Towards Good Practices for Video Object Segmentation

Dongdong Yu†, Kai Su†, Hengkai Guo, Jian Wang, Kaihui Zhou, Yuanyuan Huang, Minghui Dong, Jie Shao and Changhu Wang
ByteDance AI Lab, Beijing, China

Abstract

Semi-supervised video object segmentation is an interesting yet challenging task in machine learning. In this work, we conduct a series of refinements with the propagation-based video object segmentation method and empirically evaluate their impact on the final model performance through ablation study. By taking all the refinements, we improve the space-time memory networks to achieve a Overall of 79.1 *on the Youtube-VOS Challenge* 2019.

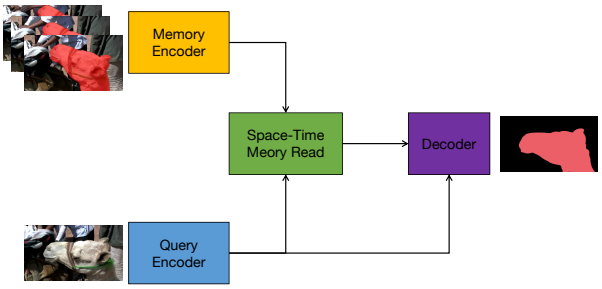

Figure 1. Overview of the Space-Time Memory Networks.

1. Introduction

In recent years, video object segmentation has attracted much attention in the computer vision community [18, 12, 6, 1, 9, 15, 17]. For a given video, video object segmentation is to classify the foreground and the background pixels in all frames, which is an essential technique for many tasks, such as video analysis, video editing, video summarization and so on. However, video object segmentation is far from a solved problem, both quality and speed are extremely vital for it.

The tremendous development of deep convolution neural networks bring huge progress in many areas, including image classification [5, 13], human pose estimation [16] and video object segmentation [18, 12, 6, 1, 9, 15]. These works can be divided into two classes: propagation-based methods [18, 12, 6] and detection based methods [1, 9, 15]. Propagation based methods, learn a convolution neural network to leverage the temporal coherence of object motion and propagate the mask of the previous frame to current frame. However, there exists some challenging cases, such as occlusions and rapid motion, which cannot be well addressed by the propagation methods. In addition, the propagation error can be accumulated. Detection-based methods, learn the appearance of the target object from a given annotated frame, and perform a pixel-level detection of the target object at each frame. However, they often fail to adapt to appearance changes and have difficulty separating object instances with similar appearances.

Space-Time Memory Networks [10] (STMN) is one of the propagation-based methods, which explores and computes the spatio-temporal attention on every pixel in multiple frames to segment the foreground and the background pixels. By using multi-frame information, it can relieve the bad performance caused by appearance changes, occlusions, and drifts. In our paper, we follow STMN and examine a collection of training procedure and model architecture refinements which affect the video object segmentation performance. First, we explore the segmentation performance of the pre-training stage with different pre-training datasets. Second, we do some ablation study to decide which backbone (including ResNet-50, Refine-50) should be selected for the encoder. Finally, we validate some testing augmentation tricks, including flip-testing, multi-scale testing and model ensemble, to improve the segmentation performance.

2. Method

The chart of Space-Time Memory Networks is shown in Figure 1. During the video processing, the previous frames with object masks are considered as the memory frames and the current frame without the object mask as the query frame. The encoder extracts the appearance information with the memory frames and query frame. The Space-time Memory Read Module will compute the spatio-temporal attention between the query frame and memory frames. Then, the decoder will output the final segmentation result for the

† Equal contribution.

query frame.

Pre-training The STMN is first pre-trained on a simulation dataset generated from static image data, then fine-tuned for real-world videos through the main training. Similar to STMN, we used image datasets with instance object masks (Pascal VOC [3, 4], COCO[8], MSRA10K[14], ECSSD [2], and Youtube-VOS) to simulate training samples. We find that add the Youtube-VOS into the simulation datasets can significantly improve the segmentation performance.

Backbone: The STMN use the ResNet-50 as the backbone of the encoder and decoder. In our work, we propose a new backbone, named Refine-50, which can well handle the scale variant cases.

Testing Tricks: In order to improve the segmentation performance, we use the flip-testing and multi-scale testing for a single model. For ensemble experiments, we average the object probability from ResNet-50 and Refine-50.

3. Experiments

In this section, we first briefly introduce the Youtube-VOS [19] dataset and corresponding evaluation metrics, then we evaluate a series of refinements through ablation studies. Finally, we report the final results in the Youtube-VOS Challenge.

3.1. Datasets and Evaluation Metrics

Youtube-VOS [19] is the latest large-scale dataset for video object segmentation. The training set consists of 3471 videos, and we further split the training set into 3321 offline-training set and 150 offline-validation set. We adopt the offline-validation set to select the model from different epochs. For evaluation, we measure the region similarity J and contour accuracy F. The results of validation set and test set are evaluated through the online CodaLab server.

3.2. Training Details

Our model is implemented in Pytorch [11]. For the training, we $4V100$ GPUs on a server are used. Adam [7] optimizer is adopted. The learning rate is set to $1e-5$. The input size for the network is made to a fixed 384×384. The cross-entropy loss is used. The batch size on each GPU is set to 4.

3.3. Testing Details

Follow [10], we simply save a memory frame every 5 frames. And the input size of the network for inference is set to an integer multiple of 16. Moreover, we adopt the multi-scale testing to boost the performance.

Table 1. The results of Pre-training, Main-training and Full-training with ResNet-50 on YouTube-VOS validation set.

Training Method	$Overall$
Pre-training only (w Youtube-VOS)	0.617
Pre-training only (w/o Youtube-VOS)	0.667
Main-training only	0.681
Full-training	0.766

Table 2. The results of different backbones with pre-training only on YouTube-VOS validation set.

Backbone	$Overall$
ResNet-50	0.667
Refine-50	0.708

Table 3. The results of flip and multi-scale testing with ResNet-50 and full-training on YouTube-VOS validation set.

Flip	Multi-Scale	$Overall$
		0.761
√		0.766
√	√	0.777

3.4. Refinements during Training and Testing Phases

In this section, we evaluate the effectiveness of a series of refinements during the training and testing phases.

3.4.1 Pre-training on images

We evaluate the performance of different training methods in this experiment. As shown in Table 1, pre-training only achieved perfermance close to main-training only, without adopting any real videos for training. Without the pre-training phase, the performance drops from 0.766 to 0.681. Therefore, diverse appearance of different objects during the pre-training stage significantly boost the generalization of our model.

3.4.2 Different Backbones

We evaluate the effectiveness of different backbones in this experiment. As shown in Table 2, by adopting our stronger refine-50 backbone, the results improve from 0.667 to 0.708.

3.4.3 Multi-Scale Testing

We evaluate the effectiveness of flip and multi-scale testing in this experiment. We adopt the multi-scale with $0.75, 1.0$. As shown in Table 3, when adopting the flip testing, the performance improve from 0.761 to 0.766. With multi-scale testing involved, we further boost the performance, from 0.766 to 0.777.

Table 4. Ranking results on the YouTube-VOS test set.

Team Name	$Overall$	J_seen	J_unseen	F_seen	F_unseen
zszhou	0.818	0.807	0.773	0.847	0.847
theodoruszq	0.817	0.800	0.779	0.833	0.855
zxyang1996	0.804	0.794	0.759	0.833	0.831
swoh	0.802	0.788	0.759	0.825	0.835
Jono	0.714	0.703	0.680	0.736	0.740
andr345	0.710	0.699	0.667	0.732	0.740
Ours (youtube_test)	0.791	0.779	0.747	0.815	0.822

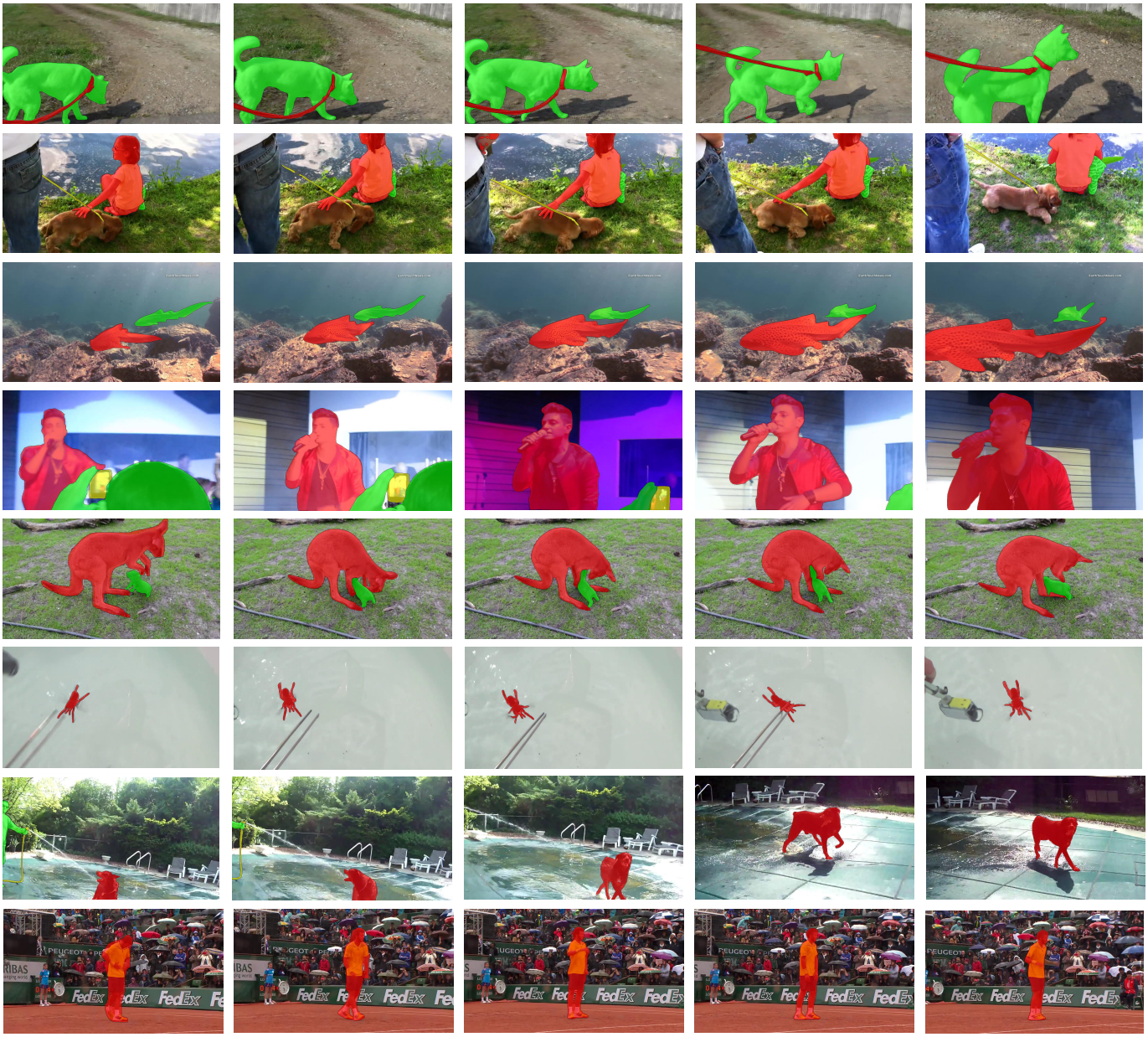

Figure 2. Qualitative results of our model on the YouTube-VOS test set.

3.5. Results on Youtube-VOS Challenge

Finally, we ensemble the model with ResNet-50 and Refine-50, and achieved 0.791 on the Youtube-VOS test set. The qualitative results of the final model are shown in Figure 2.

4. Discussions

During our experiments, we find two main problems. Firstly, the results on validation set of the model with different epochs vary seriously. Secondly, the results on validation set and test set for the model with same epoch show a large difference.

5. Conclusion

In this work, we conduct a series of refinements with the Space-Time Memory Networks and empirically evaluate their impact on the final model performance through ablation study. Finally, we achieve a $Overeall$ of 79.1 on the Youtube-VOS Challenge 2019.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.

[2] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[4] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *The DAVIS Challenge on Video Object Segmentation*, 2017.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[9] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018.

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

[11] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[12] 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.

[13] Wei Shen, Mu Zhou, Feng Yang, Dongdong Yu, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 61:663–673, 2017.

[14] J Shi, Q Yan, L Xu, and J Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2016.

[15] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2167–2176, 2017.

[16] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5674–5682, 2019.

[17] Jia Sun, Dongdong Yu, Yinghong Li, and Changhu Wang. Mask propagation network for video object segmentation. *arXiv preprint arXiv:1810.10289*, 2018.

[18] 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.

[19] 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.