SegFlow: Joint Learning for Video Object Segmentation and Optical Flow

Jingchun Cheng[1,2] Yi-Hsuan Tsai[2,4] Shengjin Wang[1*] Ming-Hsuan Yang[2,3]
[1]Tsinghua University [2]University of California, Merced
[3]NVIDIA Research [4]NEC Laboratories America
[1]chengjingchun@gmail.com, wgsgj@tsinghua.edu.cn
[2]{ytsai2, mhyang}@ucmerced.edu

1. Contents

This supplementary material provides additional results and analysis for both optical flow estimation and foreground object segmentation. In the following, we provide:

- Details of training data for optical flow on the KITTI [5] and MPI Sintel [1] datasets in Section 2
- Per-class evaluation of segmentation on DAVIS [9] in Section 3.
- Example results of optical flow (Figure 1-3) and object segmentation (Figure 4-8).

2. Optical Flow Estimation

In this section, we describe more details of training process on KITTI and Sintel in Table 3 of the manuscript.

KITTI. We finetine our model (SegFlow+ft) and FlowNetS [4] (FlowNetS+ft*) with the KITTI training set without data augmentation, and select the best model with 10-fold cross validation for comparisons.

Sintel. Similarly, we finetune our model (SegFlow+ft) and FlowNetS [4] (FlowNetS+ft*) on the Sintel training set using only original images and their flips, and select the best model using the validation set as in [4].

We show example results for comparisons between *SegFlow* and FlowNetS in Figure 1. In addition, we show visual comparisons of optical flow on DAVIS in Figure 2 and 3, in which our method generates more complete optical flow within the object corresponding to our segmentation results.

3. Video Object Segmentation

Table 1 presents the per-sequence evaluation ($Jmean$) on DAVIS compared to other state-of-the-art methods, including semi-supervised and unsupervised ones. we improve the $Jmean$ by considering the prediction of the image and its flipping one, and averaging both outputs to obtain the final result, where we refer to as Ours^2. Without adding much computational cost, we further boost the performance with 1.3% in $Jmean$ as shown in Table 1. We also present the results of MSK [6] with only using the image as the input (MSK-flo), and show that our method without flow performs better (Ours-flo v.s MSK-flo).

More comparisons between $SegFlow$ and state-of-the-art methods are shown in Figure 4-8. To summarize the results in Table 1, we find that:

- *SegFlow* outperforms state-of-the-art unsupervised methods in most sequences.
- Online training is helpful for sequences with various appearance changes (Ours^2 v.s Ours-ol), such as non-rigid objects (e.g., camel, cows and soapbox), especially for sequences with dynamic backgrounds (e.g., 48.8% and 13.8% improvement for breakdance and dance-twirl respectively).
- Optical flow branch improves segmentation results (Ours^2 v.s Ours-flo) in most sequences (e.g., bmx-trees, breakdance, goat and libby), especially on the ones with large motion changes (e.g., 25% improvement for motocross-jump).

*Corresponding Author

Table 1. Per-sequence results on DAVIS validation Set.

Sequence	Semi-Supervised							Unsupervised			
	Ours[2]	Ours	Ours-flo	OSVOS [2]	MSK [6]	MSK-flo [6]	OFL [10]	Ours-ol	FST [8]	NLC [3])	KEY [7]
blackswan	0.920	0.904	0.904	0.942	0.903	0.919	**0.947**	**0.903**	0.732	0.875	0.842
bmx-trees	0.457	0.450	0.437	0.555	**0.575**	0.321	0.149	**0.437**	0.180	0.212	0.193
breakdance	0.682	0.660	0.561	0.708	**0.762**	0.594	0.496	0.194	0.467	**0.673**	0.549
camel	0.791	0.782	0.760	0.851	0.801	0.804	**0.867**	**0.760**	0.562	0.768	0.579
car-roundabout	0.857	0.857	0.875	0.953	**0.960**	0.828	0.900	**0.874**	0.808	0.509	0.640
car-shadow	**0.945**	0.902	0.902	0.937	0.935	0.903	0.846	**0.902**	0.698	0.645	0.589
cows	0.906	0.894	0.888	**0.946**	0.882	0.919	0.910	0.727	0.791	**0.883**	0.337
dance-twirl	0.734	0.730	0.683	0.670	**0.844**	0.678	0.567	**0.596**	0.453	0.347	0.380
dog	**0.930**	0.923	0.912	0.907	0.909	0.868	0.897	**0.918**	0.708	0.809	0.692
drift-chicane	0.378	0.360	0.541	0.835	**0.862**	0.005	0.175	0.090	**0.667**	0.324	0.188
drift-straight	**0.899**	0.897	0.826	0.676	0.560	0.460	0.314	**0.860**	0.682	0.473	0.194
goat	0.861	0.854	0.844	**0.880**	0.845	0.858	0.865	**0.836**	0.554	0.010	0.705
horsejump-high	0.760	0.752	0.732	0.780	0.817	0.784	**0.862**	0.678	0.578	**0.834**	0.370
kite-surf	0.587	0.569	0.552	0.686	0.600	0.587	**0.702**	0.525	0.272	0.453	**0.685**
libby	0.700	0.686	0.655	**0.808**	0.775	0.788	0.594	**0.670**	0.507	0.635	0.611
motocross-jump	**0.839**	0.835	0.589	0.816	0.685	0.690	0.594	**0.714**	0.602	0.251	0.288
paragliding-launch	0.581	0.580	0.554	0.625	0.620	0.589	**0.637**	0.580	0.506	**0.628**	0.559
parkour	0.849	0.840	0.791	0.856	**0.882**	0.853	0.861	0.813	0.458	**0.902**	0.410
scooter-black	0.699	0.692	0.694	0.711	**0.825**	0.649	0.765	**0.660**	0.522	0.162	0.502
soapbox	0.837	0.789	0.779	0.812	**0.899**	0.861	0.689	0.737	0.410	0.634	**0.757**
mean	0.761	0.748	0.724	**0.798**	0.797	0.698	0.680	**0.674**	0.558	0.551	0.498

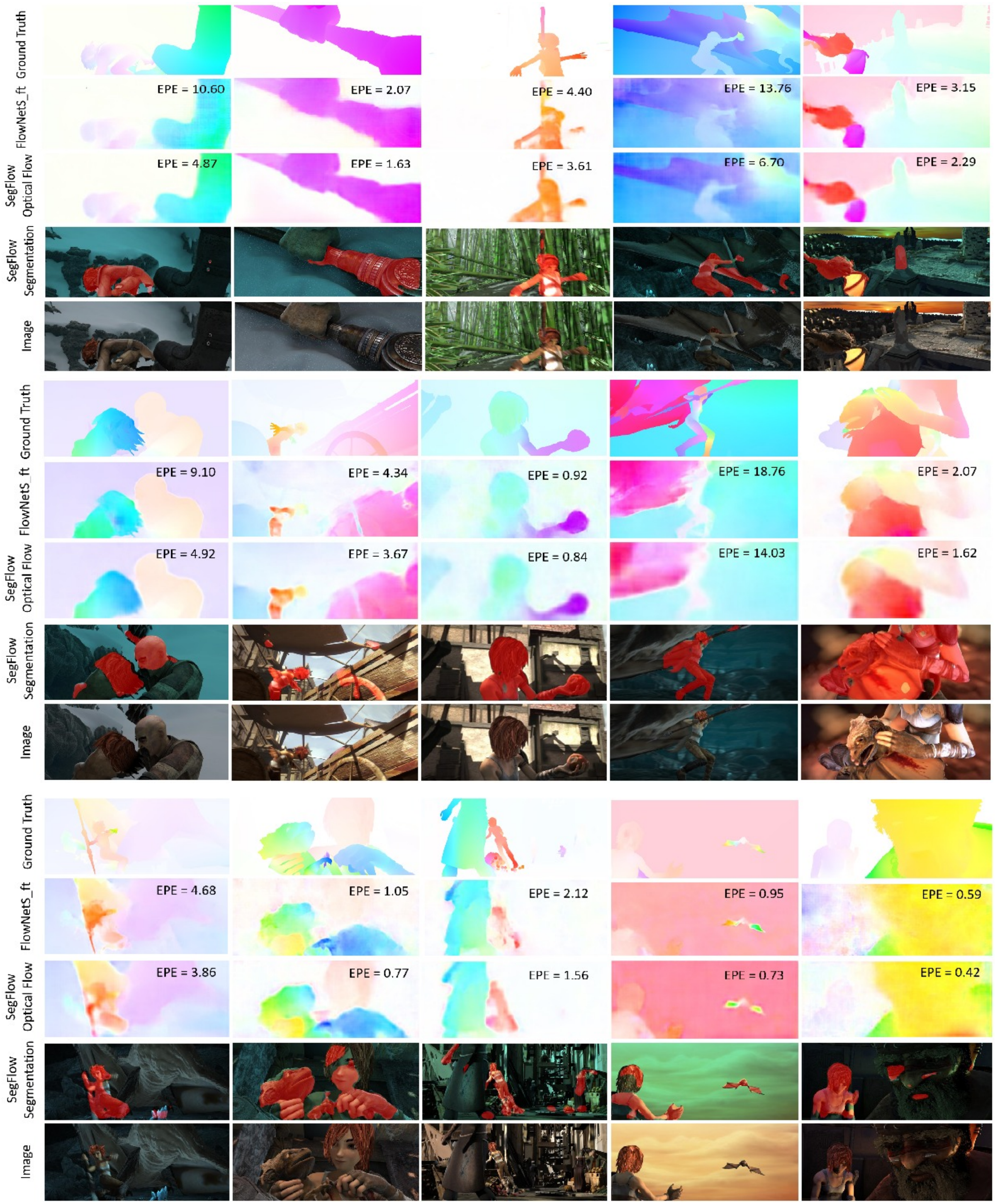

Figure 1. Example results on Sintel. For each set of results, row one to four shows the ground truth, optical flow predicted by FlowNetS+ft* (see Section 5.4 in paper for details), *SegFlow* and object segmentation generated by *SegFlow*, respectively.

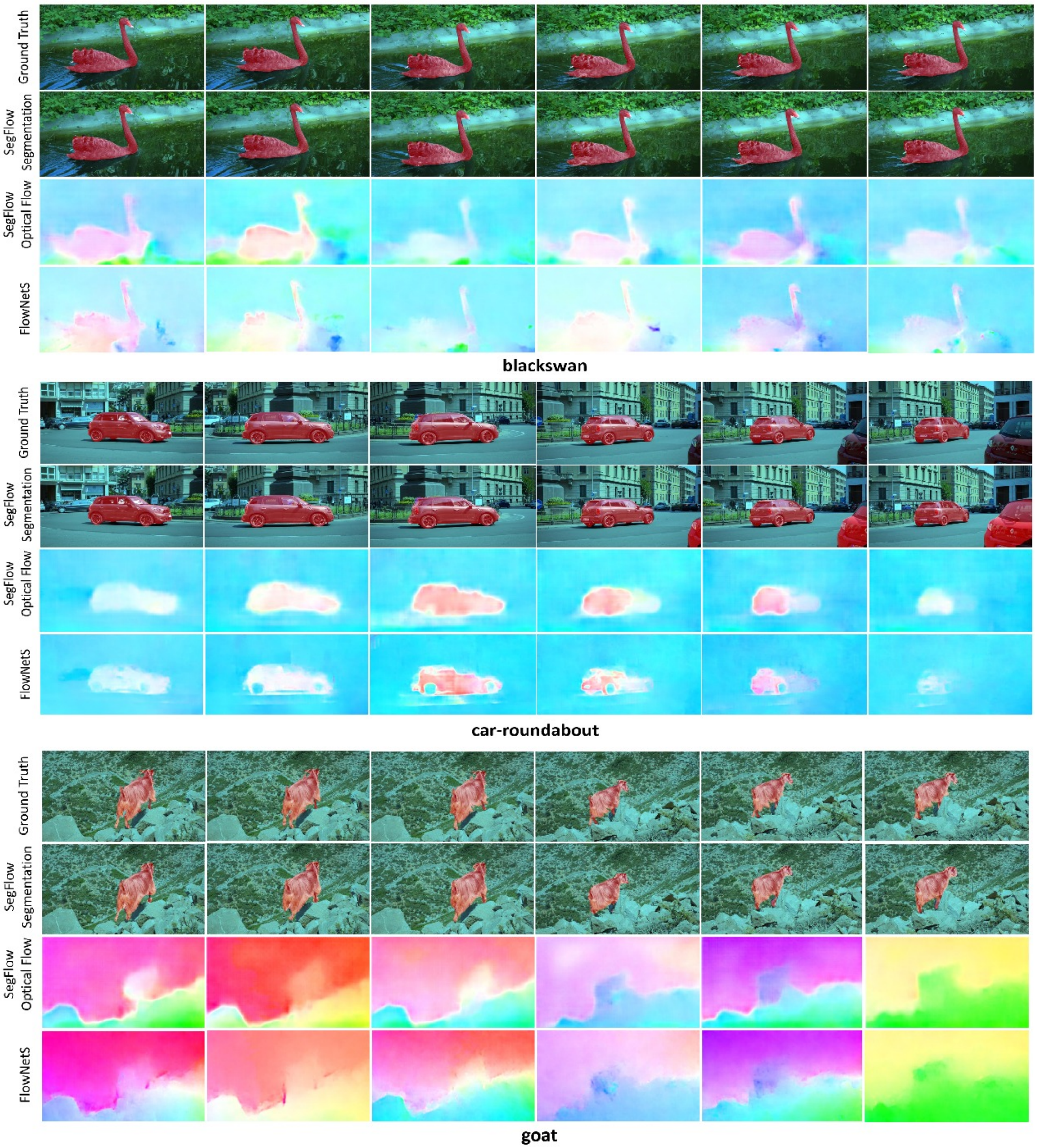

Figure 2. Example results on DAVIS. Row one to four of each sequence shows the annotations, our object segmentation and optical flow predicted by *SegFlow* and optical flow produced by FlowNetS [4], respectively.

Figure 3. Example results on DAVIS. Row one to four of each sequence shows the annotations, our object segmentation and optical flow predicted by *SegFlow* and optical flow produced by FlowNetS [4], respectively.

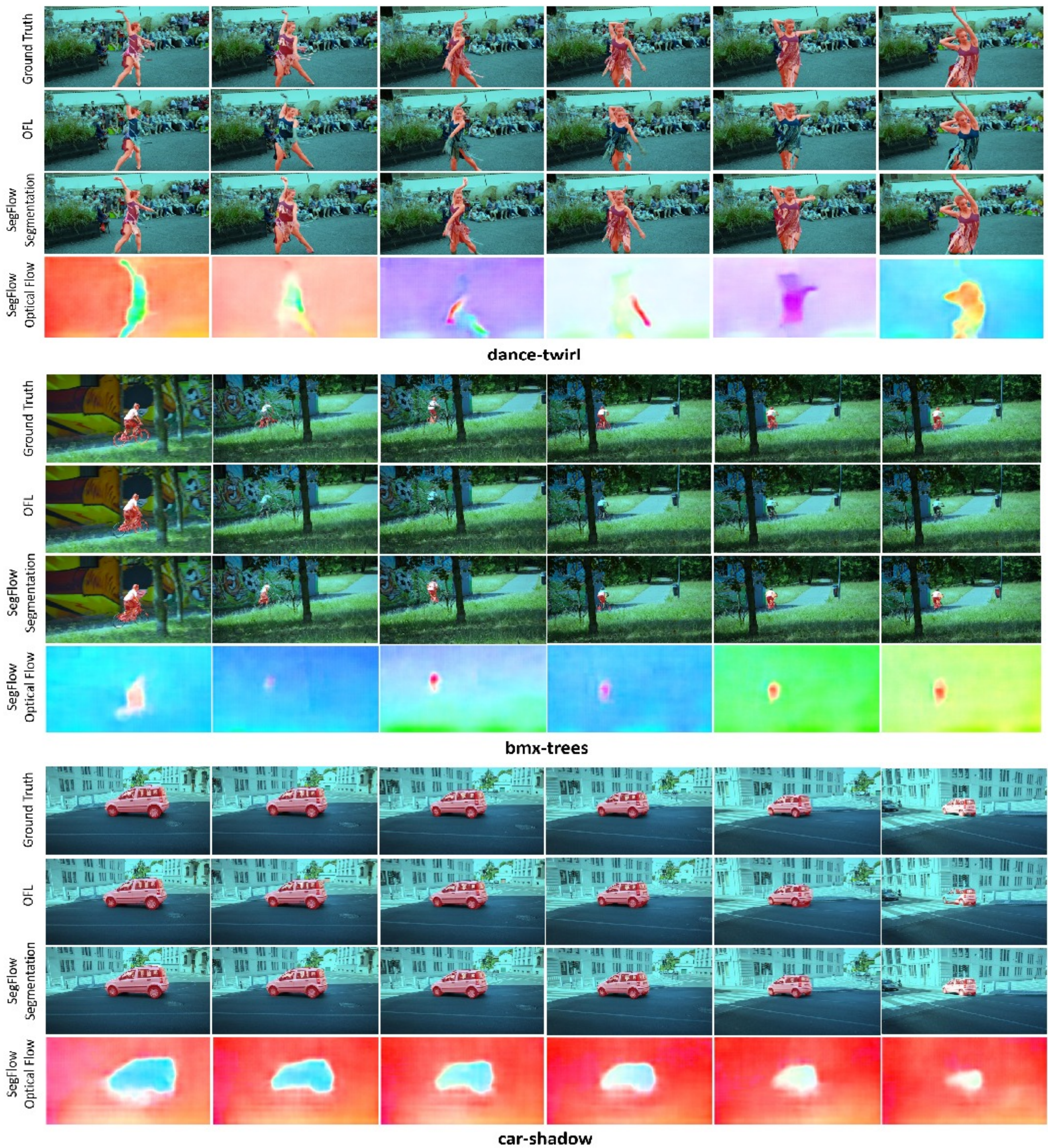

Figure 4. Example results on DAVIS. Row one to four of each sequence shows the annotations, object segmentation by OFL [10], object segmentation by *SegFlow*, and optical flow prediction by *SegFlow* respectively.

Figure 5. Example results on DAVIS. Row one to four of each sequence shows the annotations, object segmentation by OFL [10], object segmentation by *SegFlow*, and optical flow prediction by *SegFlow* respectively.

Figure 6. Example results on DAVIS dataset. Row one to four of each sequence shows the annotations, object segmentation by MSK [6], object segmentation by *SegFlow*, and optical flow prediction by *SegFlow* respectively.

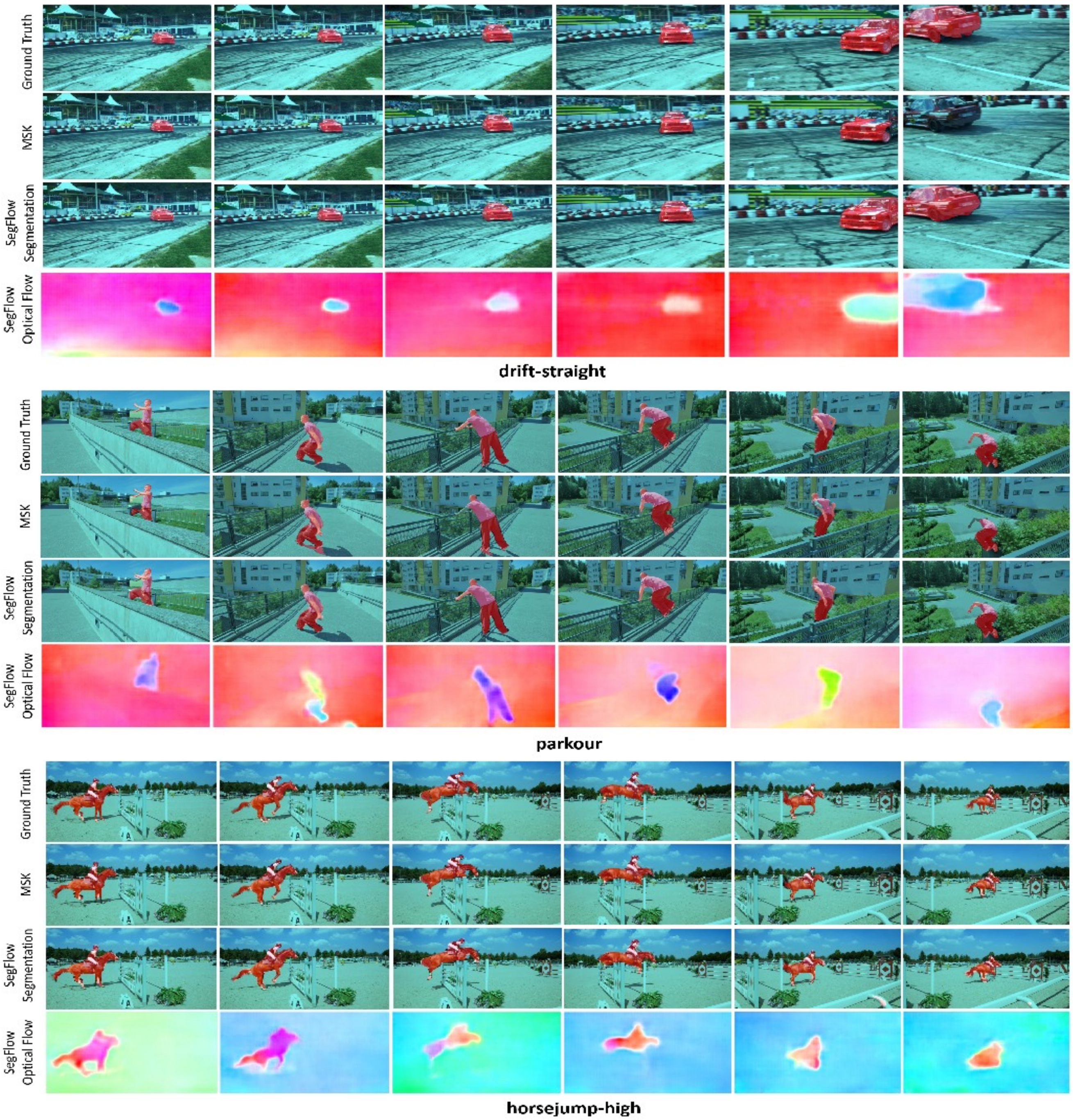

Figure 7. Example results on DAVIS. Row one to four of each sequence shows the annotations, object segmentation by MSK [6], object segmentation by *SegFlow*, and optical flow prediction by *SegFlow* respectively.

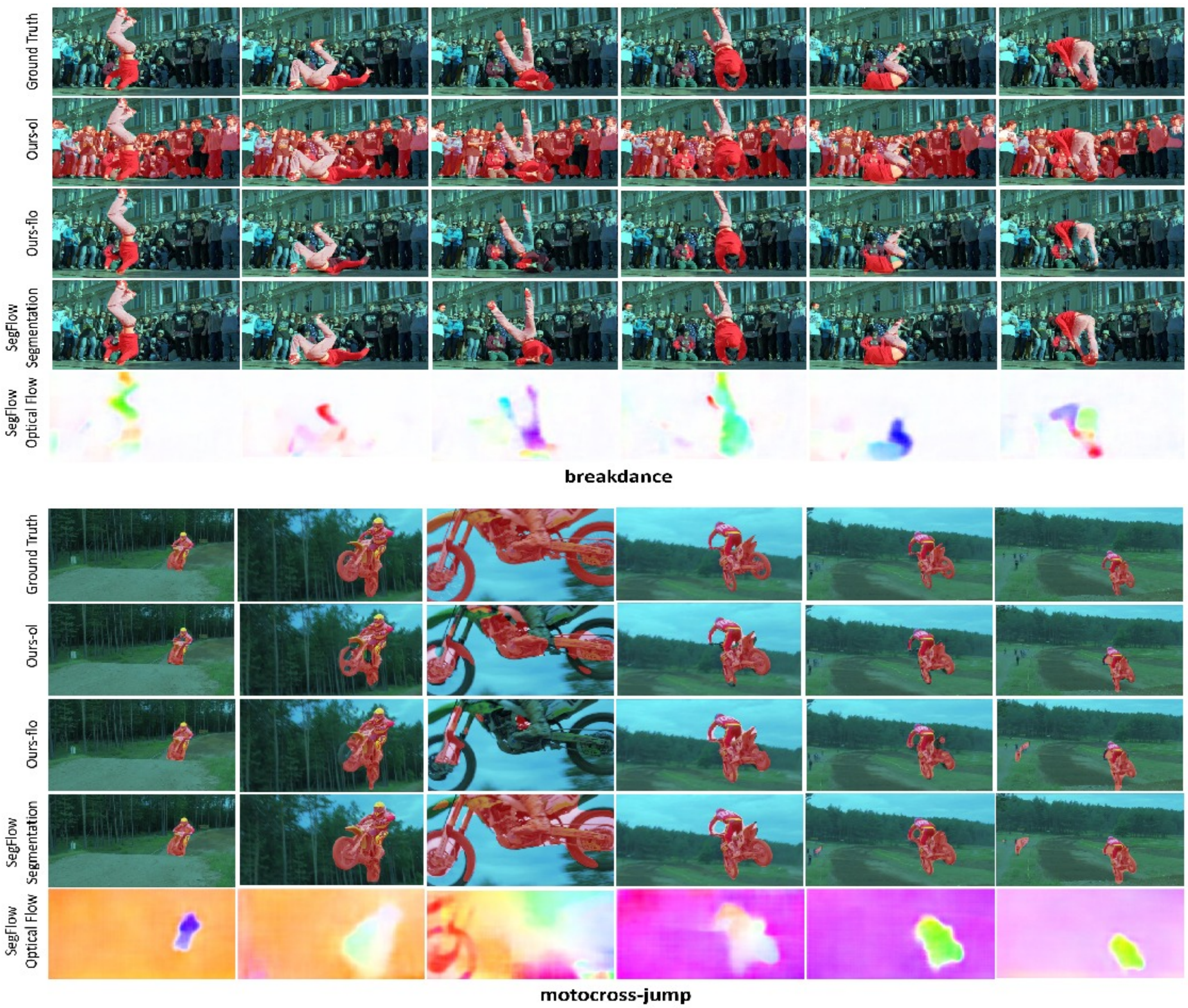

Figure 8. Example results on DAVIS. Row one to four of each sequence shows the annotations, object segmentation by *SegFlow* without online training (Ours-ol), *SegFlow* without optical flow branch (Ours-flo), *SegFlow*, and optical flow prediction by *SegFlow* respectively.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1

[2] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 2

[3] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014. 2

[4] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1, 4, 5

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[6] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1, 2, 8, 9

[7] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 2

[8] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013. 2

[9] 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

[10] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, 2016. 2, 6, 7