SiVOS: Simulated Interactive Video Object Segmentation

Tan-Cong Nguyen[1,4†], Gia-Han Diep[2,4†], Hung V. Tran[2,4], and Minh-Triet Tran[2,3,4*]

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

Abstract

We present a novel approach dealing with unsupervised video multi-object segmentation via automatically simulated interactions (SiVOS) with raw videos as the only input. This approach leverages previous work, but adding attention to objects tracking (both forward and backward) in parallel with simulated interactive segmentation for refining masks and clipping overlapped areas. Here, backward-tracking serves as additional information for forward-tracking to track objects which may be missed in the active tracking due to occlusion, speed, or deformation, causing the tracking flow of the same object decay into disjointed track-runs. And, the simulated interactions serve as a guide for refining the proposal masks and their shared regions. We evaluate our approach on the DAVIS 2020 Video Object Segmentation Benchmark and obtain the 4^{th} rank on the DAVIS 2020 Unsupervised Challenge, with the overall J&F Mean score of 43.9 having used only pretrained models.

1. Introduction

Image processing with video as input has always come with challenging problems as it often relates to multiple image processing techniques but with the former knowledge of the previous image frame. As for Video Object Segmentation (VOS) problem, it is involved with not only object detection, segmentation, and tracking, etc. but also the task of assigning consistent ID numbers for each object instance. VOS, in general, can be categorized into three tasks: semi-supervised, interactive, and unsupervised [9, 8, 1].

The semi-supervised task is to provide instance segmentation given a sequence of image frames and the ground truth mask for the first frame. Previously proposed frameworks dealing with this one-shot video segmentation are: the combined framework known as Proposal-generation, Refinement and Merging for Video Object Segmentation (PReMVOS) [6, 4, 5] or Guided Instance Segmentation (GIS) with bi-directional adaptive attention for each target instance [12].

The interactive task with the inputs are a set of scribbles for the first random frame requires the considered algorithm to segment the target objects (defined with the scribbles) for the remaining frames, allowing the algorithm to ask for correction of some unsure frames as this process of amendment continues until a predefined limitation (a number of times or a time limit [8]).

Among the algorithms aiming to solve this problem, Rethinking Backpropagating Refinement for Interactive Segmentation (f-BRS) successfully offers a framework for guided mask segmentation or refinement given sets of points as positive or negative seeds [11].

Inspired by the idea from GIS, f-BRS and Unsupervised Offline Video Object Segmentation and Tracking (UnOVOST) technique in dealing with overlapping regions [7], we decided to build a framework that set up the unsupervised VOS environment as a guided VOS environment and named it 'Simulated interactive Video Object Segmentation' or SiVOS.

The basic idea is to (1) obtain proposal masks with Mask Scoring RCNN [3] pretrained on COCO (common objects in context) dataset [2], refine masks and overlapping areas using f-BRS with seeds [11, 7], (2) build and merge track-runs (both forward and backward [10]) with Deeper and Wider Siamese Networks for Real-Time Visual Tracking (SiamDW) [13], (3) select, then refine the track-runs using PReMVOS guided with the previous frame [12].

The baseline for this approach is evaluated on the DAVIS benchmark dataset [8]. The remainder of this paper is to detail the proposed approach (Section 2) with experiment and discussion provided in Section 2.4 and 3, respectively.

*Corresponding author. Email: tmtriet@fit.hcmus.edu.vn
† These authors contributed equally.

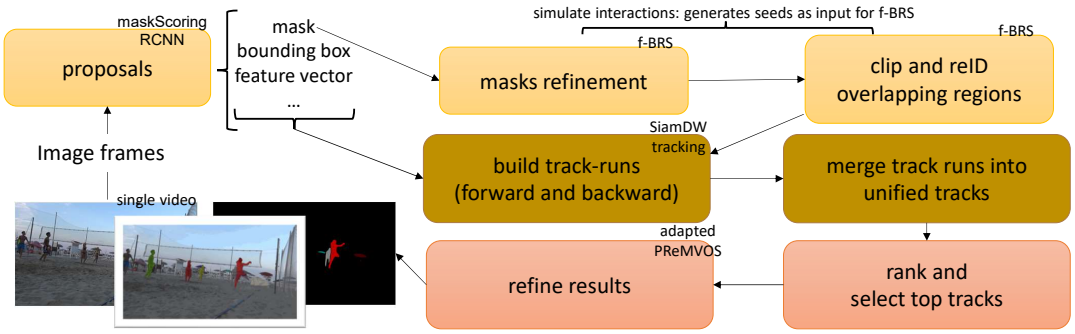

Figure 1. Overview of Simulated Interactive Video Object Segmentation (SiVOS) framework.

2. Proposed approach (SiVOS)

As shown in Figure 1, the proposed SiVOS framework composes of seven steps, grouped into three stages (with different colors). The first stage, generally, is to extract and refine proposal masks. We use a multi-instance object segmentation model to extract mask proposals together with other related information. Then we refine the proposed masks and clip the overlapping regions. The second stage is to build and merge track-runs. We use a visual tracking model to connect the instances of the same object across adjacent frames for the purpose of building track-runs. This work is performed in both forward and backward directions. Then a trajectory merging algorithm is applied to links the disjointed track-runs of the same object into a unified track. And lastly, we choose only the highest probability tracks and refine this output with guided-pretrained PReMVOS.

2.1. Mask proposal extraction and refinement

We use MaskScoring RCNN, pretrained on COCO, to extract mask proposals together with related information, such as bounding boxes, confident scores, and feature vectors in the region of interest (ROI) masking layer; These feature vectors are considered as ReID embedding vectors for mask proposals. We later refine the proposed masks as well as clip the overlapping regions and decide which masks they belong to (Figure 2).

In the segmentation mask refinement task, for each mask proposal in each frame, we refine the contours of the mask by using f-BRS interactive segmentation. Our algorithm calculates the distance map based on the mask contours. Then randomly select a number of low-value points in the distance map. With this mechanism, Our algorithm can detect some good central seed points of the mask contour and uses f-BRS to produce the refined foreground mask.

In clipping overlapping region task, for each pair of mask proposals have an overlap region. With the output mask of F-BRS given by the central seed points of the overlap region, the algorithm calculates the intersection over the union (IoU) score of the output mask and both mask proposals. The mask with higher IoU score holds the overlap regions, and the remaining mask is clipped.

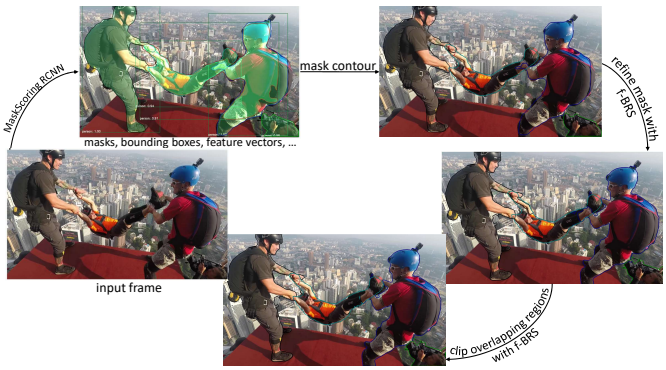

Figure 2. Stage 1: simulate interactions with f-BRS

2.2. Building and merging track-run

In building track-run task, we extend the usage of SiamDW, originally used for single object forward tracking, to multi-object segmentation combined with appearing (or re-appearing) object detection in both forward and backward directions. Specifically, SiamDW which is responsible for tracking a single object from one frame to the next. When the instance is being tracked in the current frame, SiamDW predicts the bounding box of that instance in the next frame. If the confidence score of the tracking exceeds

Figure 3. Stage 2: build and merge track-runs

a given threshold, then the mask proposal in the next frame has the highest IoU with predicted bbox is treated as the same instance and is contained the same track-run. For each instance on the next frame which cannot link to the current track-runs, a new track-run is initialized for this instance and starts the tracking process. The algorithm proceeds to build trace-runs in the forward direction from the first frame to the last frame and also do the same work in the backward direction (see the first and second rows of Figure 3).

For the next step, our algorithm merges track-runs of the same instance into a unified track. First, the algorithm (algorithm 1) pre-processes the track-runs into distinct track-runs. This task is done by eliminating duplicated track-runs, merging those with overlapping detected instances (ranking using the IoU of the instances), as illustrated in the third row of Figure 3, and clipping trace-runs that have intersection area but also contain two different instances at the same frame.

Second, because the objects may disappear and reappear during the tracking process, after obtaining the distinct track-runs at the pre-processing step, the algorithm proceeds to concatenate track-runs of the same instance together into a unified track. For example, in Figure 4, if the two track-runs are compatible in chronological order, each track-run computes the average feature vector based on their masks at each frame. Later, if the similarity score of their two feature vectors exceeds a given threshold, the two track-runs are then concatenated together.

Algorithm 1 Pre-processing the track-runs

// R_{fw} is the forward track-runs set in the sequence S
// R_{bw} is the backward track-runs set in the sequence S
// $r[t]$ means the instance mask in track-run r at frame t.
$R = R_{\text{fw}} \cup R_{\text{bw}}$
for track-run r in R **do**
  for track-run r' in $R \backslash \{r\}$ **do**
    if $r \cap r' \neq \emptyset$ **then**
      if $r \subset r'$ **then**
        $R = R \backslash \{r\}$ //remove r from R
      else if $\nexists t \in S(r[t] \neq r'[t])$ **then**
        $r' = r' \cup r$ //merge r and r'
        $R = R \backslash \{r\}$
      else
        $r_{\text{intersect}} = r' \cap r$
        $r_{\text{left}}, r_{\text{right}} \leftarrow$ split r by $r_{\text{intersect}}$
        ${r'}_{\text{left}}, {r'}_{\text{right}} \leftarrow$ split r' by $r_{\text{intersect}}$
        $r' = r_{\text{left}} \cup r_{\text{intersect}} \cup {r'}_{\text{right}}$
        $R = R \backslash \{r\}$
        $R = R \cup \{r_{\text{right}}, {r'}_{\text{left}}\}$
$ouput = R$ //return a set of the distinct track-runs

Figure 4. Merge track-runs: an example

2.3. Ranking track and refining with PReMVOS

The ranking is considered based on two criteria. The focus object is the one that often appears in the sequence. And each mask in the sequence should have the high confidence given by both the mask detector model and the tracking model. So we rank the track-runs with the probability scores calculated based on track length and the average confident score of masks in the track.

The ranking score of the track Tr in the sequence S is described in the equation 1 where l denotes the length of considered track or sequence, $conf$ denotes the confident score based on the mask detector via MaskScoring RCNN or the tracking algorithm using SiamDW, and m is the considered mask in the track Tr.

$$S_{Tr,S} = \frac{l_{Tr}}{l_S^2} \sum_{m}^{Tr} conf_{seg}(m) .conf_{track}(m) \qquad (1)$$

We choose only the highest probability tracks and refine the masks in this final output with guided PReMVOS on the proposal refinement task. For each instance in each video, we use the pretrained PReMVOS to refine the instance in frame t but also focus on its segmented result in frame $t-1$, as shown in Figure 5.

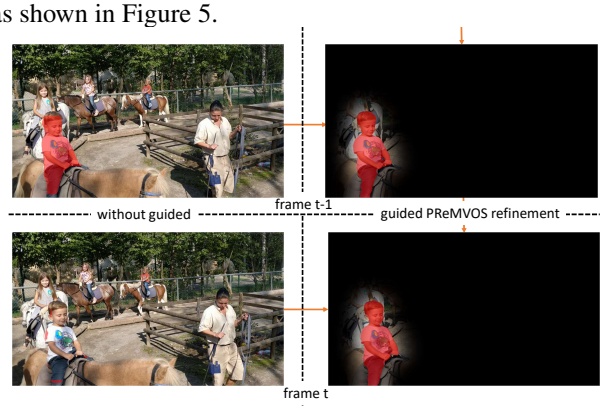

Figure 5. Stage 3: select track-runs and refine with PReMVOS

2.4. Experiment

We evaluate our proposed method on the DAVI 2020 Unsupervised benchmark dataset [1]. Figure 6 illustrates the first six-frame from three of our results. Our baseline result is also evaluated using $\mathcal{J}\&\mathcal{F}$ metric on the test-challenge set, for more details on these metrics can be found in [9]. Here, the result is obtained using the previously mentioned method thanks to the pretrained models provided by MaskScoring RCNN, SiamDW, f-BRS, and PReMVOS (table 1). This result gained the 4^{th} position with an overall $\mathcal{J}\&\mathcal{F}$ Mean score of 43.9.

Figure 6. Example results

Table 1. Ranking results in the DAVIS 2019 Challenge. The rankings in each categories are placed in parentheses. Our results are marked in **boldfaced blue**.

#	Team	Global	Region $\mathcal{J}$			Boundary $\mathcal{F}$		
		Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
1	Team Phoenix	61.6	58.4	65.0	-1.6	64.7	71.1	0.5
2	IIAI	55.6	53.1	60.0	-0.5	58.2	62.5	1.6
3	BLIIT	52.3	50.2	57.5	-5.0	54.4	58.9	-2.5
4	**Ours**	**43.9**	**40.2**	**45.7**	**-0.6**	**47.5**	**50.1**	**4.0**

3. Conclusion

This method (SiVOS) contributes two key ideas: improving object tracking by combining object tracking information in both directions and refining mask with simulated interactive segmentation. However, the method still struggles with several limitations: dealing with unknown objects (which are not provided in the pretrained models), tracking loss in some sudden frames, which may lead to failed reID for some cases. However, these ideas are up-and-coming for future development. We are currently training more concepts as an attachment for the pretrained models. Also, we are adding more attention to the guided refinement and overlapping mask clipping using f-BRS. One such solution is to provide more positive and negative seeds (from the proposal masks) for the framework.

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, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019.

[2] Caesar, Holger, J. Uijlings, Ferrari, and Vittorio. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[3] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang. Mask Scoring R-CNN. In *CVPR*, 2019.

[4] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, Refinement and Merging for the DAVIS Challenge on Video Object Segmentation 2018. *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2018.

[5] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, Refinement and Merging for the YouTube-VOS Challenge on Video Object Segmentation 2018. *The 1st Large-scale Video Object Segmentation Challenge - ECCV Workshops*, 2018.

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

[7] J. Luiten, I. E. Zulfikar, and B. Leibe. Unovost: Unsupervised offline video object segmentation and tracking, 2020.

[8] 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.

[9] 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.

[10] F. M. Porikli, X. Mei, and D. Brinkman. Method for tracking objects in videos using forward and backward tracking. 2010.

[11] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. *arXiv preprint arXiv:2001.10331*, 2020.

[12] 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.

[13] Z. Zhang and H. Peng. Deeper and wider siamese networks for real-time visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.