The 2018 DAVIS Challenge on Video Object Segmentation - CVPR 2018 Workshops

Fast User-Guided Video Object Segmentation by Deep Networks

Seoung Wug Oh, Yonsei University
Joon-Young Lee, Adobe Research
Ning Xu, Adobe Research
Seon Joo Kim, Yonsei University

Abstract

We present a deep learning method for the interactive video object segmentation. Our method is built upon two core operations, interaction and propagation, and each operation is conducted by a Convolutional Neural Network (CNN). The two networks are trained jointly to adapt to each other, which reduces unstable behaviors between the two operations. At the testing time, our method produces high-quality results and also runs fast enough to work with users interactively. Our method took first place on the interactive track of DAVIS Challenge on Video Object Segmentation 2018 with an AUC of 0.641 and a J@60s of 0.647.

1. Introduction

Video object segmentation is a task of separating a foreground object from a video sequence. This problem is often investigated through either a fully-automatic (*i.e*. unsupervised foreground object segmentation [14]) or a semi-supervised (*i.e*. ground-truth object masks are given on few frames [2, 10]) approach. However, both solutions have an intrinsic performance issue as they have a limitation in reflecting the user's intention and refining incorrect estimations. Interactive video segmentation can potentially resolve this issue by allowing user intervention given in a user-friendly form such as scribbles [15, 13, 1].

In [3], a workflow for interactive segmentation introduced that can minimize user effort is proposed. In this scenario, the user provides annotations on a selected frame and an algorithm computes segmentation maps for all the frames in the video in a batch process. The user can provide additional annotations and an algorithm need to refine the object masks of all the frames. The authors present the Interactive track on the 2018 DAVIS Challenge on Video Object Segmentation to automatically evaluate interactive methods with a robot that simulates user interactions [3].

In this paper, we present a fast and accurate method for interactive video object segmentation. We define two core operations, *interaction* and *propagation*, and construct two deep networks dedicated to each operation (see Fig. 1). The interaction network takes the user annotation (*e.g*. scribbles) to segment the foreground object. The propagation network transfers the object mask computed in the source frame to other neighboring frames. These two networks are internally connected using our feature aggregation module and are also externally connected so that each of them takes the other's output as their input.

The two networks are jointly trained through two stages: pre-training on synthetic static images and fine-tuning on real video data [9]. At the testing time, the inference of our networks is fast enough to interact with users in real-time. Our method took first place on the interactive track of DAVIS Challenge on Video Object Segmentation 2018 [3] with an AUC of 0.641 and a J@60s of 0.647.

2. Method

Network structure. We have two networks, interaction and propagation, and both networks are constructed as an encoder-decoder structure. We adopt the ROI align before the encoder to make our networks to pay attention to the region of interest (area near the target object) [5]. We take ResNet50 [6] pre-trained on ImageNet as the encoder network and modify it to be able to take additional input channels (*e.g*. scribbles and the previous masks) by implanting additional filters at the first convolution layer [10, 16].

The decoder takes the output of the encoder and produces an object mask. We make modifications to the feature pyramid networks [8, 11] by adding residual blocks [7] and use it as the building block of our decoder, as shown in Fig. 1 (d),(e).

Multi-object segmentation. For the multi-object scenario where scribbles for each object are given, we first estimate masks for each object then merge the masks to get the multi-object mask using the soft aggregation proposed in [9]:

$$p_{i,m} = \sigma\big(l(\hat{p}_{i,m})\big) = \frac{\hat{p}_{i,m}/(1-\hat{p}_{i,m})}{\sum_{j=0}^{M}\hat{p}_{i,j}/(1-\hat{p}_{i,j})}, \quad \text{s.t.} \quad \hat{p}_{i,0} = \Pi_{j=1}^{M}(1-\hat{p}_{i,j}), \quad (1)$$

where σ and l represent the softmax and logit functions respectively, $\hat{p}_{i,m}$ is the network output probability of the object m at the pixel location i, m=0 indicates the background, and M is the total number of objects.

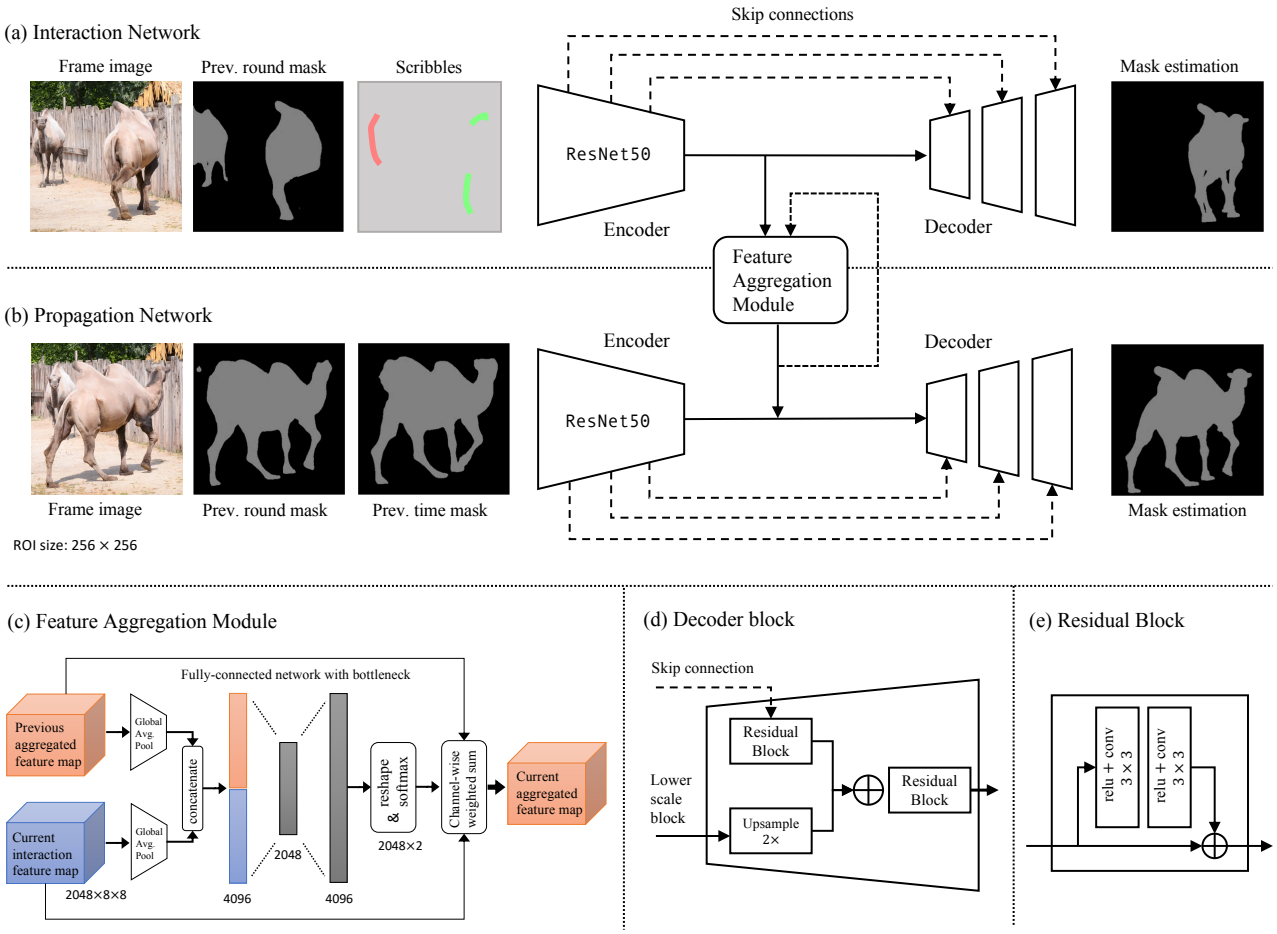

Figure 1: The overall network structure. We have two deep networks dedicated each to (a) interaction and (b) propagation tasks. The two networks are internally connected by (c) our feature aggregation module.

Training. We use videos from DAVIS [12], GyGo [4], and a new dataset[1] to train our networks. Our networks are first pre-trained on synthetic image data following the data simulation method in [9], and then are fine-tuned on real video data.

Testing. At the beginning of each round, a user selects a frame and draws scribbles on it. The interaction network produces a segmentation result for the selected frame. Then, the propagation network successively propagates the object mask both forward and backward until it reaches the end. A number of rounds can be repeated until the user stops providing additional scribbles. We update previous round's masks with newly computed estimation by weighted average where the weighting factor is inversely proportional to the propagated distance. The object mask is propagated until we reach a frame in which user annotations were given in any previous rounds.

[1]This is under review. We use the dataset only and there is no overlap in the technical contribution with our submission.

3. DAVIS Challenge

To fairly compare our method against the state-of-the-art methods [3], we test our model on the interactive track benchmark at 2018 DAVIS Challenge. In the challenge, each method can interact with a robot agent up to 8 times and is expected to compute masks within 30 seconds per object for each interaction. The performance of each method is evaluated using two metrics: area under the curve (AUC) and Jaccard at 60 seconds (J@60). AUC is designed to measure the overall accuracy during the evaluation. J@60 measures the accuracy with a limited time budget (60 seconds). Our method took first place on this challenge with an AUC of 0.641 and a J@60s of 0.647 and the leaderboard is available at https://davischallenge.org/challenge2018/interactive.html.

References

[1] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. In *ACM Transactions on Graphics (ToG)*, volume 28, page 70. ACM, 2009. 1

[2] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 1

[3] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 1, 2

[4] I. Friedman, I. Chemla, E. Smolyansky, M. Stepanov, I. Afanasyeva, G. Sharir, S. Nadir, and S. Rorlich. Gygo: an e-commerce video object segmentation dataset by visualead, 2017. 2

[5] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, pages 770–778, 2016. 1

[7] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 1

[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[9] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim. Fast video object segmentation by reference-guided mask propagation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[10] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[11] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016. 1

[12] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 2

[13] B. L. Price, B. S. Morse, and S. Cohen. Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 779–786. IEEE, 2009. 1

[14] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[15] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen. Interactive video cutout. In *ACM Transactions on Graphics (ToG)*, volume 24, pages 585–594. ACM, 2005. 1

[16] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 373–381, 2016. 1