanations per model configuration.

	Component			Boundary Scoring (BS)						
	p^s, p^e	p^{se}	p^c	AUC	PG	AR@30	AR@50	AR@80	AR@100	implementation of boundary prediction
1	baseline-BS			65.63	-	64.74	69.13	72.50	73.85	directly use the enhanced feature **F**
2	A	-	-	67.01	+1.38	65.94	70.49	74.12	75.70	only use actionness RNNs' hidden states;
3	A+B	-	-	68.21	+2.58	67.58	71.56	74.93	76.28	use RNNs of actionness & background;
4	A+B	✓	-	68.88	+3.25	68.26	72.37	75.68	77.12	use RNNs of actionness & background; with proposal-level probability p^{se}
5	A+B	-	✓	68.84	+3.21	68.34	72.31	75.33	76.51	use RNNs of actionness & background; with proposal-level probability p^c
6	A+B	✓	✓	**69.47**	+3.84	**68.74**	**73.26**	**76.27**	**77.66**	use RNNs of actionness & background; with proposal-level probabilities p^{se} & p^c

Table 1: Ablation study of boundary scoring mechanism on ActivityNet-1.3 validation split. PG: performance gain on AUC; A/B: carry out the boundary prediction using the temporal context of actionness/background.

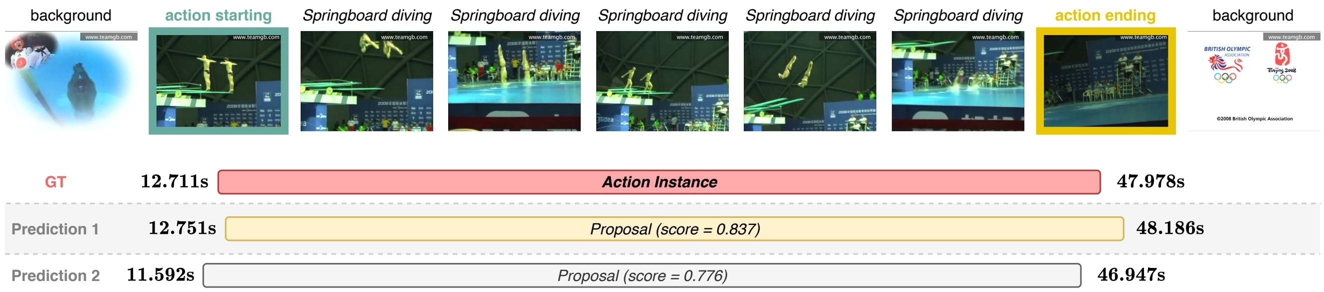


(a) Reduction of false-positive boundary predictions by CPN on the video with *Tumbling* action.



(b) Reduction of false-positive boundary predictions by CPN on the video with *Calf roping* action.

Figure 1: Effect visualization of our model compared to BSN [9]. For each subfigure, the top images are centered frames of corresponding snippets on the video, and the bottom four charts are composed of p^s on the left and p^e on the right. Each chart plots the predicted boundary probabilities (y-axis) over the snippet dimension (x-axis).



(a) The qualitative results of CPN tackling the video of one single action instances.

(b) The qualitative results of CPN tackling the video of multiple action instances.

Figure 2: Qualitative results of our CPN on ActivityNet-1.3 dataset. For each subfigure, the top images are corresponding frames of the selected snippets covered by the proposals generated by CPN. The bottom bars per subfigure represent the ground-truth action instances, top-scored predictions (yellow), and inferior predictions.

actionness r_a^n, action starting r_s^n, and action ending r_e^n as

$$r_a^n = \left[\hat{t}_s^n, \hat{t}_e^n\right] , \tag{1}$$

$$r_s^n = \left[\hat{t}_s^n - d^n/k, \hat{t}_s^n + d^n/k\right] , \tag{2}$$

$$r_e^n = \left[\hat{t}_e^n - d^n/k, \hat{t}_e^n + d^n/k\right] , \tag{3}$$

where $d^n = \hat{t}_e^n - \hat{t}_s^n$ and $k = 40$. Furthermore, for each temporal location t_n within a region of $[t_n - d_t/2, t_n + d_t/2]$, we separately calculate the maximum overlap ratio IoR at t_n with r_a^n, r_s^n and r_e^n as the corresponding label for actionness, action starting, and action ending, where $d_t = t_n - t_{n-1}$. The label of actionness and background is complementary to 1. Additionally, the label of bi-directional actionness (background), namely $\overrightarrow{p}^a$ and $\overleftarrow{p}^a$ ($\overrightarrow{p}^b$ and $\overleftarrow{p}^b$), are the same. Here we define the proposal-level target values. Given an action proposal (t_s^n, t_e^n), we compute the maximum Intersection-over-Union (IoU) with all ground-truth actions Ψ_g as the label of confidence probability p^c. Next, we simply multiply the i-th target value in p^s and the j-th target value in p^e as the label of boundary-relation probability p^{se}.

References

[1] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *ECCV*, 2020.

[2] 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*, pages 1130–1139, 2018.

[3] Jiyang Gao, Kan Chen, and Ram Nevatia. CTAP: complementary temporal action proposal generation. In *ECCV*, pages 70–85, 2018.

[4] Lianli Gao, Tao Li, Jingkuan Song, Zhou Zhao, and Heng Tao Shen. Play and rewind: Context-aware video temporal action proposals. *Pattern Recognit.*, 107:107477, 2020.

[5] 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*, pages 8328–8335, 2019.

[6] Jun Li, Xianglong Liu, Zhuofan Zong, Wanru Zhao, Mingyuan Zhang, and Jingkuan Song. Graph attention based proposal 3d convnets for action detection. In *AAAI*, 2020.

[7] 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. In *AAAI*, pages 11499–11506, 2020.

[8] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *ICCV*, 2019.

Method	Reference	mAP@0.1	mAP@0.2	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	Average
CTAP [3]	ECCV'18	-	-	-	-	29.9	-	-	-
BSN [9]	ECCV'18	-	-	53.5	45.0	36.9	28.4	20.0	-
TAL-Net [2]	CVPR'18	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1
MGG [11]	CVPR'19	-	-	53.9	46.8	37.4	29.5	21.3	-
GTAN [12]	CVPR'19	69.1	63.7	57.8	47.2	38.8	-	-	-
BMN [8]	ICCV'19	-	-	56.0	47.4	38.8	29.7	20.5	-
P-GCN [18]	ICCV'19	69.5	67.8	63.6	57.8	49.1	-	-	-
DBS [5]	AAAI'19	56.7	54.7	50.6	43.1	34.3	24.4	14.7	39.8
DBG [7]	AAAI'20	-	-	57.8	49.4	39.8	30.2	21.7	-
FC-AGCN-P-C3D[6]	AAAI'20	59.3	59.6	57.1	51.6	38.6	28.9	17.0	44.6
PBRNet [10]	AAAI'20	-	-	58.5	54.6	51.3	**41.8**	**29.5**	-
G-TAD [17]	CVPR'20	-	-	54.5	47.6	40.2	30.8	23.4	-
G-TAD [17]+P-GCN	CVPR'20	-	-	66.4	60.4	51.6	37.6	22.9	-
Zhao's model+P-GCN [19]	ECCV'20	71.8	70.3	66.3	61.0	50.1	-	-	-
BC-GNN [1]	ECCV'20	-	-	57.1	49.1	40.4	31.2	23.1	-
Gao's model [4]	PR'20	71.0	69.6	66.4	58.4	48.8	36.7	25.5	53.8
BSN++ [14]	AAAI'21	-	-	59.9	49.5	41.3	31.9	22.8	-
TCANet [13]	CVPR'21	-	-	60.6	53.2	44.6	36.8	26.7	-
SSTAP [16]	CVPR'21	-	-	58.4	51.5	42.3	32.8	22.8	-
RTD-Net [15]	ICCV'21	-	-	58.5	53.1	45.1	36.4	25.0	-
RTD-Net+PGCN [15]	ICCV'21	-	-	**68.3**	**62.3**	51.9	38.8	23.7	-
CPN+P-GCN	-	**74.0**	**71.8**	68.2	62.1	**54.1**	41.5	28.0	**57.1**

Table 2: Temporal action detection results on THUMOS-14 testing split. Supplemented results are shown in gray region.

[9] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV*, pages 3–21, 2018.

[10] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *AAAI*, 2020.

[11] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granularity generator for temporal action proposal. In *CVPR*, pages 3604–3613, 2019.

[12] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, pages 344–353, 2019.

[13] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. 2021.

[14] 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 *AAAI*, 2021.

[15] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. 2021.

[16] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. 2021.

[17] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020.

[18] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019.

[19] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wan, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, 2020.