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

Lazy Instance Segmentation Framework with Fine-tuned Bidirectional Propagation for Interactive Video Segmentation

Thanh-An Nguyen *[1,4], Hai-Dang Nguyen[1,4], Khanh-An C.Quan[2,4], and Minh-Triet Tran[1,3,4]

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

Abstract

Video Instance Segmentation is a hot topic in the computer vision research community in recent years. In this paper, we propose a Lazy Instance Segmentation (LIS) framework with Fine-tuned Bidirectional Propagation (FBP) for improving interactive video instance segmentation on the aspect of processing speed. Our proposed pipeline comprises three state-of-the-art deep neural networks used for different purposes, including instance segmentation, object tracking, and object boundary refinement. Strikingly, the "laziness" of our framework is expressed through a proactive-waiting strategy to gather information from the user. From that, we can leverage parallel computation to process parts of a video sequence simultaneously and efficiently. Our proposed method achieves 0.468 and 0.473 in terms of area under the curve (AUC) and $\mathcal{J}\&\mathcal{F}$ value at the 60th second ($\mathcal{J}\&\mathcal{F}$@60s), respectively on the interactive DAVIS 2020 dataset, rank 4th in the challenge. The experimental results indicate that there are some potential directions for further improvement in terms of accuracy.

1. Introduction

Thanks to the development of computing resources, machine learning, and deep learning are widely applied in different branches in computer vision. Video Instance Segmentation (VIS) is one of the well-known topics in this field, in which massive experiments with deep learning methods are being performed by top scientists around the world.

VIS aims to classify each pixel in a single video frame to a particular instance or background and assign a consistent ID to the instance over the video sequence. It is potential for a wide range of practical applications, for instance, autonomous vehicle [1], action recognition [3], video summarization [6], object tracking [9], and scene understanding [11].

In general, the VIS system alone without any support from human has certain challenges and limitations. Interactive Video Instance Segmentation (IVIS) involves human-in-the-loop to provide feedback for the VIS system, which helps constantly improving the performance of the segmentation process.

In this work, we propose Lazy Instance Segmentation (LIS) to tackle the challenging problem of Interactive Video Instance Segmentation, which targets certain instances and segments them thanks to given scribbles. Our proposed framework is constructed on top of the three following subprocesses.

- Feature back-propagating refinement scheme using f-BRS [5]
- Fast online object tracking and segmentation with SiamMask [9]
- Instance segmentation refinement with CascadePSP [2]

In the following sections, we discuss the related works, our proposed method, the results, and the conclusion. Particularly, Section 2 mentions the related works that we refer to. Sections 3 describes our framework and its implementation in detail. The experimental results are reported and discussed in Section 4. Finally, Section 5 presents our comments on performance in conjunction with limitations, and we discuss some potential directions in the future.

*Corresponding author. Email: ntan@selab.hcmus.edu.vn

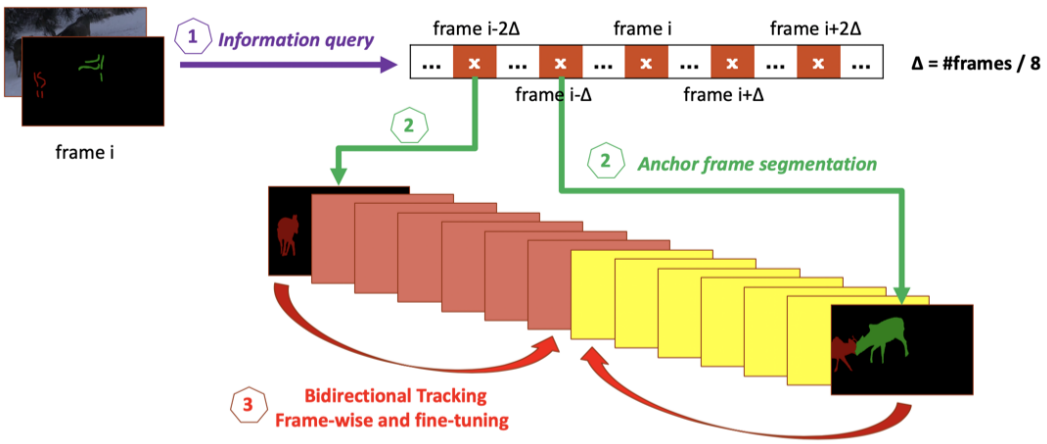

Figure 1. Overview of Lazy Instance Segmentation (LIS) pipeline.

2. Related Works

First, the approach used to segment instances in a particular frame is a feature back-propagating refinement scheme (f-BRS) [5]. This is an interesting method for interactive instance segmentation, published by Samsung AI Center in Moscow in 2020. For a single image, the users provide clicks, including positive on the instance of interest and negative clicks for the false-positive region (the click points are considered as control points). This method segments the instance of interest based on these control points.

SiamMask [10] is the second sub-process in our pipeline. It is the state-of-the-art approach for object tracking and segmentation in 2019, which is best known for efficiency in terms of speed. From an initial bounding box of an object, the dubbed SiamMask model tracks the object frame by frame in the sequence, and it also segments the object. The main idea of this method is the fully-convolutional Siamese network.

The last related work that we employed to fine-tune our results is the Cascade-PSP [2]. It is a novel approach that helps refines and corrects local object boundaries whenever possible. Besides, Point-Rend [4] from Facebook AI Research is also a method worth being applied to produce accurate boundaries.

3. Lazy Video Instance Segmentation with Fine-tuned Bidirectional Propagation

3.1. Overview

In the context of this interactive track, we propose a Lazy Instance Segmentation (LIS) framework with Fine-tuned Bidirectional Propagation (FBP). Figure 1 illustrates the three main modules in our framework. The first module named Information query, in which we propose a way to pre-process the sequence by identifying the anchor frames and request information (the annotations) for the anchor frames. Second, our framework perform instance segmentation phase for these anchor frames to get the masks. Finally, the segmented masks are bidirectionally propagated (backward and forward) on the sequence. The refinement process is considered as part of the third module. With these modules, the instances in the sequences are segmented.

Figure 2. The visualization results of our framework after applying CascadePSP for refinement. Imperfect segmentation masks (top row) and refined masks (bottom row).

3.2. Implementation

First, our framework proactively waits for sufficient information from the first seven interactions, by sending requests to the server seven anchor frames. Equation 1 indicates the set of frame indices F that are chosen as anchor frames, where n represents the total number of frames in the current sequence, and k is the maximum number of interaction.

$$F = \{\frac{n \times i}{k} | i \in \mathbb{Z} \text{ and } 0 < i \leq k\} \quad (1)$$

Table 1. Ranking results in the Interactive DAVIS 2020 Challenge. Our results are marked in **boldfaced blue**.

#	Participants (Organization)	AUC	$\mathcal{J}\&\mathcal{F}$@60s
1	Yuk Heo (Korea University)	77.2	78.7
2	Quoc-Cuong TRAN (University of Science, VNU-HCM)	76.7	75.9
3	Chen Liang (Tongji University)	75.4	76.2
4	Nguyen Thanh An (University of Science, VNU-HCM)	46.8	47.3

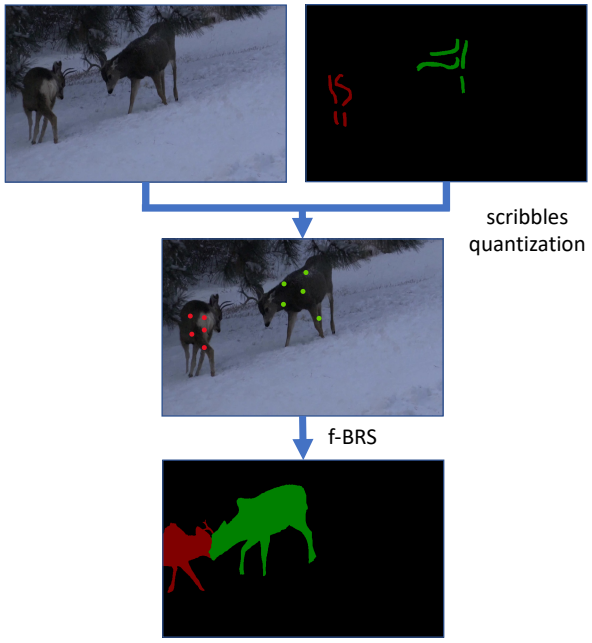

Figure 3. The process of instance segmentation at each anchor frame.

After gathering enough information on the eight scribbles, our LIS framework starts to execute simultaneously at the eight annotated frames of the current sequence. This idea benefits us substantially on the execution time and reduces the consequences of losing tracking. It means that instead of performing bidirectional tracking from a single anchor frame, in which the bidirectional tracking prone to fail at frames far from the anchor frame, we require more anchor frames so that the tracking distance is much less. This "lazy" strategy leads to a significant improvement in our framework during the object tracking phase.

At each of the anchor frames, we segment the instances from the given scribbles using f-BRS. In particular, we quantize the scribbles of each instance into control-points. The f-BRS model considers the instance of interest as positive, while the other instances are considered as negative, which are used to reduce false positive mask. This phase is illustrated in Figure 3.

After having the segmented masks at anchor frames, we identify the region of interest of that instance by padding the size of the segmented mask. At this stage, SiamMask is used for bi-directional tracking.

In the last step, we fine-tune the boundary segmented masks by using Cascade-PSP. The refinement module takes a single frame with its corresponding imperfect segmentation mask to produce a refined mask. Figure 2 illustrates some samples of the result of this step. This process is performed throughout the whole sequence, from the beginning to the end. We realize that this process becomes the bottleneck of our framework. However, it could be integrated to SiamMask tracker to process parts of a sequence simultaneously.

4. Results On Interactive DAVIS 2020 Challenge

We evaluate our proposed method on the DAVIS2017 dataset [8]. The result is measured by two factors, which are Area Under The Curve (AUC) and $\mathcal{J}\&\mathcal{F}$ at the 60th second ($\mathcal{J}\&\mathcal{F}$@60s). By applying the CascadePSP refinement process, our result is significantly improved in both metrics of AUC and $\mathcal{J}\&\mathcal{F}$@60s. In the end of the contest, we achieve 0.468 and 0.473 in terms of AUC and $\mathcal{J}\&\mathcal{F}$@60s, respectively on the interactive DAVIS 2020 dataset, rank 4th in the challenge (as shown in Table 1). Figure 4 shows some sample sequences output from our framework.

5. Conclusion

In conclusion, we introduce the novel LIS framework with fine-tuned bidirectional propagation for interactive instance segmentation in videos. In particular, our strategy consists of three different deep neural networks for different purposes including instance segmentation, object tracking, and object boundary refinement. Besides, although there are some limitations in accuracy, the proposed framework is optimal in terms of speed thanks to the "lazy" strategy with parallel processing. Throughout the experiments, our proposed framework achieved rank 4th in the DAVIS 2020 challenge. In the future, we will resolve the bottleneck of our method, especially in the tracking module. Furthermore, Space-Time Memory Networks (STM) [7] and Point-Rend [4] are two approaches that we consider to use for refinement in our framework to increase the accuracy.

Figure 4. Output masks from our framework. Name of sequences from the first to the last row: “gym”, “subway”, and “people-sunset”.

Acknowledgements

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We also gratefully acknowledge AIOZ Pte Ltd for the support of GPU and computing infrastructure.

References

[1] B. D. Brabandere, D. Neven, and L. V. Gool. Semantic instance segmentation for autonomous driving. In *CVPR Workshops*, 2017.

[2] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, 2020.

[3] J. Ji, S. Buch, A. Soto, and J. C. Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *ECCV*, 2018.

[4] A. Kirillov, Y. Wu, K. He, and R. Girshick. PointRend: Image segmentation as rendering. 2019.

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

[6] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 114(1), 2015.

[7] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9226–9235, 2019.

[8] 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 preprint arXiv:1704.00675*, 2017.

[9] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.

[10] 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.

[11] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019.