Depth Estimation by Combining Binocular Stereo and Monocular Structured-Light (Supplementary Material)

A1. Implementation Details

We implement the models in PyTorch on two NVIDIA RTX 2080Ti GPU. We pre-train the models on FlyingThings3D cleanpass dataset for 65k iterations, then finetune the model by mixing the FlyingThings3D dataset and IRS dataset for another 170k iterations. The crop size of the training dataset is set to 512×256. In addition, following CRL [3], the stereo pairs with more than 25% of their disparity values larger than 300 are removed.

All modules are initialized from scratch with random weights. During training, we use the AdamW optimizer [2] with one-cycle learning rate scheduler [4]. For PSMNet, the batch size is set to 6, while the maximum learning rate is set to 1e-3, and mixed precision is used for training to avoid out Of memory (OOM). We train RAFT with 12 disparity-field updates, and 20 for evaluation. The batch size is set to 16, while the maximum learning rate is set to 4e-4.

We perform photometric augmentation by randomly perturbing brightness, contrast, saturation, and Gaussian blur. Each augmentation module has a 50% chance performed to each of the images independently. We also perform spatial augmentation by randomly rescaling in the range [0.53,1], and y-disparity augmentation [5], both with probability of 50%. In order to make the proposed system compatible with both indoor and outdoor scenes, there is 50% chance of using external guidance for the stereo networks during training.

Calibration of the MSL system: In the original Fig. 3, we assume the reference plane is perpendicular to the axis of the camera. Under this assumption, Eq. (1) holds. In practice, we fix the 3D camera with a clamp and adjust the clamp so that its surface is parallel to a white wall with a laser rangefinder.

A2. Supplementary Results

In Figure 1, the point clouds of the person in the original Fig. 8 are shown.

In Figure 2, we show two IR images Kinect in indoor scene and outdoor scene respectively. In outdoor scene, the projected speckles are seriously interfered by sun light, which leads to unstable depth estimation.

In Figure 3, we show the qualitative comparison between the proposed fusion method and the passive stereo method. Obviously, the proposed method can obtain higher quality disparity map.

In Figure 4, the results of MSG [1] (a depth completion method) are also shown (with RGB and MSL depth as input).

In experiment, we find that increasing the number of guidance points does not improve the accuracy, as shown in Table 1. However, if the guidance is sampled from the ground truth disparity maps, the smaller errors can be obtained, as shown in Table 2.

More qualitative comparisons in outdoor scenes are shown in Figure 6.

References

[1] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2020. 1, 4

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[3] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. 1

[4] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. 1

[5] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Con-*

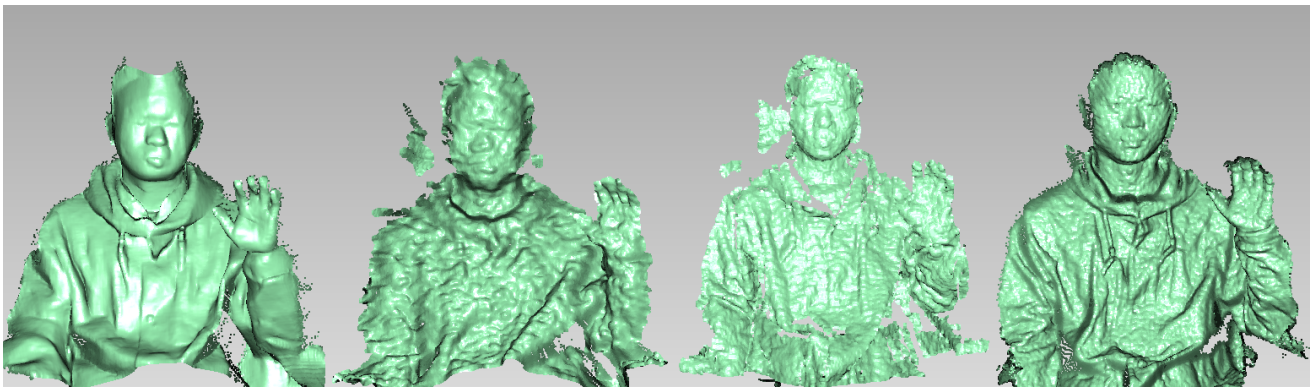

Figure 1. Point-cloud visualization. From left to right: our system, Intel D435, Kinect V1 and Kinect Azure (TOF camera). The distance from the target to the camera is about 65cm

(a) Indoor (b) Outdoor

Figure 2. IR images of Kinect in indoor and outdoor scenes. In outdoor scene, the projected speckles are seriously interfered by sun light, which leads to unstable depth estimation.

ference on Computer Vision and Pattern Recognition, pages 5515–5524, 2019. 1

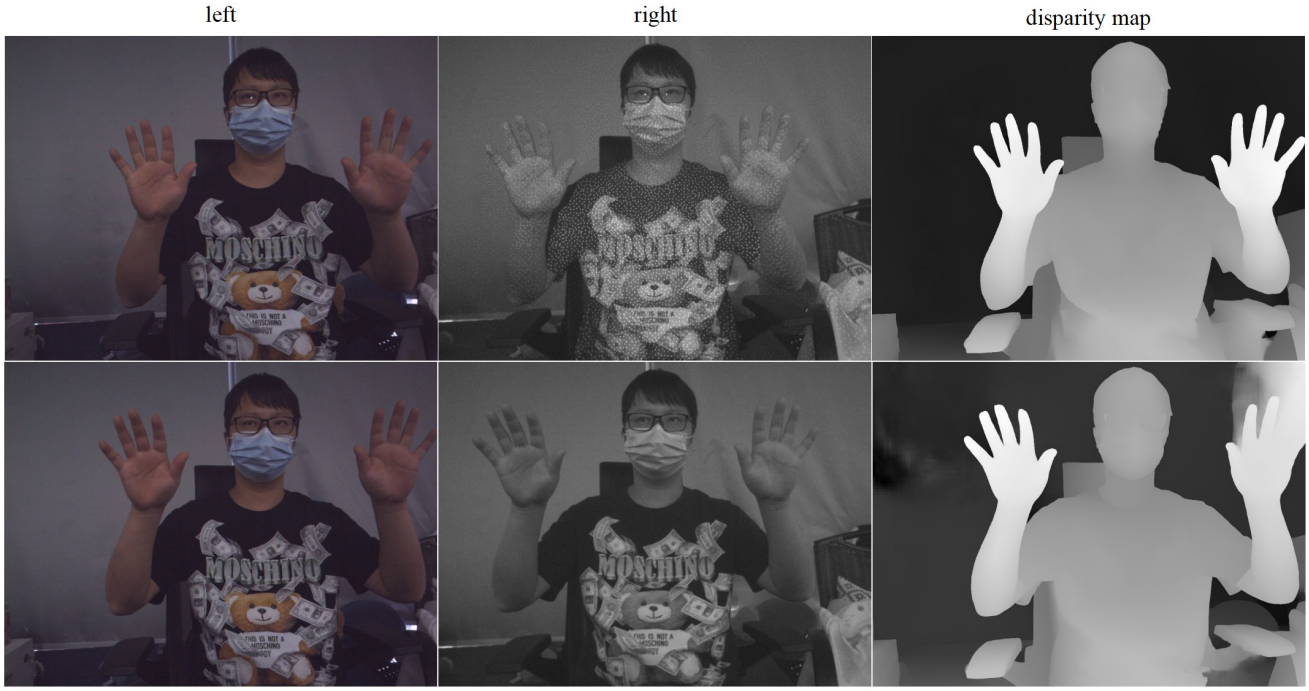

Figure 3. Qualitative comparison of our fusion method and the passive stereo method. The first column shows the left images, the second column shows the right images, and the last column shows the disparity maps. The first row shows the results of our fusion method (RAFT-OM-G is used), and the second row shows the results of the passive stereo method (RAFT-O is used).

Sampling rate (%)	EPE	Bad0.5 (%)	Bad1.0 (%)	Bad2.0 (%)
10	**0.811**	**45.13**	**16.08**	**3.59**
20	0.856	46.75	16.18	3.77
30	0.836	46.65	15.91	3.62
50	0.865	47.89	17.33	3.86

Table 1. Errors of RAFT-OM-G with different sampling rate of guidance pixels, where the guidance is sampled from the depth maps generated by the monocular structured light subsystem. The same sampling rate is used for both training and testing of the network.

Sampling rate (%)	EPE	Bad0.5 (%)	Bad1.0 (%)	Bad2.0 (%)
10	0.448	12.94	4.94	2.00
20	0.425	13.06	4.89	1.88
30	0.413	13.25	4.97	1.87
50	0.406	12.89	4.90	1.79

Table 2. Errors of RAFT-OM-G with different sampling rate of guidance pixels, where the guidance is sampled from the ground truth.

Figure 4. Comparison with MSG [1].

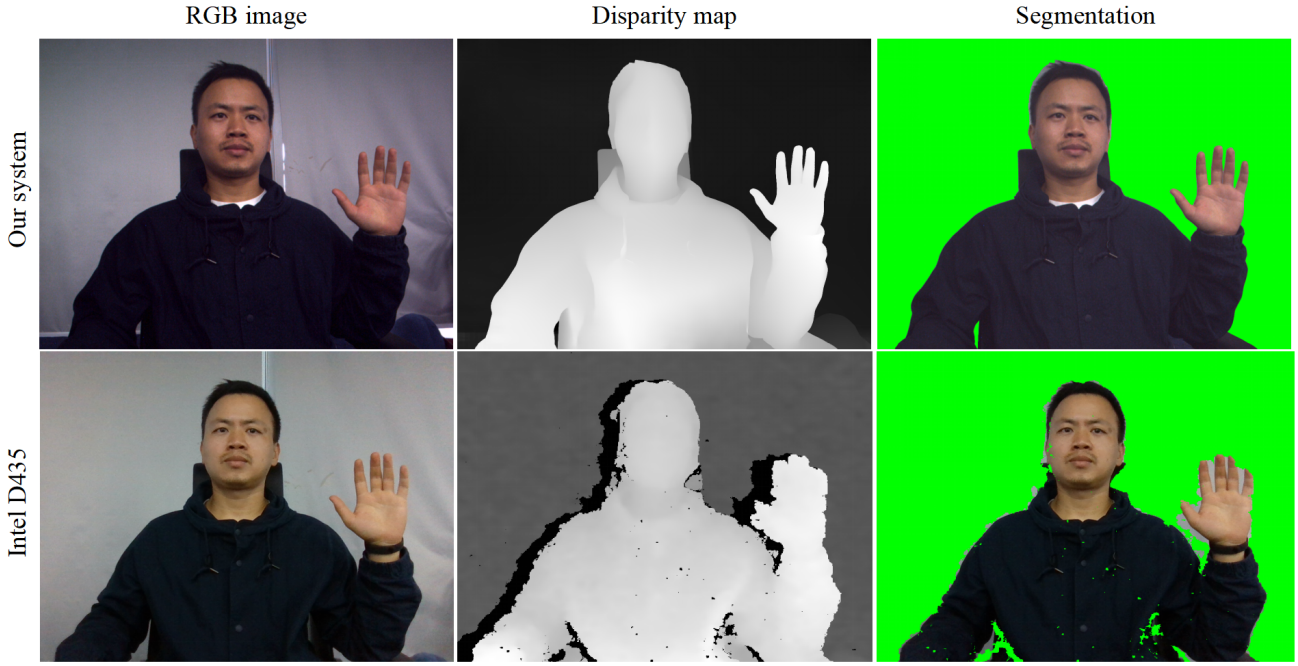

Figure 5. Human segmentation using depth information.

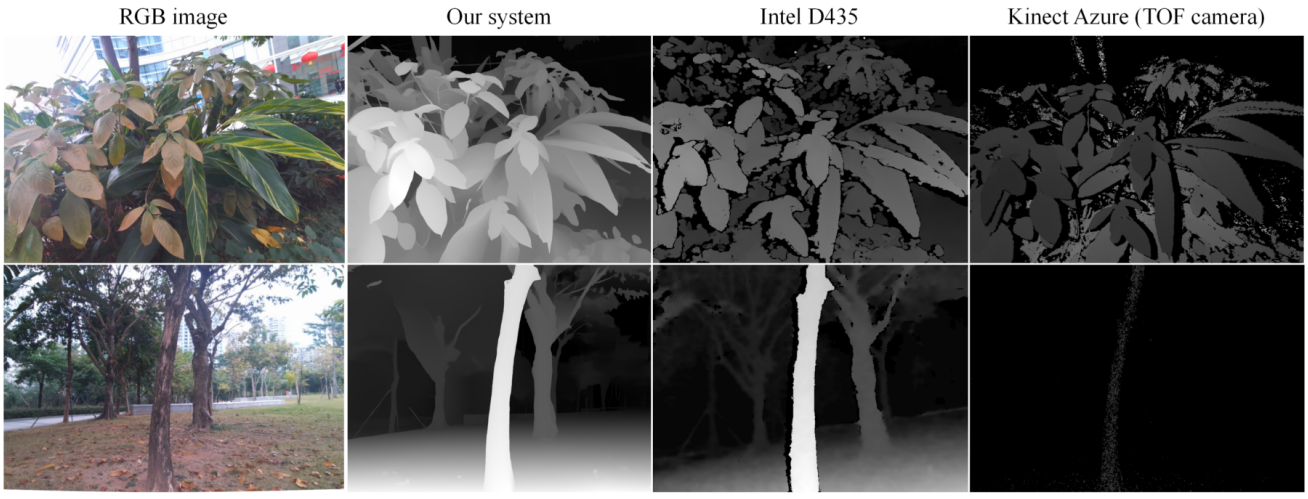

Figure 6. Qualitative comparison in outdoor scenes. Note that, the results here and in Fig. 1 of our system are generated with the same network model (RAFT-OM-G).