Video Object Segmentation with Memory Augmentation and Multi-Pass Approach

The-Anh Vu-Le[1,3], Hong-Hanh Nguyen-Le[1,3], E-Ro Nguyen[1,3], Minh N. Do[4], 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
[4]University of Illinois at Urbana-Champaign, U.S.A.

Abstract

We propose to leverage the existing memory-based models and enhance their capability in the task of human-guided video object segmentation by adding pre-processing and post-processing steps. Specifically, for the pre-processing stage, we suggest (1) to reinforce the memory using examples generated from the transformation of the first frame and its ground-truth segmentation, in complement to intermediate frames and their unreliable predicted segmentation; and (2) to generate a localisation mask for better object tracking and use it to guide the segmentation process by masking potentially unrelated background. For the post-processing stage, we suggest to redo the segmentation process on the same image using the predicted mask from the previous pass as a way to refine previous predictions; this can be done in both ways: after the prediction of a single frame or after the prediction of the entire video. The proposed additional steps achieved 73.3 and 78.7 for region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) respectively on the DAVIS 2017 challenge dataset, ranking 6th in the Semisupervised Track of The 2020 DAVIS Challenge.

1. Introduction

Semi-supervised video object segmentation (or human-guided video object segmentation) is the task of providing segmentation of the entire video based on the segmentation provided for a reference frame, usually the first frame of that video. The annotated segmentation of the reference frame will specify the objects of interest and encode identification information. It is required to keep both of them consistent throughout the entire video. [1, 5].

To approach this problem, we put our focus on existing memory-based models. Upon inspection of a particular model of this family, we found that it is possible to make modifications to the inference process to solve several problems. The first problem is the lack of reliable segmentation masks to fill the memory pool. We propose to solve it by a Memory Augmentation module, which produces a transformed frame from the given annotated reference frame. The second problem is the localisation of objects of interest. We propose a solution that uses a tracking-based segmentation method to produce a coarse segmentation which serves as a localisation mask to obscure unnecessary background.

We also observed that it is possible to use this reference-based segmentation method as a refinement module. Therefore, we propose the Multi-pass Approach, in which we perform segmentation multiple times with reference to previous predictions of the same frame or other frames.

The outline of this report is as follows. Section 2 reviews different approaches for video object segmentation. We propose additional modules to improve the performance of the model in Section 3. Section 4 showcases qualitative results and the result in the DAVIS 2020 Challenge. Section 5 identifies possible directions for future work.

2. Related Work

Existing methods to approach this problem come in different styles. One approach is the segmentation of individual frames with a ranking-based tracking system. PReMVOS [3] is an example of this method, where it uses image detection and image segmentation models to perform segmentation on frames individually before merging them by ranking relatedness of the predicted objects to objects predicted in previous frames. Another approach is to consider the annotated frame and previously predicted frames as references to perform segmentation on a query frame. Space-Time Memory Networks (STM) [4] and FEELVOS

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

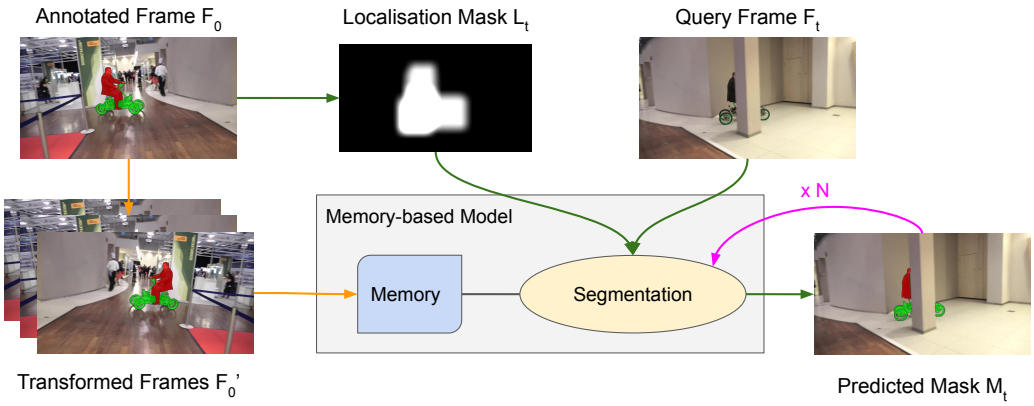

Figure 1: Overview of the proposed method. The orange arrows denote the pipeline of Memory Augmentation, the green arrows denote the pipeline of Guided Segmentation, and the pink arrows denote the Multi-pass Approach.

[7] are examples of this kind of method. While FEELVOS specifically uses only the annotated frame and the previous frame, STM uses as many frames as possible as references, limited by the resource required to store all previous frames.

Several other works pay attention to the localisation of objects of interest. SiamMask [9] takes a bounding box as input to specify an object of interest, then tracks and segments it from subsequent frames. BoLTVOS [8] extends Mask-RCNN by using the bounding box from the reference frame as an additional input to the bounding box regression module, which helps to achieve better localisation of the object of interest. Guided Instance Segmentation [6] proposes a forward and backward guidance method to hides unrelated background, which confuses the inference process.

Other methods focus on different aspects. BubbleNet [2] proposed a method to select better frames to provide human annotation, instead of the usual first frame.

3. Proposed Method

3.1. Pipeline

Figure 1 shows an overview of our proposed method. At the core is a memory-based segmentation model. Its mechanism is to maintain a memory which stores frames and their corresponding masks, whether it is the reliably annotated mask (for the first frame) or the predicted mask (for other frames). It then can segment any input image with reference to the aforementioned memory pool. The model of our choice is the Space-Time Memory Networks [4].

For each sequence, we use the annotated reference frame in two ways. First, it is used in the Memory Augmentation module to enrich the memory. Second, it is used by a tracking algorithm to generate localisation masks in all the subsequent frames, which are used in the Guided Segmentation module. We use SiamMask [9] for tracking.

For each query frame, the segmentation model will utilize the generated localisation mask to perform guided segmentation and produce an initially predicted mask. This initial mask can be further refined by the multi-pass approach to generate the final predicted mask.

3.2. Memory Augmentation

With memory-based models, it is crucial to supplement the memory pool with reliable frames. However, we only have one annotated frame for each video. All other frames are just predictions, making it more prone to errors, and these errors are accumulated as we segment to the end of the sequence.

To counter this problem, taken inspiration from the test-time augmentation method, we reinforce the memory pool with reliable annotations that are the transformed first frame. The set of possible transformations includes flipping, rotation, blurring, etc. This way, the memory pool can be reliable and, at the same time, keep its diversity. Figure 2 shows examples of memory augmentation.

3.3. Guided Segmentation

Another improvement we suggest is the guided segmentation, inspired by [6]. A problem with memory-based models is that it predicts instances on the pixel level. Besides,

Figure 2: Examples of memory augmentation using transformations. On the left side of the arrow is the original image, on the right side is the transformed image by horizontal flipping (left) and motion blurring (right).

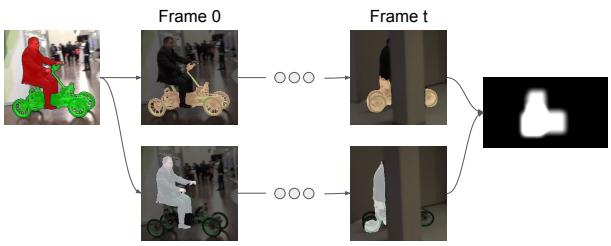

Figure 3: Procedure of guided segmentation.

although taking previous frames in memory into account, it is not tracking-based, which means there is no enforcement on the structure of the object and the temporal continuity.

We propose to counter this problem by utilizing other tracking methods to localize the objects and mask out potentially unrelated background.

The procedure to produce the localisation masks is shown in Figure 3 and is as follows: the mask of the reference frame is split into individual masks, each for an object in the frame. Each of the mask is then independently fed into the tracking algorithm to produce tracking masks of that object for the entire video. Each of the resulting mask is then dilated and Gaussian blurred out to create a saliency mask. We merge these individual masks into a single localisation mask for all objects.

As the tracking algorithm is also another predictive algorithm, it can make errors that negatively affect the result, for example, by masking out objects of interest in the frame. We can mitigate this problem by a heuristic: through the comparison between two consecutive frames, we can determine the reliability of the tracking process to that point and decide whether or not to continue using the produced localisation mask.

For instance, if the current mask does not overlap with the previous mask, the tracking is assume to have failed and that the algorithm has switched to tracking other objects. In this case, we propose to discard all masks from that point on. Figure 4 shows a visualisation of this approach.

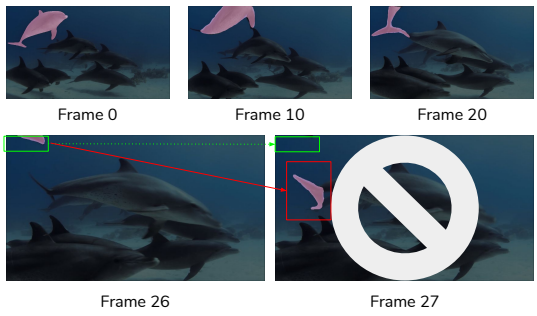

Figure 4: Heuristic to minimize tracking error.

3.4. Multi-pass approach

Our final suggestion is based on the observation that the memory-based model can also serve as a refinement module. If it memorizes the prediction to re-perform segmentation on the same frame, it basically refines that prediction. In actuality, we conjecture that this helps push the probability in the same direction as previously predicted, making it possible to fix under- and over-segmentation.

We also observe that the original inference process is one-directional with one forward pass, and memory is accumulated in the progress. However, memory-based models, specifically those that use attention-based methods, do not assume the order or temporal consistency. Therefore, it is possible to reference to both past and future frames. Thus another pass is necessary to accumulate future frames.

In conclusion, there are two different interpretations of the multi-pass approach: to use it either after the prediction of each frame, or after the entire video has been processed. In the first interpretation, the frame to be segmented is refined by referencing to its previous predictions. In the latter interpretation, the frame in question is segmented with reference to both the prediction of past and future frames.

4. Experimental Results

Table 1 shows **DAVIS 2020 Challenge Leaderboard**. We achieve 73.3 and 78.7 for region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) respectively on the DAVIS 2017 challenge dataset, ranking 6th in the Semisupervised Track of The 2020 DAVIS Challenge.

To demonstrate the benefits of our proposed techniques to enhance the segmentation quality, we compare the results between using (right column) and not using (left column) Figure 6 demonstrates the comparison between results obtained with and without applying certain techniques. From top to bottom, we show that (1) Memory Augmentation can help avoid avoid referencing from a highly unreliable frame, such as the misidentified turtle early into the videos

Table 1: DAVIS 2020 Challenge Leaderboard

Measure	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
AMIND	**84.1**	**81.5**	**86.7**
ReLER	83.8	81.1	86.5
Hongje	79.5	77.0	82.1
HCMUS-UD-NII-UIUC	79.3	76.5	82.1
DSVOS	76.9	74.4	79.5
Vltanh (ours)	**76.0**	**73.3**	**78.7**
JingshanXu	75.0	72.2	77.7
Bytedance	72.2	69.8	74.6
Mingmingdiii	69.9	67.6	72.2
DeepDream	64.9	62.5	67.3
Birdman	50.6	47.5	53.8

Figure 5: Examples of the final result. From top to bottom: sampled frames from sequences `speed-kating`, `juggle`, `running`.

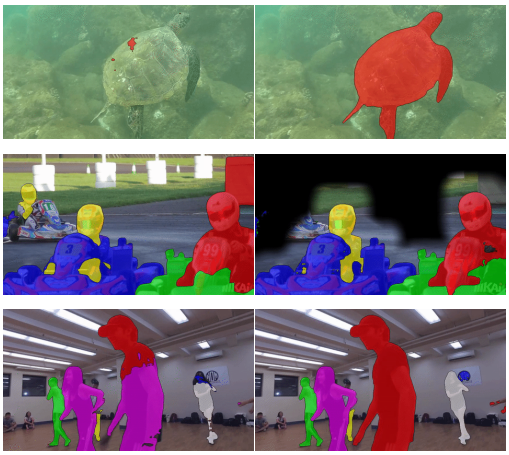

Figure 6: Result comparison between using (right) and not using (left) memory augmentation (row 1), guided segmentation (row 2), and multi-pass approach (row 3).

(2) Guided segmentation can help in prevent the model from mis-identifying the block as the man in the top-right corner because the tracking algorithm determines those regions as unrelated to the objects of interest (3) the multi-pass approach corrects the identification of the person on the right.

5. Conclusion

In conclusion, we present an extension to memory-based models by three main ideas: memory augmentation, guided segmentation, and multi-pass approach. For future directions, we suggest looking into developing better strategies for memory augmentation and finding a way to incorporate the idea of guidance into the architecture design in an end-to-end fashion.

Acknowledgements

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd 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] B. A. Griffin and J. J. Corso. Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] J. Luiten, P. Voigtlaender, and B. Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 2018.

[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.-T. Tran, T.-N. Le, T. V. Nguyen, V. Ton-That, T.-H. Hoang, N.-M. Bui, T.-L. Do, Q.-A. Luong, V.-T. Nguyen, D. A. Duong, et al. Guided instance segmentation framework for semi-supervised video instance segmentation. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[7] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] P. Voigtlaender, J. Luiten, and B. Leibe. Boltvos: Box-level tracking for video object segmentation. *CoRR*, abs/1904.04552, 2019.

[9] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.