Guided Instance Segmentation Framework for Semi-supervised Video Instance Segmentation

Minh-Triet Tran *[1], Trung-Nghia Le[2], Tam V. Nguyen[3],
That-Vinh Ton[1], Trung-Hieu Hoang[1], Ngoc-Minh Bui[1], Trong-Le Do[1], Quoc-An Luong[1],
Vinh-Tiep Nguyen[4], Duc Anh Duong[4], and Minh N. Do[5]

[1]University of Science, VNU-HCM, Vietnam
[2]University of Tokyo, Japan
[3]University of Dayton, U.S.
[4]University of Information Technology, VNU-HCM, Vietnam
[5]University of Illinois at Urbana-Champaign, U.S.

Abstract

In this paper, we propose a novel Guided Instance Segmentation (GIS) framework to tackle the challenging problem of semi-supervised video instance segmentation. To improve the accuracy for instance segmentation, we propose to perform fined-grained segmentation on a non-rectangular region of interest (ROI). The natural-shaped ROI is generated by applying guided attention from the neighbor frames of the current one. By this way, our method can reduce the ambiguity in the segmentation of different instances, especially those of the same category, in a regular rectangular region. GIS first performs the normal forward mask propagation as in other instance segmentation methods. Then the backward mask propagation is executed to further restore missing instance fragments. This proposed idea is motivated by the scenarios where an instance reappears in a video sequence: it is initially small due to the far distance, then gradually increases in terms of size. The re-appreared instance can be detected and segmented when it is large enough. By using mask back-propagation, GIS can restore small instance fragments before it is large enough for detection and segmentation. Our proposed GIS achieved 0.724, 0.784, and 0.754 in terms of region similarity (J), contour accuracy (F), and global score, respectively on DAVIS 2019 Challenge dataset, rank 3rd in the challenge. Our method achieved the best scores in Decay of all metrics.

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

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], action recognition [5], video summarization [7], object tracking [11], and scene understanding [12].

In this work, we Guided Instance Segmentation (GIS), a novel framework, to tackle 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.

First, to segment an instance in a region of interest (ROI), we use guided segmentation based on attention to eliminate complex background inside the ROI for performance improvement. We transform a regular rectangular ROI to a non-rectangular ROI by applying guided attention inferred from neighbor frames and object flow estimation. We then perform fine-grained segmentation on this guided natural-shaped ROI. Our proposed guided segmentation outperforms the standard segmentation, which is mostly applied in rectangular ROIs.

Second, we propose bi-directional strategies to construct adaptive attention for guided segmentation. Particularly, initial segments from neighbor frames are used as references for segmentation at the current frame. Attention is computed in two strategies sequentially, *i.e.*, forward propagation and back-propagation, in specific ways adapting the context. Forward propagation strategy, where attention is

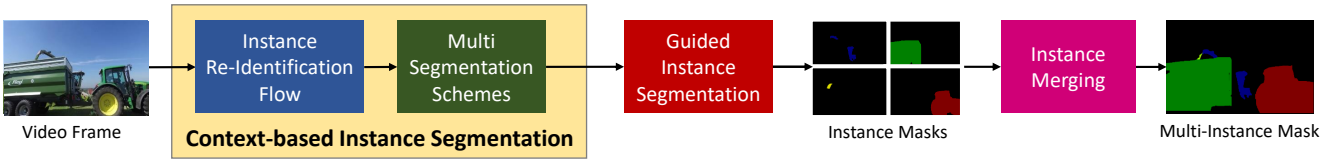

Figure 1. Overview of Guided Instance Segmentation (GIS) framework.

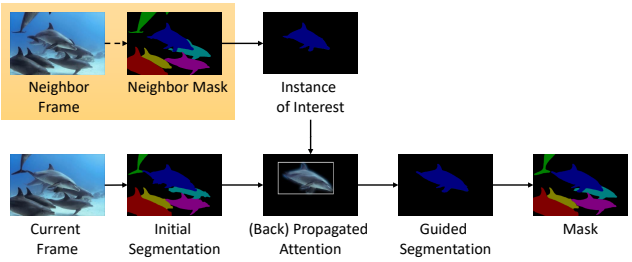

Figure 2. Propagation process of our GIS.

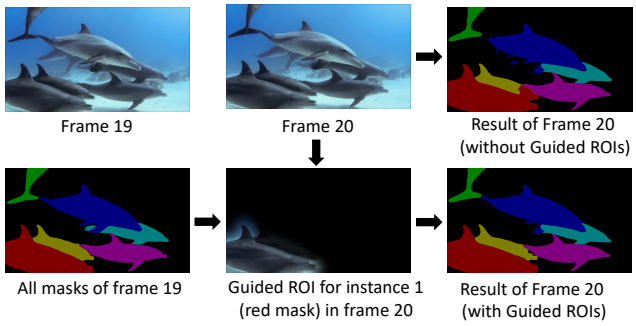

Figure 3. Visualization of forward propagation.

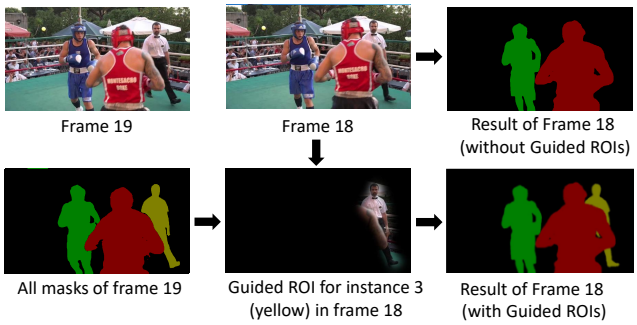

Figure 4. Visualization of back-propagation.

referenced from initial segments of previous frames, can correct missing segmentation due to dense objects in a ROI. Meanwhile, back-propagation strategy, where attention is referenced from initial segments of next frames, can recover missing instances due to fast motion, occlusion, or heavy deformation (size changing from tiny to large or vice versa).

Our framework can adapt to any existing segmentation methods. We adopt Context-based Instance Segmentation (CIS) [10] for initial segmentation, use Deep Grabcut [13] for fine-grained segmentation. Finally, instances in each frame are heuristic merged similarly to [10].

Our proposed GIS achieved 0.724, 0.784, and 0.754 in terms of region similarity (J), contour accuracy (F), and global score, respectively on DAVIS 2019 Challenge dataset. 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 GIS further improves the original CIS framework [10] up to 9.1% in global score.

The remainder of this paper is organized as follows. Our proposed methods are presented in Sections 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. Guided Instance Segmentation

2.1. Overview

Traditional Fully Convolutional Networks (FCNs) consider the entire rectangular region of interest (ROI) as the input to segment objects inside the ROI. This can lead to incorrect boundary segmentation due to complex background and concave hull of the object. To overcome this limitation, we aims to transform a rectangular ROI to a non-rectangular ROI across the object boundary to eliminate complex background inside the ROI. Inspired by skeleton-guided segmentation [10], we utilize referral information from extra frames to identify the shape of the instance of interest inside the ROI of the current frame. We propose to apply guided attention to construct the non-rectangular ROI and then perform fine-grained segmentation on this guided non-rectangular ROI.

We propose bi-directional strategies to construct adaptive attention for guided segmentation. Particularly, initial segments from neighbor frames are used as references for segmentation at the current frame (Fig. 2). Attention is computed in two strategies sequentially, *i.e.*, forward propagation and back-propagation , in specific ways adapting the context. Forward propagation strategy, where attention is referenced from initial segments of previous frames, can correct excessed segmentation due to dense objects in a ROI (cf. Fig. 3). Meanwhile, back-propagation strategy, where

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 J						Boundary F					
		Mean ⇑		Mean ⇑		Recall ⇑		Decay ⇓		Mean ⇑		Recall ⇑		Decay ⇓	
1	ZX_VIP	76.7	(1)	72.7	(2)	81.5	(3)	19.5	(3)	80.6	(1)	87.3	(2)	22.0	(3)
2	Jono	76.2	(2)	72.9	(1)	81.7	(1)	16.3	(2)	79.4	(2)	86.7	(3)	19.5	(2)
3	**Ours**	**75.4**	**(3)**	**72.4**	**(4)**	**81.7**	**(2)**	**11.0**	**(1)**	**78.4**	**(4)**	**87.6**	**(1)**	**12.9**	**(1)**
4	swoh	75.2	(4)	72.6	(3)	81.0	(4)	21.2	(5)	77.7	(4)	84.9	(4)	24.5	(5)
5	H2VISION	73.1	(5)	70.1	(5)	77.3	(6)	24.8	(8)	76.1	(5)	84.0	(6)	28.3	(9)
6	savor-123	71.3	(6)	67.7	(7)	74.8	(7)	24.7	(7)	75.0	(6)	81.2	(7)	27.5	(8)
7	dolfers	70.6	(7)	68.5	(6)	78.1	(5)	20.3	(4)	72.8	(7)	84.2	(5)	24.0	(4)
8	ByteCV	69.2	(8)	66.0	(8)	73.4	(9)	28.5	(10)	72.3	(8)	80.4	(8)	31.1	(10)
9	sourf	68.9	(9)	65.9	(9)	74.8	(8)	21.4	(6)	71.9	(9)	80.0	(9)	25.6	(6)
10	AGAMErs	64.3	(10)	61.2	(10)	69.4	(10)	26.4	(9)	67.4	(10)	77.7	(10)	27.3	(7)

Table 2. The performance of our methods on the DAVIS 2019 Challenge dataset. Our final results are marked in **boldfaced blue**.

Method	Global	Region J			Boundary F		
	Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
IRIF [6] (Pass 1)	63.8	61.5	68.6	17.1	66.2	79.0	17.6
CIS [10] (Passes 1, 2)	66.3	64.1	75.0	11.7	68.6	80.7	13.5
GIS (Passes 1, 2, 3)	**75.4**	**72.4**	**81.7**	**11.0**	**78.4**	**87.6**	**12.9**

attention is referenced from initial segments of next frames, can recover missing instances due to fast motion, occlusion, or heavy deformation (size changing from tiny to large or vice versa) (cf. Fig. 4).

2.2. Implementation

Figure 1 illustrates our proposed method, which consists of three passes. For initial segmentation, we adopt Context-based Instance Segmentation (CIS) [10]. Particularly, in the first pass, Instance Re-Identification Flow [6] (IRIF) is applied to generate the preview mask sequence, and then extract different contextual properties from each instance (*i.e.*, human/non-human, known/unknown category, and rigid/deformable). In the second pass, we implement multiple segmentation schemes corresponding to properties, adapting to the contextual properties of each instance, *i.e.*, the category and visual properties. The third pass is our proposed Guided Instance Segmentation (GIS) for fine-grained segmenting instance masks from the second pass. Finally, instances in each frame are heuristic merged similarly to [10].

In order to construct a guided non-rectangular ROI, we expand mask of the interest instance at neighbor frames and then transfer and combine them at the current frame. This guarantees that the ROI can cover the entire interest instance. We do not apply mask propagation to reduce complexity of computation. Then, we create a smooth transition region (by applying blurred mask to remove background) for the guided ROI to avoid a clear border between the ROI and background (cf. Fig. 2). It is essential to make segmentation method focus on the interest instance and avoid inaccurate segmentation due to clear border. We remark that the range of boundary expansion and transition smooth is computed based on the intensity of movement of the instance. Both two propagation strategies are performed adaptively if initial segments of the interest instance at the current frame is much different (in appearance or size) from those at neighbor frames or the instance re-appears. On the other hand, we only refine the interest instance at the current frame to save the cost.

We adopt Deep Grabcut [13] for fine-grained segmentation in guided non-rectangular ROIs. We used an off-the-shelf FCN implemented by Luiten and Voigtlaender [4]. This network is DeepLab3+ [2] with Xception-65 [3] backbone, and was trained on MS-COCO [8] and Mapillary [9] datasets.

3. Results On DAVIS 2019 Challenge

As shown in Table 1, we obtain very competitive results. Our proposed GIS achieved 0.724, 0.784, and 0.754 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 note that our GIS is in top 3 over 4 teams achieving 0.75 in terms of global score in the last three years.

Furthermore, we highlight contribution of our GIS as shown in Table 2. Our proposed GIS (using all three passes) outperforms using only two passes (CIS [10]) or a pass (IRIF [6]). Particularly, our proposed GIS improves CIS up to 9.1% in global score. Figure 5 visualizes segmentation results. From top row to bottom row, we can observe the first video frame, and a triple of processed video

Figure 5. The visualization results on the DAVIS 2019 Challenge. From top to bottom: the first video frame with the ground-truth label followed by results of IRIF [6], CIS [10], and our 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.

frames (IRIF [6], CIS [10], and our GIS, respectively). Our final GIS results successfully track and segment the key instances. More visual results can be found from our website[1].

4. Conclusion

In this paper, we introduce the novel GIS framework for multiple instances segmentation in videos. In particular, we propose guided segmentation based on attention to eliminate complex background inside the region of interests for performance improvement. Throughout the experiments, our proposed framework surpasses our previous performance and achieves a competitive result among the leading submissions. In the future, we plan to incorporate flow warping, deform mask, and learning attention for better results.

Acknowledgements

This work is in part granted by the research fund of University of Science, VNU-HCM for Software Engineering Laboratory and the STEM Catalyst Grant of University of Dayton. We gratefully acknowledge NVIDIA and AIOZ Pte Ltd for the support of GPU and computing infrastructure.

[1] https://sites.google.com/view/ltnghia/research/vos

References

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

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[3] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

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

[5] 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.

[6] 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. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

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

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[9] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[10] 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. *DAVIS Challenge - CVPR Workshops*, 2018.

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

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

[13] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep grabcut for object selection. *BMVC*, 2017.