High-Resolution Optical Flow from 1D Attention and Correlation
Supplementary Material

Haofei Xu[1*] Jiaolong Yang[2] Jianfei Cai[3] Juyong Zhang[1] Xin Tong[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia
[3]Department of Data Science and AI, Monash University
{xhf@mail., juyong@}ustc.edu.cn {jiaoyan, xtong}@microsoft.com jianfei.cai@monash.edu

In this supplementary document, we first present additional evaluations of our proposed method. Then we provide more visual results on 4K (2160×3840) resolution images from DAVIS dataset and real-world scenes captured by a mobile phone. Finally, we present additional visual results on Sintel test set and more implementation details.

A. Additional Evaluations

In the main paper, we have analyzed the role of each 3D cost volume plays and the evaluation results on Sintel (train, clean) shows that the performance of horizontal or vertical flow is coupled with the correlation direction. Here we present additional evaluations on Sintel (train, final) and KITTI (train) datasets and observe consistent results: horizontal cost volume is mainly responsible for the horizontal flow estimation, and similarly for the vertical cost volume. Concatenating these two cost volumes gives the network necessary information for estimating both horizontal and vertical flow components.

Cost volume	Sintel (train, clean)			Sintel (train, final)			KITTI (train)		
	EPE	EPE (x)	EPE (y)	EPE	EPE (x)	EPE (y)	EPE	EPE (x)	EPE (y)
y attn, x corr	3.10	1.66	2.12	4.59	2.76	2.92	10.39	7.87	5.15
x attn, y corr	4.05	3.55	1.13	5.66	4.75	2.03	14.37	13.71	2.84
concat both	**1.98**	**1.48**	**0.94**	**3.27**	**2.35**	**1.73**	**6.69**	**6.00**	**2.16**

Table 1: Analysis on horizontal (x) and vertical (y) cost volumes. EPE (x) and EPE (y) represent the end-point-error of the horizontal and vertical flow component, respectively.

B. More Results on 4K Resolution

We provide additional visual results on 4K resolution (2160×3840) images from DAVIS dataset in Fig. 1, 2, 3, 4, and real-world scenes captured by a mobile phone in Fig. 5, 6, 7, 8.

C. Visual Results on Sintel

We further show the visual comparison results with PWC-Net+ [2] and MaskFlowNet [4] on Sintel test set in Fig. 9.

D. Implementation Details

We use the same dataset schedule and hyper-parameters as RAFT [3] when training on FlyingChairs and FlyingThings3D datasets. For Sintel, we mix FlyingThings3D, KITTI 2015, HD1K and Sintel training set for additional fine-tuning. We random crop 368×960 resolutions as input and train for 100K iterations with a batch size of 6. For KITTI, we perform additional fine-tuning on KITTI 2015 training set for 50K iterations with a batch size of 6. The random crop size is 320×1024.

*Work primarily done while interning at MSRA mentored by JY

For training on very high-resolution images, we mix FlyingThings3D, Sintel, HD1K and Slow Flow [1] datasets for additional fine-tuning from Sintel weights. The Slow Flow dataset is created with high-speed camera and optimization is used to produce the 'pseudo ground truth' flow. The resolutions of this dataset include 720×1280, 1024×1280 and 576×1024. 3448 image pairs in this dataset are used for training. To help our method generalize on 4K resolution images, we use larger crop size for training. Specifically, we random crop images in every mini-batch so that the resolutions are uniformly distributed between 640×1080 and 896×1792. For training images that are smaller than the crop size, we upsample them to the desired resolution, and the ground truth flow is upsampled accordingly. We train for 150K iterations with a batch size 2. All training is conducted on a single 32G V100 GPU.

References

[1] Joel Janai, Fatma Guney, Jonas Wulff, Michael J Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3597–3607, 2017. 2

[2] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1408–1423, 2019. 1, 11

[3] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1

[4] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020. 1, 11

Figure 1: Optical flow prediction results on 4K resolution (2160×3840) images from DAVIS dataset.

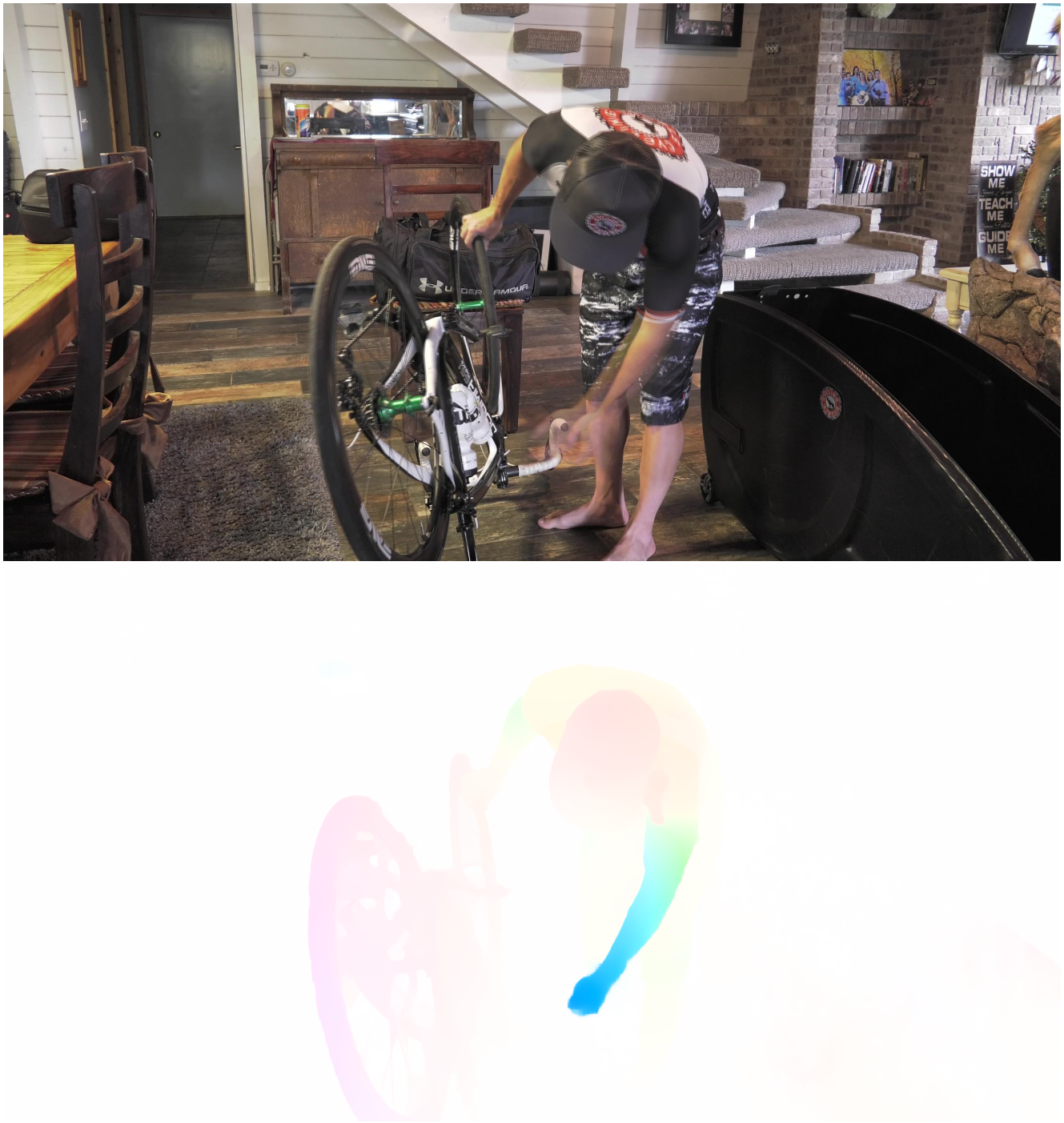

Figure 2: Optical flow prediction results on 4K resolution (2160×3840) images from DAVIS dataset.

Figure 3: Optical flow prediction results on 4K resolution (2160×3840) images from DAVIS dataset.

Figure 4: Optical flow prediction results on 4K resolution (2160×3840) images from DAVIS dataset.

Figure 5: Optical flow prediction results on real-world 4K resolution (2160×3840) images captured by a mobile phone.

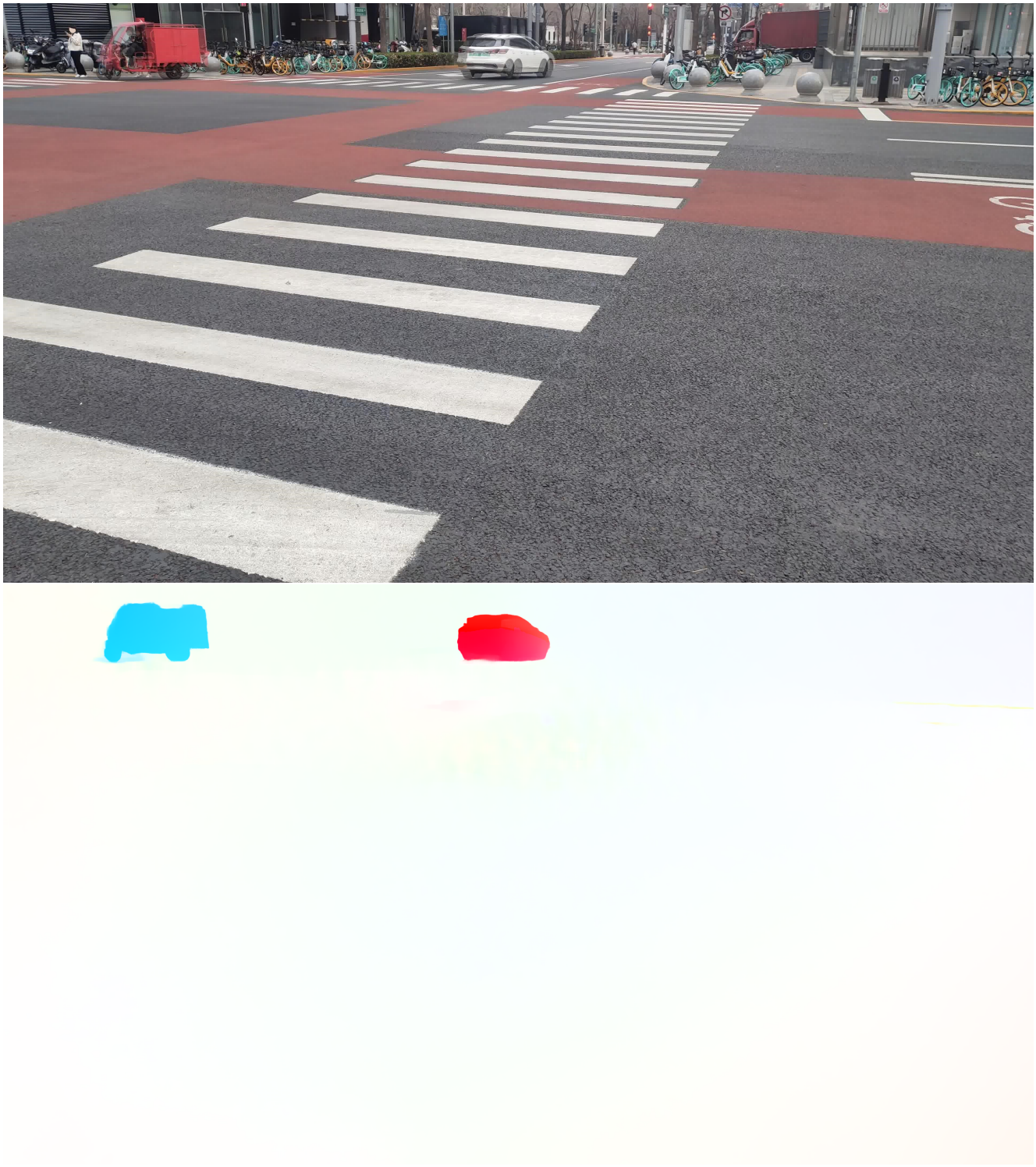

Figure 6: Optical flow prediction results on real-world 4K resolution (2160×3840) images captured by a mobile phone.

Figure 7: Optical flow prediction results on real-world 4K resolution (2160×3840) images captured by a mobile phone.

Figure 8: Optical flow prediction results on real-world 4K resolution (2160×3840) images captured by a mobile phone.

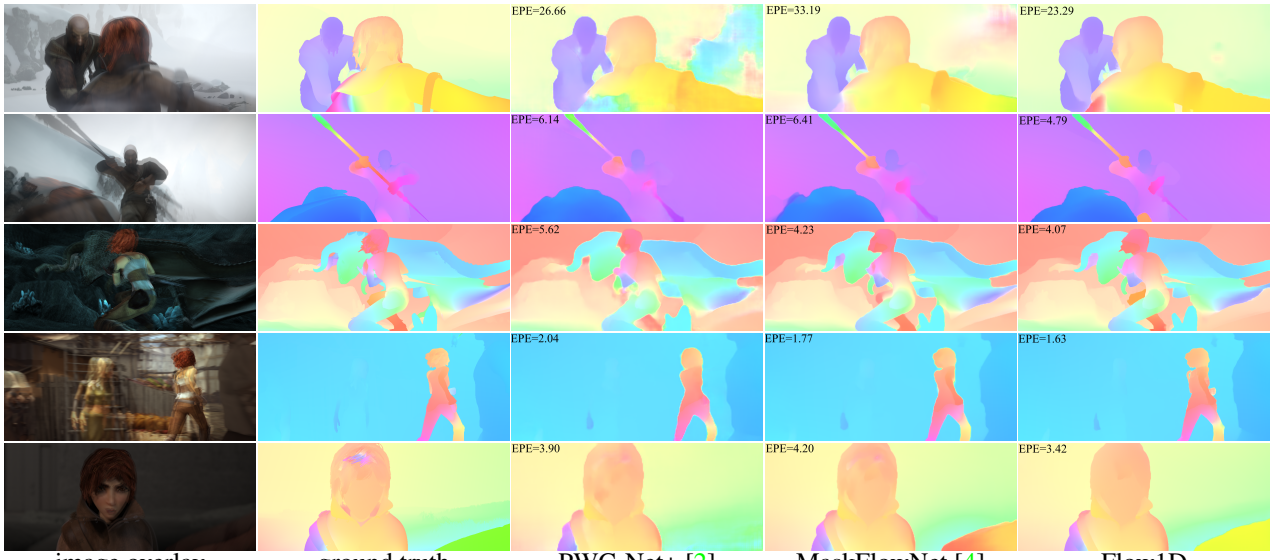

Figure 9: Visual results on Sintel test set.