ACCLVOS: Atrous Convolution with Spatial-Temporal ConvLSTM for Video Object Segmentation

[1]Muzhou Xu, Shan Zhong, Chunping Liu, *Shengrong Gong, Zhaohui Wang, Yu Xia

[1]School of Computer Science and Technology, Soochow University
[2]Changshu Institute of Technology

ABSTRACT

Semi-supervised video object segmentation aims at segmenting the target of interest throughout a video sequence when only the annotated mask of the first frame is given. A feasible method for segmentation is to capture the spatial-temporal coherence between frames. However, it may suffer from mask drift when the spatial-temporal coherence is unreliable. To relieve this problem, we propose an encoder-decoder-recurrent model for semi-supervised video object segmentation. The model adopts a U-shape architecture that combines atrous convolution and ConvLSTM to establish the coherence in both the spatial and temporal domains. Furthermore, the weight ratio for each block is also reconstructed to make the model more suitable for the VOS task. We evaluate our method on two benchmarks, DAVIS-2017 and Youtube-VOS, where state-of-the-art segmentation accuracy with a real-time inference speed of 21.3 frames per second on a Tesla P100 is obtained.

OBJECTIVES

- We propose an encoder-decoder architecture that combines ConvLSTM with atrous convolution on both the spatial domain and the temporal domain to establish spatiotemporal coherence for segmenting target.
- In order to maximize the performance of atrous convolution and ConvLSTM, we design a new network configuration to increase the role of deep features in the establishment of spatiotemporal coherence.
- We evaluate our method on two benchmarks, YoutubeVOS and DAVIS-2017, and obtain a competitive segmentation result compared to state-of-the-art methods while achieving a real-time segmentation speed.

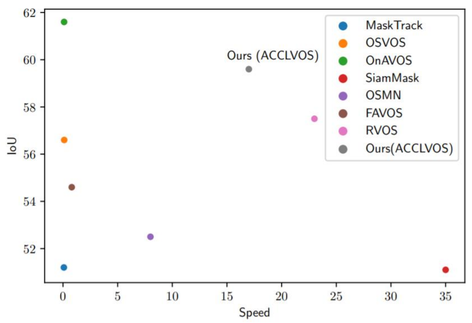

Fig. 1. Inference speed and IoU performance on DAVIS-2017 validation set [12]. Results of state-of-the-art methods, including OSVOS [1], OnAVOS [16], FAVOS [5], MaskTrack [10], OSMN [22], SiamMask [17], RVOS [15]. IoU refers to the intersection over union between the inference mask and the ground truth. Speed refers to the inference speed and the evaluation indicator is frames per second.

METHODS

The network consists of two parts, an encoder for feature extraction, and a decoder for establishing the spatiotemporal coherence.

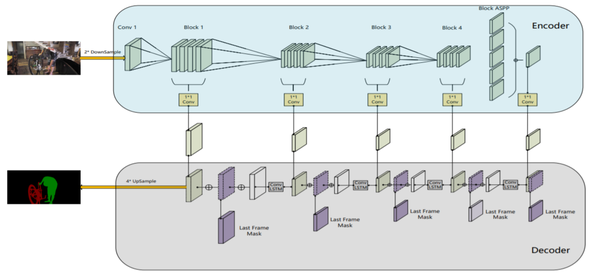

Fig. 2. Our proposed end-to-end architecture at time step t. $\oplus$ refers to concatenate. We use bilinear interpolation to transform the feature maps.

our encoder consists of two parts, (i) the feature extraction network. It is based on ResNet-101 with different atrous rates $r \in \{r1 = 1, r2 = 1, r3 = 1, r4 = 2\}$, where r represents the block of ResNet-101. (ii) The multi-scale feature extraction network (ASPP).

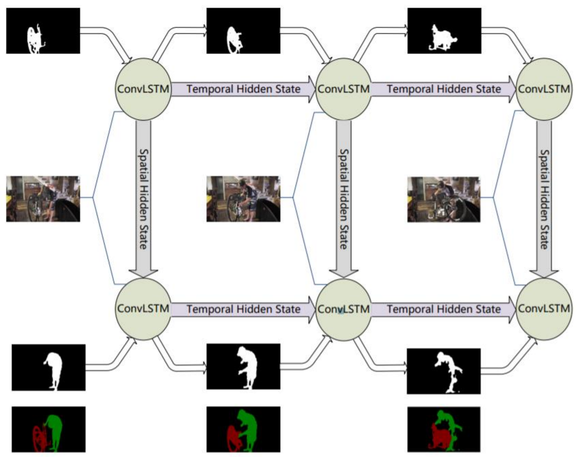

Fig. 4. Our proposed decoder architecture in a video sequence.

Our decoder architecture is depicted in Figure 4. Similar with Ventura et al., in terms of the decoder, we utilize ConvLSTM to establish the spatial-temporal coherence on both the spatial and temporal domains. In terms of the temporal domain, the coherence of the same target in different frames is established by ConvLSTM. As for the spatial domain, ConvLSTM is used to establish coherence among different targets.

RESULTS

We evaluate our model on two benchmarks, DAVIS-2017 and Youtube-VOS(2018). In section §4.3, we perform an ablation experiment to investigate the effectiveness of each component and different video scenes. The evaluation measures are mainly the following two aspects, (i) the region similarity J, and (ii) the contour accuracy F.

On both of two datasets, we achieve the state-of-the-art performance.

TABLE I
COMPARISON AGAINST STATE-OF-THE-ART TECHNIQUES FOR ONE-SHOT VIDEO OBJECT SEGMENTATION, 'OL' REFERS TO MAKING USE OF ONLINE LEARNING. THE TABLE IS DIVIDED INTO TWO PARTS, DEPENDING ON WHETHER THE TECHNIQUES USING ONLINE LEARNING OR NOT.

DAVIS-2017 (Validation Set) [12]								
	OL	*J*-Mean	*J*-Recall	*J*-Decay	*F*-Mean	*F*-Recall	*F*-Decay	Speed(fps)
MaskTrack [10]	✓	51.2	59.7	28.3	57.3	65.5	29.1	0.08
OSVOS [1]	✓	56.6	63.8	**26.1**	63.9	73.8	27.0	**0.11**
OnAVOS [16]	✓	**61.6**	**67.4**	27.9	**69.1**	**75.4**	**26.6**	0.1
SiamMask [17]		51.1	60.5	**-1.1**	55.0	64.3	**1.9**	**35.0**
OSMN [22]		52.5	60.9	21.5	57.0	66.1	24.3	8.0
FAVOS [5]		54.6	61.1	14.1	61.8	72.3	18.0	0.8
RVOS [15]		57.5	65.2	24.9	63.6	73.8	27.0	23.0
Ours(ACCLVOS)		**59.6**	**70.1**	22.4	**65.0**	**78.6**	26.1	17.0

TABLE II
PERFORMANCE COMPARISON OF OUR APPROACH WITH STATE-OF-ART METHODS ON YOUTUBE-VOS TEST SET, 'OL' REFERS TO ONLINE LEARNING

Youtube-VOS one-shot (Test Dev Set) [21]			
	OL	*J*-Mean	*F*-Mean
OnAVOS [16]	✓	51.2	57.3
MaskTrack [10]	✓	56.6	63.9
OSVOS [1]	✓	**61.6**	**69.1**
OSMN [22]		52.5	57.0
S2S [20]		60.5	63.3
Ours (ACCLVOS)		**61.3**	**64.5**

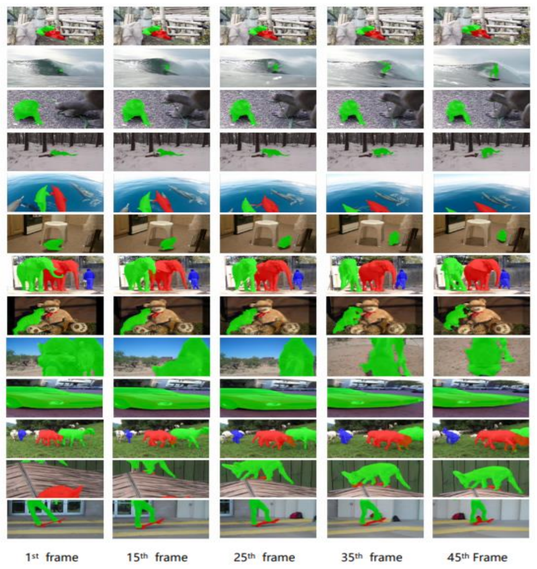

TABLE IV
THE ABLATION EXPERIMENT OF THE DAVIS-2017 VALIDATION SET. 'TRANSFER LEARNING' REFERS TO USING THE TRAINED MODEL FROM THE YOUTUBE-VOS DATASET FOR TRANSFER LEARNING. 'ASPP' REFERS TO USING THE 'ASPP' MODULE.

DAVIS-2017 (Validation Set) [12]					
Transfer Learning	ASPP	*J*-Mean	*J*-Recall	*F*-Mean	*F*-Recall
		42.0	44.0	47.5	49.2
	✓	44.0	45.5	52.1	55.7
✓		57.6	67.3	63.4	73.6
✓	✓	**59.6**	**70.1**	**65.0**	**78.6**

CONCLUSION

This paper proposes a recurrent method based on ConvLSTM and atrous convolution, named ACCLVOS, to learn the coherence in the video object segmentation. Compared to other recurrent models, the proposed model combines the ConvLSTM and atrous convolution in both the spatial and temporal domains. Furthermore, the proportion of each part is reconstructed to increase the impact of spatial details on coherence establishment. By combining atrous convolution and ConvLSTM, our method not only recognizes the target appearance better but also establishes more reliable coherence. Therefore, our method achieves a good balance in segmentation accuracy and speed.

The model has been evaluated on two benchmarks, YoutubeVOS and DAVIS-2017. Because our method establishes more reliable coherence and learns more accurate target appearance. Compared with other recurrent model, it reduces mask drift and segments target more accurately. Furthermore, from the experimental conclusions, it can be found that even if our method dose not use online-learning, it can achieve competitive segmentation results, thus greatly improving the segmentation speed. since our method does not need to introduce online-learning, the segmentation speed is improved greatly. In the future, we would like to explore the effect of our model in practical application scenarios, In addition, we will attend to introduce global coherence to alleviate the misclassification problem.

ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (NSFC Grant No.61972059, 61702055)cn Natural Science Foundation of Jiangsu Province under Grant (BK20191474, BK20161268). Research and Innovation Fund of the Science and Technology Development Center of the Ministry of Education(2018A01007), and Ministry of Education Science and Technology Development Center Industry-University Research Innovation Fund(2018A02003), and Humanities and Social Sciences Foundation of the Ministry of Education under Grant 18YJCZH229.