The 2019 DAVIS Challenge on Video Object Segmentation - CVPR 2019 Workshops

Interactive Video Object Segmentation Using Sparse-to-Dense Networks

Yuk Heo
Korea University
yukheo@mcl.korea.ac.kr

Yeong Jun Koh
Chungnam National University
yjkoh@cnu.ac.kr

Chang-Su Kim
Korea University
changsukim@korea.ac.kr

Abstract

An interactive video object segmentation algorithm, which takes scribble annotations for target objects as input, is proposed in this work. First, we develop a sparse-to-dense network, called SDI-Net, to yield segmentation results in an annotated frame where scribbles are given. Then, we generate points within the segmented regions and propagate them to adjacent frames using optical flow vectors. Second, we design another sparse-to-dense network, called SDP-Net, to achieve segmentation in the adjacent frames using the propagated points. SDP-Net yields dense segmentation results, while refining propagation errors due to unreliable optical flow vectors. We perform this propagation process sequentially from annotated frames to all frames to obtain segment tracks of objects. The proposed algorithm ranks 2nd on the Interactive Scenario of the DAVIS Challenge 2019 on Video Object Segmentation with the performances of 0.647 AUC and 0.609 J&F@60s using precomputed optical flow and 0.639 AUC and 0.560 J&F@60s using online optical flow.

1. Introduction

Video object segmentation (VOS) aims at separating objects in a video sequence. There are semi-supervised VOS techniques [1, 7–9] and unsupervised VOS techniques [10, 11]. Semi-supervised VOS takes user annotations for target objects at the first frame, while unsupervised VOS detects and segments out objects automatically. However, semi-supervised methods require time-consuming pixel-level annotations at the first frame. On the other hand, unsupervised methods may fail to separate multiple objects in a video sequence. Therefore, as an alternative approach, interactive VOS can be considered; a work-flow to achieve interactive VOS was presented in the 2019 DAVIS challenge [2].

We propose a novel interactive VOS algorithm using scribble annotations. We develop neural networks, which transform sparse annotations (scribbles or points) into dense segmentation results. Specifically, we propose two sparse-to-dense networks, called SDI-Net and SDP-Net. First, SDI-Net yields segmentation results in an annotated frame, where scribbles are given. Then, we generate points within the segmented regions and propagate them to next and previous frames using optical flow vectors. Second, SDP-Net produces dense segmentation masks in the adjacent frames, while refining propagation errors due to unreliable optical flow vectors. We sequentially perform this propagation process from annotated frames to all frames to obtain segment tracks of target objects. Then, the scribble interaction is repeatedly performed to refine inaccurate segmentation regions according to the work-flow in [2]. Experimental results demonstrate that the proposed algorithm yields competitive interactive VOS performances.

2. Proposed Algorithm

We propose two sparse-to-dense networks, SDI-Net and SDP-Net, to perform interaction and propagation, respectively. In the first segmentation round, given a scribble for each target object, SDI-Net yields a segmentation mask for the corresponding target object. Then, SDP-Net propagates the segmentation masks temporally to obtain segment tracks for the target objects. In the next round, we find the frame that has the worst segmentation results and then provide additional scribbles to correct the inaccurate results. Specifically, we extract positive and negative scribbles by comparing the segmentation results in the selected frame with the ground-truth, and use those scribbles as input to SDI-Net. Then, we propagate the refined segmentation regions to adjacent frames via SDP-Net. This interaction process is repeated 8 times, including the first round.

2.1. SDI-Net

Figure 1 shows the architecture of SDI-Net, which infers a segmentation results in an annotated frame with user interaction. The user interaction has two types according to iteration rounds. In the first round, only a scribble for a target object is given. In contrast, in subsequent rounds, both positive and negative scribbles are extracted by comparing segmentation results in the previous rounds with the ground-truth. More specifically, SDI-Net takes the annotated frame and one scribble interaction map for the target

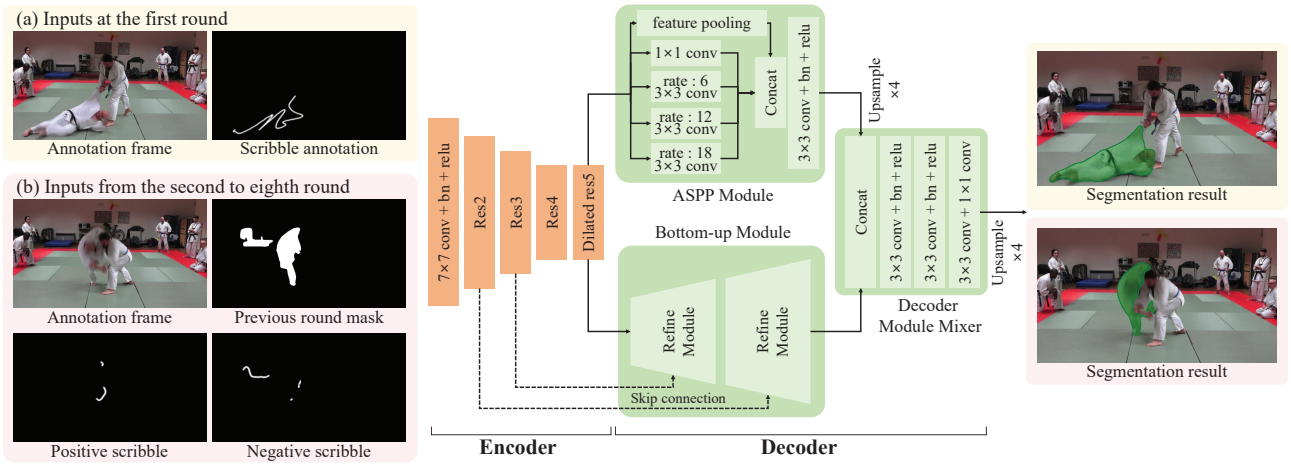

Figure 1: An overview of SDI-Net. The "Judo" sequence is used as an example. Segmented regions are depicted in green.

object in the first round. In the remaining rounds, SDI-Net takes the annotated frame, the segmentation mask map from the previous round, and two scribble interaction maps (positive and negative).

SDI-Net has the encoder-decoder architecture in Figure 1. We adopt ResNet-50 [5] as the encoder to extract features and employ the skip connections for low-level and high-level features. Then, we use two parallel decoder modules, ASPP module [3] and a bottom-up module. ASPP analyzes multi-scale context features using various convolutional layers. The bottom-up module has two Refine modules in [13] and exploits both low-level and high-level features via the skip connections. The output signals of the two decoder modules are concatenated and transformed into a probability map of a target object through convolutional layers, ReLUs, and batch normalization in the decoder module mixer.

2.2. SDP-Net

After obtaining the segmentation mask in the annotated frame, we propagate the mask forwardly and backwardly to yield the segment track for the target object in the entire video sequence. To this end, we generate points within the segmentation mask in the annotated frame and then propagate the generated points to next and previous frames using optical flow vectors. Specifically, we randomly pick 50 points from the segmented regions and dilate them using a Gaussian filter. We warp the dilated points to the adjacent frames bidirectionally using forward and backward flow vectors. Then, SDP-Net transforms the propagated sparse points into dense segmentation masks.

Figure 2 illustrates SDP-Net. It has two encoders based on the Siamese structure. The first encoder takes the annotated frame and the scribble interaction map, while the second encoder takes the current frame t and propagated points. The output features of the two encoder are combined by convolutional layers and squeeze-and-excitation (SE) module [6] in a feature mixer. Then, the combined features and the features of the second encoder are fed into ASPP module and the bottom-up module, respectively. Finally, the decoder module mixer yields a probability map for the target object as done in SDI-Net.

2.3. Training Strategy

To train the proposed networks, we use three datasets SBD [4], DAVIS2017 [15], and YouTube-VOS [16]. SDI-Net and SDP-Net use different datasets and training strategies to overcome the over-fitting problem. We adopt the pixel-wise cross-entropy loss between a predicted probability map and the ground-truth mask. We use the Adam optimizer with the learning rate decay 1/10 in every 20 epochs.

For SDI-Net, we use the whole SBD dataset and the training set in DAVIS2017. We first pre-train the network using the SBD dataset. Since there are no scribble annotations in SBD, we generate points from the ground-truth masks and substitute scribbles with the generated points for the pre-training. We then fine-tune SDI-Net using 180 scribble annotations in DAVIS2017.

For SDP-Net, we use YouTube-VOS that is the largest VOS dataset. For each video, we randomly select two frames that contain an identical object and obtain a point map by extracting points from the ground-truth mask of the object in one of the selected frames. To mimic segmentation errors, we transform the point map through rotation, scaling, translation, and erosion.

2.4. Inference

In the interactive scenario [2], scribble annotations for target objects in a randomly selected frame are provided in the first round. SDI-Net produces segmentation results

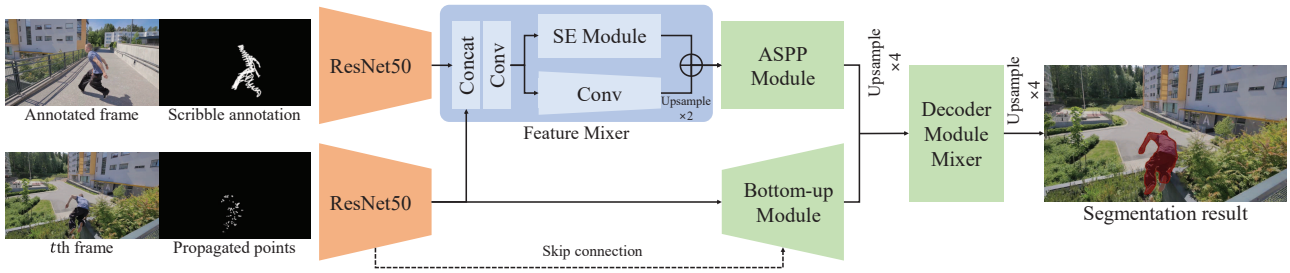

Figure 2: An overview of SDP-Net. The "Parkour" sequence is used as an example. Segmented regions are depicted in red.

Table 1: Comparison of the proposed algorithm with the conventional algorithms on the validation set in DAVIS2017. The best result and the second best result are boldfaced and underlined, respectively.

Algorithm	AUC	J@60s
Najafi *et al*. [12]	<u>0.702</u>	0.548
Oh *et al*. [14]	0.691	**0.734**
Proposed	0.698	0.691
Proposed*	**0.704**	<u>0.725</u>

* Performance with precomputed optical flow

using these scribbles. Then, SDP-Net propagates the segmentation masks forwardly and backwardly to the ends of the sequence. In the second round, a user selects a frame that has the worst segmentation result, and provides positive and negative scribbles for SDI-Net to improve the result. Then the refined mask is bi-directionally propagated again by SDP-Net. This is repeated iteratively.

3. Experimental Results

We assess the proposed algorithm on the DAVIS2017 dataset [15] with the metrics of area under the curve (AUC), Jaccard at 60 seconds (J@60s), and joint Jaccard and boundary at 60 seconds (J&F@60s). Table 1 compares the proposed algorithm with conventional interactive VOS algorithms [12, 14] on the validation set in DAVIS 2017. With precomputed optical flow, the proposed algorithm yields 0.704 AUC and 0.725 J@60s, which are the best and the second best results, respectively. Figure 3 shows the Jaccard performances according to time (s) and number of interaction rounds, respectively. The performances increase quickly and saturate at around 40s or 3rd round.

The DAVIS2017 test-dev dataset was adopted for the interactive scenario of the DAVIS Challenge 2019 on Video Object Segmentation [2]. In this challenge, as listed in Table 2, the proposed algorithm ranks 2nd with the performances of 0.647 AUC and 0.609 J&F@60s using precomputed optical flow. Also, the proposed algorithm yields 0.639 AUC and 0.560 J&F@60s using online optical flow.

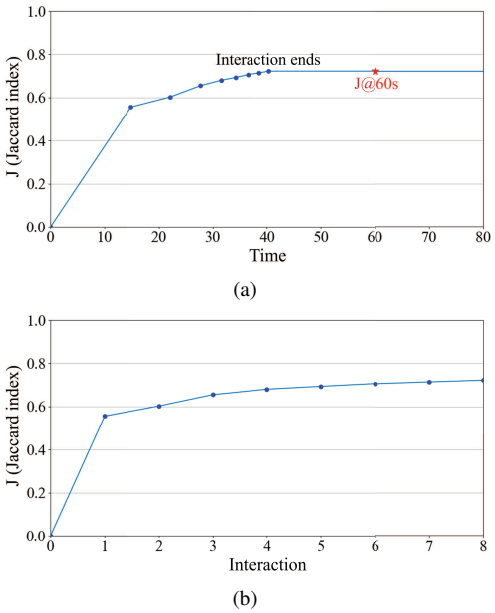

Figure 3: Jaccard performances on the validation set in DAVIS 2017 according to (a) time (s) and (b) number of interaction rounds. A red star in (a) indicates the Jaccard performance at 60s, *i.e*. J@60s.

Table 2: Comparison on the DAVIS2017 test-dev dataset. The best result and the second best result are boldfaced and underlined, respectively.

Participant	AUC	J&F@60s
S. W. Oh	**0.783**	**0.791**
Y. Heo (Proposed*)	<u>0.647</u>	<u>0.609</u>
Z. Lin	0.621	0.601
YK_CL	0.589	0.491

* Performance with precomputed optical flow

Figure 4 shows qualitative segmentation results of the "Pigs" sequence. As the interaction round goes on, segmentation masks are refined more accurately using additional positive and negative scribbles.

4. Conclusions

We proposed a novel algorithm to segment out objects in a video sequence in an interactive manner. We developed two sparse-to-dense networks, called SDI-Net and SDP-Net. First, SDI-Net yields segmentation results in an annotated frame, where scribbles are given. Then, points are generated within the segmented regions and then are propagated to adjacent frames. Second, SDP-Net produces dense segmentation masks from the point map, while refining propagation errors. This propagation process is sequentially performed from the annotated frame to all frames in the video sequence to yield a segment track of the target object. Experimental results showed that the proposed algorithm outperforms conventional interactive VOS algorithms on the validation set in DAVIS2017. Also, the proposed algorithm ranked 2nd on the Interactive Scenario of the DAVIS Challenge 2019 on Video Object Segmentation.

Acknowledgement

This work was supported by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government (MSIT) (No. GK18P0200, Development of 4D reconstruction and dynamic deformable action model based hyperrealistic service technology), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2018R1A2B3003896).

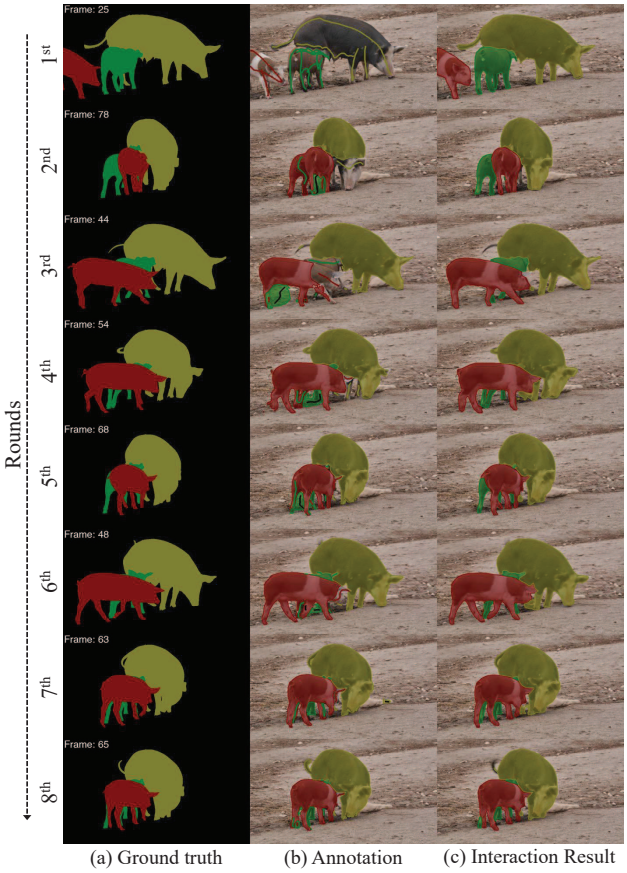

Figure 4: Interactive VOS results of the "Pigs" sequence according to the scribble interaction rounds.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 1

[2] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. V. Gool, F. Perazzi, and J. Pont-Tuset. The 2018 DAVIS Challenge on Video Object Segmentation. In *arXiv:1803.00557*, 2018. 1, 2, 3

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[4] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[6] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[7] W.-D. Jang and C.-S. Kim. Semi-supervised video object segmentation using multiple random walkers,. In *BMVC*, 2016. 1

[8] W.-D. Jang and C.-S. Kim. Streaming video segmentation via short-term hierarchical segmentation and frame-by-frame markov random field optimization,. In *ECCV*, 2016. 1

[9] W.-D. Jang and C.-S. Kim. Online video object segmentation via convolutional trident network. In *CVPR*, pages 5849–5858, 2017. 1

[10] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, pages 3442–3450, 2017. 1

[11] Y. J. Koh, Y.-Y. Lee, and C.-S. Kim. Sequential clique optimization for video object segmentation. In *ECCV*, 2018. 1

[12] M. Najafi, V. Kulharia, T. Ajanthan, and P. H. S. Torr. Similarity learning for dense label transfer. In *CVPR workshops*, 2018. 3

[13] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018. 2

[14] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Fast user-guided video object segmentation by interaction-and-propagation networks. In *CVPR*, 2019. 3

[15] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. In *arXiv:1704.00675*, 2017. 2, 3

[16] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. YouTube-VOS: a large-scale video object segmentation benchmark. In *arXiv:1809.03327*, 2018. 2