Supplementary Material for Seeing Farther Than Supervision: Self-supervised Depth Completion in Challenging Environments

A Network Architecture

The proposed method consists of DCNet and FlowNet. Both networks use almost the same network architecture as in [1]. For DCNet, we employ ResNet-18 [2] as an encoder and DispNet [3] as a decoder. Since the input is an RGB-D image, the input channel of the first convolution layer is changed to 4. We employ PWCNet [4] for FlowNet. Similar to DCNet, the input channels of the first layer of FlowNet are 4.

B Loss Functions

As described Sec.2.2, the loss function for FlowNet consists of three terms. The first term is the photometric loss $\mathcal{L}_{fp}$. Using input image I and warped image I', this is defined as follows:

$$\mathcal{L}_{fp} = \alpha \frac{1 - \mathrm{SSIM}(I, I')}{2} + (1 - \alpha)|I - I'|_1 \quad (1)$$

where SSIM is the structural similarity [5]. We set α to 0.85 in our experiments. The second term is the flow smoothness loss $\mathcal{L}_{fs}$. Using optical flow F, this is defined as follows:

$$\mathcal{L}_{fs} = \sum_p |\nabla F(p)| \cdot (\mathrm{e}^{|\nabla I(p)|})^T \quad (2)$$

where p is a pixel in image I. The third term is the forward-backward flow consistency loss $\mathcal{L}_{fc}$. Let $\Delta F(p_t)$ be the flow difference computed by forward-backward consistency check at pixel p_t in I. The forward-backward flow consistency loss is calculated as follows:

$$\mathcal{L}_{fc} = \sum_{p_t} \delta(p_t) \cdot |\Delta F(p_t)|_1 \quad (3)$$

where

$$\delta(p_t) = \begin{cases} 1 & (|\Delta F(p_t)|_2 < \max\{\alpha, \beta |\Delta F(p_t)|_2\}) \\ 0 & (otherwise) \end{cases} \quad (4)$$

We set (α, β) to $(3.0, 0.05)$ in our experiments.

As described in Sec. 2.3, the loss function for DCNet consists of five terms. Here, we will explain three losses: photometric loss, depth smoothness loss, and dense reprojection loss. The photometric loss is same as Eq. 1. The depth smoothness loss is the same as in Eq. 2, replacing the optical flow with the depth.

$$\mathcal{L}_{ds} = \sum_p |\nabla D(p)| \cdot (\mathrm{e}^{|\nabla I(p)|})^T \quad (5)$$

We employ the dense reprojection loss proposed by Zhao et al. [1]. It is computed using depth and optical flow estimates. Let two images be I_a and I_b, the corresponding depth estimates be D_a and D_b, camera pose be $T_{a\to b}$, and optical flow be $F_{a\to b}$. The reprojection loss based on optical flow $F_{a\to b}$ is defined as follows:

$$p_{bd} = \phi(K T_{a\to b} D_a(p_a) K^{-1} [p_a 1]^T) \quad (6)$$

$$p_{bf} = p_a + F_{a\to b}(p_a) \quad (7)$$

$$\mathcal{L}_{pf} = \frac{1}{|M_r|} \sum_{p_a} M_r(p_a) |p_{bd} - p_{bf}| + |D_{epi}| \quad (8)$$

where p_a is a pixel in image I_a, K is a camera intrinsic matrix, ϕ is a transformation from homogeneous to Euclidean space, M_r is an inlier score map, and D_{epi} is a distance map from each pixel to the corresponding epipolar line. The depth reprojection is defined as follows:

$$\mathcal{L}_{pd} = \frac{1}{|M_o M_r|} \sum_{p_a} M_o(p_a M_r(p_a) |1 - \frac{D_b^a(p_{bd})}{D_b^s}| \quad (9)$$

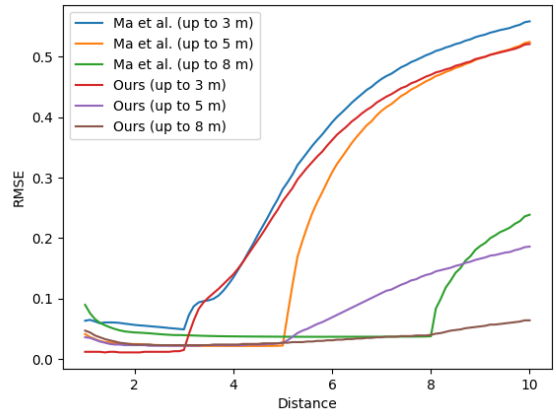
where D_b^a is the reprojected depth map by D_a and $T_{a\to b}$, D_b^s is the interpolated depth map of D_b, and M_o is the occlusion mask from optical flow. Finally, the dense reprojection loss is formulated as follows:

$$\mathcal{L}_{dr} = \lambda_1 \mathcal{L}_{pf} + \lambda_2 \mathcal{L}_{pd} \quad (10)$$

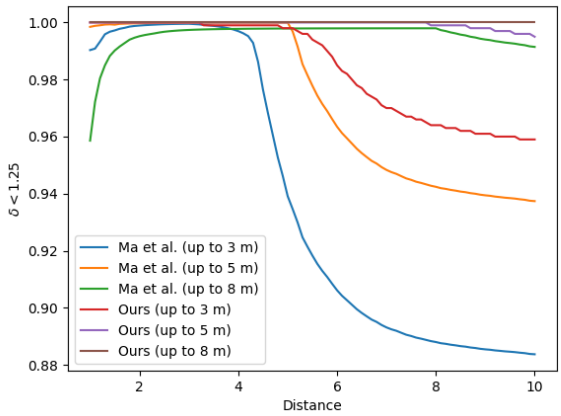
C Comparison with Existing Methods

We provide a quantitative comparison of the proposed method with existing methods in Sec.4.2. Since this work aims to complement long distances that sensors cannot acquire, we verify the effectiveness of the proposed method at such distances. Fig. 1 shows the performance at different criteria ranges, i.e., from the front of the camera to a specific distance. For most of the distance ranges, the proposed method outperforms the previous self-supervised depth completion method [6]. Both methods are very accurate up to the input depth of the sensor. When the measurable range of the sensor is exceeded, both methods have larger errors and lower accuracy. The results show that the proposed method has less performance degradation over distance than the conventional method.

We provide more qualitative results on the NYU dataset [7] in Fig. 2.



(a) RMSE



(b) Accuracy ($\delta < 1.25$)

Figure 1: Performance at different criteria ranges.

References

[1] W. Zhao, S. Liu, Y. Shu, and Y. Liu: "Towards better generalization: Joint depth-pose learning without posenet," in *CVPR*, pp.9148–9158, 2020.

[2] K. He, X. Zhang, S. Ren, and J. Sun: "Deep residual learning for image recognition," in *CVPR*, pp.770–778, 2016.

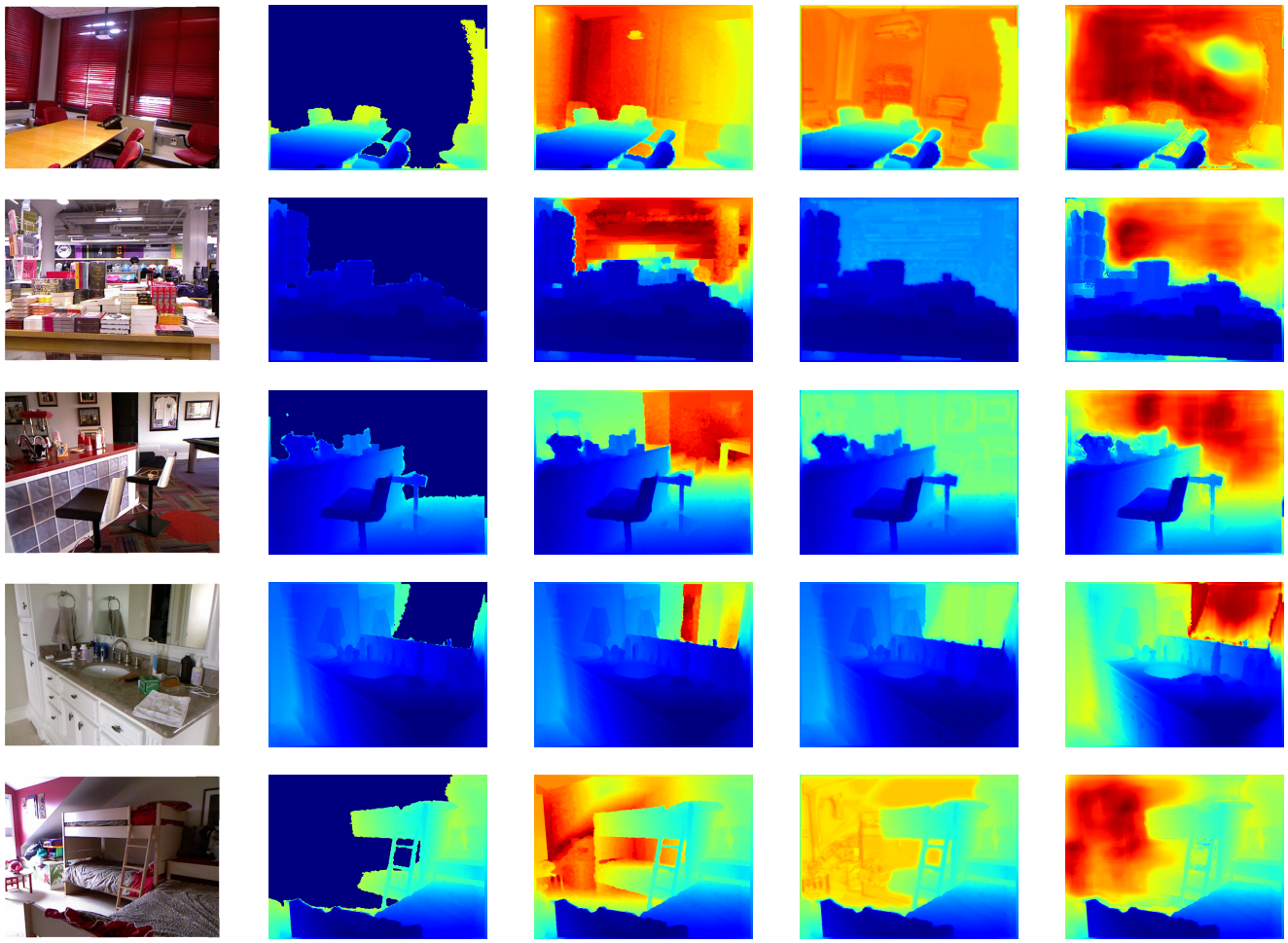
[3] C. Godard, O. M. Aodha, and G. J. Brostow: "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, pp.6602–6611, 2017.

[4] D. Sun, X. Yang, M. Liu, and J. Kautz: "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *CVPR*, pp.8934–8943, 2018.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli: "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol.13, no.4, pp.600–612, 2004.

[6] F. Ma, G. V. Cavalheiro, and S. Karaman: "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *ICRA*, pp.3288–3295, 2019.

[7] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus: "Indoor segmentation and support inference from RGBD images," in *ECCV*, pp.746–760, 2012.



(a) RGB (b) Short-range depth (c) Ground truth (d) Ma et al. [6] (e) Ours

Figure 2: Qualitative comparison of the proposed method with the recent self-supervised depth completion method [6].