Supplementary Material for Hierarchical Normalization for Robust Monocular Depth Estimation

Chi Zhang[1], **Wei Yin**[2], **Zhibin Wang**[1], **Gang Yu**[1]*, **Bin Fu**[1], **Chunhua Shen**[3]

[1]Tecent PCG, China [2]DJI Technology, China [3]Zhejiang University, China

[1] {johnczhang, brianfu, skicyyu}@tencent.com; [2] yvanwy@outlook.com; [3] Chunhua@icloud.com

A Introduction

In our supplementary material, we present more experiment results to validate the effectiveness our designs. The content is organized as follows:

- In Section B.1, we combine our hierarchical normalization strategy with another state-of-the-art monocular depth estimation method, Leres [6], which also relies on the instance-level normalization.
- In Section B.2, we conduct cross-domain experiments on synthesized and object-centric datasets.
- In Section B.3, we provide more visualization examples to qualitatively evaluate our model.

B Experiments and Results

B.1 Combination with Leres

In the main body of our paper, the proposed HDN is developed based on the scale-and-shift invariant (SSI) loss in Midas [3]. Here we combine our HDN with another loss proposed in Leres [6], where the main difference is the normalization strategy. We reuse the notations in Section 3.2 of our paper. The normalization operations $\mathcal{N}_{u_i}(d_i)$ and $\mathcal{N}_{u_i}(d_i^*)$ in Leres[6] are defined as follows:

$$\mathcal{N}_{u_i}(d_i) = \frac{d_i - \text{median}_{u_i}(\mathbf{d})}{\text{std}_{u_i}(\mathbf{d})}, \;\; \mathcal{N}_{u_i}(d_i^*) = \frac{d_i^* - \text{median}_{u_i}(\mathbf{d}^*)}{\text{std}_{u_i}(\mathbf{d}^*)} \tag{A}$$

, where mean_{u_i} and std_{u_i} operators compute the mean and the standard deviation of trimmed depth representations in u_i, with the neatest and farthest 10% pixels removed. The other difference is that they adopt tanh normalization for computing the loss:

$$\mathcal{L}_i^{\text{Leres}} = |\mathcal{N}_{u_i}(d_i) - \mathcal{N}_{u_i}(d_i^*)| + |\tanh(\mathcal{N}_{u_i}(d_i)/100) - \tanh(\mathcal{N}_{u_i}(d_i^*)/100)|. \tag{B}$$

Similar to HDN with SSI, we only change the contexts for normalization while keeping the rest the same. We follow the experiment set-ups in our ablation study and the result is shown in Table A. As wen can see, our proposed HDNs still outperform the instance-level normalization baseline remarkably, which validates the effectiveness and flexibility of our proposed designs.

B.2 Cross Domain Experiments

We next evaluate different methods with cross-domain experiments, where models are tested on datasets with large domain gaps. We first evaluate models on the object-centric Replica-GSO dataset

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Norm. Method	DIODE	ETH3D	KITTI	NYU	ScanNet	Mean Improv.
			AbsRel↓			
Instance-level [6]	47.3	24.7	14.9	10.2	11.6	-
HDN-S	38.6 (−18%)	16.1 (−35%)	12.2 (−18%)	8.9 (−12%)	9.6 (−17%)	(−20%)
HDN-DR	30.5 (−35%)	16.1 (−35%)	13.8 (−7%)	9.1 (−11%)	10.7 (−7%)	(−19%)
HDN-DP	31.8 (−33%)	14.8 (−40%)	14.1 (−5%)	10.1 (−0%)	11.9 (−3%)	(−16%)

Table A: The combination of our HDNs with the instance-level normalization in Leres [6]. Our normalization strategies significantly outperform the instance-level baseline.

Norm.Method	Replica+GSO		Blended-MVS	
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
Instance	24.0	64.8	16.6	82.1
HDN-S	22.9	67.0	15.0	83.5
HDN-DR	23.6	65.4	15.4	84.4
HDN-DP	23.4	65.8	14.3	84.7

Table B: Cross-domain experiments. The proposed methods outperform the baseline on synthesized and object-centric datasets.

processed from Omnidata [2], which contains synthesized indoor scenes [4] with scanned objects [1] scattered in the scene . The challenges in this dataset are the close distance between cameras and objects and the large domain gaps between real-world training images and synthesized testing images. We also evaluate the models on Blended-MVS [5] processed from Omnidata [2], which contains rendered images of 3D models, including, cities, architectures, sculptures, and small objects. We sample 500 images from each dataset for evaluation. The comparison of different normalization methods is shown in Table B. Our proposed models outperform the baseline on both two cross-domain datasets, which shows better generalization capability of our models.

B.3 Visualization

We provide more visualization examples in this part. Fig. B provides more comparisons of the predicted depth maps, which is an extension of Fig. 4. We also visualize the predicted depth maps of web images and reconstructed point clouds based on them in Fig. A. As we can see, our learned depth estimator generalizes well to images of dynamic scenes, even including animation and art photos. Due to its powerful generalization capability, we can deploy our depth estimator to estimate unrealistic images.

References

[1] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arXiv preprint arXiv:2204.11918*, 2022.

[2] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2021.

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[4] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[5] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1790–1799, 2020.

[6] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.

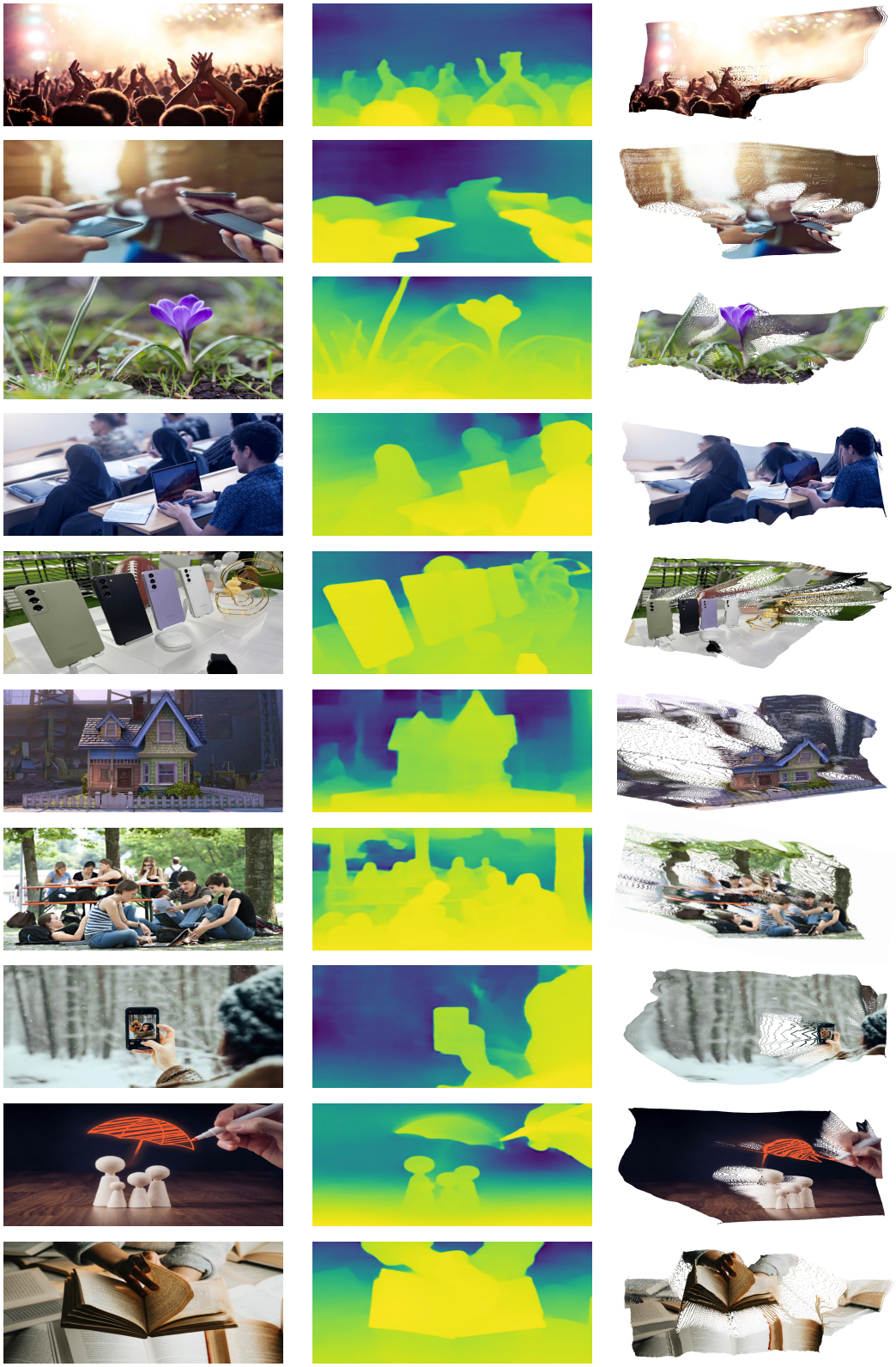

Figure A: Predictions of web images. Please zoom in to see details of depth maps and point clouds.

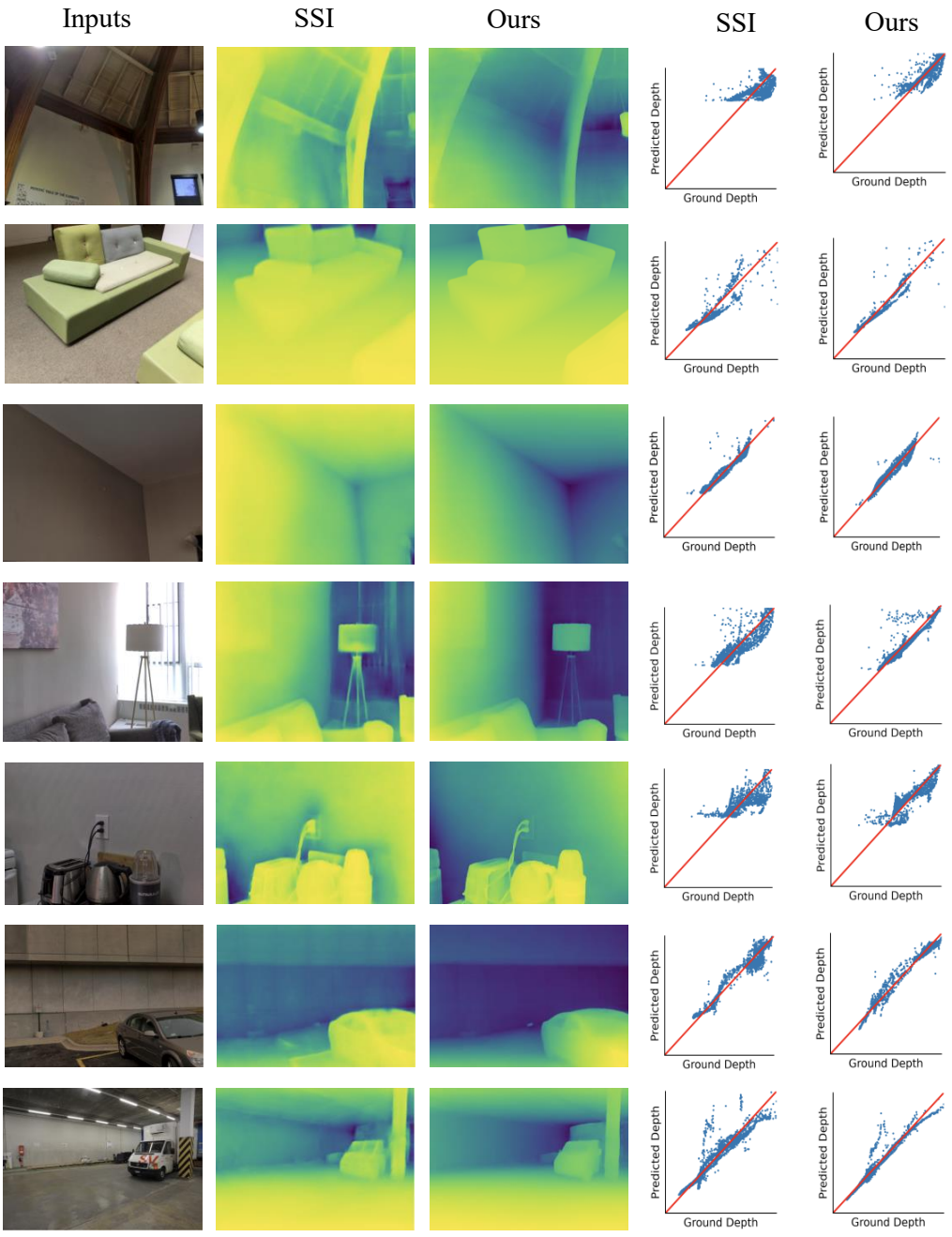

Figure B: More examples for qualitative comparisons between our method and the baseline relying on instance-level normalization (SSI [3]).