Fast and Accurate Online Video Object Segmentation via Tracking Parts

Jingchun Cheng[1,2] Yi-Hsuan Tsai[3] Wei-Chih Hung[2] Shengjin Wang[1*] Ming-Hsuan Yang[2]
[1]Tsinghua University [2]University of California, Merced [3]NEC Laboratories America

1. Overview

We present additional results and analysis for the proposed algorithm in the supplementary material and the videos. In the following, we provide:

- Per-sequence evaluation on the DAVIS 2016 dataset [7] and example comparisons of video object segmentation methods with strong online applicability in Section 2.
- Analysis and sample results of part tracking in Section 3.

2. Video Object Segmentation

Table 1 presents the per-sequence evaluation (J mean) of state-of-the-art algorithms on the DAVIS 2016 dataset [7], including methods with *strong*, *weak* and *no* online applicability. We show that the proposed algorithm performs best among the methods with strong online applicability, with a margin of 12.2% compared to the VPN method [3]. More comparisons among fast online video object segmentation methods are shown in Figure 1-2 and the supplementary videos. Some sample results for the proposed algorithm with or without refinement (**Ours-ref** v.s. **Ours-part**) are shown in Figure 3-4.

The experimental results show that: 1) The proposed algorithm achieves state-of-the-art results in most sequences, and performs competitively with methods that need significantly longer pre-processing time. 2) The refinement stage of our method is able to recover details (see the first sequence in Figure 3), but may cause minor noises (see the second sequence in Figure 3).

3. Part Tracking

We present sample results of some high-scored representative parts and their tracking results on the DAVIS 2016 dataset [7] with single instance (Figure 3-4) and the DAVIS 2017 dataset [8] with multiple instances (Figure 5-6). Figure 3-6 show that the object parts selected by our algorithm keep tracking of the target instance without the need of time-consuming pre-processing. A video for comparisons of tracking via parts and entire object by the SiaFC method [1] is included for more visual illustrations.

*Corresponding Author

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. 1

[2] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017. 2

[3] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *CVPR*, 2017. 1, 2, 3, 4

[4] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 2

[5] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, 2017. 2

[6] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *CVPR*, 2016. 2, 3, 4

[7] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1

[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:1704.00675*, 2017. 1

[9] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 2

[10] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 2

Table 1. Per-sequence J mean on the DAVIS 2016 validation set.

Online Ability	No		Weak			Strong			
Method	LVO [9]	ARP [5]	SFL [2]	MSK [4]	OnAVOS [10]	BVS [6]	VPN [3]	**Ours-part**	**Ours-ref**
Runtime Speed	-	-	7.9s	12s	13s	0.84s	0.63s	0.60s	1.8s
blackswan	74.1	**88.1**	92.0	90.3	**96.3**	**94.3**	92.5	89.2	94.0
bmx-trees	**49.9**	**49.9**	45.7	57.5	**58.1**	38.2	33.5	55.4	**58.1**
breakdance	37.1	**76.2**	68.2	**76.2**	70.9	50.0	46.5	62.9	**67.3**
camel	88.1	**90.3**	79.1	80.1	**85.4**	66.9	75.9	**87.6**	87.5
car-roundabout	**88.6**	81.6	85.7	96.0	**97.5**	85.1	83.2	**93.9**	93.5
car-shadow	**92.0**	73.6	94.5	93.5	**96.9**	57.8	81.3	94.0	**93.5**
cows	90.2	**90.8**	90.6	88.2	**95.5**	89.5	89.9	91.8	**92.7**
dance-twirl	**81.0**	79.8	73.4	**84.4**	**84.4**	49.2	62.8	77.6	**82.1**
dog	**88.7**	71.8	93.0	90.9	**95.6**	72.3	88.6	92.0	**93.7**
drift-chicane	63.9	**79.7**	37.9	86.2	**89.2**	3.3	24.3	42.4	**73.2**
drift-straight	**84.9**	71.5	89.9	56.0	**94.4**	40.2	62.9	85.8	**87.8**
goat	**82.3**	77.6	86.1	84.5	**91.3**	66.1	82.2	**87.5**	87.4
horsejump-high	82.4	**83.8**	76.0	81.7	**90.1**	80.1	**81.8**	79.6	81.4
kite-surf	**64.6**	59.1	58.7	60.0	**69.1**	42.5	62.3	63.0	**66.5**
libby	**69.0**	65.4	70.0	77.5	**88.4**	77.6	72.6	77.7	**80.5**
motocross-jump	80.5	**82.3**	**83.9**	68.5	82.3	34.1	72.8	82.6	**86.5**
paragliding-launch	**62.2**	60.1	58.1	62.0	**64.3**	**64.0**	61.4	61.9	62.9
parkour	**84.9**	82.8	84.9	88.2	**93.6**	75.6	87.3	88.4	**90.1**
scooter-black	71.8	**74.6**	69.9	82.5	**91.1**	33.7	60.5	81.0	**83.4**
soapbox	81.3	**84.6**	83.7	**89.9**	88.5	78.9	81.8	77.1	**86.6**
mean	75.9	**76.2**	76.1	79.7	**86.1**	60.0	70.2	78.6	**82.4**

Figure 1. Sample results on the DAVIS 2016 dataset. For each set of results, we show the ground truth, segmentation results predicted by BVS [6], VPN [3] and the proposed method, respectively.

Figure 2. Sample results on the DAVIS 2016 dataset. For each set of results, we show the ground truth, segmentation results predicted by BVS [6], VPN [3] and the proposed method, respectively.

Figure 3. Sample results on the DAVIS 2016 dataset. We show the ground truth, high-scored parts via tracking, our segmentation results without refinement (Ours-part) and with refinement (Ours-ref), respectively.

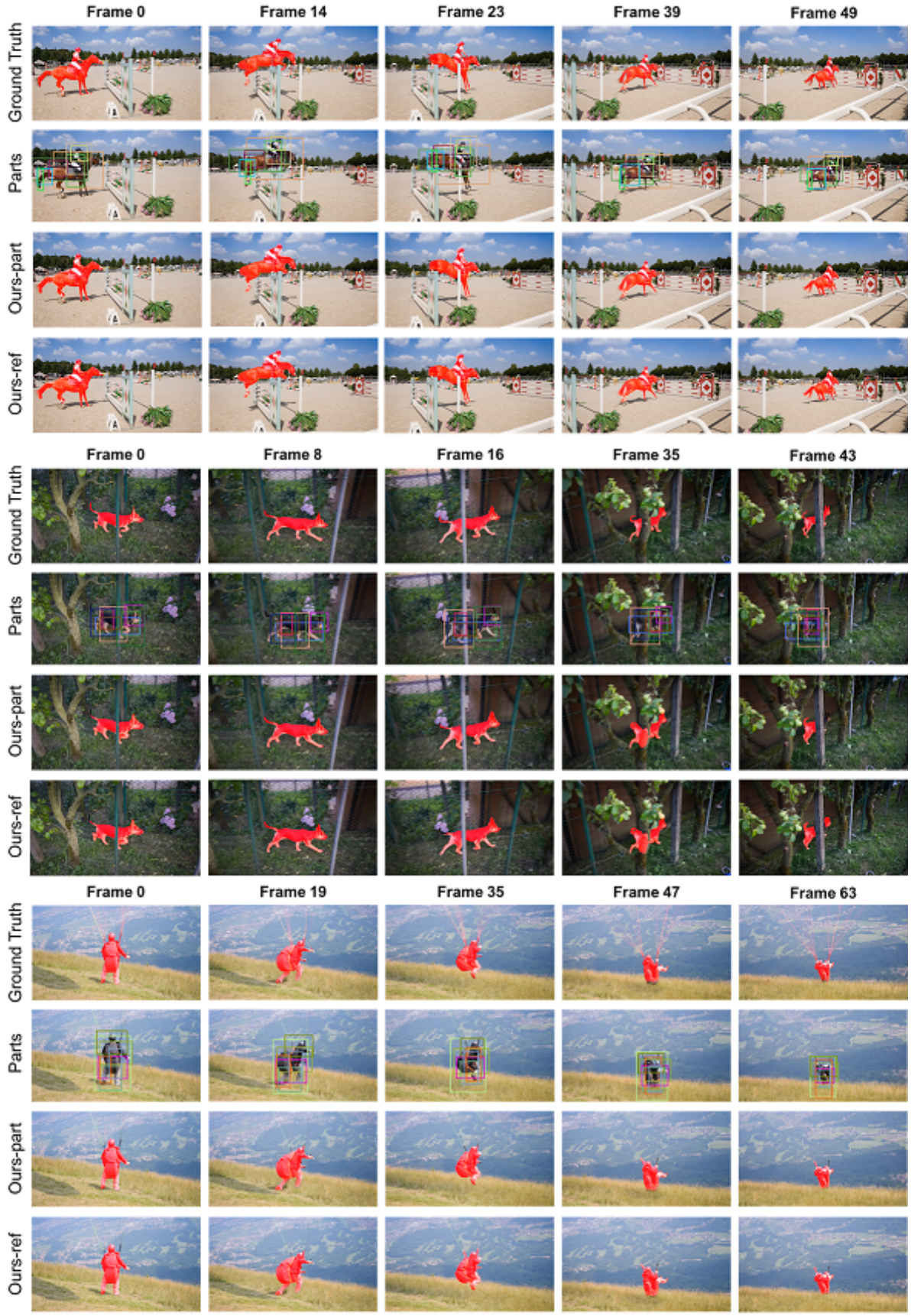

Figure 4. Sample results on the DAVIS 2016 dataset. We show the ground truth, high-scored parts via tracking, our segmentation results without refinement (Ours-part) and with refinement (Ours-ref), respectively.

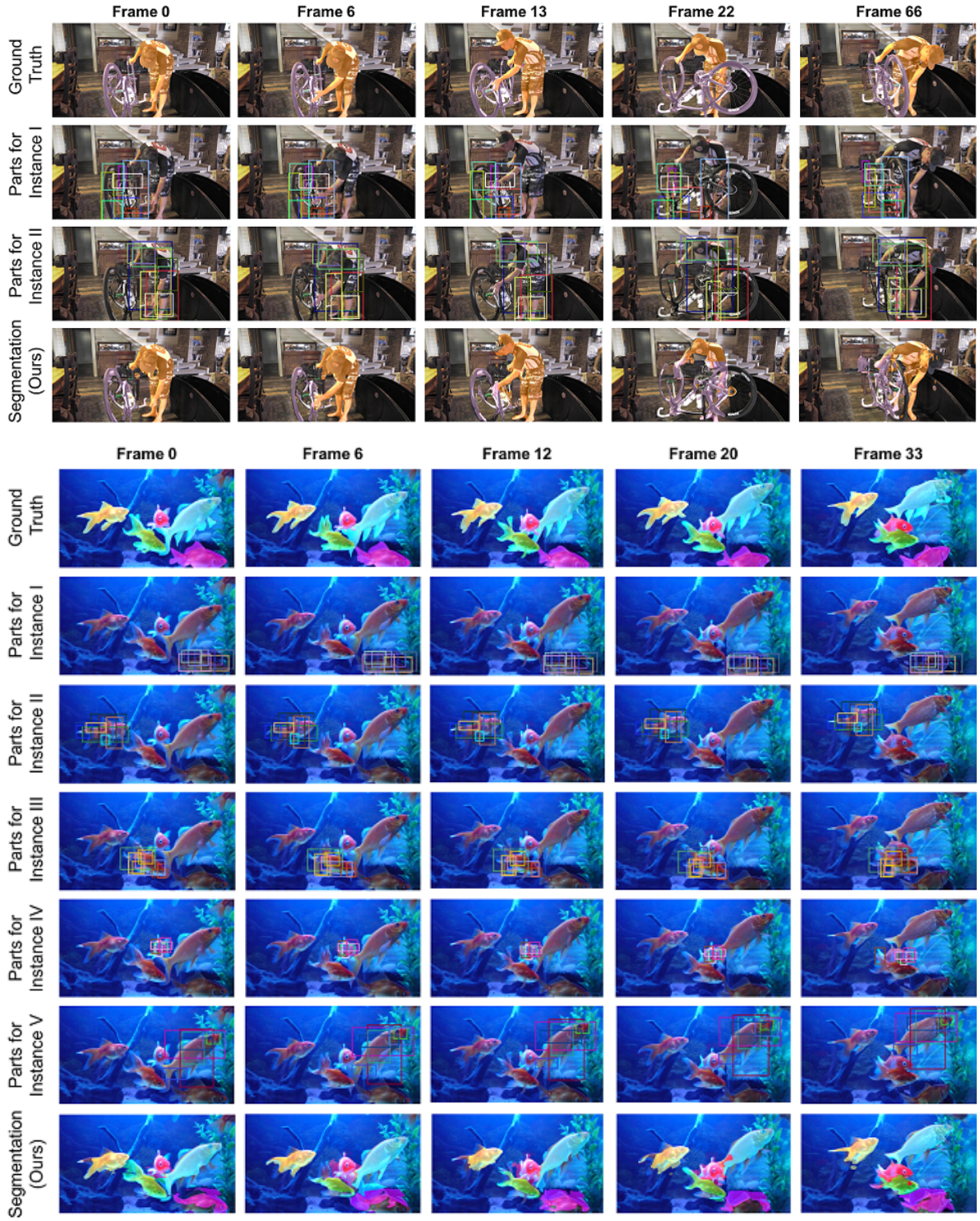

Figure 5. Sample results on the DAVIS 2017 dataset. The first row and the last row show the ground truth and segmentation results generated by the proposed method, while results in the middle rows show high-scored parts and their tracking results for each instance.

Figure 6. Sample results on the DAVIS 2017 dataset. The first row and the last row show the ground truth and segmentation results generated by the proposed method, while results in the middle rows show high-scored parts and their tracking results for each instance.