Supplementary Material for Video Propagation Networks

Varun Jampani[1], Raghudeep Gadde[1,2] and Peter V. Gehler[1,2]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, Tübingen, Germany
{varun.jampani,raghudeep.gadde,peter.gehler}@tuebingen.mpg.de

1. Parameters and Additional Results

In this supplementary, we present experiment protocols and additional qualitative results for experiments on video object segmentation, semantic video segmentation and video color propagation. Table 1 shows the feature scales and other parameters used in different experiments. Figures 1, 2 show some qualitative results on video object segmentation with some failure cases in Fig. 3. Figure 4 shows some qualitative results on semantic video segmentation and Fig. 5 shows results on video color propagation.

Experiment	**Feature Type**	**Feature Scale-1, Λ_a**	**Feature Scale-2, Λ_b**	α	**Input Frames**	**Loss Type**
Video Object Segmentation	(x, y, Y, Cb, Cr, t)	(0.02,0.02,0.07,0.4,0.4,0.01)	(0.03,0.03,0.09,0.5,0.5,0.2)	0.5	9	Logistic
Semantic Video Segmentation with CNN1 [5]-NoFlow	(x, y, R, G, B, t)	(0.08,0.08,0.2,0.2,0.2,0.04)	(0.11,0.11,0.2,0.2,0.2,0.04)	0.5	3	Logistic
with CNN1 [5]-Flow	$(x+u_x, y+u_y, R, G, B, t)$	(0.11,0.11,0.14,0.14,0.14,0.03)	(0.08,0.08,0.12,0.12,0.12,0.01)	0.65	3	Logistic
with CNN2 [3]-Flow	$(x+u_x, y+u_y, R, G, B, t)$	(0.08,0.08,0.2,0.2,0.2,0.04)	(0.09,0.09,0.25,0.25,0.25,0.03)	0.5	4	Logistic
Video Color Propagation	(x, y, I, t)	(0.04,0.04,0.2,0.04)	No second kernel	1	4	MSE

Table 1. **Experiment Protocols.** Experiment protocols for the different experiments presented in this work. **Feature Types**: Feature spaces used for the bilateral convolutions, with position (x, y) and color $(R, G, B$ or $Y, Cb, Cr)$ features $\in [0, 255]$. u_x, u_y denotes optical flow with respect to the present frame and I denotes grayscale intensity. **Feature Scales** (Λ_a, Λ_b): Validated scales for the features used. α: Exponential time decay for the input frames. **Input Frames**: Number of input frames for VPN. **Loss Type**: Type of loss used for back-propagation. "MSE" corresponds to Euclidean mean squared error loss and "Logistic" corresponds to multinomial logistic loss.

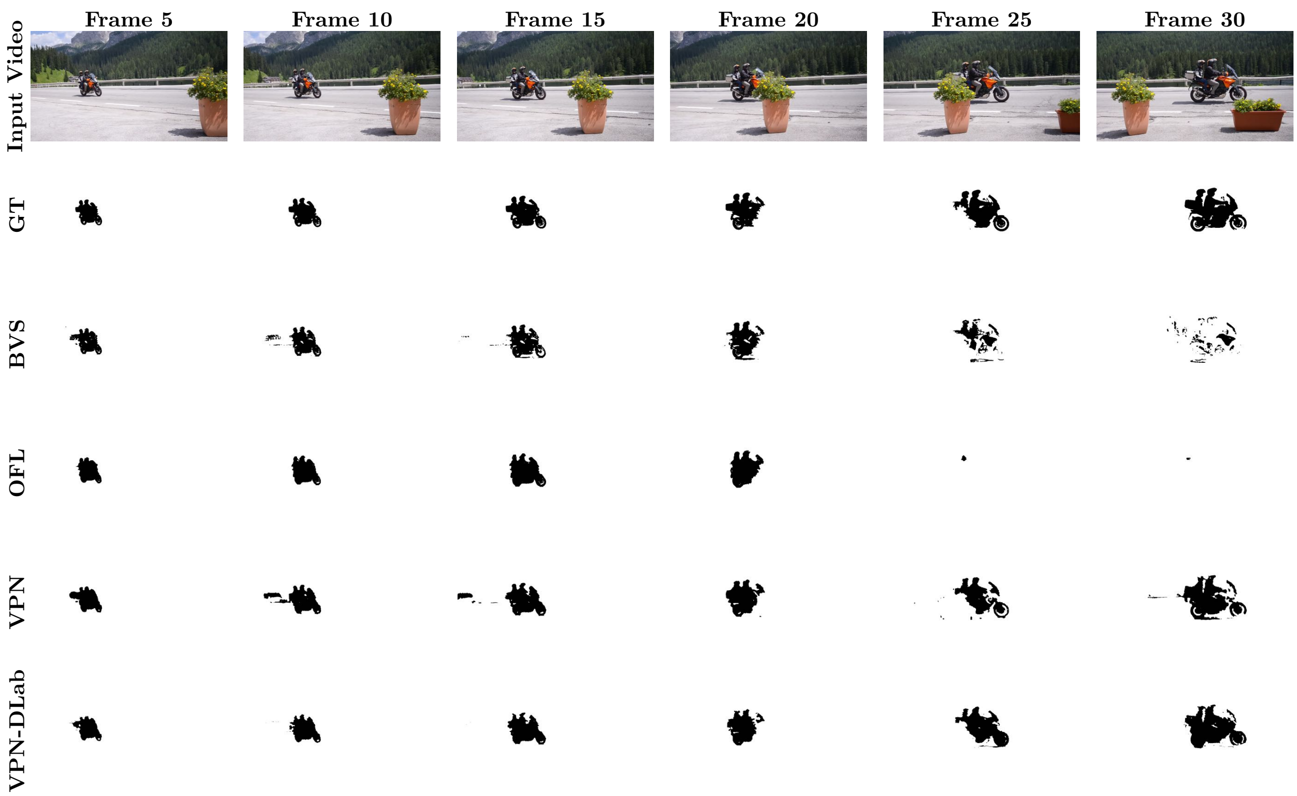

Figure 1. **Video Object Segmentation.** Shown are the different frames in example videos with the corresponding ground truth (GT) masks, predictions from BVS [2], OFL [4], VPN (VPN-Stage2) and VPN-DLab (VPN-DeepLab) models.

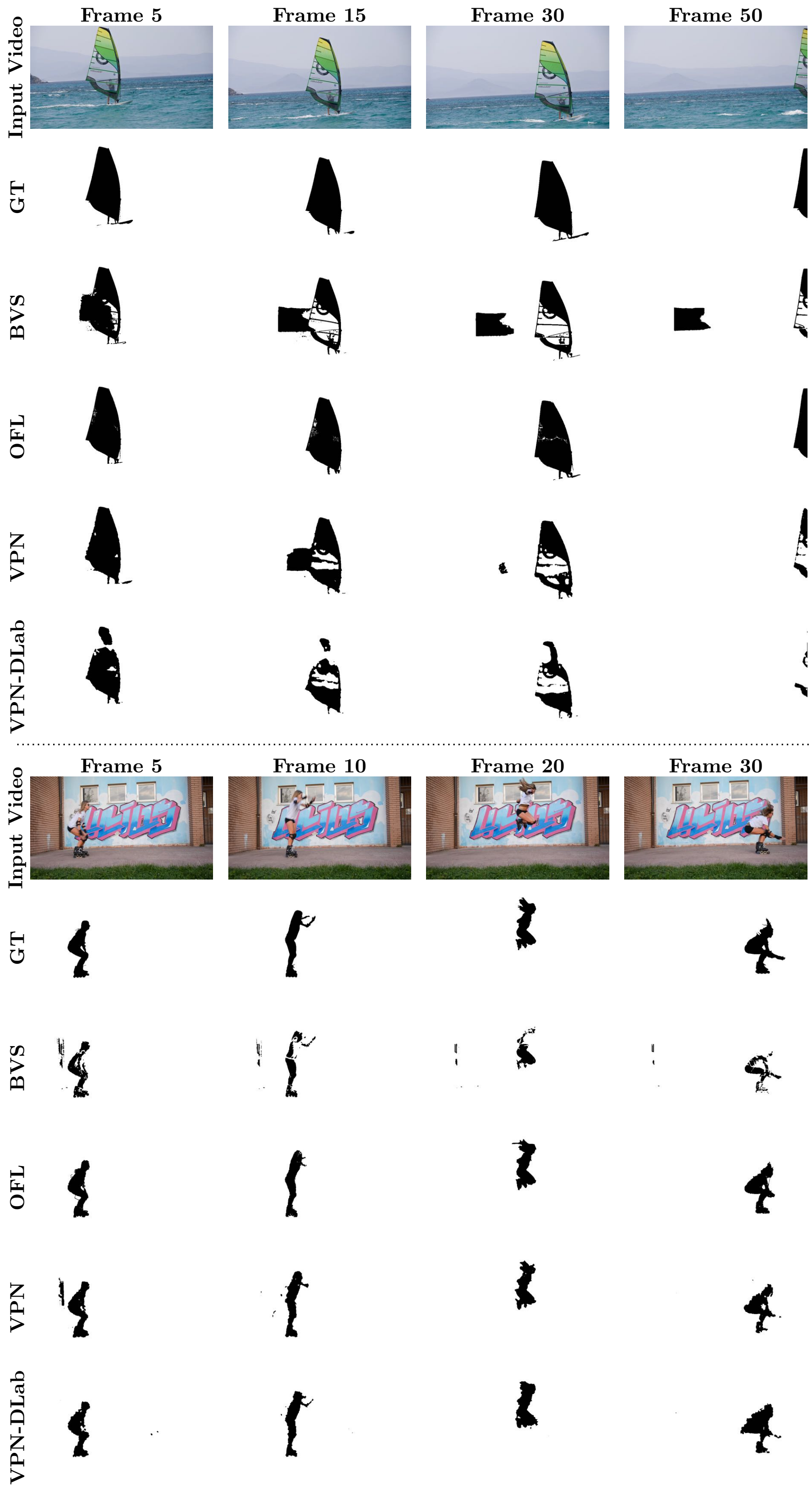

Figure 2. **Video Object Segmentation.** Shown are the different frames in example videos with the corresponding ground truth (GT) masks, predictions from BVS [2], OFL [4], VPN (VPN-Stage2) and VPN-DLab (VPN-DeepLab) models.

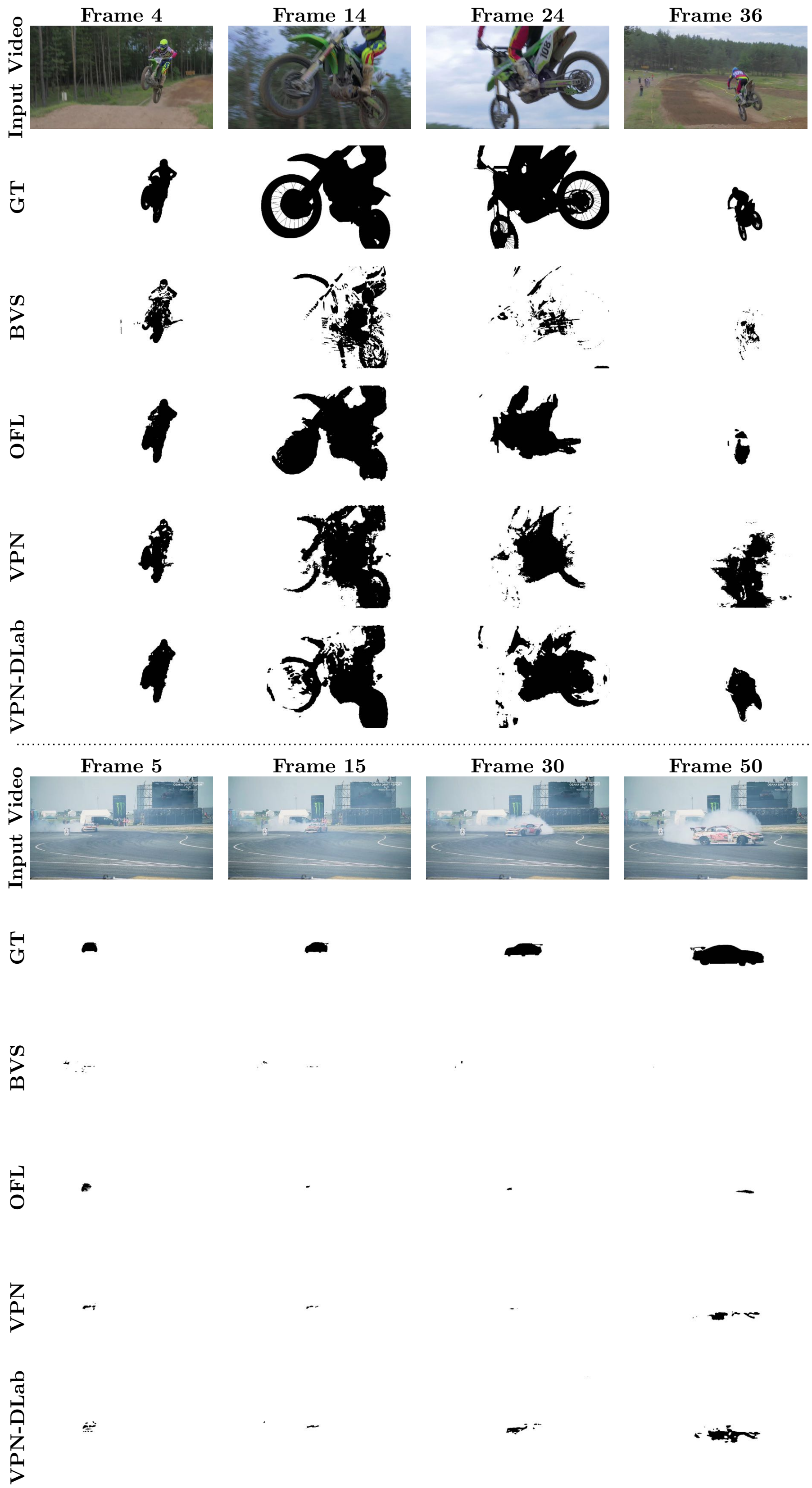

Figure 3. **Failure Cases for Video Object Segmentation.** Shown are the different frames in example videos with the corresponding ground truth (GT) masks, predictions from BVS [2], OFL [4], VPN (VPN-Stage2) and VPN-DLab (VPN-DeepLab) models.

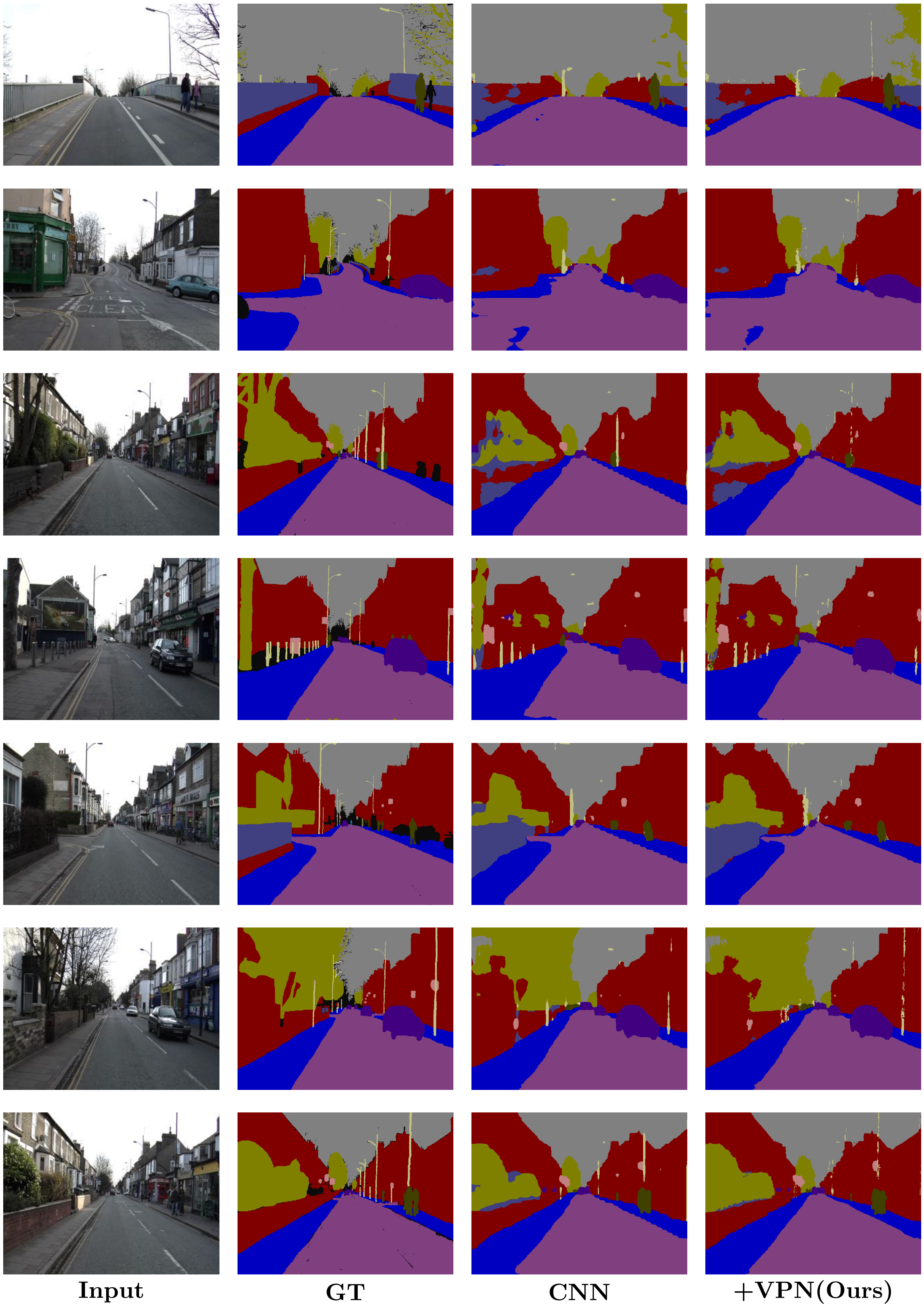

Figure 4. **Semantic Video Segmentation.** Input video frames and the corresponding ground truth (GT) segmentation together with the predictions of CNN [5] and with VPN-Flow.

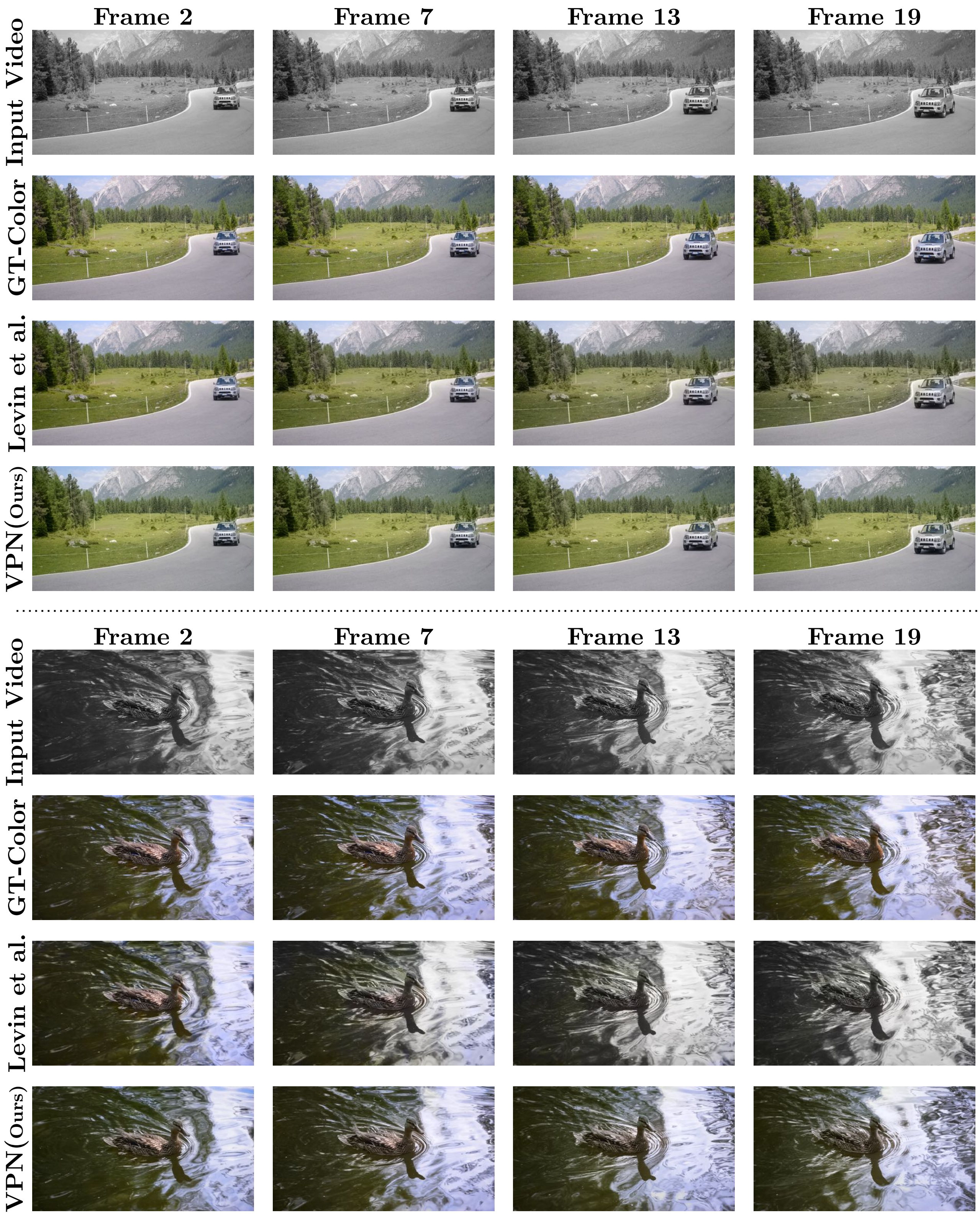

Figure 5. **Video Color Propagation.** Input grayscale video frames and corresponding ground-truth (GT) color images together with color predictions of Levin et al. [1] and VPN-Stage1 models.

References

[1] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Transactions on Graphics (ToG)*, 23(3):689–694, 2004. 6

[2] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 743–751, 2016. 2, 3, 4

[3] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1

[4] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2016. 2, 3, 4

[5] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations*, 2016. 1, 5