Quality-aware and Selective Prior Enhancement Memory Network for Video Object Segmentation

Yong Liu[1]†, Ran Yu[1], Xinyuan Zhao[2], Yujiu Yang[1]
[1] Shenzhen International Graduate School, Tsinghua University
[2]Huawei Technologies
{liu-yong20,yu-r19}@mails.tsinghua.edu.cn,
zhaoxinyuan1@huawei.com,
yang.yujiu@sz.tsinghua.edu.cn

Abstract

Several spatial-temporal memory network-based methods have recently proven that using more intermediate frames with predicted masks helps segment objects in the current frame. However, low-quality frames will also be memorized in these methods, which takes detrimental effects on the segmentation of subsequent frames. What's more, they also do not fully use long-term and short-term dependency on video tasks. Thus, this paper proposes a quality-aware and selective prior enhancement memory network (QPM) for Video Object Segmentation. In QPM, a quality assessment branch evaluates the accuracy of each frame's segmentation results. Besides, knowing the quality of the previous adjacent frame, the model can determine whether long-term dependency or short-term dependency will be utilized to enhance the representation since they may be noise information in some cases. And then, a prior enhancement module will strengthen this dependency for the current frame. Based on these improvements, without multi-scale testing, our method achieves 84.2% overall score on the YouTube-VOS test set, which is the 4th on the Youtube-VOS Challenge 2021.

1. Introduction

Given a video and the annotations of single or multiple objects of the first frame, the task of semi-supervised video object segmentation (Semi-VOS) is to segment these target objects in subsequent frames. It is one of the most challenging tasks in computer vision with many potential applications, including interactive video editing, augmented reality, and autonomous driving. Early methods such as MaskRNN [2] and PReMVOS [3] refine masks from previous frames with a fully convolutional network. But, mask propagation usually leads to error accumulation especially in the case of occlusion and drifting.Recently, as a promising solution to Semi-VOS, matching-based methods have achieved increasing attention. FEELVOS [6] and CFBI [8] perform global and local matching between pixels in the current frame and ones in the first or the previous adjacent frame, which does not perform well when objects disappear and reappear. STM [4], KMN [5], and MiVOS [1] utilize a memory network storing intermediate frames to make use of more frames, which has been proved to be effective. However, these methods memorize frames without taking the quality of their segmentation result into account, which is unreasonable since low-quality frames will take detrimental effects on the segmentation of subsequent frames. From Figure 1, we can see that for fast motion, truncation, and similar objects, the algorithm will predict inaccurate masks. If a memory network can make sense of low-quality frames, it can selectively memorize high-quality frames to reduce the influence of noise information. Besides, there is a mismatch between these methods and the nature of the Semi-VOS task. STM [4] treats each memory frame equally, while the first frame whose annotations are given and the previous adjacent frame, which is the most similar to the current frame, are more significant than others.

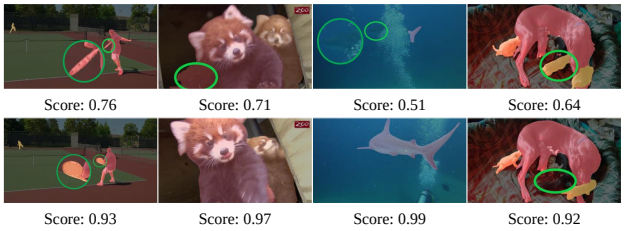

Figure 1. Visualization of segmented results with different quality scores.

†This work was done during an internship at Huawei Technologies.

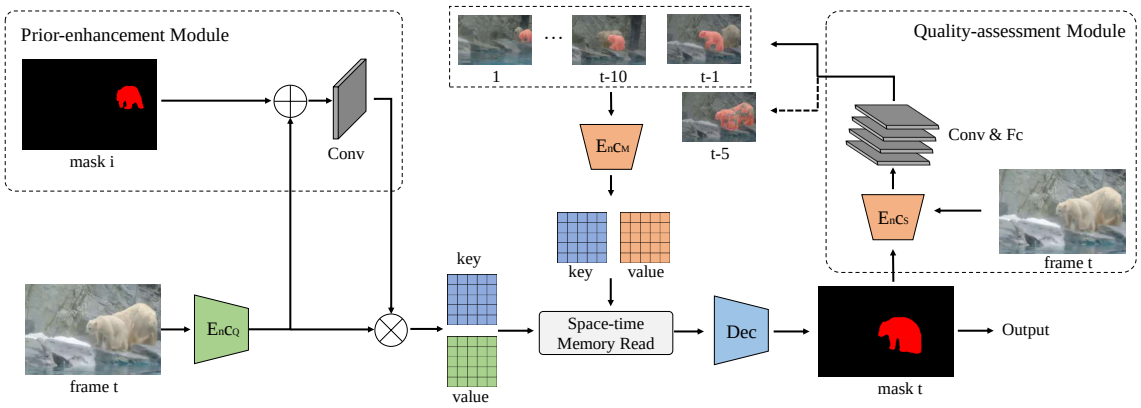

Figure 2. An overview of QPM. $\oplus$ and $\otimes$ denote concatenate operator and element-wise product, respectively. Based on STM [4], the network takes both query (current frame) and memory (memorized intermediate frames) as input. Our method introduces two new modules, the prior enhancement module and the quality assessment module. The quality assessment module takes the prediction mask, and the current frame as input to predict a score for each segmentation result and determines whether an intermediate frame should be memorized. According to the quality score, the prior enhancement module selectively takes the mask that belongs to the first frame or the previous adjacent frame as input.

To alleviate the problems above, we propose a quality-aware and selective prior enhancement memory network, which is obtained by adding a quality assessment module and a prior enhancement module on the basis of MiVOS [1]. According to quality scores predicted by the quality assessment module, there are two effective improvements for the network. The first is all memory frames can be guaranteed to be high-quality frames which means that noise information is significantly filtered. The second is that the prior enhancement branch will selectively enhance the first frame or the previous adjacent frame. The necessity for selective enhancement is that although the first frame is annotated, it does not provide a good prior for all frames since the object is deforming and the scene is changing. And if the previous adjacent frame is reliable, it is able to provide the most accurate prior for the current frame. If not, it means that the model's prediction is wrong. To avoid error accumulation, we use the long-term dependency of the first frame to search globally at this time. By filtering low-quality memory frames and strengthening the prior of the most significant frames, our method achieves 84.2% $J\&F$ on the YouTube-VOS 2021 test set [7].

2. Method

The structure of our quality-aware and prior enhancement memory network (QPM) is shown in Figure 2. As in STM [4], during the video processing, previous frames with segmentation masks are considered memory, and the current frame is considered query. Query encoder and memory encoder extract query feature and memory feature, respectively. A prior enhancement module firstly enhances the query feature to strengthen long-term dependency or short-term dependency selectively. Then the enhanced query feature and memory features are fed into convolution layers to generate corresponding key and value maps. Afterward, the spatial-temporal memory read block performs pixel-level matching between query key and memory keys and outputs the weighted sum of all value maps to the decoder. After segmentation in the decoder, the quality assessment module predicts a score indicating the mask's accuracy. If the quality score is low, which means that the segmentation result is poor, this frame will not be stored in memory frames, as shown by the dotted line branch in Figure 2. It is worth noting that in our method, the memory network stores the first frame, the previous adjacent frame, and every 5 intermediate frames with high-quality segmentation results.

2.1. Spatial-temporal Memory Network

The concept of memory networks has been used in many fields. Since it fits the nature of the video tasks, the researchers adapted this idea to resolve the semi-supervised video object segmentation task by the Space-Time Memory Network (STM) [4]. Rather than only the first frame and the previous adjacent frame, STM [4] calculates the similarity between the current frame and all memory frames in the space-time memory read block. This procedure can be summarized as:

$$\mathbf{y}_i = \mathbf{v}_i^Q \oplus \frac{1}{Z} \sum_{\forall j} f(\mathbf{k}_i^Q, \mathbf{k}_j^M) \mathbf{v}_j^M, \tag{1}$$

where i and j are the index of the query (Q) and the memory (M) location. $\mathbf{k}$ and $\mathbf{v}$ denote key map and value map, respectively. $Z = \sum_{\forall j} f(\mathbf{k}_i^Q, \mathbf{k}_j^M)$ is the normalizing factor and $\oplus$ denotes concatenation. And the similarity function f is $f(\mathbf{k}_i^Q, \mathbf{k}_j^M) = \exp(\mathbf{k}_i^Q \circ \mathbf{k}_j^M)$, in which $\circ$ denotes dot-product.

2.2. Quality Assessment Module

Taking the quality of memory into account, in order to reduce the negative impact of poorly segmented memory frames, we construct a quality assessment module to evaluate the quality of segmentation results for each frame in videos. The module is composed of a score encoder that is the same as the memory encoder, four 3×3 convolution layers, and two fully connected layers. The score encoder takes both the current frame and segmentation mask as input to acquire more information. The output of this module is a quality score whose target value is the **Mask IoU** between the prediction and ground truth.

Due to the difficulty of processing different videos is various, there are differences in the scores of different videos. Therefore, in order to better measure the relative quality of predicted masks in a video, we make the final score of each segmentation mask to be the result that its initial score divided by the first frame's initial score. Back to Figure 1, it shows the consistency between the evaluation score and the segmentation quality. The process of predicting scores can be expressed as:

$$m_t = E_n(t_{in} \oplus t_{out}); \quad score_t = \frac{A(m_t)}{score_1} \tag{2}$$

where t_{in} and t_{out} denote input frame and output prediction mask respectively. E_n is the score encoder used to get the feature m_t. A is the score prediction function formed by convolution layers and fully connected layers.

With the quality score, we can only save the intermediate frames whose scores are higher than $\sigma_m = 0.7$, which can be summarized as:

$$M_t = \begin{cases} M_{t-1} \oplus F_t, & \text{if } score_t > \sigma_m \\ M_{t-1}, & \text{otherwise} \end{cases} \tag{3}$$

where M denotes the memory network. F_t is the current frame.

By filtering low-quality frames, the model is able to merely perform matching between the current frame and those accurate previous segmented frames. Thus, the segmentation for the current frame can make use of intermediate frames without considering the possible bad effects of incorrectly segmented pixels.

2.3. Prior Enhancement Module

We propose a prior enhancement module to emphasize the importance of short-term dependency or long-term dependency, which are not considered by the STM series methods. Firstly, a prior mask is concatenated with the current frame's embedding to get a prior feature map. Then a convolution layer is adopted to produce a prior-enhancement map. Finally, it performs an element-wise product between the prior-enhancement map and the current frame's embedding to get the prior-enhanced feature.

Based on the quality score of the previous adjacent frame, there are two enhancement options for this module. Due to the continuity of video frames, the previous adjacent frame is usually very similar to the current frame. In contrast, although the first frame has been annotated, it can not always provide precise guidance or even bring a bad influence for object deformation and scene changes. Therefore, if the segmentation result of the previous adjacent frame is relatively accurate, we tend to strengthen the short-term dependency. Otherwise, we choose to enhance long-term dependency and utilize global information provided by the first frame to avoid error accumulation. This process can be expressed as the following equation:

$$f_{en} = h(f_t \oplus mask_i) \otimes f_t, \tag{4}$$

where h represents convolution and sigmoid operation. f_t is the current frame's feature extracted by the encoder, and f_{en} is the enhanced feature. If the quality score of the previous adjacent frame is higher than $\sigma_{en} = 0.65$, $i = t - 1$, otherwise $i = 1$.

3. Experiment

3.1. Training Details

Following the training setting in MiVOS [1], we take a three-stage training strategy: our model is pre-trained on static image datasets (stage 0) and BL30K dataset (stage 1). During pre-training, each image is expanded into a pseudo video of 3 frames through data augmentation. Then the pre-trained model is fine-tuned on DAVIS and YouTube-VOS training sets with randomly sampling 3 frames of each video for training (stage 2).

We randomly crop 384×384 patches from images for training. Using two Tesla V100 GPU, the batch size is set to 14 (stage 0) and 8 (stage 1&2) each GPU. All encoders take ResNet-50 as the backbone. And we minimize the BCE loss for segmentation and MSE loss for evaluation score with Adam optimizer. Their weights are the same.

3.2. Results

As shown in Table 1, our method achieves an overall score of 84.2% on the YouTube-VOS Challenge 2021 test set (Semi-VOS track) and ranks the fourth place without flip and multi-scale testing.

3.3. Ablation Study

We analyze the effectiveness of our proposed modules on the YouTube-VOS 2019 validation set. As we can see in Table 2, both the assessment branch and enhancement branch show significant performance improvement. Without any tricks, the performance is boosted from 82.4% to

Figure 3. Visual comparison of MiVOS (previous SOTA method) and our QPM. Each row demonstrates five frames sampled from a video sequence. (a) and (b) show the experimental results of MiVOS and QPM, respectively. We can see that MiVOS wrongly recognizes similar background object as target object and cannot correct.

Table 1. Ranking results on the YouTube-VOS 2021 test set. "seen" and "unseen" indicate whether the categories of tracking instances appeared in the training set or not. Our results are highlighted in bold.

Team	Overall	$\mathcal{J}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{seen}$	$\mathcal{F}_{unseen}$
wenhaowang	0.856	0.836	0.811	0.888	0.889
hkchengrex	0.854	0.828	0.814	0.883	0.893
testing-gg	0.854	0.836	0.806	0.888	0.885
Ours	**0.842**	**0.816**	**0.799**	**0.870**	**0.881**
cncyww	0.839	0.823	0.788	0.874	0.871
cheng321284	0.836	0.809	0.798	0.859	0.877
PixelKitty	0.835	0.814	0.793	0.866	0.868

84.0% only by applying these two novel branches. Besides, Figure 3 shows the visual improvement of our method compared to MiVOS [1] which is state-of-the-art before. We can see that when in challenging situations, MiVOS may make mismatches and cannot correct these mistakes. In contrast, by strengthening the critical frame's prior and only storing high-quality memory frames, the proposed method achieves satisfactory predictive performance.

Table 2. Ablation study of the components on YouTube VOS 2019 validation set

Assessment branch	Enhancement branch	Ensemble	Overall
			0.824
✓			0.836
✓	✓		0.840
✓	✓	✓	**0.852**

4. Conclusion

In this paper, we propose a quality aware and prior enhancement memory network (QPM) for Semi-VOS. Compared to the STM-based methods, QPM only memorizes the intermediate frames with high-quality segmentation results, which effectively alleviates the negative impact from wrongly segmented pixels. Besides, in order to utilize more appropriate prior information, QPM also selectively emphasizes long-term dependency or short-term dependency for different situations. Based on these improvements, our method achieves 84.2% overall score on the YouTube-VOS test set.

References

[1] H. K. Cheng, Y. Tai, and C. Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. *CoRR*, abs/2103.07941, 2021. 1, 2, 3, 4

[2] Y. Hu, J. Huang, and A. G. Schwing. Maskrnn: Instance level video object segmentation. In *Annual Conference on Neural Information Processing Systems 2017*, pages 325–334, 2017. 1

[3] J. Luiten, P. Voigtlaender, and B. Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer vision Part IV*, pages 565–580. Springer, 2018. 1

[4] S. W. Oh, J. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV2019*, pages 9225–9234. IEEE, 2019. 1, 2

[5] H. Seong, J. Hyun, and E. Kim. Kernelized memory network for video object segmentation. In *Computer Vision - ECCV 2020 - 16th European Conference, Part XXII*, pages 629–645. Springer, 2020. 1

[6] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L. Chen. FEELVOS: fast end-to-end embedding learning for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9481–9490. Computer Vision Foundation / IEEE, 2019. 1

[7] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 2

[8] Z. Yang, Y. Wei, and Y. Yang. Collaborative video object segmentation by foreground-background integration. In *Computer Vision - ECCV 2020 - 16th European Conference, Part V*, pages 332–348. Springer, 2020. 1