Supplementary file for "G2I: Transitioning a Generalized Monocular Depth Estimation Model to In-Domain Metric Depth Prediction"

Chao Ning[1,2] Naoto Yokoya[1,2*]
[1]The University of Tokyo [2]RIKEN AIP
{6575088851,yokoya}@edu.k.u-tokyo.ac.jp,

1. Visual Comparison of In-Domain Evaluation

We visually compare our G2I with the finetuned Metric3Dv2 baseline. On the indoor NYU-Depth-v2 dataset, the qualitative results in Fig. 1 show that G2I preserves large regions of highly accurate depth predictions. For example, the wall, bed, and bookshelf contain noticeably more high-accuracy pixels than Metric3Dv2. On the outdoor KITTI dataset, the comparison in Fig. 2 demonstrates that G2I maintains accurate predictions even at long ranges, with many pixels remaining within 1 m error. Moreover, the proportion of high-accuracy pixels is consistently higher than that of Metric3Dv2. These observations suggest that simply finetuning a generalized metric depth estimation model is insufficient to correct architectural and depth-regression design limitations, whereas G2I can effectively adapt a generalized MDE model to a fixed domain and satisfy the need for high-accuracy metric depth predictions.

2. Comparison of Generalization Performance

G2I achieves state-of-the-art metric depth prediction after finetuning on a fixed dataset. In this section, we investigate whether G2I, when finetuned only on the indoor NYU-Depth-v2 dataset, can still maintain robust metric depth prediction within the indoor domain.

We compare G2I against three state-of-the-art generalized metric depth estimation models. For each method, we combine the predicted metric depth with the RGB image to obtain a 3D point cloud reconstruction; the results are shown in Fig. 3. We observe an interesting phenomenon during pseudo-label learning. Even though we only reuse the encoder of Metric3Dv2, our model can inherit the generalization ability of Metric3Dv2 from pseudo labels, and further leverage the ground-truth labels of NYU-Depth-v2 to obtain high-quality metric depth predictions in indoor scenes. For example, on the IBims-1 dataset, our zero-shot error maps clearly exhibit smaller errors than Metric3Dv2, and the 3D reconstruction results in the second row demonstrate that our depth predictions are more favorable for recovering detailed 3D structures.

On the DIODE and SUN RGB-D datasets, the 3D reconstruction results also show that, compared to DepthPro and UniDepthV2, our metric depth predictions yield overall 3D structures that are better aligned with the ground truth.

3. Evaluation Metrics

Following common practice in monocular depth estimation, we evaluate the predicted depth $\hat{d}$ against ground-truth depth d on the set of valid pixels

$$\Omega = \{i \mid d_i > d_{\min},\, d_i < d_{\max}\},$$

where $d_{\min} = 10^{-3}$ and $d_{\max} = 80$ in our experiments.

Threshold accuracy. For each valid pixel $i \in \Omega$, we define

$$\delta_i = \max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right).$$

The threshold accuracies are then given by

$$\delta_1 = \frac{1}{|\Omega|}\left|\{i \in \Omega \mid \delta_i < 1.25\}\right|. \tag{1}$$

Error metrics. We report the following depth error measures:

$$\text{AbsRel} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\frac{|d_i - \hat{d}_i|}{d_i}, \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega}(d_i - \hat{d}_i)^2}, \tag{3}$$

$$\text{RMSE}_{\log} = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega}\left(\log d_i - \log \hat{d}_i\right)^2}. \tag{4}$$

*Corresponding author

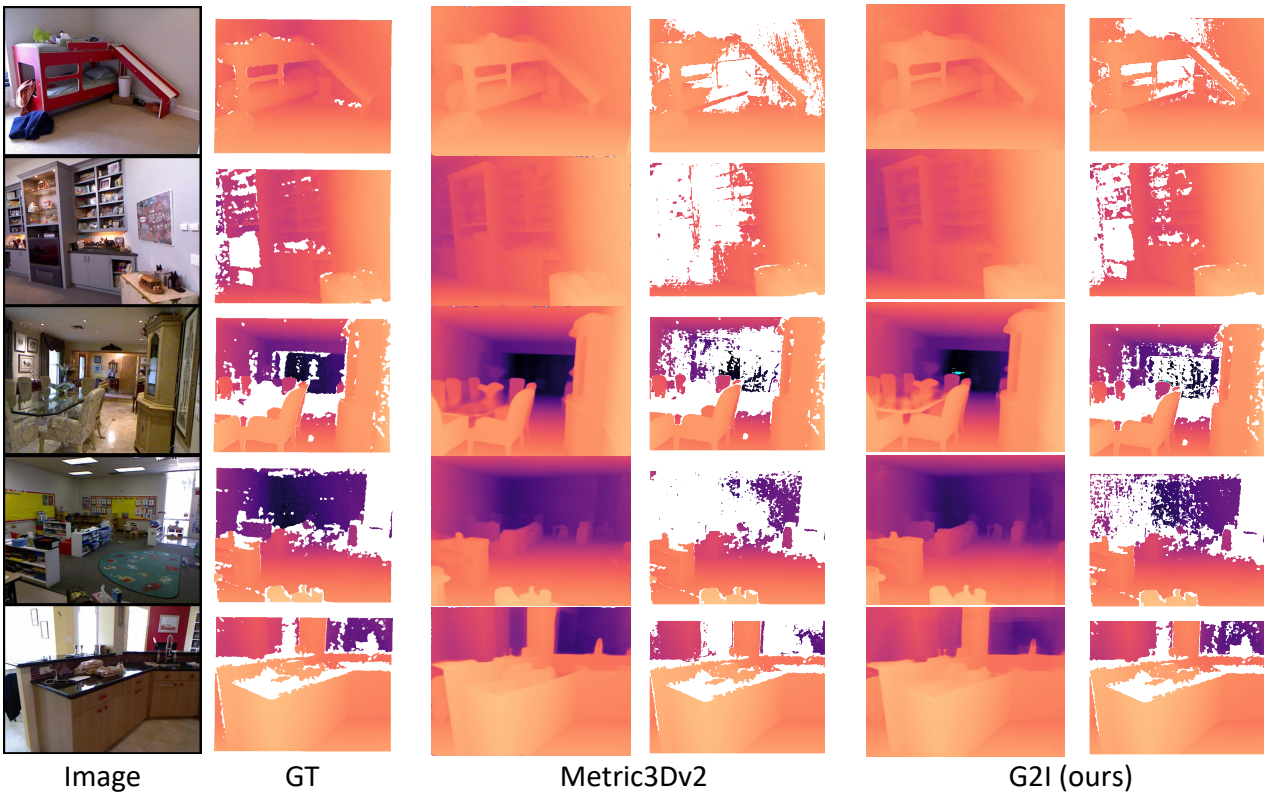

Figure 1. **Qualitative comparison on the NYU-Depth-V2 dataset.** For ease of observation, we provide predictions that have an error within 0.5m of the GT.

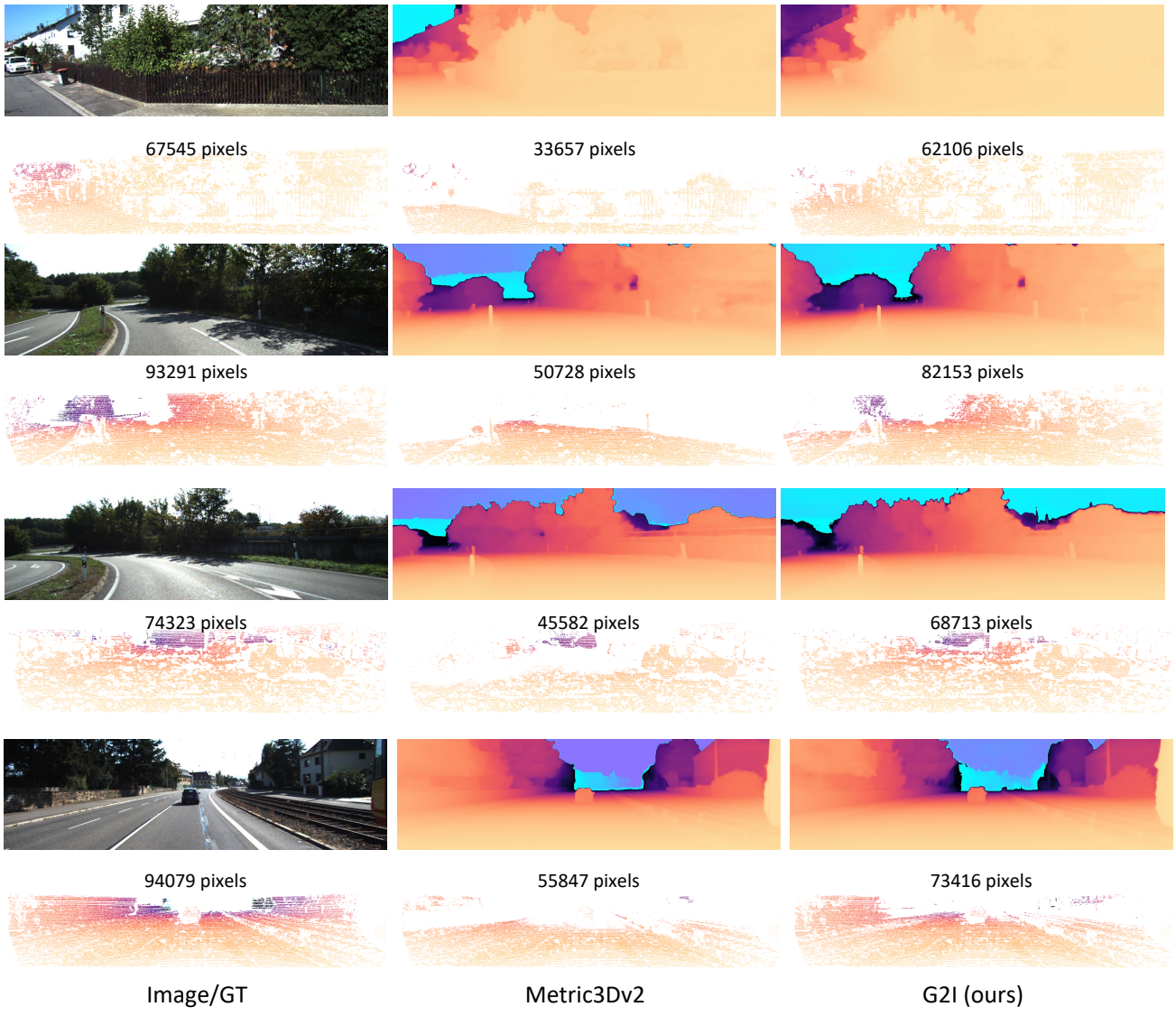

Figure 2. **Qualitative comparison on the KITTI dataset between direct in-domain finetuning and G2I.** For ease of observation, we provide predictions that have an error within 1m of the GT and record the number of valid pixels.

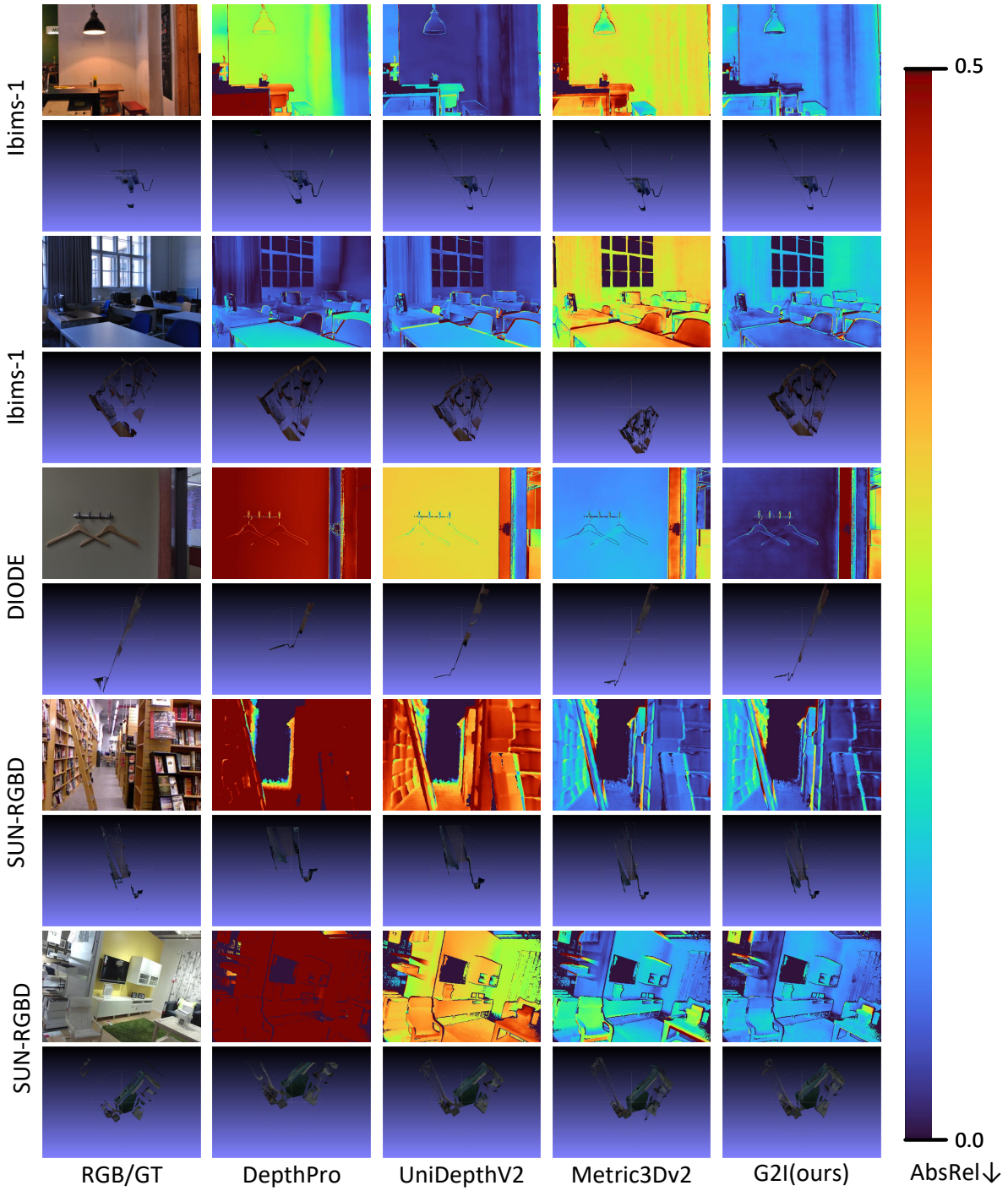

Figure 3. **Zero-shot performance comparison.** Odd rows show the RGB images and the corresponding depth prediction error maps, while even rows show the 3D reconstruction results.