Modeling Motion with Multi-Modal Features for Text-Based Video Segmentation
(Supplementary Materials)

Wangbo Zhao[1,2,3] Kai Wang[1] Xiangxiang Chu[2] Fuzhao Xue[1] Xinchao Wang[1] Yang You[1*]
[1] National University of Singapore [2] Meituan Inc. [3] Northwestern Polytechnical University
wangbo.zhao96@gmail.com, kai.wang@comp.nus.edu.sg, chuxiangxiang@meituan.com,
f.xue@u.nus.edu, xinchao@nus.edu.sg, youy@comp.nus.edu.sg

Abstract

In supplementary materials, we first demonstrate the loss function of our method in detail in Section 1. Then, we compare the computational overhead of our method with previous methods in Section 2. We visualize some representative samples from A2D Sentences to compare our method with ACGA in Section 3. We replace the BERT with Bi-LSTM as the linguistic encoder to extract linguistic features in Section 4. In Section 5, we change the appearance encoder from the ResNet101 to the lighter ResNet50. We conduct experiments to explore the best choice of the number of LGFF modules in Section 6. Section 7 shows the performance of our method on Ref-YouTubeVOS. We explore the impact of changing the optical flow estimation in our model in Section 8. Finally, we visualize some examples when the optical flow estimation fails in Section 9.

1. Loss Function

In this section, we illustrate our loss function during training in detail. Since the training set of A2D Sentences [3] only contains 3 to 5 frames with pixel-level annotation in each video sequence, we view the frame with annotation as the target frame and its previous and next frames as reference frames, which means that we **can only calculate the loss on the target frame**. This is the same as previous methods. We incorporate the traditional binary cross with our multi-modal alignment loss to train our whole model, which is denoted as "B+M+T+L+A" in Table 3 in the main paper.

For the feature f^1 belonging to the target frame, we adopt two convolutional layers followed by a sigmoid activation function to generate the final prediction map. Then, we upsample it to the original size and obtain $\hat{P}$. The binary cross entropy loss can be defined as:

*Corresponding author.

Table 1. Comparison of computational overhead.

Name	Input Size	GFLOPs	mAP 0.5:0.95
ACGA[10]	$16 \times 512 \times 512$	630.83	27.4
CMDY[9]	$16 \times 512 \times 512$	>600	33.3
CSTM[5]	$8 \times 320 \times 320$	213.06	39.9
Our	$3 \times 320 \times 320$	181.47	**41.9**
B+T+L	$3 \times 320 \times 320$	116.34	40.1

$$L^{bce} = -\sum p^i \log \hat{p}^i + (1 - p^i) \log(1 - \hat{p}^i), \quad (1)$$

where $\hat{p}^i \in \hat{P}$ represents an element in $\hat{P}$. p^i is the label of $\hat{p}^i$.

In Section 3.4 of the main paper, we obtain the multi-modal alignment loss L^{align} for the target frame. The total loss function can be formulated as:

$$L = L^{bce} + L^{align}. \quad (2)$$

2. Computational Overhead

In Table 1, we compare the computational overhead of our method with previous methods. Previous works depend on 3D CNNs *e.g.* I3D [1] to extract temporal and implicit motion information from many reference frames, which leads to huge computational overhead. Our method achieves the best performance while with less FLOPs. We also illustrate the result from "B+T+L", which does not contain the motion branch. It can also surpass previous methods with about 2× less FLOPs than CSTM[5]. These results verify the effectiveness and efficiency of our method.

3. Qualitative Comparison

In Figure 2, we visualize some representative samples from A2D Sentences[3] to compare our method with ACGA[10]. In (a)(b)(c)(d)(e)(f)(g), the 1th, 2th, and 3th column are the ground-truth masks, results from our method, and results from ACGA[10], respectively. The colored text describes the object with the same color mask. We

Figure 1. Examples when the optical flow estimation fails.

Table 2. Comparison of the linguistic encoder. "Our" is the same as "B+M+T+L+A" in the main paper.

Name	mAP	IoU	
	0.5:0.95	Overall	Mean
ACGA[10]	27.4	60.1	49.0
CMDY[9]	33.3	62.3	53.1
CSTM[5]	39.9	66.2	**56.1**
Our–LSTM	41.2	66.6	55.4
Our	**41.9**	**67.3**	55.8

can find that, our method can generate more accurate and complete segmentation masks *e.g.* (a)(b) (d) (e). In addition, our method can distinguish the target object described by the text from other objects *e.g.* (c) (f). These results further verify the effectiveness of incorporating the motion features with appearance and linguistic features.

4. Linguistic Encoder

We also conduct experiments to verify the effectiveness of BERT [2] as the linguistic encoder. Given a text with L words, we first replace the BERT in "B+M+T+L+A" with a Bi-LSTM[4] to extract the feature of each word, and obtain $\mathcal{L} \in \mathbb{R}^{L \times C_{\mathcal{L}}}$. Then we average the feature of all words and obtain $f_{\mathcal{L}}$, which can be viewed as the representation of the whole sentence. This model is denoted as "Our-LSTM" and we adopt "Our" to represent "B+M+T+L+A" in the main paper. From Table 2, we can find that the performance of "Our-LSTM" is worse than "Our", although it can still surpass state-of-the-art methods by a large margin in most metrics.

5. Appearance Encoder

We change the appearance encoder from the ResNet101 to the lighter ResNet50. From Table 3, we find that this model still achieves similar performance, *e.g.* overall IoU 67.1 vs 67.3, and surpass prior methods. This demonstrates that, with effective multi-modal fusion, our model does not heavily rely on a strong backbone to achieve good performance.

Table 3. Our model with different backbones.

Methods	mAP	IoU	
	0.5:0.95	Overall	Mean
Our-ResNet50	41.4	67.1	55.3
Our-ResNet101	41.9	67.3	55.8

Table 4. The number of LGFF modules in our model.

Number of LGFF	mAP	IoU	
	0.5:0.95	Overall	Mean
0	37.6	63.5	51.6
1	39.7	65.7	53.7
2	40.6	66.1	54.7
3	41.1	66.8	54.8

Table 5. Results on Ref-YouTubeVOS [6]

Name	Overall	$\mathcal{J}$	$\mathcal{F}$
URVOS [6]	-	41.3	-
Our	45.2	44.1	46.2

Table 6. The Effectiveness of optical flow estimation method.

Name	mAP	IoU	
	0.5:0.95	Overall	Mean
Our-RAFT	41.9	67.3	55.8
Our-PWC	41.8	67.6	55.8

6. The number of LGFF modules in our model

We conduct experiments to explore the best choice of the number of LGFF modules in our model. We gradually replace the concatenation fusion strategy with our "LGFF" from the high level to the low level. The results are demonstrated in Table 4. The model with 0,1,2 and 3 LGFF modules achieve 63.5, 65.7, 66.1, and 66.8 respectively on the overall IoU. Note that, when the number of LGFF is 0 and 3, the model equals "CAT" and "LGFF" in Table 5 in the main paper, respectively. Results show that all LGFF modules in our model are essential.

7. Experiments on Ref-YouTubeVOS

Ref-YouTubeVOS [6] is a recent proposed large dataset for text-based video segmentation. Compared with A2D Sentences [3] and J-HMDB Sentences [3], it contains more object categories and challenging sceneries. To further verify the generalization ability of our method, we conduct experiments on this dataset, whose results are shown in Table 5. Our model can surpass URVOS [6] without multiple iterations by a large margin.

8. Change the optical flow estimation method.

We change the optical flow estimation method from RAFT [8] to PWC [7] to generate optical flow maps. The results are shown in Table 6. We find that both "Our-RAFT" and "Our-PWC" achieve similar performance, which verifies the robustness of our method on optical flow estimation models.

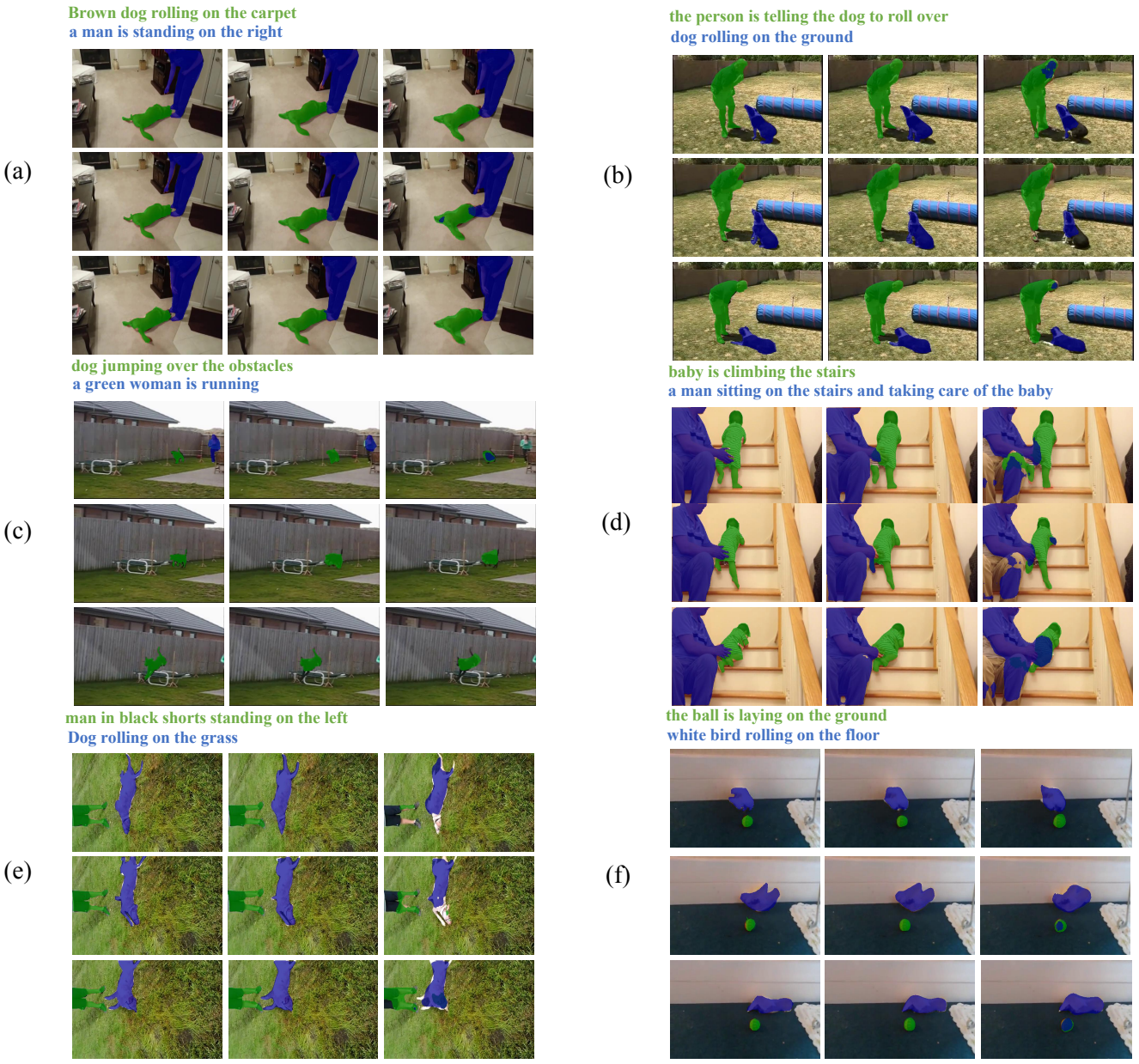

Figure 2. Qualitative results comparison on A2D Sentences. In (a)(b)(c)(d)(e)(f)(g), the 1th, 2th, and 3th column are the ground-truth mask, results from our method, and results from ACGA[10], respectively.

9. Examples when the optical flow estimation fails

There are two cases when the optical flow estimation fails. The one is that the flow of the object is incomplete *e.g.* the car in Figure 1 (a). The other is that the target object is not distinctive in the flow map *e.g.* Figure 1 (b)(c). Figure 1 shows that our method can handle these two situations. This further proves the robustness of our method.

References

[1] 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. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[3] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision*

and Pattern Recognition, pages 5958–5966, 2018. 1, 2

[4] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 2

[5] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2021. 1, 2

[6] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 2

[7] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2

[9] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12152–12159, 2020. 1, 2

[10] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948, 2019. 1, 2, 3