Multi-Referenced Guided Instance Segmentation Framework for Semi-supervised Video Instance Segmentation

Minh-Triet Tran *[1,2,3], Trung-Hieu Hoang[1,3], Tam V. Nguyen[4], Trung-Nghia Le[5], E-Ro Nguyen[1,3], Minh-Quan Le[1,3], Hoang-Phuc Nguyen-Dinh[1,3], Xuan-Nhat Hoang[1,3], and Minh N. Do[6]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]John von Neumann Institute, VNU-HCM, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]University of Dayton, U.S.A.
[5]National Institute of Informatics, Japan
[6]University of Illinois at Urbana-Champaign, U.S.A.

Abstract

In this paper, we propose a novel Multi-Referenced Guided Instance Segmentation (MR-GIS) framework for the challenging problem of semi-supervised video instance segmentation. Our proposed method consists two passes of segmentation with mask guidance. First, we quickly propagate an initial mask to all frames in a sequence to create an initial segmentation result of the instance. Second, we re-propagate masks with reference to multiple extra samples. We put high confidence reliable frames in the memory pool for reference, namely Reliable Extra Samples. To enhance the consistency of instance masks across frames, we search for mask anomaly in consecutive frames and correct them. Our proposed MR-GIS achieves 76.5, 82.1, and 79.3 in terms of region similarity (J), contour accuracy (F), and global score, respectively, on DAVIS 2020 Challenge dataset, rank 4th in the challenge on semi-supervised task.

1. Introduction

Video instance segmentation aims to label each video frame pixel to instances or the background region, and then assign consistent IDs to these instances over the video sequence. Instance segmentation in videos is beneficial in a wide range of practical applications, *i.e*., autonomous vehicle [1, 6], action recognition [3], video summarization [7], object tracking [13], and scene understanding [4, 14].

In this paper, we propose a novel Multi-Referenced Guided Instance Segmentation (MR-GIS) Framework for the challenging problem of semi-supervised video instance segmentation, which targets certain objects whose ground-truth mask for the first video frame is given. Our proposed method consists of two key ideas as below.

In literature, Nguyen *et al.* [8, 9] proposed *"you should look twice"* in the task of object detection. Along with this key idea, we propose two-pass guided segmentation, *i.e.*, mask propagation based on single-source first for coarse segmentation and then multi-source for fine segmentation.

Particularly, we propose to propagate masks with reference to multiple Reliable Extra Samples. We put all high confidence reliable frames in the memory pool for reference. Each reference frame influences the results of the frames in its neighbors, and each frame usually depends on its nearest reference frame.

To reduce the ambiguity between masks of different instances, we create a guided region [11] in each frame by expanding the segmented mask in a previous frame with reference to the current motion of the instance. If the instance is not segmented in the guided region, we restart object detection and re-identification in that frame and its succeeding frames.

Second, to enhance the consistency of instance masks across frames, we find mask anomaly in consecutive frames. An anomaly occurs when a mask accidentally disappears or re-appears in a short period, or its size and shape change significantly. To handle an anomaly case, we use both forward and backward mask propagation to restore missing segments, and to correct sudden changes in mask size and shape.

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

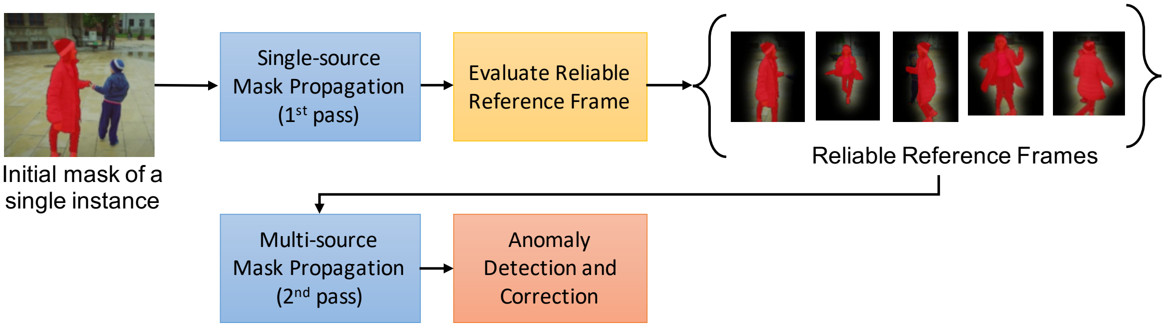
Initial mask of a single instance
Single-source Mask Propagation (1st pass)
Evaluate Reliable Reference Frame
Reliable Reference Frames
Multi-source Mask Propagation (2nd pass)
Anomaly Detection and Correction

Figure 1. Overview of Multi-Referenced Guided Instance Segmentation (MR-GIS) framework.

Our proposed MR-GIS achieves 76.5, 82.1, and 79.3 in terms of region similarity (J), contour accuracy (F), and global score, respectively, on DAVIS 2020 Challenge dataset, rank 4th in the challenge on semi-supervised task. We remark that among submissions, our method is the most stable because our method achieved the best scores in Decay of all metrics. Furthermore, our proposed MR-GIS further improves the original GIS framework [11] up to 4% in global score.

The remainder of this paper is organized as follows. Our proposed methods are presented in Section 2. Experimental results are then reported and discussed in Section 3. Finally, Section 4 draws the conclusion and paves the way for future work.

2. Multi-Referenced Guided Instance Segmentation

2.1. Overview

In this paper, we propose a novel Multi-Referenced Guided Instance Segmentation (MR-GIS) Framework for the challenging problem of semi-supervised video instance segmentation. In general, we process each instance independently, then heuristic merge all instance masks similarly to [5, 12, 11], such as from far to near based on estimated depth with high priority for rare/tiny objects.

Figure 1 shows the pipeline of our proposed Multi-Referenced Guided Instance Segmentation (MR-GIS) framework for each instance in the video. We utilize the structure of each object, together with its movement flow and deformable transformation, to create better segmentation for that instance.

In particular, there are three main steps in the process of each instance. First, we quickly propagate an initial mask to all frames in a sequence to create an initial segmentation result of the instance. Next, we re-propagate masks with reference to multiple extra samples. Finally, to enhance the consistency of instance masks across frames, we find mask anomaly in consecutive frames and correct them.

2.2. Reliable Extra Samples

We propose various strategies to select Reliable Extra Samples for an instance to create a set of high confidence reference frames for the instance of interest:

- Removing blurry images or tiny segments.
- Removing frames with sudden changes in instance mask (anomaly).
- Removing too similar frames (redundant frames).
- Removing low-confidence segmentation results.
- Using weak alignment for checking topological consistency between instances in a pair of frames to remove frame with low-confidence instance matching.

2.3. Two-Pass Mask Guidance Segmentation

2.3.1 Single-Source Mask Propagation

We quickly propagate an initial mask to all frames in a sequence to create an initial segmentation result of the instance. To improve segmentation, we use our previous works, *i.e.*, multi-scheme segmentation [12] followed by guided segmentation [11]. We integrate IRIF [5], STM [10], PremVOS [2] into the multi-scheme segmentation process to enhance object tracking and re-identification.

2.3.2 Multi-Source Mask Propagation

We re-propagate masks with reference to multiple extra samples (c.f. Fig. 2). We put all reliable frames in the memory pool for reference. However, we further propose to propagate mask from multiple reference frames, instead of from a single initial frame. Each reference frame influ-

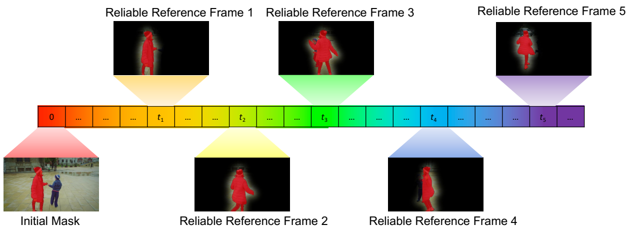
Reliable Reference Frame 1
Reliable Reference Frame 3
Reliable Reference Frame 5
0
t_1
t_2
t_3
t_4
t_5
Initial Mask
Reliable Reference Frame 2
Reliable Reference Frame 4

Figure 2. Multi-source mask propagation.

Table 1. Ranking results in the DAVIS 2020 Challenge. The rankings in each categories are placed in parentheses. Our results are marked in **boldfaced blue**.

#	Team	Global		Region J						Boundary F					
		Mean ⇑		Mean ⇑		Recall ⇑		Decay ⇓		Mean ⇑		Recall ⇑		Decay ⇓	
1	Alibaba-Vision	84.1	(1)	81.5	(1)	89.1	(1)	14.2	(2)	86.7	(1)	92.9	(2)	16.1	(2)
2	ReLER	83.8	(2)	81.1	(2)	88.4	(2)	18.1	(5)	86.5	(2)	93.3	(1)	17.8	(4)
3	Hongje	79.5	(3)	77.0	(3)	85.7	(3)	14.9	(3)	82.1	(3)	89.6	(4)	17.1	(3)
4	**MR-GIS (Ours)**	**79.3**	**(4)**	**76.5**	**(4)**	**85.1**	**(4)**	**10.1**	**(1)**	**82.1**	**(3)**	**90.7**	**(3)**	**12.0**	**(1)**
5	DSVOS	76.9	(5)	74.4	(5)	82.7	(5)	20.5	(6)	79.5	(5)	86.9	(6)	23.6	(7)
6	Vltanh	76.0	(6)	73.3	(6)	81.9	(6)	16.6	(4)	78.7	(6)	87.2	(5)	19.6	(5)
7	Bytedance	72.2	(7)	69.8	(7)	76.7	(7)	20.7	(7)	74.6	(7)	83.0	(7)	23.0	(6)
8	DeepDream	64.9	(8)	62.5	(8)	72.1	(8)	24.2	(8)	67.3	(8)	76.7	(8)	27.7	(8)

Table 2. The performance of our methods on the DAVIS 2020 Challenge dataset. Our latest results are marked in **boldfaced blue**.

Method	Global	Region J			Boundary F		
	Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
IRIF [5]	63.8	61.5	68.6	17.1	66.2	79.0	17.6
CIS [12]	66.3	64.1	75.0	11.7	68.6	80.7	13.5
GIS [11]	75.4	72.4	81.7	11.0	78.4	87.6	12.9
MR-GIS	**79.3**	**76.5**	**85.1**	**10.1**	**82.1**	**90.7**	**12.0**

ences the results of the frames in its neighbors, and each frame usually depends on its nearest reference frame.

Similarly to [11], to reduce the ambiguity between masks of different instances, we create a guided region in each frame from its previous frame. The guided region is created by expanding the segmented mask in a previous frame with reference to the current motion of the instance, and the less-potential regions in the frame are filtered out. If the instance could not be segmented in the guided region, we restart object detection and re-identification in that frame and its following frames.

2.4. Mask Anomaly Detection and Correction

To enhance the consistency of instance masks across frames, we find mask anomaly in consecutive frames. An anomaly occurs when a mask accidentally disappears or reappears in a short period, or its size and shape change significantly. To handle an anomaly case, we use both forward and backward mask propagation to restore missing segments, and to correct sudden changes in mask size and shape (c.f. Fig. 3).

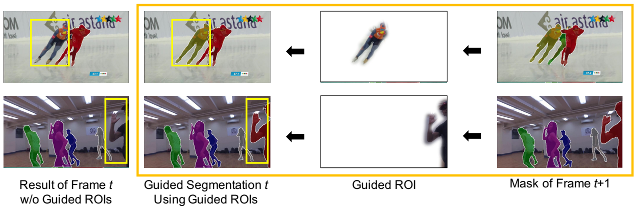

Figure 3. Anomaly mask detection and correction.

3. Results On DAVIS 2020 Challenge

As shown in Table 1, we obtain very competitive results. Our proposed MR-GIS achieved 76.5, 82.1, and 79.3 in terms of region similarity (J), contour accuracy (F), and global score, respectively on DAVIS 2019 Challenge dataset. Our method achieved the best scores in Decay of all metrics. Furthermore, we highlight the improvement of our research on DAVIS competition in Table 2. Our proposed MR-GIS outperforms our previous work (IRIF [5], CIS [12], and GIS [11]). Particularly, our newly proposed MR-GIS based on multi-source mask propagation improves GIS, which is based on single-source mask propagation, up to 4% in global score.

Figure 4 visualizes segmentation results. From top to bottom row, we show the first video frame, and a triple of processed video frames (IRIF [5], CIS [12], GIS [11], and our MR-GIS, respectively). Our final MR-GIS results successfully track and segment the key instances. More visual results can be found from our websites[1].

4. Conclusion

In this paper, we introduce the novel MR-GIS framework for multiple instances segmentation in videos. Our proposed method consists two passes of segmentation with mask guidance. To enhance the consistency of instance masks across frames, we find mask anomaly in consecutive frames and correct them. Throughout the experiments, our proposed framework surpasses our previous performance and achieves a competitive result among the leading submissions.

[1] https://www.selab.hcmus.edu.vn/vos or https://sites.google.com/view/ltnghia/research/vos

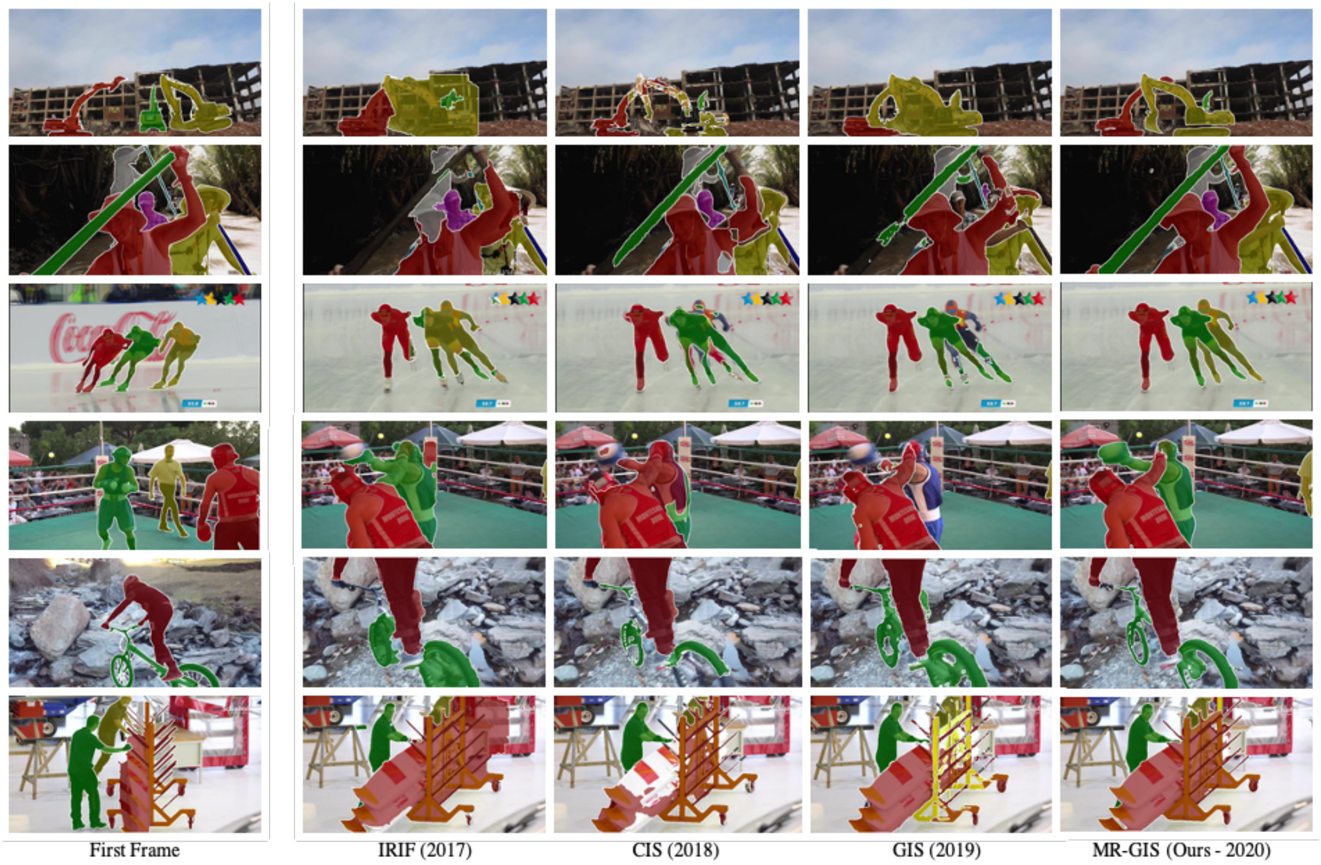

Figure 4. The visualization results on the DAVIS 2020 Challenge. From left to right: the first video frame with the ground-truth label followed by results of IRIF [5], CIS [12], GIS [11], and our MR-GIS. The ground-truth of the certain video frame is not publicly available. Our final results significantly track and segment the instances of interest as annotated in the first frame.

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

[2] B. L. J. Luiten, P. Voigtlaender. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation. *CVPR Workshops*, 2018.

[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] T. Le and A. Sugimoto. Semantic instance meets salient object: Study on video semantic salient instance segmentation. In *WACV*, 2019.

[5] T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, V. Ton-That, T.-A. Nguyen, X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A. D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identification flow for video object segmentation. *CVPR Workshops*, 2017.

[6] T.-N. Le, A. Sugimoto, S. Ono, and H. Kawasaki. Attention r-cnn for accident detection. In *Intelligent Vehicles Symposium*, 2020.

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

[8] K. Nguyen, K. Nguyen, D. Le, D. A. Duong, and T. V. Nguyen. YADA: you always dream again for better object detection. *Multim. Tools Appl.*, 78(19):28189–28208, 2019.

[9] K. Nguyen, K. Nguyen, D. Le, D. A. Duong, and T. V. Nguyen. You always look again: Learning to detect the unseen objects. *J. Vis. Commun. Image Represent.*, 60:206–216, 2019.

[10] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *ICCV*, October 2019.

[11] 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, and M. N. Do. Guided instance segmentation framework for semi-supervised video instance segmentation. In *CVPR Workshops*, 2019.

[12] M.-T. Tran, V. Ton-That, T.-N. Le, K.-T. Nguyen, T. V. Ninh, T.-K. Le, V.-T. Nguyen, T. V. Nguyen, and M. N. Do. Context-based instance segmentation in video sequences. *CVPR Workshops*, 2018.

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

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