The 2020 DAVIS Challenge on Video Object Segmentation - CVPR 2020 Workshops

Spatial Consistent Memory Network for Semi-supervised Video Object Segmentation

Peng Zhang, Li Hu, Bang Zhang, and Pan Pan

Alibaba DAMO academy

Abstract

Semi-supervised video object segmentation is a fundamental yet challenging problem in computer vision. Deep learning based methods have achieved promising results by exploiting the guidance information of past frames. Despite its superior performance, these works exhibit distinct shortcomings, especially the false predictions caused by similar appearance instances, even they could readily be distinguished with spatial guidance. Moreover, they suffer from object's appearance variations and error drifts. In order to mitigate the shortcomings, we propose Spatial Consistent Memory Network with an enhanced segmentation head. We introduce a spatial constraint module that takes the previous prediction to generate a spatial prior for current frame, which helps to disambiguate appearance confusion and eliminate false predictions. Additionally, a segmentation head with Atrous Spatial Pyramid Pooling(ASPP) module and a refinement module are adopted to handle scale variance and improve segmentation quality. Furthermore, we propose a training strategy to minimize the gap between training and testing. Finally, the proposed method can achieve the J&F mean score of 84.1% for the DAVIS semi-supervised VOS test-challenge dataset.

1. Introduction

Video Object Segmentation (VOS) applies widely in video editing, video composition, autonomous driving, *etc*. Given a video and the ground truth object mask of the first frame, semi-supervised VOS predicts the segmentation masks of the objects specified by the ground truth mask in all the following frames. The task is challenging due to the appearance variations, occlusions and error drifts. Furthermore, the target object may disappears in some frames of the video and may confuse with similar instances of same categories.

Many VOS methods are proposed over the past few years for this challenge. In general, they can be divided into propagation-based methods and matching-based methods. Propagation-based methods[4, 10, 8] rely on the segmentation mask from the previous frames. MaskTrack[4] predicts mask with the guidance of the previous predicted mask, and it can train from static images which are deformed to simulate videos frames. These methods suffer from occlusions and fast motions between sequential frames. Matching-based methods[2, 7, 11] learn pixel-level features via pixel-level matching. But they fail when the appearance of objects change dramatically. STM[3] further develops the matching based method by leveraging a memory network to read relevant information from all the past frames. STM performs dense matching in the feature space to retrieve useful information, which allows it to handle challenges such as appearance changes and occlusions. However, the matching based mechanism does not take spatial consistency into account. Model sometimes suffers from false predictions when there are similar objects entering into view. Further more, the model may performs worse in case of large scale variances.

We extend the STM model with a spatial constraint module and an enhanced segmentation head. In segmentation head, we exploit a ASPP[1] module to handle scale variability in videos. A refinement module inspired by semantic segmentation method BASNet[6] is adopted to boost the performance especially at the boundary. The spatial constraint module is similar to [13] but performed on the final embedding rather than the key feature map and we learn the spatial prior directly rather than the weight matrix W and bias matrix b. It takes the previous predict and current embedding generated by the encoder and ASPP module, to learn a spatial prior map. This spatial prior map serves as a rough constraint for guiding the model to filter out confusing instances of similar appearance. Finally, we adopt a similar training strategy as [3], but make improvements to reduce the gap between training and testing.

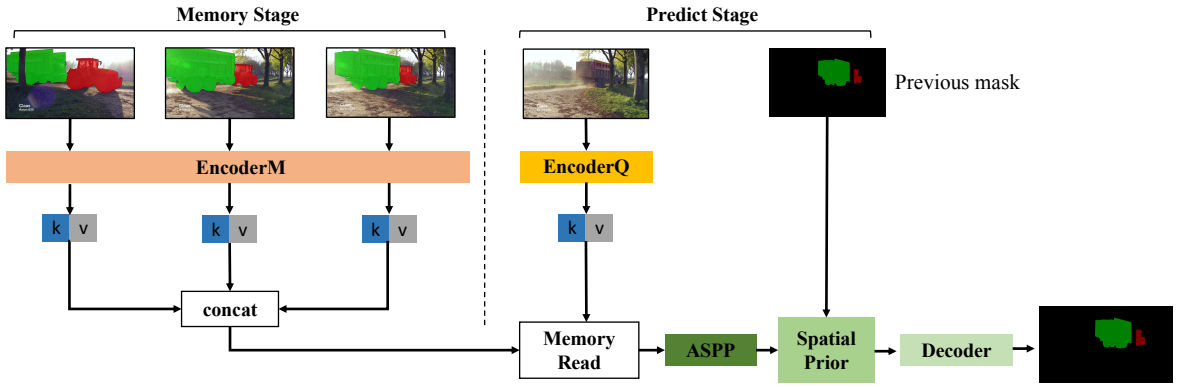

Figure 1. Framework of the proposed method. The previous mask is used to guide the predict of current frame. In test phase, previous predict is used.

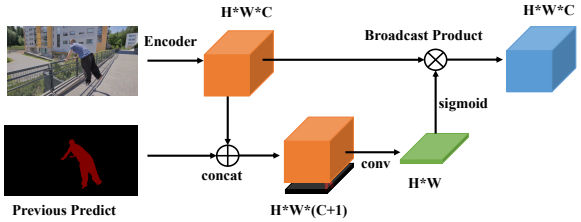

Figure 2. The proposed spatial constraint module

2. Approach

Our approach is developed base on the STM[3] framework. The structure is illustrated in Fig.1. We introduce the proposed spatial constraint module in Sec.2.1. The detail of segmentation head is explained in Sec.2.2. Finally, we introduce the training strategy in Sec.2.3.

2.1. Spatial Constraint Module

We introduce a spatial constraint module(Fig.2) to ensure the spatial consistency between adjacent frames and to disambiguate appearance confusion and eliminate false predictions caused by similar instances of same category. The predicted mask of previous frame is a 0-1 mask in shape of $H \times W$. It is concatenated with current frame embedding $(H \times W \times C)$ to get a feature map of shape $H \times W \times (C+1)$. A convolution layer with kernel size 3×3 and the sigmoid function are adopted to produce a spatial prior, which is a feature map of shape $H \times W$. The prior are multiplied with the current frame embedding.

We visualized some frames and the corresponding spatial priors in Fig.3. The spatial priors roughly capture the location of specific objects.

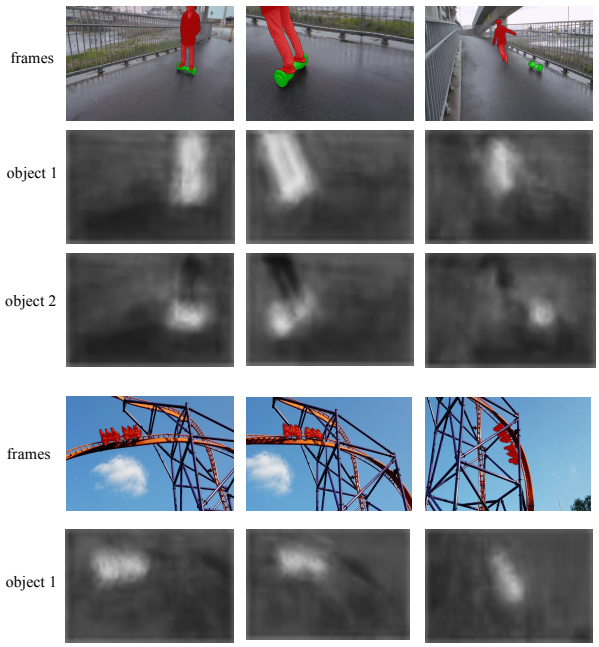

Figure 3. Visualization of the spatial constraint module. The values of prior are mapped from 0-1 to 0-255 for better visualization

2.2. Segmentation Head

The spatial constraint module aims to correctly capture the target objects, but it is not enough to get a high-quality results. As the objects come in various sizes in multi-object cases and they often change in size during sequential frames, previous methods suffer from the scale variance problem. In this section we propose our segmentation head to tackle this problem and improve the segmentation quality as illustrated in Fig.4. Inspired by the general semantic segmentation, we apply Atrous Spatial Pyramid

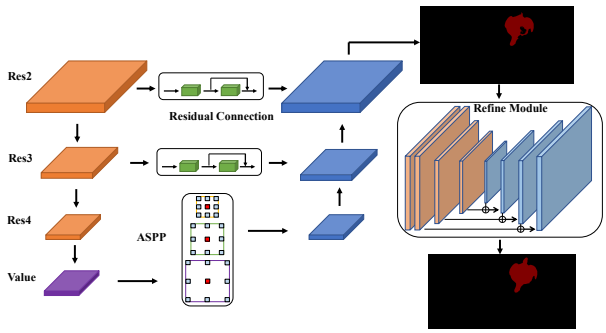

Figure 4. Structure of the proposed decoder.

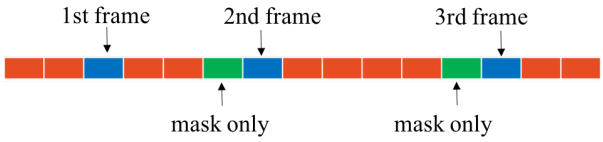

Figure 5. Sampling strategy in training stage

Pooling(ASPP)[1] module after the memory read operation to make different receptive fields. We use three parallel dilated convolution layers and set the dilation rate as 2, 4 and 8 to adapt to the encoder. Then a decoder is attached as the upsampling operation with residual skip connection proposed in [6]. We apply soft aggregation[3] to merge the multi-object prediction and the loss function is Cross Entropy Loss.

To further boost the performance especially at the boundary, we apply a refine module based on the encoder-decoder structure. We take the feature map before soft aggregation as the input of the refine module. And we re-downsample it with 3x3 convolution layer and ReLU function, then it is upsampled to the origin resolution and merged by soft aggregation again. Cross Entropy Loss and IoU Loss are applied as the loss function for the refined result.

2.3. Training Strategy

Due to the limitation of GPU memory, for an iteration we select three frames from a certain video[3]. we additionally sample the ground truth masks of the previous frame for last two frames to serve as a prior, as illustrated in Fig.5. Predictions and losses are only computed for the last two frames, while the first frame serves as reference frame. To reduce the gap between training and testing, we decrease the maximum skip number between sampled frames in the last few epochs. Thus we can use the predicted mask for memory network and spatial constraint module.

3. Experiments

We evaluate our method on the DAVIS 2017 dataset[5].

3.1. Training Details

We follow the two stage training settings in [3]. The same static image datasets are used for pretraining. In main training stage, both DAVIS and Youtube-VOS 2019[9] are used. During pretraining, 384×384 patches are randomly cropped from images. In main training, we randomly resize the shortest edge of image to $[384, 1080]$ and crop a 640×384 patch around all the maximum bounding box of all instances in three frames. Other data augmentation like flip, affine transform *etc*. are also applied.

3.2. Experimental Results

As shown in Table 1, our proposed method finally achieves 84.1 J&F score on the DAVIS test-challenge and ranks the first place in the Semi-supervised video object segmentation track. The performance of our method surpasses the winner in DAVIS 2019 by 7.4.

3.3. Ablation Study

In this section we study the contribution of our proposed method and how we achieve the final result as shown in Table 2. Following experiments are conducted on DAVIS 2017 test-dev and J&F score is reported. The baseline is the original space-time memory network and our reimplementation with resnet-50 backbone achieves 72.2 on DAVIS 2017 test-dev.

We apply a powerful ResNeST-101[12] backbone, which achieves SOTA performance in many other computer vision tasks. And in our experiment, it brings an improvement of 2.9. Then we demonstrate the effectiveness of our proposed segmentation head, spatial constraint module and the training strategy. Based on a strong baseline of 75.1, it still achieves the J&F score of 79.7, with a significant improvement of 2.8, 1.0 and 0.8 respectively, surpassing the winner in DAVIS 2019 and all other existing methods without bells and whistles.

Then we further boost the performance with other tricks. We apply flip and multi-scale testing with a gain of 1.5. We apply online finetune which is widely used in early methods and then ensemble these methods. Finally we achieve 83.2 on DAVIS test-dev.

4. Conclusion

In this paper, we propose Spatial Constraint Memory Network for semi-supervised video object segmentation including a spatial constraint module, a enhanced segmentation head and a new training strategy. Our approach achieves J&F score of 84.1 on DAVIS test-challenge, rank-

Team	Global Mean	J Mean	J Recall	J Decay	F Mean	F Recall	F Decay
Ours	84.1	81.5	89.1	14.2	86.7	92.9	16.1
ReLER	83.8	81.1	88.4	18.1	86.5	93.3	17.8
Hongje	79.5	77.0	85.7	14.9	82.1	89.6	17.1
HCMUS-UD-NII-UIUC	79.3	76.5	85.1	10.1	82.1	90.7	12.0
DSVOS	76.9	74.4	82.7	20.5	79.5	86.9	23.6
Vltanh	76.0	73.3	81.9	16.6	78.7	87.2	19.6
Bytedance	72.2	69.8	76.7	20.7	74.6	83.0	23.0
DeepDream	64.9	62.5	72.1	24.2	67.3	76.7	27.7

Table 1. Final results on DAVIS 2020 semi-supervised challenge

Methods	Global Mean	Boost
Baseline(reimplementation STM)	72.2	-
+ResNeST101 Backbone	75.1	+2.9
+Segmentation Head	77.9	+2.8
+Spatial Constraint	78.9	+1.0
+Training Strategy	79.7	+0.8
+Flip and Multi-scale Testing	81.2	+1.5
+Online Finetune and Model Ensemble	83.2	+2.0

Table 2. Ablation study on DAVIS test-dev

ing the first place on DAVIS 2020 semi-supervised video object segmentation challenge.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[2] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018.

[3] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.

[4] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2663–2672, 2017.

[5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[6] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019.

[7] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019.

[8] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2018.

[9] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[10] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018.

[11] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. *arXiv preprint arXiv:2003.08333*, 2020.

[12] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[13] Qiang Zhou, Zilong Huang, Lichao Huang, Yongchao Gong, Han Shen, Wenyu Liu, and Xinggang Wang. Motion-guided spatial time attention for video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.