Supplementary Material of Asymmetric Cross-Guided Attention Network for Actor and Action Video Segmentation From Natural Language Query

Hao Wang[1], Cheng Deng[1,2]*, Junchi Yan[3], Dacheng Tao[4]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of CSE, and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[4]UBTECH Sydney AI Centre, School of Computer Science, FEIT, University of Sydney, Australia
{haowang.xidian, chdeng.xd}@gmail.com, yanjunchi@sjtu.edu.cn, dacheng.tao@sydney.edu.au

In this supplementary material, we show quantitative comparison of model complexity, quantative analysis of ablation studies, and qualitative analysis of segmentation results between our proposed approach and state-of-the-art method [1].

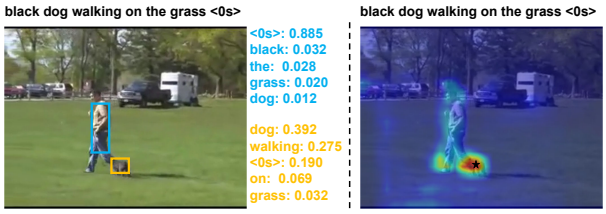
1. Model Complexity

We compare the model complexity of our approach with state-of-the-art method [1] in terms of the total number of model parameters to be learned, as depicted in Table 1. The multi-resolution fusion scheme and weighted loss for foreground pixels we adopted can boost the segmentation performance by a large margin with negligible computation cost. Our approach trained on RGB inputs obtains state-of-the-art performance even though it has much fewer parameters to be optimized than the method [1] trained on RGB and optical flow inputs.

2. Quantative Analysis

More detailed ablation studies is showed in Table 2. To verify the effectiveness of each attention component, we conduct experiment by using each component independently. The results show that they are complementary and the combination of them achieves state-of-the-art performance. As there are similar methods in relevant tasks [3, 4], we also consider them for comparison. We replace the attention mechanism with the component from Key-Word-Aware [3] and validate it on both datasets. But it is unfair to compare with MAttNet [4] which depends on complicated data preprocessing and pre-trained object detection on image feature while our method is trained on video feature in an end-to-end way. For sentence feature encoder, the 1D CNN [2] obtains similar performance with LSTM encoder on A2D sentences. However, it shows strong generalization ability on JHMDB sentences. For input modality, we obtain further improvement by intergrating motion information into our method. We take 16 frames clip as in [1] for fair comparison and also validate it with different frames, such as 8 and 24 frames. Finally, the performance of using the fusion with different resolutions and various foreground weights are demonstrated in the table.

Figure 1. Left: the attention score of words (vision-guided language attention) in sentence for each box. Right: the spatial attention heatmap conditioned on the sentence (language-guided vision attention) for the query position (black star). <0s> means padded zeros and the input sentence is above the image.



3. Qualitative Analysis

To show how the proposed attention works, we visualize the attention results in Figure 1. Given the description above the frame, the vision guided language attention overweighs the relevant words to the pixels. For example, the pixels of "dog" focus on the words of "dog" and "walking". On the contrary, the "man" is not relevant to the description, so the vision guided attention module almost assigns padded zeros to it. The language guided vision attention takes the query-focused pixels into consideration. For example, given the query point (denoted by black star on the right side of Figure 1), most pixels of the "dog" are incorporated as global context for the query point, which shows the effectiveness of language guided vision attention for such

*Corresponding author

Table 1. Comparison of model parameters between our proposed approach and state-of-the-art method [1]. Multi-resolution Fusion, Weighted Binary Cross Entropy with logits, Attention model are abbreviated as "MRF", "WBCE" and "ATT", respectively.

Method	**mAP**	**IoU**		**Model Parameters**
	0.5:0.95	Overall	Mean	
Gavrilyuk *et al*. [1] (RGB)	19.8	53.6	42.1	18,945,652
Gavrilyuk *et al*. [1] (RGB + Flow)	21.5	55.1	42.6	**37,891,304**
Baseline (RGB)	20.6	52.8	44.1	22,749,071
Baseline + MRF (RGB)	23.1	57.7	45.5	22,749,099
Baseline + MRF + WBCE (RGB)	25.5	57.4	47.5	22,749,099
Baseline + MRF + WBCE + ATT (RGB)	**27.4**	**60.1**	**49.0**	25,613,063

Table 2. Experimental results of ablation studies in terms of attention mechanism, sentence feature encoding, input modality, input frame length, fusion of different resolution and weight of foreground pixel.

Method	**Dataset**	**mAP@0.5:0.95**	**Overall IoU**	**Mean IoU**
F_{TA}	A2D	26.4	59.1	48.3
F_{VA}	A2D	26.5	**60.2**	48.3
$F_{TA} + F_{VA}$	A2D	**27.4**	60.1	**49.0**
F_{TA}	JHMDB	26.0	53.7	55.6
F_{VA}	JHMDB	26.5	55.9	56.3
$F_{TA} + F_{VA}$	JHMDB	**28.9**	**57.6**	**58.4**
Key-Word-Aware [3]	A2D	24.9	56.6	47.2
Ours	A2D	**27.4**	**60.1**	**49.0**
Key-Word-Aware [3]	JHMDB	27.5	**58.0**	57.7
Ours	JHMDB	**28.9**	57.6	**58.4**
Conv1D	A2D	**27.4**	**60.1**	49.0
LSTM	A2D	27.1	**60.1**	**49.1**
Conv1D	JHMDB	**28.9**	**57.6**	**58.4**
LSTM	JHMDB	27.7	56.5	57.2
RGB	A2D	27.4	60.1	49.0
Flow	A2D	18.2	48.9	39.9
RGB+Flow	A2D	**28.7**	**60.6**	**50.3**
RGB	JHMDB	28.9	57.6	58.4
Flow	JHMDB	14.2	41.2	43.8
RGB+Flow	JHMDB	**29.5**	**57.9**	**59.1**
Length=8	A2D	26.4	58.5	47.9
Length=16	A2D	**27.4**	**60.1**	**49.0**
Length=24	A2D	27.0	**60.1**	48.2
Res3	A2D	24.6	58.5	46.6
Res3 + Res1	A2D	27.1	**60.1**	48.4
Res3 + Res2	A2D	25.9	59.1	47.5
Res3 + Res1 + Res2	A2D	**27.4**	**60.1**	**49.0**
Weight=1.0	A2D	26.4	59.3	47.6
Weight=1.5	A2D	**27.4**	**60.1**	**49.0**
Weight=2.0	A2D	27.2	59.4	**49.0**

conditional segmentation task.

We then conduct the segmentation of the frames containing multiple actors in Figure 2. We observe that the separated different actors and their actions are segmented well via our proposed approach. However, it is still challenging for our model to handle the instances among the same actor

Figure 2. Segmentation results of the frames containing multiple actors and actions. Multi-resolution Fusion, Weighted Binary Cross Entropy with logits, Attention model are abbreviated as "MRF", "WBCE" and "ATT", respectively.

and action by sentence descriptions.

Finally, we present segmentation results of different actors with the same action (*i.e.*, rolling) in Figure 3. Compared with the method [1], our proposed approach can produce fine-grained segmentation because of the multi-resolution fusion scheme. Besides, our model can segment small size object with the help of weighted loss for foreground pixels, *e.g.*, "ball rolling" in the last video. In addition with the proposed asymmetric cross-guided attention, our approach effectively learns the correlation between visual and linguistic features.

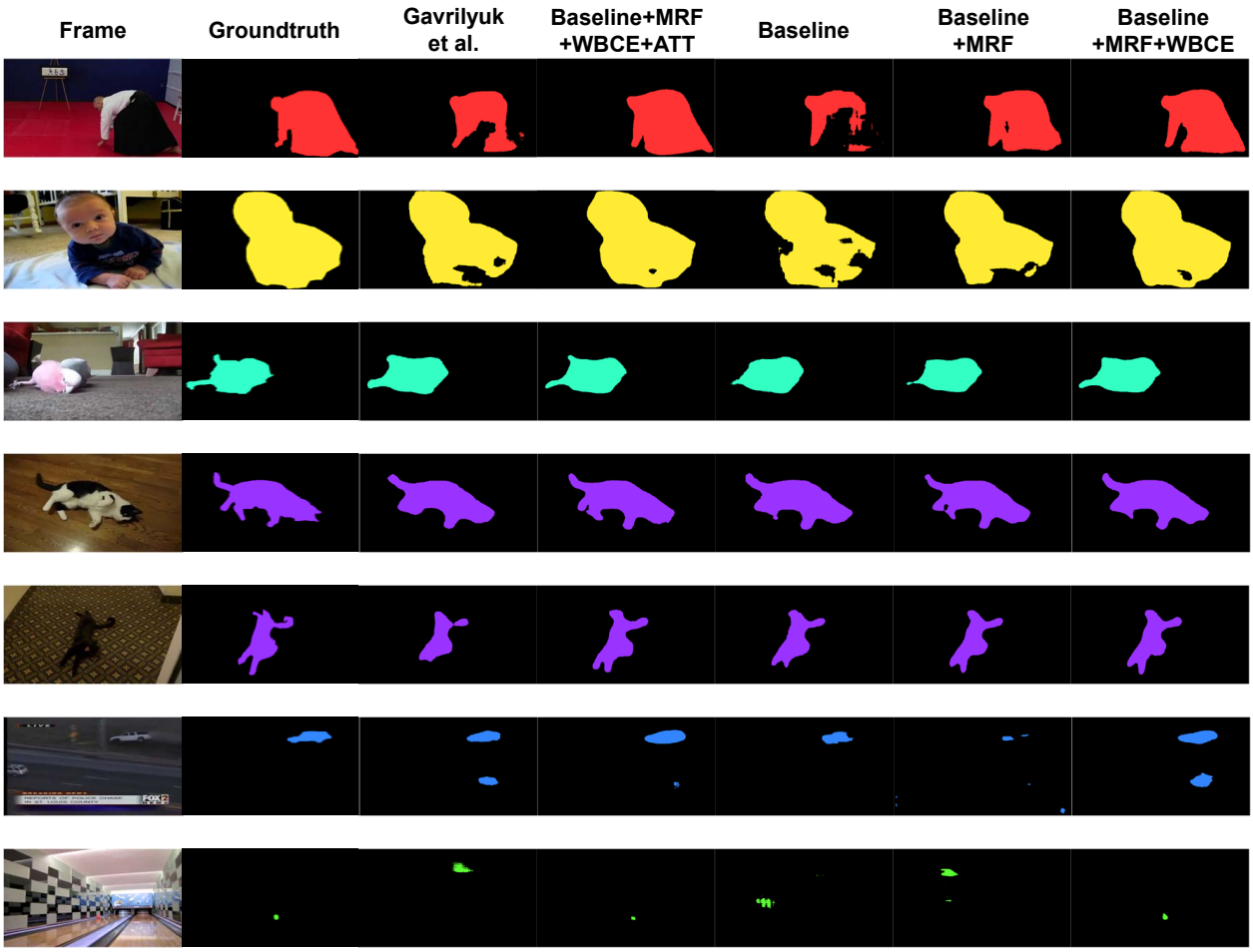


Figure 3. Segmentation results of different actors (*i.e.*, adult, baby, bird, cat, dog, car and ball) with the "rolling" action on A2D Sentences. Columns from left to right are frame to segment, groundtruth segmentation, output of Gavrilyuk *et al.*, output of Baseline+MRF+WBCE+ATT, output of Baseline, output of Baseline+MRF and output of Baseline+MRF+WBCE, respectively. Multi-resolution Fusion, Weighted Binary Cross Entropy with logits, Attention model are abbreviated as "MRF", "WBCE" and "ATT".

References

[1] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018.

[2] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv:1408.5882*, 2014.

[3] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, pages 38–54, 2018.

[4] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018.