Enhanced Memory Network for Video Segmentation

Zhishan Zhou[1]*, Lejian Ren[2], Pengfei Xiong[3]†, Yifei Ji[4], Peisen Wang[3], Haoqiang Fan[3], Si Liu[5]

[1]Beijing University of Post and Telecommunications [2]Institute of Information Engineering, CAS
[3]Megvii Research [4]TsingHua University [5]Beihang University

Abstract

This paper proposes an Enhanced Memory Network (EMN) for semi-supervised video object segmentation. Space-Time Memory Networks[10] has proven the effectiveness of the abundant use of guidance information. To further improve the accuracy of unknown and small targets, we propose to perform fined-grained segmentation based on the correlation attention map. We introduce a siamese network to obtain the semantic similarity and relevance between the tracking objects and the whole image. The feature map extracted from the siamese network on the cropped image is multiplied onto the whole feature map as the attention of proposal objects. Also, an ASPP module is employed to increase the semantic receptive filed to further improve the segmentation accuracy on different scale. Based on the multi-object combination and multi-scale ensemble, the proposed algorithm achieves the first place on the YouTube-VOS 2019 Semi-supervised Video Object Segmentation Challenge with a JF mean score of 81.8%.

1. Introduction

Video object segmentation has revealed increasing importance in very recent years and proved to be crucial in many industrial applications. Currently, there are two popular video object segmentation benchmarks in the community, DAVIS, and YouTube-VOS, which introduce real-world complex scenarios such as small, unknown, deformed, overlapping objects.

Instead of segmenting all the predefined objects in a given image, semi-supervised video object segmentation requires the model to only segment a target object, which the annotated mask in the first frame is provided as guidance. This problem becomes more complicated when multiple target objects is required. The inherent characteristics give semi-supervised video segmentation the ability of better describing user intention in practical applications since it allows user to decide which object to be followed.

Many VOS methods are proposed over the past few years for this challenge. Part of them [6][9][10] relied on temporal continuity and propagates the segmentation mask from the first frame to the next. MaskTrack[6] introduced optical flow combining with the semantic branch in an end-to-end manner to achieve this goal. However, the historical results in these methods seriously damage the accuracy of current frame. They have difficulties in handling large object translation and moving objects drifting. Others separately segmented video frames and concatenated each frame in series with classification-based method. PReMVOS[9] employed Mask RCNN[4] and Deeplab v3+[2] for object segmentation, then merged the objects with ReID features[7] and optical flow features[5]. Limited by the accuracy of Mask RCNN, these algorithms are difficult to improve the accuracy of a single object.

In this paper, we propose a novel network for semi-supervised video object segmentation. We consider VOS from both tracking and segmentation perspectives. Inspired by Spatial Temporal Memory Network (STMN) [10], we design an Enhanced Memory Network (EMN) to keep the multi-frames information. The accuracy depends on the segmentation quality of the current frame and the guidance information hidden in the encoder which is different from previous methods. The proposed method adopts a two-step flow. Firstly, encoders are used to extract semantic feature maps of multiple frames, which are randomly selected and discontinuous. Then, these feature maps are applied for the segmentation of individual objects, together with the guidance of historical frames. By doing so, our model can handle fast pose and scale changes caused by the movements of the objects themselves and the camera view. Next, to perform fined-grained segmentation, We introduce a siamese network to extract the feature map of the cropped target patch and multiply it onto the whole feature map as

*This work is done when Zhishan Zhou, Lejian Ren, Yifei Ji are interns at Megvii Research.

†Corresponding author. Email: xiongpengfei@megvii.com

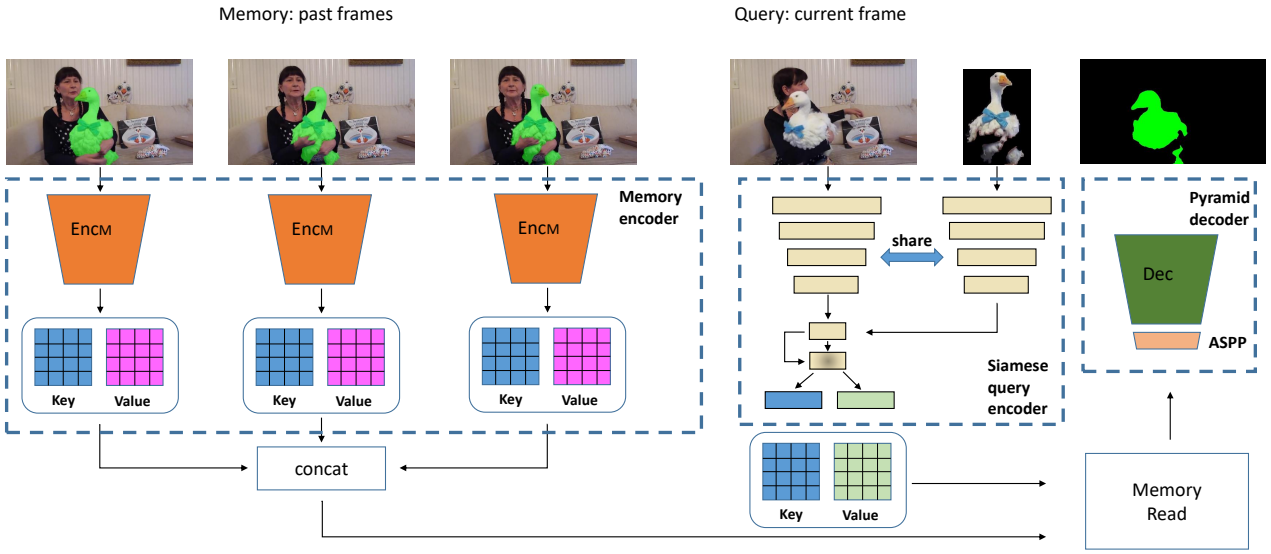

Figure 1: Overview of Enhanced Memory Network. On the basis of STMN, our method introduces two new modules, siamese query encoder and pyramid decoder. The siamese query encoder makes the generated key-value relevant to the target object while the pyramid decoder increases the robustness to object size.

the attention of the target object. Furthermore, to improve the segmentation accuracy on small objects, we employ an ASPP module to increase the semantic receptive filed.

2. Method

2.1. Revisit Space-Time Memory Network

Memory network was first introduced in the NLP research. The information of former steps is treated as memories which are then used to guide the prediction of the current step. This idea was adapted to VOS by the Space-Time Memory Network (STMN) [10]. The STMN is composed of three components, memory encoder, memory reader and query encoder & decoder, respectively. At step t, the image I_t is regarded as query image while the past frames $[I_0, ..., I_{t-1}]$ and masks $[m_0, ..., m_{t-1}]$ (given ground truth for the first frame, otherwise, the predicted mask) are memories. A key-value pair $\{k_t^Q, v_t^Q\}$ is encoded by the query encoder. Similarly, memory encoder encodes the past frames and additional masks into $[\{k_0^M, v_0^M\}, ..., \{k_{t-1}^M, v_{t-1}^M\}]$. The keys are used to address the relevant frames since not every past frame contains useful information to the current frame and the values store the information about this frame. Next, the current frame information is combined with past frames through memory reader. This process can be expressed as follow:

$$f_t = [v_t^Q, \frac{1}{Z}\sum_{i=0}^{t-1} R(k_t^Q, k_i^M)v_i^M], \tag{1}$$

where Z denotes normalizing factor and $R(\cdot)$ is the correlation function to calculate the similarity between current frame and memories:

$$Z = \sum_{i=0}^{t-1} R(k_t^Q, k_i^M) \tag{2}$$

$$R(k_t^Q, k_i^M) = exp(k_t^Q \cdot k_i^M). \tag{3}$$

The f_t is then fed into query decoder to obtain final mask of current frame.

2.2. Enhanced Memory Network

Siamese query encoder: The original encoder consists of a feature extractor F and two encoding heads H_k and H_v for key and value. Formally, for a given query image I_t,

$$feat_t = F([I_t]) \tag{4}$$

$$key_t = H_k(feat_t) \tag{5}$$

$$value_t = H_v(feat_t). \tag{6}$$

However, unlike the memory encoder, the key-value generated by the query encoder contains redundant information

due to the lack of guidance, which leads to inaccurate encodings. To solve this problem, we expect the query encoder to incorporate the target information. Inspired by the correlation filter[12] in the tracking research, we introduce a siamese query encoder in our network. Specifically, we crop the target object from the first frame using the bounding box of ground truth mask. This target patch P_0 is resized to $\mathbb{R}^{3\times\frac{H}{4}\times\frac{W}{4}}$ and fed into the feature extractor to obtain target feature:

$$feat_p = F(P_0). \tag{7}$$

Note that this feature extractor shares the parameters with the query encoder. The output target feature works as a convolution filter to highlight the most correlated region in the query feature. We then concatenate the original and the target aggregated feature:

$$\hat{feat}_t = [feat_t, feat_p * feat_t], \tag{8}$$

where the $*$ denotes convolution.

Pyramid decoder: We also consider the VOS from the segmentation perspective. After we capture the target object in current frame, the next thing we need to do is to obtain a delicate mask. We notice that in YouTube-VOS dataset, a common decoder works well in most cases. However, the performance drops when the target objects are extremely small. A simple way to deal with this problem is to use larger resolution inputs. But it brings another problem that the predictions on big objects get worse due to shrinking of receptive field. We solve this problem by adding an ASPP[15] module to the decoder. ASPP is composed of convolutions with different scales, which can be used effectively in increasing the receptive field and recognizing small objects.

3. Experiments

3.1. Training Details

Similar to [10][13], we use a two-stage training strategy. Firstly, Our model is pre-trained on image datasets. Several saliency and semantic segmentation datasets are used in this stage, such as MSRA10K[3], ECSSD[11], COCO[8]. In our experiments, more datasets lead to slightly better results. The pre-trained model is then fine-tuned on YouTube-VOS 2019 training set with multi-frame in the second stage. Unlike [10], we do not use the multi-object training method. Instead, we randomly select one instance from multiple instances for training so that we can use a larger batch size.

During pre-training, 384 x 384 patches are randomly cropped from all the image datasets. While in main training, we adopt randomly cropped 384x640 patches from the 3471 training videos following the official split. We found that maintaining the aspect ratio of the input image has an advantage to the training. The data augmentation, like rotation, flip, saturation, are applied to increase the data diversity. The batch size is set to 24 using four NVIDIA GeForce 1080 Ti GPUs (6 per GPU). We minimize the cross-entropy loss with Adam optimizer using "poly" policy[1] as learning rate schedule. The base learning rate is set to 5e-5 and power to 0.9. Pre-training stage takes about 2 days and main training stage takes about 1.5 days.

Method	Overall
-	0.763
Winner-take-all	0.768
softmax aggregation	0.783

Table 1: Multiple objects fusion methods comparison.

Siamese query encoder	ASPP	ensemble	Overall
			0.783
✓			0.791
✓	✓		0.802
✓	✓	✓	0.820

Table 2: Ablation study of our proposed modules on YouTube-VOS validation set.

3.2. Components analysis

We evaluate our model on YouTube-VOS 2019 [14] validation set. Region similarity J and the contour accuracy F are used as metrics, following the official test scripts.

The decoder of our model can only predict the foreground and background probabilities for a given object. To obtain a multi-object prediction, we need to run our model several times to get object probabilities separately. Then these results need to be combined. To this end, we first generate a mask for each class without competition among classes. Since each pixel can only belong to a single instance, we then use the method described in [13], the *softmax-aggregation*. Unlike STMN [10], we only use this method for testing and do not track the background. Another combination approach is the *winner-take-all*, it simply set the non-maximum instance probabilities to zeros. Table 1 shows the results comparison of different multi-object combination methods.

Furthermore, we analyze the effectiveness of our proposed modules. As can be seen in Table 3, the accuracy of the original Enhanced Memory Network (EMN) is 0.783. Our siamese query encoder brings a 0.8% improvement and the performance is further boosted by 1.1% using ASPP module. The final results on the validation benchmarks achieve 0.802. The ablation study shows the effectiveness of our proposed modules.

Team	Overall	Seen		Unseen	
		J	F	J	F
Ours	0.818 (1)	0.807 (1)	0.773 (2)	0.847 (1)	0.847 (2)
theodoruszq	0.817 (2)	0.800 (2)	0.779 (1)	0.833 (2)	0.855 (1)
zxyang1996	0.804 (3)	0.794 (3)	0.759 (4)	0.833 (3)	0.831 (4)
swoh	0.802 (4)	0.788 (4)	0.759 (3)	0.825 (4)	0.835 (3)
youtube_test	0.791 (5)	0.779 (5)	0.747 (5)	0.815 (5)	0.822 (5)
color94	0.779 (6)	0.775 (6)	0.726 (6)	0.810 (6)	0.804 (6)
Jono	0.714 (7)	0.703 (10)	0.680 (7)	0.736 (10)	0.740 (8)
andr345	0.710 (8)	0.699 (11)	0.667 (8)	0.732 (11)	0.740 (7)
hthieu	0.688 (9)	0.707 (8)	0.619 (9)	0.742 (9)	0.685 (9)
JLU_thunder	0.687 (10)	0.713 (7)	0.610 (10)	0.750 (7)	0.673 (10)

Table 3: Results in YouTube-VOS 2019 test set. Our method ranks first on J score of both seen and unseen objects and achieves an overall first place.

3.3. Model ensemble

We also utilize the model ensemble which consists of 3 models with different hyperparameter settings to promote model performance. Simply, we average the object probabilities of different models. After the model ensemble, we achieve 82.0% global mean on YouTube-VOS 2019 validation set. The results of the same model are submitted onto the test server an obtain a global mean of 81.8%. As shown in Table 3, our method ranks first on YouTube-VOS 2019 semi-supervised video object segmentation challenge on both seen and unseen objects.

4. Conclusion

In this paper, we propose an enhanced memory network for semi-supervised video object segmentation. Our approach achieves an overall score of 0.817, ranking first place on YouTube-VOS 2019 semi-supervised video object segmentation challenge.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[3] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[6] Anna Khoreva, Federico Perazzi, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3491–3500, 2016.

[7] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018.

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[9] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018.

[10] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. *CoRR*, abs/1904.00607, 2019.

[11] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:717–729, 2016.

[12] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach, 2018.

[13] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018.

[15] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.