The 2020 DAVIS Challenge on Video Object Segmentation - CVPR 2020 Workshops

Interactive Video Object Segmentation with Multiple Reference Views, Self Refinement, and Guided Mask Propagation

Quoc-Cuong Tran[1,3], The-Anh Vu-Le[1,3], and Minh-Triet Tran *[1,2,3]

[1]University of Science, VNU-HCM, Vietnam
[2]John von Neumann Institute, VNU-HCM, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

In this work, we propose a fast and efficient method for interactive video object segmentation that takes scribbles as iterative input to specify the objects of interest and outputs the segmentation for those objects. Our approach contains three key modules: Control-point-based Scribbles-to-Mask, Self-Reference Refinement, and Guided Mask Propagation. It also maintains for each sequence a memory pool containing reliable frame-mask pairs, which is global across interactions, called Global Memory Pool. Additionally, we propose a strategy for the enrichment of this memory pool using more high-quality memory fragments, called Multiple Reference Views. Our proposed method achieved 0.767 and 0.759 in terms of Area Under the Curve (AUC) and $\mathcal{J}\&\mathcal{F}$ at 60 seconds ($\mathcal{J}\&\mathcal{F}$@60s), respectively, ranking 2nd in the Interactive Scenario of The DAVIS Challenge 2020.

1. Introduction

Video object segmentation aims to label pixel-level objects or background for a sequence of video frames. In an interactive scenario, a user gives iterative refinement inputs to the algorithm, in this case, in the form of scribbles, to segment the objects of interest.

In the first interaction of each sequence of frames, the server chooses a particular frame and provides a human-annotated scribble for each object in this frame. Based on these scribbles, the user's model has to predict segmentation masks of those objects for all the frames. After that, the user submits the predicted masks to the server. In each of the subsequent interactions, from a list of frames specified by the user, the server chooses the frame with the worst prediction and provides a new set of human-simulated scribbles in this frame. These scribbles point out the false positive and false negative regions. The user's model uses these human-simulated scribbles to refine its previous prediction masks. This procedure is repeated until a maximum number of interactions or a timeout is reached. The server measures the time needed to perform each interaction, these timings are combined to compute the final results [1].

Methods for tackling interactive video object segmentation challenge have to satisfy multiple goals, including being fast, generating a good initial result, and improving accuracy after interactions.

Our method view scribbles as a set of points. A Control-point-based Scribbles-to-Mask module is used to yield the mask of the scribbled frame from this set. The propagation phase usually is the bottle-neck on the runtime of video object segmentation methods. By using a memory-based model, our method can be fast, and the result improves by interactions. We use this memory-based model in the Self-Reference Refinement module for the mask achieved from the Control-point-based Scribbles-to-Mask module. To improve the memory pool of memory-based model to have a better reference, we propose a strategy for enriching the memory pool with reliable frames and Multiple Reference Views. We also implement a Guided Mask Propagation strategy to eliminate complex background for performance improvement.

The remainder of this paper is organized as follows. We show related work in Section 2. Our proposed methods are presented in Section 3. Experimental results are reported in Section 4. Finally, Section 5 concludes and paves the way for future work.

2. Related Work

In the Interactive Scenario of The DAVIS Challenge 2019, proposed methods from participants achieved promising results. The 1st place - Oh et al. [4], proposed

*Corresponding author. Email: tmtriet@fit.hcmus.edu.vn

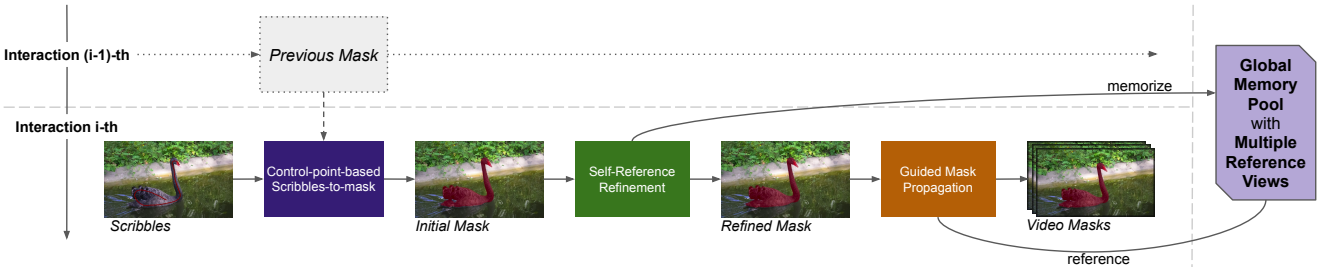

Figure 1. Overview of proposed method.

the Space-Time Memory Networks, which uses scribbled frames to form an external memory. This method is fast and shows the potential to improve the result after interactions. The 2nd place - Heo et al. [2], proposed a method of using two sparse-to-dense networks to yield the mask of scribbled frames and propagate points.

We implemented and extended the work of the following modules. For each object, f-BRS [3] takes into account two sets of points: points that belong to that object as positive points and points that do not belong to that object as negative points, to yield the mask. Space-Time Memory Networks [4] is fast, efficient, and can be extended with an enrichment strategy for the memory pool. To focus on the regions of interest of each object, a guided propagation strategy is promising to improve the result is suggested by Tran et al. [6].

3. Proposed Method

3.1. Pipeline

Figure 1 illustrates our proposed method. For each interaction, our method has two stages: interactive image object segmentation and video object mask propagation.

In the first stage, we develop a Control-point-based Scribbles-to-Mask module that converts input scribbles path into an initially predicted mask. After that, we apply on this initial mask the Self-Reference Refinement strategy, which is based on a memory-based model.

In the second stage, we propagate the self-refined mask using a memory-based model. This memory-based model utilizes a memory pool, called Global Memory Pool, which is filled with all frames having scribbles in current and previous interactions. Furthermore, we propose to enrich the Global Memory Pool with the Reliably Inferred Frames, obtained in the propagation process. In this way, we have more high-quality referenced views for the objects of interest. All frames are only influenced by its nearest reliable frame. We use a Guided Mask Propagation strategy to focus on potential regions of interest in a frame, as suggested in [6].

3.2. Control-point-based Scribbles-to-Mask

The input scribbles are first converted into control points using a Control-points Extractor. These control points are used by a Control-points-to-Mask module to generate a complete multi-object segmentation mask. If it is not the first interaction, a previously inferred mask is available and combined with this mask to generate the Initial Mask necessary for the next steps. Figure 2 shows an example of how it works.

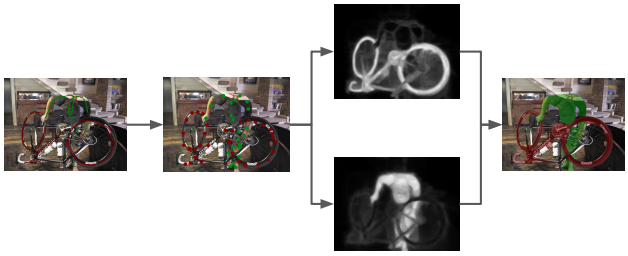

Figure 2. An example of Control-points-to-Mask.

Control-points Extractor To extract control points from a scribble path, we evenly sample points along that path. Each of these control points holds the information about its coordinates on the image and the object to which it belongs.

Control-points-to-Mask From the generated control points, we generate a probability mask for each object using a backpropagating refinement algorithm, called f-BRS [3]. This algorithm takes as input two sets of points, a positive set to localise the object position in the image, and a negative set to discriminate the object from other parts of the image. We consider control points belonging to that object positive points and control points belonging to other objects, often including the background class, as negative points. All the generated probability masks can be merged to generate the image object mask.

3.3. Global Memory Pool with Multiple Reference Views

The core algorithm used in the remaining components are based on a memory-based video object segmentation model. Memory-based models maintain a memory pool that

can be filled with frames and their corresponding masks. They perform segmentation on a frame with reference to that memory pool. We use such models in two different ways: as a refinement mechanism in the Self-Reference Refinement module and as a mask propagator in the Guided Mask Propagation module. In our work, we utilize the Space-Time Memory Networks [4] as the memory-based model.

For each sequence, our method maintains a memory pool that is global across interactions, called Global Memory Pool.

Naturally, our method relies heavily on the quality of the memory fragments stored in the Global Memory Pool. The most important fragments are the scribbled frames obtained after each interaction. Another set of frames to consider are the Reliably Inferred Frames: the predicted masks that are considered reliable obtained throughout the propagation process. These frames go through a filter to delete non-informative frames such as blurry, deformation frames, which are eventually enriched to improve memory pool quality.

3.4. Self-Reference Refinement

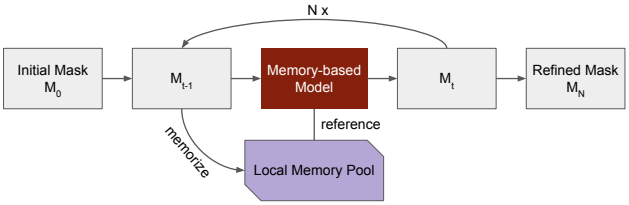

Figure 3. Self-Reference Refinement module.

As a refinement mechanism, the memory-based model maintains a Local Memory Pool that memorizes all previously inferred masks and uses that memory pool to perform segmentation on the same frame again (hence the name: Self-Reference). It is also possible to initialize the Local Memory Pool with the aforementioned Global Memory Pool. Figure 3 illustrates the Self-Refinement module.

Figure 4 shows examples where the Self-Refinement module helps refine the prediction.

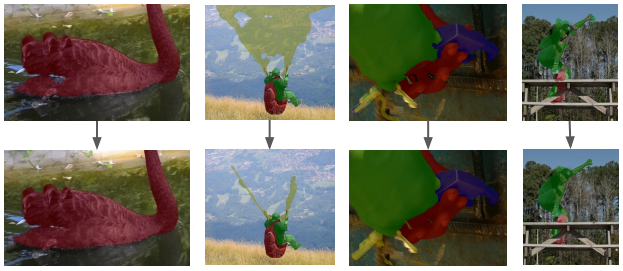

Figure 4. Cases that Self-Reference Refinement module helps improve the result: under-segmentation, over-segmentation.

3.5. Guided Mask Propagation

In each interaction, with the Refined Frame as the initiator, the propagation goes in two directions to both ends of the video sequence.

However, the process does not go all the way but instead at length calculated by considering the indices of the annotated frames in previous interactions. As can be seen from Figure 5, it stops midway between the current annotated frame index and the nearest previous annotated frame index. Considering the possible decay from the accumulated error of any propagation algorithm, this is a reasonable approach as each frame is only influenced by its nearest reliable frame.

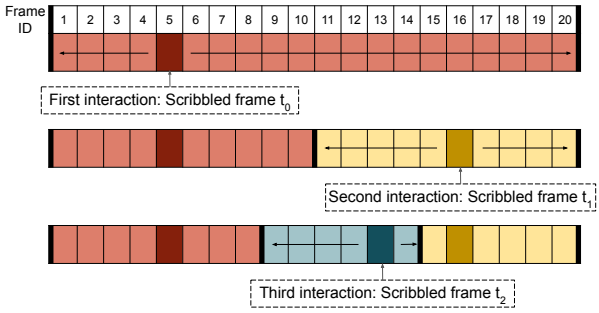

Figure 5. All frames are only influenced by its closest reliable frame.

To counter the problem of misidentification or over-segmentation, we employ a guidance mechanism in our propagation module. It uses the predicted masks of neighboring frames to generate a localisation mask. We use this mask to filter out potentially unrelated surroundings so that the segmentation algorithm can have a better focus on what to segment. Figure 6 shows an example where Guided Mask Propagation module focuses on regions of interest in a frame.

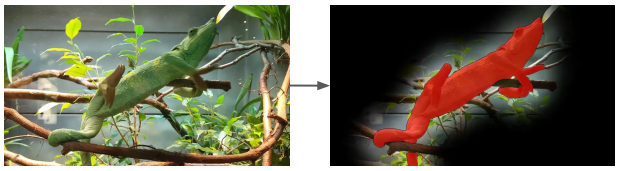

Figure 6. Region of Interest in a frame on sequence “chameleon”.

4. Experimental Results

For each sequence, more interactions mean more frames and their corresponding masks for the memory-based model to memorize, therefore we can expect the $\mathcal{J}\&\mathcal{F}$ score of our method to increase progressively after each interaction. Figure 8 shows our proposed method improves the $\mathcal{J}\&\mathcal{F}$ [5] score of sequences after interactions. About runtime, the method takes 25.72 seconds on average per interaction.

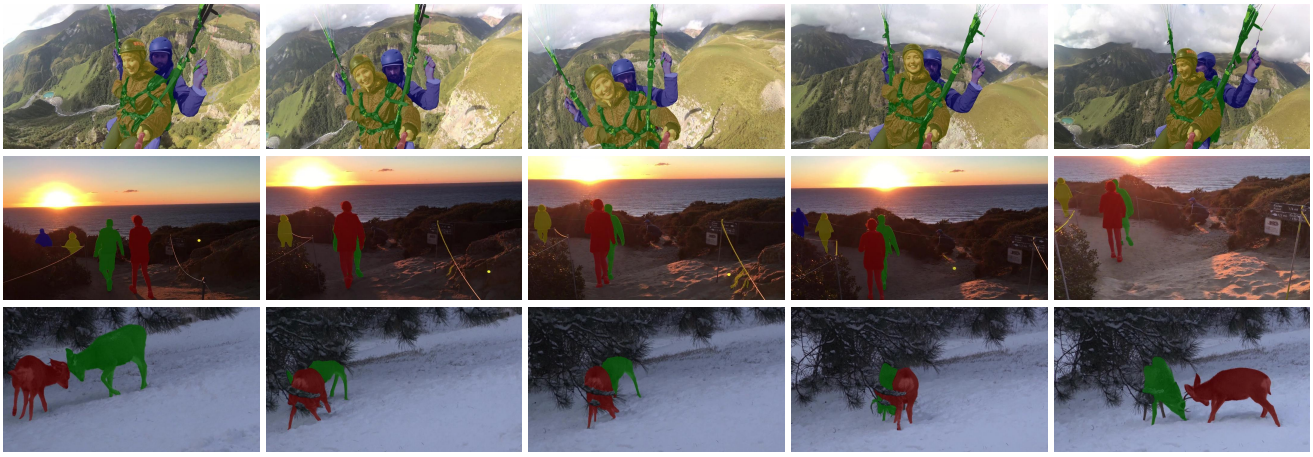

Figure 7. Final masks after the last interaction. Name of sequences from the first to the last row: "tandem", "people-sunset", and "deer".

Figure 7 shows qualitative results of our proposed method. As shown in Table 1, our method achieves 0.767 and 0.759 in terms of Area Under the Curve (AUC) and $\mathcal{J}\&\mathcal{F}$ at 60 seconds ($\mathcal{J}\&\mathcal{F}$@60s), respectively, ranking 2^{nd} in Interactive Scenario of DAVIS Challenge 2020.

Table 1. Result on the DAVIS 2020 test-dev dataset.

Position	Participant	AUC	$\mathcal{J}\&\mathcal{F}$@60s
1	Yuk-Heo	0.772	0.787
2	**Ours**	**0.767**	**0.759**
3	ChenLIANG	0.754	0.762
4	Thanh-An Nguyen	0.468	0.473

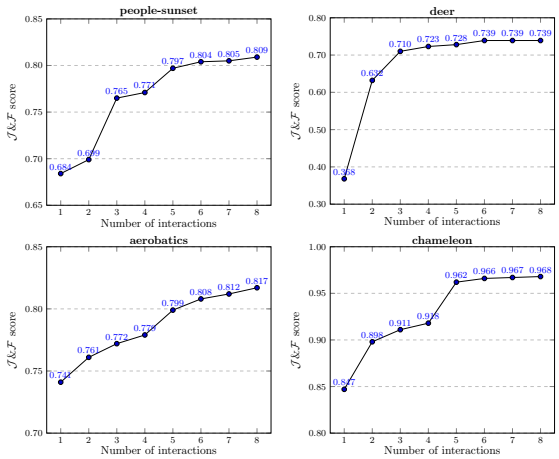

Figure 8. $\mathcal{J}\&\mathcal{F}$ score improving after interactions.

5. Conclusion

In this paper, we proposed a fast and efficient method for interactive video object segmentation that utilizes three key modules: Control-point-based Scribbles-to-Mask, Self-reference Refinement, and Guided Mask Propagation. The strong performance of our method is proven by the promising performance, both qualitatively and quantitatively, from the result of this year's challenge.

We propose some directions for further research. First, the scribbles returned from the server from the second interaction onward can be better leveraged to refine previous predictions. Second, we need a better system to evaluate and decide which frames to be added to the memory pool. Third, mask propagation should be guided by a better, flow-based tracking method.

Acknowledgements

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We also thank AIOZ Pte Ltd and the research group of Prof. Minh Do at University of Illinois at Urbana Champaign for supporting our research team with computing infrastructure.

References

[1] 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:1803.00557*, 2018.

[2] Y. Heo, Y. J. Koh, and C. Kim. Interactive video object segmentation using sparse-to-dense networks. *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2019.

[3] O. B. A. K. Konstantin Sofiiuk, Ilia Petrov. f-brs: Rethinking backpropagating refinement for interactive segmentation. *arXiv preprint arXiv:2001.10331*, 2020.

[4] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[6] M. Tran, T. Le, T. V. Nguyen, T. Ton, T. Hoang, N. Bui, T. Do, Q. Luong, V. Nguyen, D. A. Duong, and M. N. Do. Guided instance segmentation framework for semi-supervised video instance segmentation. *The 2019 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2019.