Iris: Bringing Real-World Priors into Diffusion Model for Monocular Depth Estimation

Supplementary Material

SUMMARY OF THE APPENDIX

This appendix contains additional details for CVPR2026 submission, titled *Iris: Bringing Real-World Priors into Diffusion Model for Monocular Depth Estimation*, which is organized as follows:

- §A discusses our limitations, directions of our future work, and societal impact.
- §B introduces multi-task learning for zero-shot monocular depth and normal estimation with a single model.
- §C provides more quantitative results.
- §D provides more visualizations.

A. Discussion and Outlook

A.1. Limitation and Future Work

Although Iris achieves the best overall performance among both traditional deterministic feed-forward models and diffusion-based methods, with particularly large improvements on outdoor and mixed benchmarks (*e.g.,* KITTI [4], ETH3D [9]), the gains on certain indoor datasets (*e.g.,* NYUv2 [10]) are more modest. We attribute this gap primarily to the distribution of our data source the SA-1B subset used for distillation is dominated by outdoor scenes and contains relatively few indoor scenes. As part of future work, we plan to incorporate more indoor real-image datasets and increase the scene diversity of our real-image supervision, in order to further enhance the scene generalization capability of Iris.

A.2. Social Impact

Iris is a generic monocular depth estimation framework that can be used as a building block in many 3D perception systems. By providing accurate and robust depth from a single RGB image, Iris can substantially lower the hardware and annotation cost of 3D perception. This has the potential to broaden access to 3D reconstruction in domains such as urban mapping, cultural heritage digitization, robotics, and AR/VR content creation. In robotics and autonomous navigation, improved monocular depth estimation can serve as a complementary signal to LiDAR or stereo sensors, providing redundancy in case of sensor degradation and enabling more affordable platforms that rely primarily on cameras.

However, using monocular depth in safety-critical settings such as autonomous driving also introduces risks. Rare but large depth errors, domain shift to unseen environments, or biases stemming from unbalanced training data may all lead to incorrect distance estimation and unsafe control decisions if the model is used as a primary sensor. In addition, the ability to recover dense 3D geometry from ordinary images may raise privacy concerns when applied to people or private spaces without consent.

B. Multi-task Learning

While the main paper primarily focuses on the depth estimation task, we demonstrate that *simultaneous depth and normal estimation* can be achieved with *fully shared parameters* in a single diffusion-based framework. This is realized through the integration of parameter sharing and task embedding injection. Let s_x denotes the reconstruction task switcher, and s_y denotes the switcher for dense prediction. Throughout the main text, s_y is tailored to depth estimation. However, In the context of simultaneous estimation, the dense prediction switcher s_y takes values from the set $\{s_y^{\text{depth}}, s_y^{\text{normal}}\}$, allowing the model to adapt to the specific modality. During inference, the model can seamlessly transition between depth estimation and normal prediction solely by toggling the switcher s_y. See the Fig. S2 and Fig. S1 for visualizations.

C. More Quantitative Results

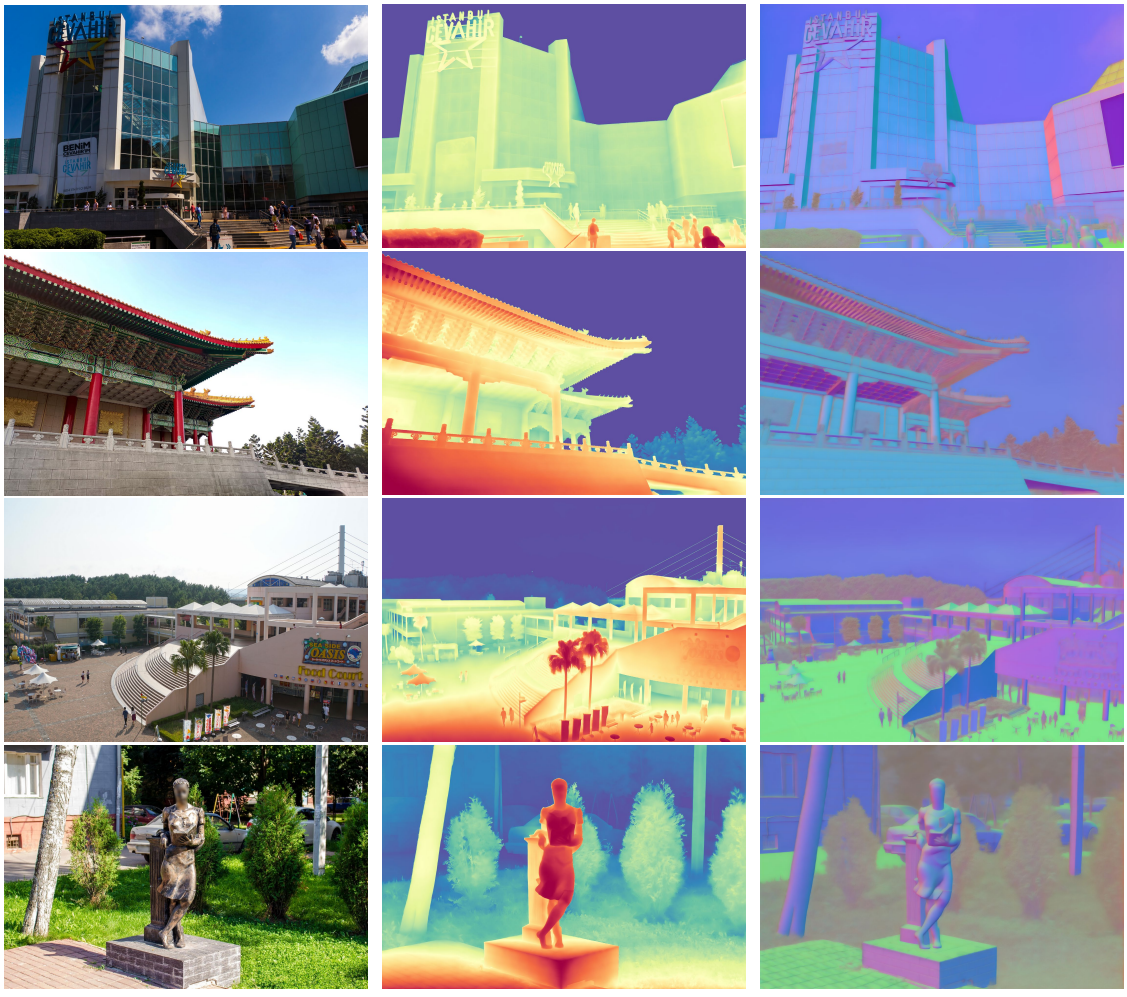
Table S1 shows the additional quantitative results on DA-2K, which is introducted by Depth Anything V2 [13]. On this benchmark, Iris outperforms all diffusion-based methods by large margins and narrows the gap to DepthAnything V2, which is trained on massive real-image corpora, demonstrating strong real-world generalization.

D. More Qualitative Results

We provide additional qualitative results on indoor scenes and paintings in Fig. S3, and on outdoor scenes in Fig. S4. Since Lotus [5] is trained only on synthetic datasets, it almost fails to produce meaningful depth on paintings (*i.e.,* Fig. S3 line 1-2). In contrast, Iris recovers plausible and detailed depth for these challenging artistic images. Across both indoor and outdoor scenes, Iris further demonstrates stronger scale awareness than Lotus, and delivers sharper object boundaries and richer fine-grained details than both Lotus and Depth Anything V2 [13].

Table S1. **Quantitative comparison on zero-shot affine-invariant depth estimation**.

Method	Training Data↓	KITTI (Outdoor)		NYUv2 (Indoor)		ETH3D (Various)		ScanNet (Indoor)		DIODE (Various)		**DA-2K**	Group Avg
		AbsRel ↓	δ_1↑	AbsRel ↓	δ_1↑	AbsRel ↓	δ_1↑	AbsRel ↓	δ_1↑	AbsRel ↓	δ_1↑	Acc(%) ↑	Ranking
DiverseDepth [14]	320K	19.0	70.4	11.7	87.5	22.8	69.4	10.9	88.2	37.6	63.1	79.3	7.5
MiDaS [7]	2M	23.6	63.0	11.1	88.5	18.4	75.2	12.1	84.6	33.2	71.5	80.6	7.2
LeRes [15]	354K	14.9	78.4	9.0	91.6	17.1	77.7	9.1	91.7	27.1	76.6	81.1	5.2
Omnidata [1]	12.2M	14.9	83.5	7.4	94.5	16.6	77.8	7.5	93.6	33.9	74.2	76.8	5.0
DPT [8]	1.4M	10.0	90.1	9.8	90.3	**7.8**	**94.6**	8.2	93.4	**18.2**	75.8	83.2	3.5
HDN [16]	300K	11.5	86.7	6.9	94.8	12.1	83.3	8.0	93.9	24.6	**78.0**	85.7	3.0
DepthAnything [12]	62.6M	7.6	**94.7**	**4.3**	**98.1**	12.7	88.2	4.3	**98.1**	26.0	75.9	88.5	**1.9**
DepthAnything V2 [13]	62.6M	**7.4**	94.6	4.5	97.9	13.1	86.5	**4.2**	97.8	26.5	73.4	**97.1**	2.5
Diffusion-E2E-FT* [3]	74K	9.6	92.1	5.4	96.5	6.4	95.9	5.8	96.5	30.3	**77.6**	83.6	4.3
GeoWizard* [2]	280K	14.4	82.0	5.6	96.3	6.6	95.8	6.4	95.0	33.5	72.3	88.1	6.8
Marigold$_{(LCM)}$* [6]	74K	9.8	91.8	6.1	95.8	6.8	95.6	6.9	94.6	30.7	77.5	86.8	6.5
Marigold* [6]	74K	9.9	91.6	5.5	96.4	6.5	95.9	6.4	95.2	30.8	77.3	85.6	5.8
Lotus-D* [5]	59K	8.1	93.1	5.1	97.2	6.1	97.0	5.5	96.5	22.8	73.8	91.2	2.7
Lotus-G* [5]	59K	8.5	92.2	5.4	96.8	5.9	97.0	5.9	95.7	22.9	72.9	88.9	3.7
GenPercept* [11]	90K	7.8	93.5	5.9	96.7	9.4	96.1	6.4	96.1	**22.8**	74.0	85.1	4.6
Iris (ours)*	59K + 100K	**7.2**	**94.5**	**4.9**	**97.4**	**5.5**	**97.6**	**5.0**	**97.1**	24.3	74.3	**94.5**	**1.5**



Input Image | **Iris (depth)** | **Iris (normal)**

Figure S1. **Visualizations of Joint depth and normal estimation.** Iris enables simultaneous depth and normal estimation with *fully shared parameters* by swapping the task switcher s_y^{depth} and s_y^{normal}. See §B for details.

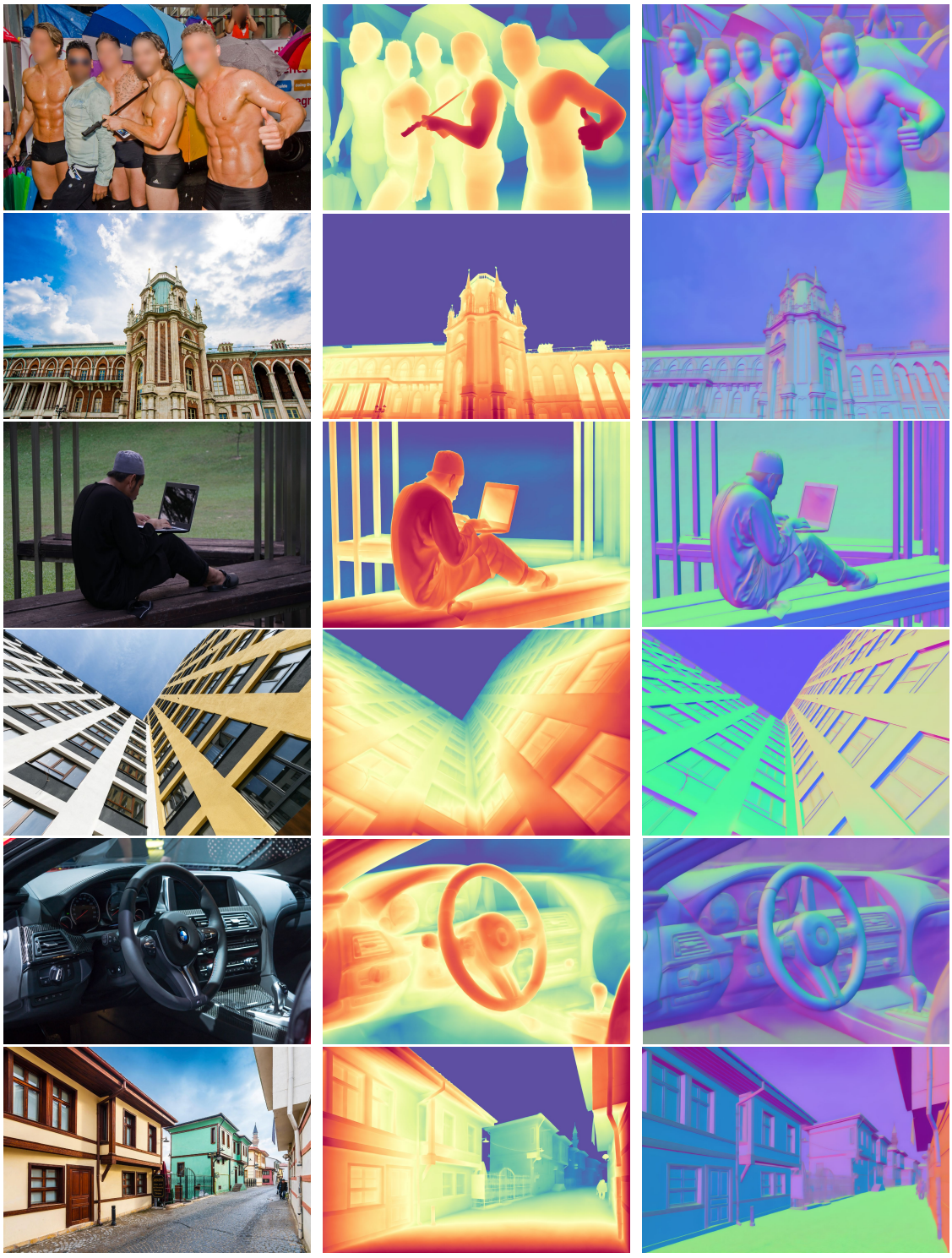


Figure S2. **Visualizations of Joint depth and normal estimation.** Iris enables simultaneous depth and normal estimation with *fully shared parameters* by swapping the task switcher s_y^{depth} and s_y^{normal}. See §B for details.

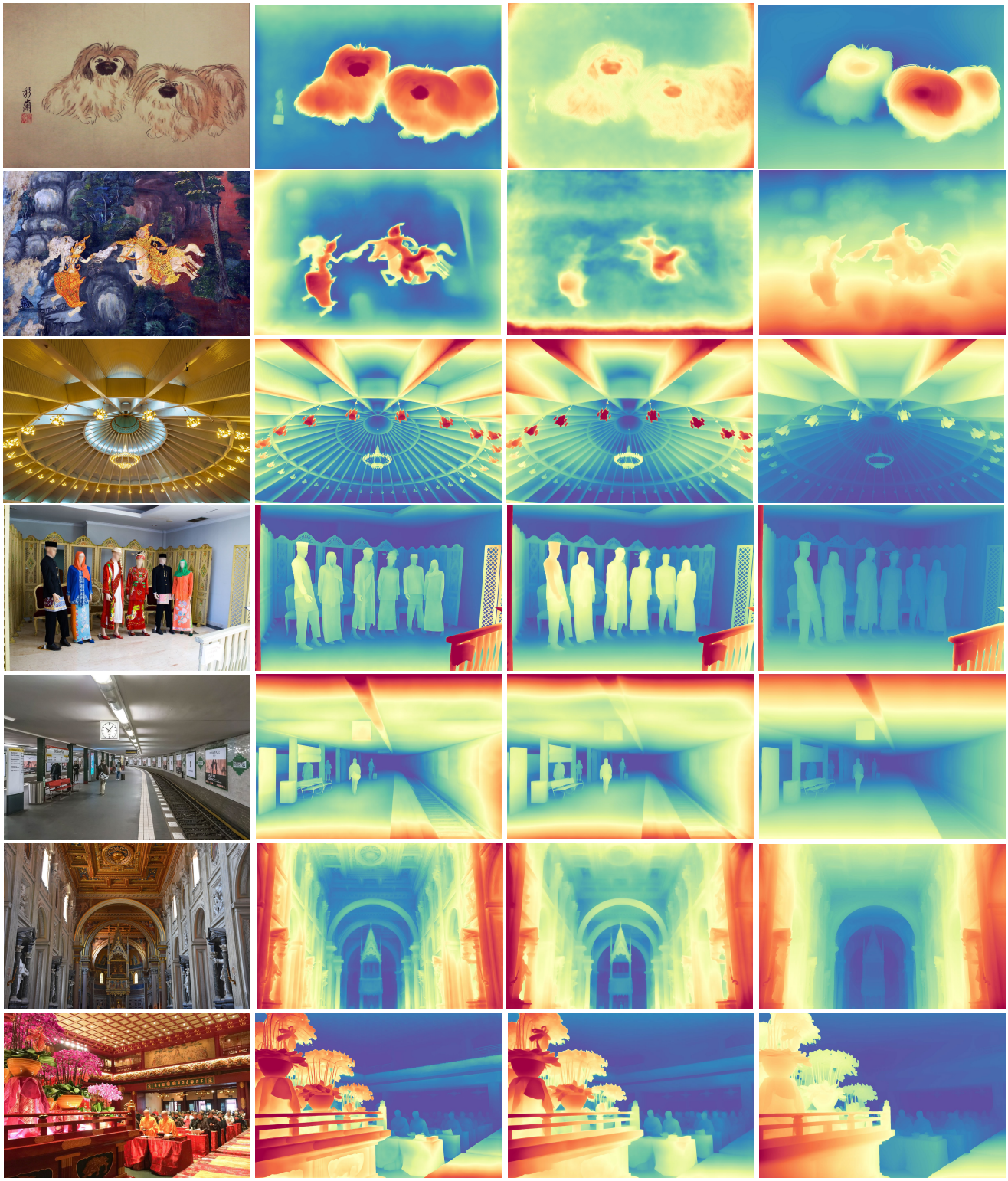
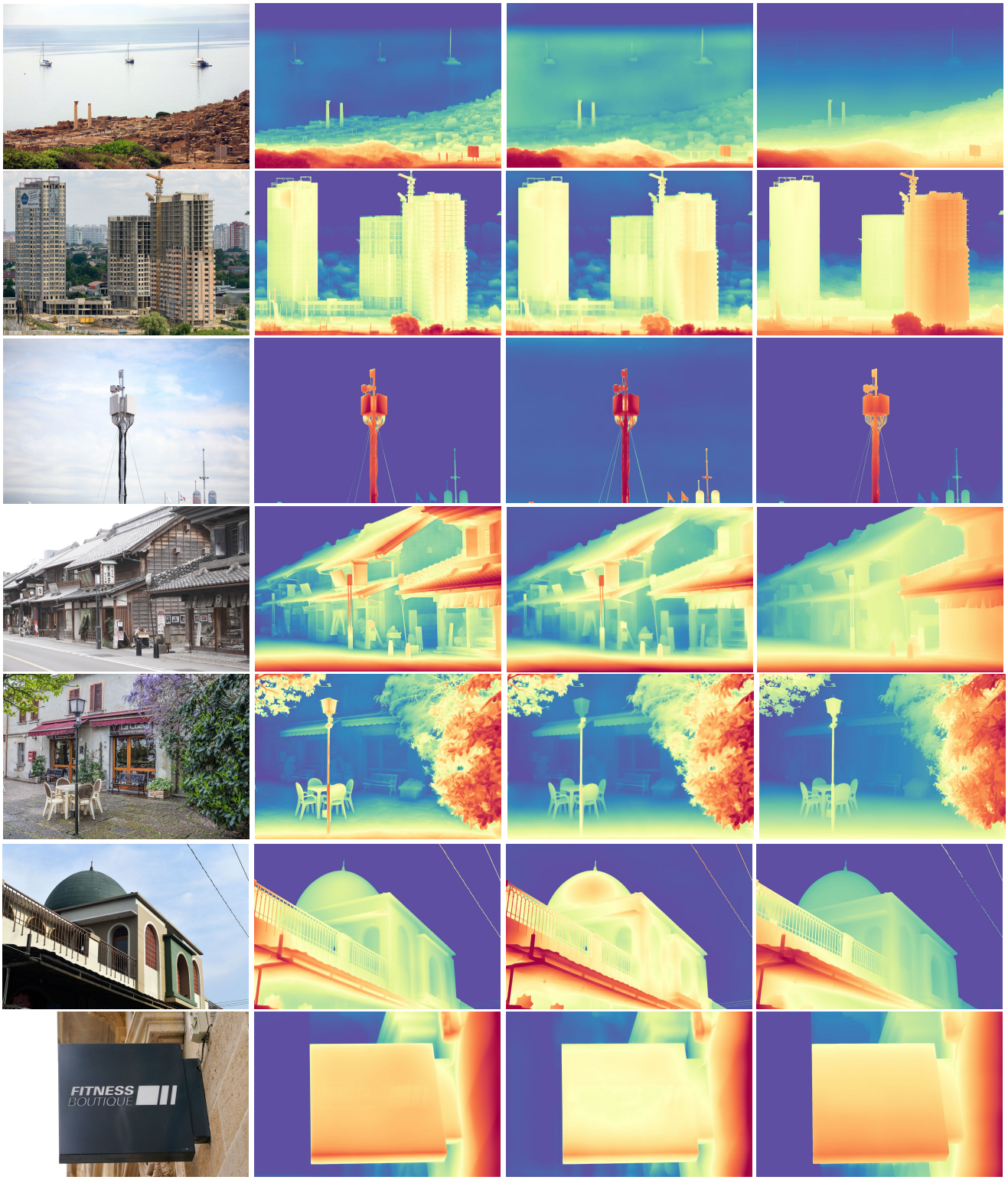


Figure S3. **More qualitative results on indoor scenes and paintings.** See §D for more details.



Input Image | **Iris (ours)** | Lotus | Depth Anything V2

Figure S4. **More qualitative results on outdoor scenes.** See §D for more details.

References

[1] 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 *ICCV*, pages 10786–10796, 2021. 2

[2] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, pages 241–258, 2024. 2

[3] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *WACV*, pages 753–762, 2025. 2

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 1

[5] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *ICLR*, 2025. 1, 2

[6] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024. 2

[7] 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 TPAMI*, 44(3):1623–1637, 2020. 2

[8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 2

[9] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 1

[10] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 1

[11] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *ICLR*, 2025. 2

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 2

[13] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, pages 21875–21911, 2024. 1, 2

[14] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2

[15] 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 *CVPR*, pages 204–213, 2021. 2

[16] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. In *NeurIPS*, pages 14128–14139, 2022. 2