A Appendix

In this supplementary, we offer more details on implementation and experiments in Appendix B and Appendix C, respectively. We also include more qualitative comparisons regarding depth and normal on zero-shot test sets and in-the-wild scenarios, 3D reconstruction, and novel view synthesis in Appendix D. Finally, we discuss limitations and potential negative impact in Appendix E.

B Implementation Details

B.1 Data Preprocessing

We standardize the resolution at 576×768 to blend samples from various scene distributions. To maintain the original aspect ratio, we resize the shorter side of a sample to 576 and randomly crop along the longer side. In the data augmentation strategy, we assign photometric distortion probabilities of 0.05, 0.1, and 0.05, and greyization probabilities of 0.1, 0.2, and 0.1 for indoor, outdoor, and object level, respectively. We set the far plane to be 80 meters in both 3D Ken Burns [12] and our own city dataset for outdoor scenes, and 5 meters in Objaverse [5] for background-free objects. We also define the normal orientation in these far (background) regions along the z-axis. In the Replica dataset [18], we exclude samples with fewer than 50 invalid pixels, designating the invalid areas to represent distant depths and background normals.

B.2 Our Synthetic Urban Dataset

We first tried to add Virtual KITTI [2] to involve more driving scenarios but ultimately decided against it, as the generated normal map is of low quality due to the limited resolution of depth map (375×1242). As an alternative, we utilize Unreal Engine to create high-resolution (1440×3840) urban samples (see Fig. S1), and derive the normal map from depth using the least square algorithm. Our synthetic data encompasses a wide variety of city entities under different environmental conditions. Since the data is clean and complete, it allows our model to learn high-quality outdoor patterns.

C Experimental Details

C.1 Limitation on Normal GT

During our zero-shot tests on traditional normal benchmarks, we discovered that a lot of normal GT maps have noise, potentially impacting measurement precision. As shown in Fig. S2, NYUv2's normal maps struggle with fine details such as book outlines, shelf edges, and folds, and even incorrectly represent the flat wall surfaces. Likewise, the normal maps from iBims-1 (limited resolution) and ScanNet (unexpected surface undulation and poor fine detail capture) are also of low quality. Thus, the quantitative comparisons presented in the main paper may only partially reflect the ground truth.

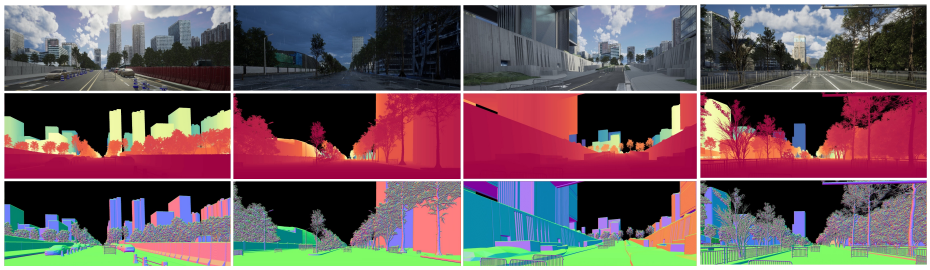

Fig. S1: Some samples of our rendered dataset. We mask the regions whose depth values are larger than 80m as black for better visualization.

C.2 More Ablation Studies

Applying Erroneous Domain Indicator When using the wrong domain-specific indicator for testing across various domains, we see a decline in both depth and normal (see Table R1), especially during zero-shot tests on indoor and outdoor benchmarks with an object indicator (w/ Object Indicator). This result makes sense since the indicator directs the model to concentrate on a specific distribution. We also observe that the geometric consistency seems to remain stable or even improved (14.7→14.4 on indoor test with an outdoor indicator), suggesting the model's adaptability and robustness when guided by an out-of-domain indicator.

Geometric Modeling We also study shared geometry embedding [11] by increasing the dimension of the input in input ('w/ Shared Geometry' in Table R1). Without the specialized geometry switcher and using the same training iterations, we observe that this alternative converges more slowly, and the overall quality of depth and normal quality both decrease (6.7→7.2, 14.8→15.3), whereas the geometric consistency remains relatively unchanged.

Model Components When removing the pyramid multi-level noise and adopting single-level noise (w/o Multi-level Noise), both the depth and normal quality decreased significantly. Compared to ϵ-pred (w/ ϵ-pred), v-pred enables better geometry results and faster convergence.

C.3 Comparison with 3D&Video Generation Methods

For object-level geometry estimation, other alternatives are to resort to per-scene optimized 3D generation methods, such as Magic123 [13] and DreamCraft3D [19], or amortized video generation methods [21]. However, they rely on multi-view images generated from diffusion models to optimize 3D representation but result in object-level 3D reconstruction with poor geometry, unstable optimization and lengthy processing. As shown in Table R2, ours, trained

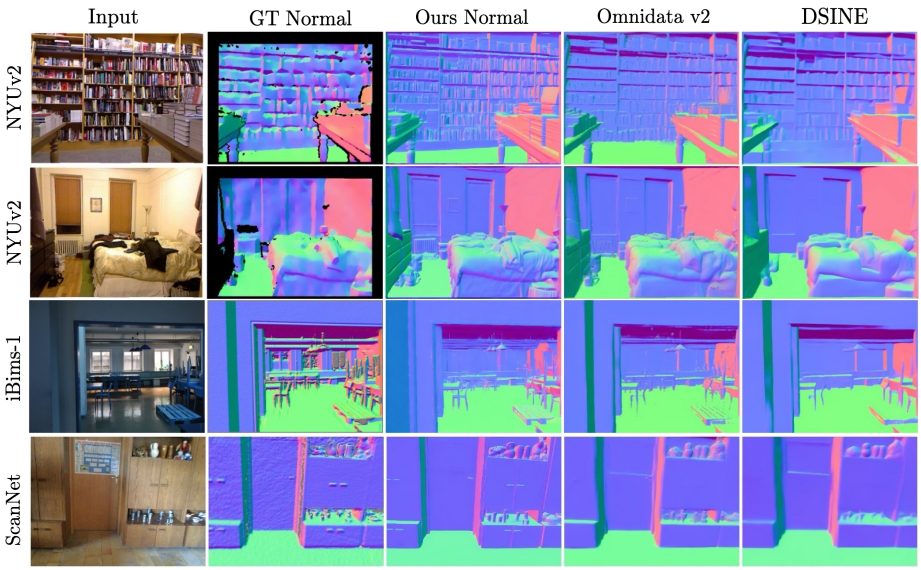

Fig. S2: Effect of noisy GT normal map. Our normal maps here display the best visual effect but are inferior in quantitative comparison with Omnidata v2 or DSINE.

Method	Indoor			Outdoor			Object			Overall		
	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓	AbsRel ↓	Mean ↓	GC ↓
w/ Indoor Indicator	**5.5**	**12.6**	14.7	10.1	22.8	23.9	3.7	15.8	17.7	6.8	15.0	16.4
w/ Outdoor Indicator	5.8	13.1	**14.4**	**9.6**	**22.1**	**23.5**	3.9	15.9	18.2	7.0	15.2	16.4
w/ Object Indicator	6.4	13.7	14.9	10.8	23.5	23.7	**3.5**	**15.4**	**17.6**	7.5	15.5	16.6
Shared Geometry [11]	6.1	13.2	14.6	10.4	23.6	23.8	3.6	16.4	17.8	7.2	15.3	16.3
w/o Multi-level Noise	7.3	13.9	15.1	10.8	24.6	24.0	4.3	16.4	18.0	8.3	17.1	16.5
w/ ϵ-pred	5.7	12.9	14.9	10.1	22.3	24.0	3.7	15.8	17.9	6.9	15.2	16.4
Full Model	**5.5**	**12.6**	14.7	**9.6**	**22.1**	**23.5**	**3.5**	**15.4**	**17.6**	**6.7**	**14.8**	**16.2**

Table R1: Quantitative ablation studies on different scene types.

on 3D data, directly produces accurate and efficient geometric representation, depth&normal. Moreover, ours can be applied beyond object scenarios, e.g., outdoor and indoor.

C.4 GeoWizard V2

We additionally train a v2-model with architecture modifications (replace image CLIP embedding with three types of text embeddings ('indoor geometry', 'outdoor geometry', and 'object geometry'). Now it can generate more realistic and three-dimensional normal maps on some rare images (e.g., cartoon style).

	Magic123 [13]	DreamCraft3D [19]	SV3D [21]+NeuS [22]	Ours
AbsRel ↓	2.9	3.4	3.7	**1.9**
$\delta 1$ ↑	25.4	23.7	27.9	**20.3**
Time Cost (min.)	54.3	45.5	10.3	**0.16**

Table R2: Quantitative comparison on Omniobject3D benchmark (50 samples).

Fig. S3: Qualitative comparison on GeoWizard v1/v2 models.

D More Qualitative Comparisons

D.1 Testset Depth and Normal

We include additional qualitative comparisons across 7 zero-shot test datasets [4, 6, 10, 15, 17, 20, 23], where our model is evaluate against Marigold [8] and DepthAnything-L [24] for depth, and agins Omnidata v2 [7] and DSINE [1] for normal. These comparisons, visualized in Fig. S4 to Fig. S10, cover both depth and normal maps. To enhance visual contrast, we initially math the inverse relative depth from DepthAnything with the inverse GT depth. Following this affine alignment, we further convert it into actual depth. Note that the GT normal maps are shown in default grey when unavailable. For outdoor scenes, the 'sky' in our normal maps is colored in pure blue [0,0,255] to denote the standard orientation [0,0,1]. In comparison on iBim-1, we mask out the erroneous parts in GT with red boxes. Overall, Geowizard consistently outperforms in generating detailed high-frequency details across all datasets, although the difference might not be as discernible in OmniObject3D due to its simplistic object structures.

D.2 In-the-Wild Depth and Normal

We collect in-the-wild images that are publicly available and allow for disclosure from the Internet, our daily life, or AI-generated pool to test the generalizability. For examples in the main paper, we carefully transform each inverse relative depth to relative depth with manually estimated scale and shift for clearer differentiation. To prevent any confusion regarding this transformation, we maintain the original color bar in the disparity depth maps in the supplementary, and this still demonstrates obvious differences in high-frequency details. As shown from Fig. S11 to Fig. S21, GeoWizard consistently produces high-fidelity details and correct spatial layout compared to baselines, i.e., Marigold and DepthAnything for depth, and Omnidata v2 and DSINE for normal.

D.3 In-the-Wild 3D Reconstruction

We provide more 3D reconstruction results as visualized in Fig. S22, comparing Ours with DSINE [1] and Omnidata v2 [7]. For a fair comparison, we exclusively use only normal maps as input for the BiNI algorithm [3]. The meshes reconstructed by GeoWizard generate enhanced high-frequency details, including hair, clothing folds, metal and wood textures, and thin handrails. Meanwhile, it delivers superior predictions of the 3D structural layout that align more closely with the original input image.

D.4 Depth-aware Novel View Synthesis

We present more novel view synthesis results as shown in Fig. S23. Our approach, GeoWizard, outperforms Midas V3.1 [14] to guide the generation of more coherent and believable structures for objects that pose challenges in monocular depth estimation, including AI generated cars, buildings with unusual shapes, slender lampposts, and white bed under sunlight. Since this method [16] takes inverse depth in pretraining, thus the manual transformation of our depth into its inverse form will cause accuracy loss. And we find the difference in the novel views generated by our model compared to DepthAnything is relatively minor.

E Limitation and Potential Negative Social Impact

GeoWizard serves as a foundation model for estimating geometry in both real-world and artificially created images. Despite its strengths, the current framework still has some limitations. First, the iterative denoising process is time-consuming when applied to large-scale collections. Since the depth and normal maps are generated from randomly initialized noise, this diffusion leads to inconsistencies when applied to video sequences. In terms of the reconstruction, the pseudo scale and shift derived from the combined depth and normal maps may exhibit accuracy issues in some cases. Meanwhile, some concerns exist when making our models publicly available. It model can be extended to create fake

but realistic 3D assets. Depth and normal maps play important roles in scene understanding, and our model could be incorporated into surveillance systems to identify regions that are not clearly distinguishable to the human eyes. To mitigate these issues, we will include stipulations in the license agreement for the code limiting its applications only to academic research.

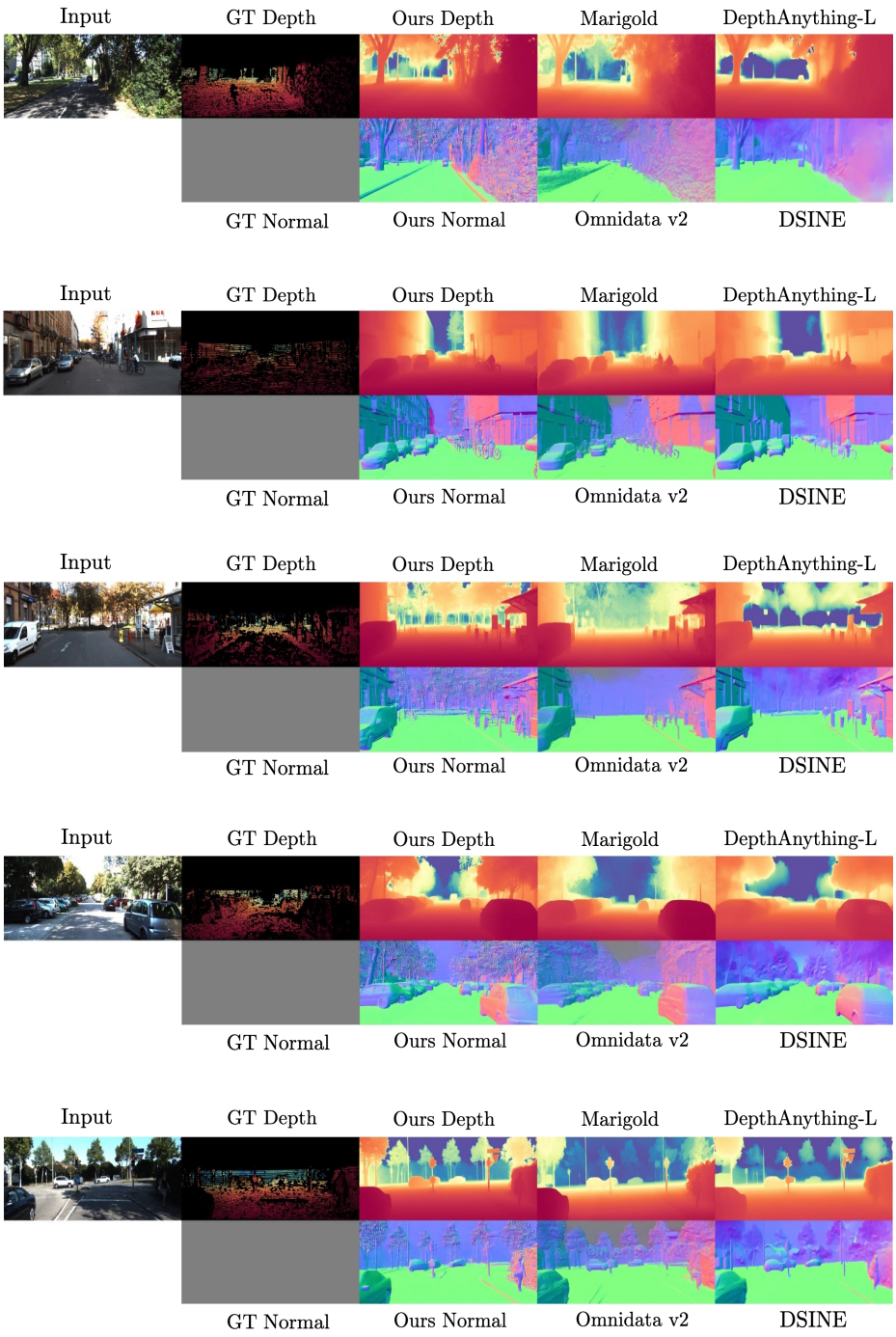

Fig. S4: Qualitative comparison on KITTI [6].

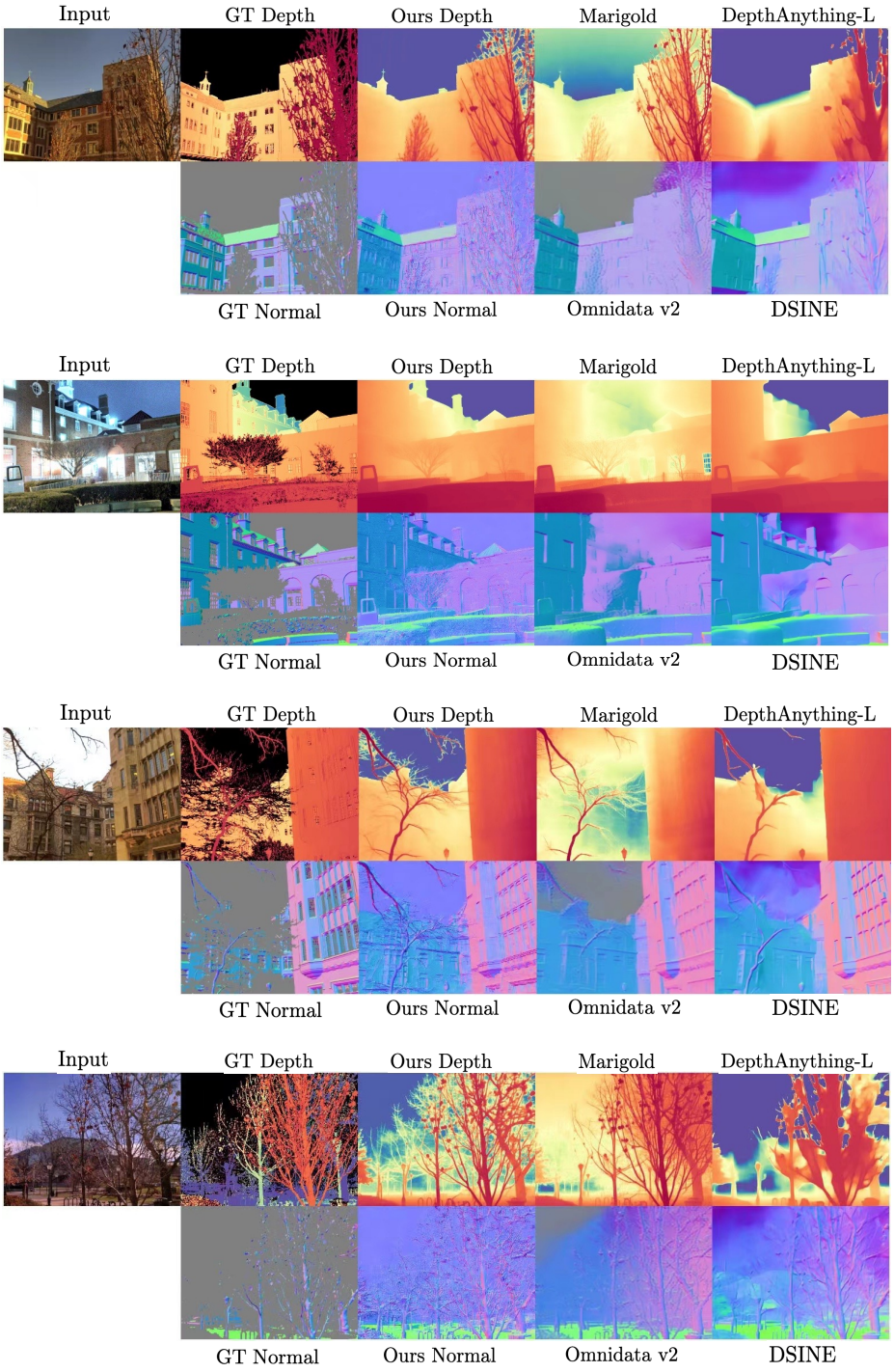

Fig. S5: Qualitative comparison on DIODIE [20].

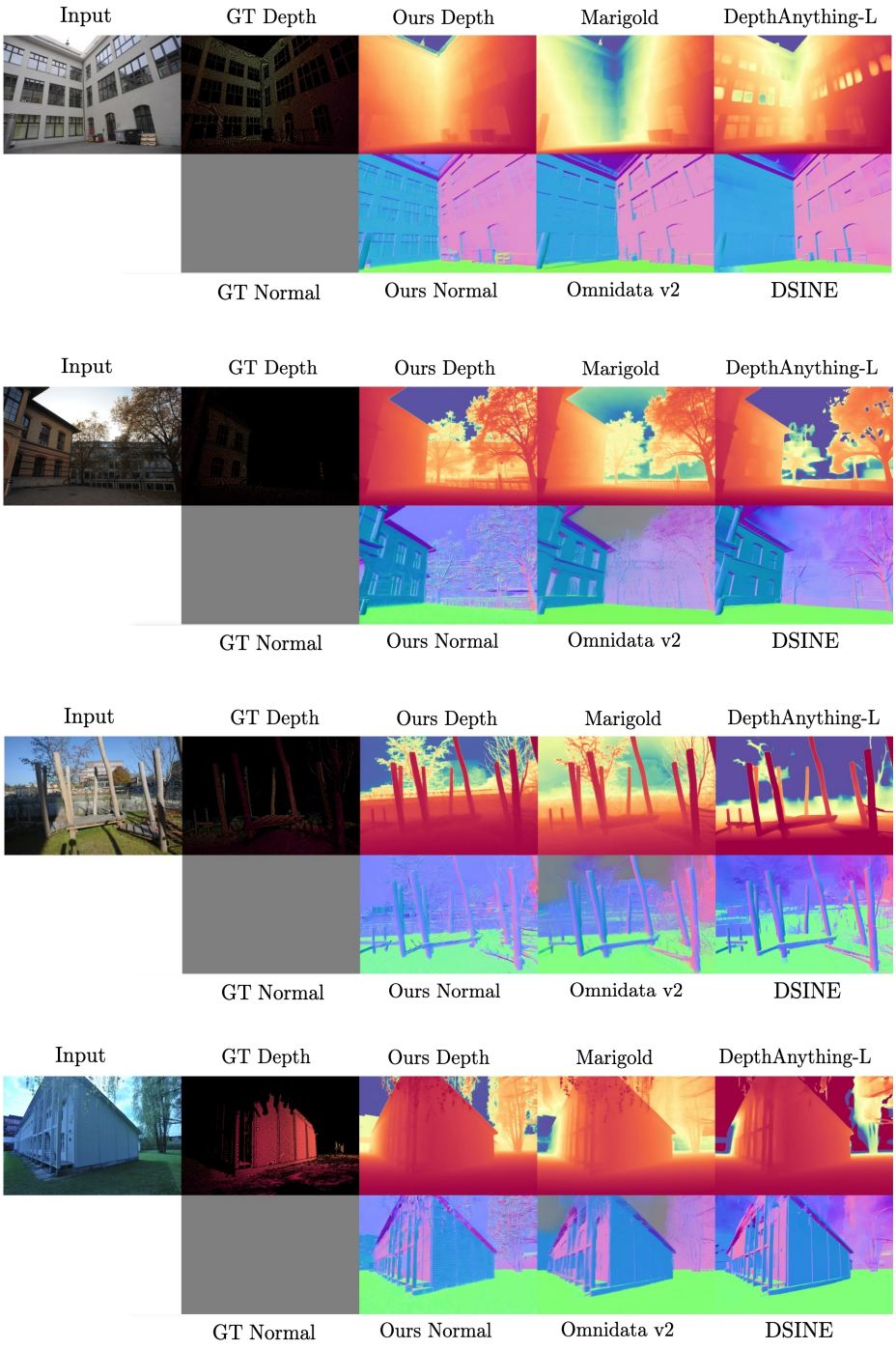

Fig. S6: Qualitative comparison on ETH3D [15].

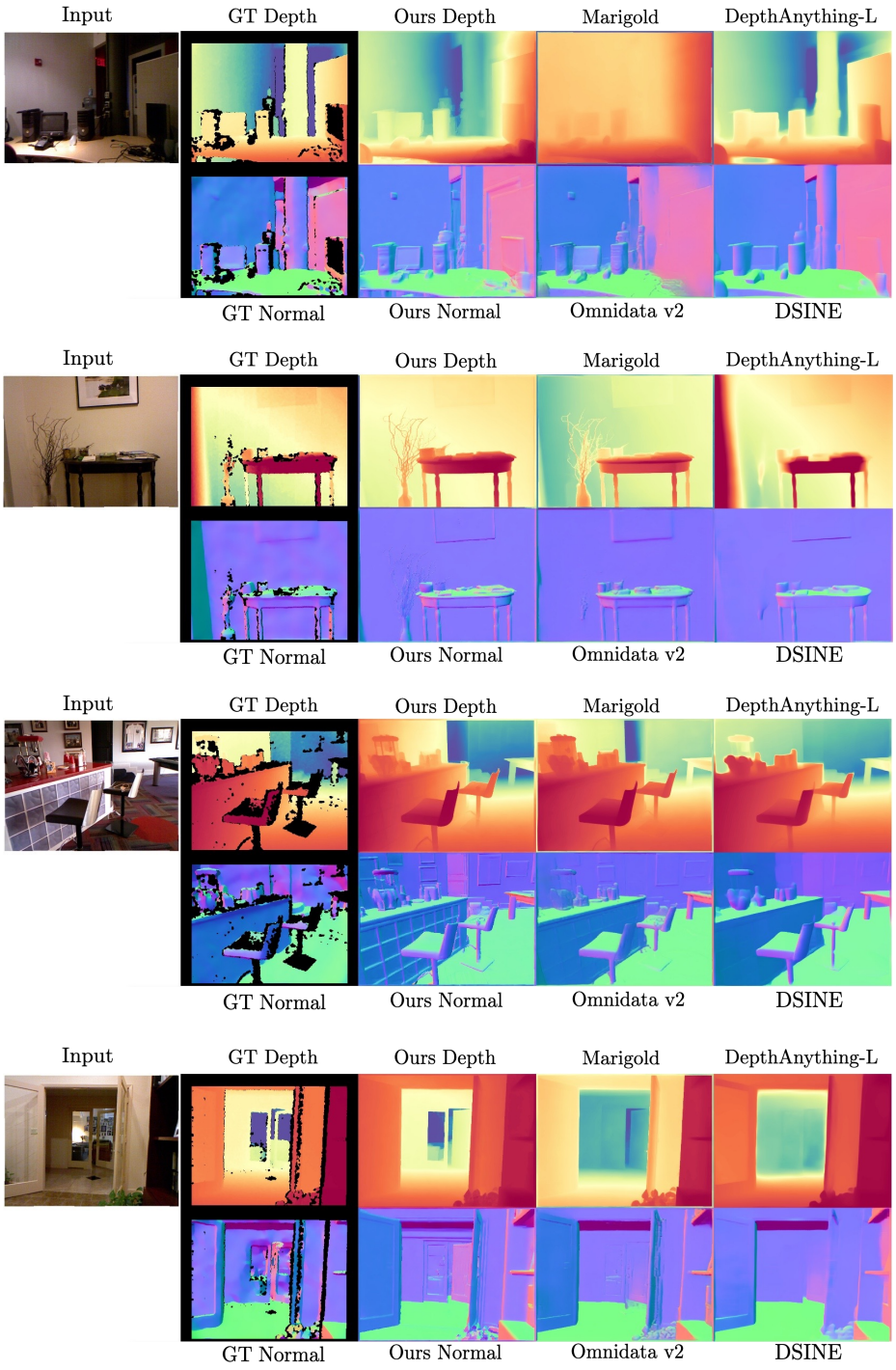

Fig. S7: Qualitative comparison on NYUv2 [17].

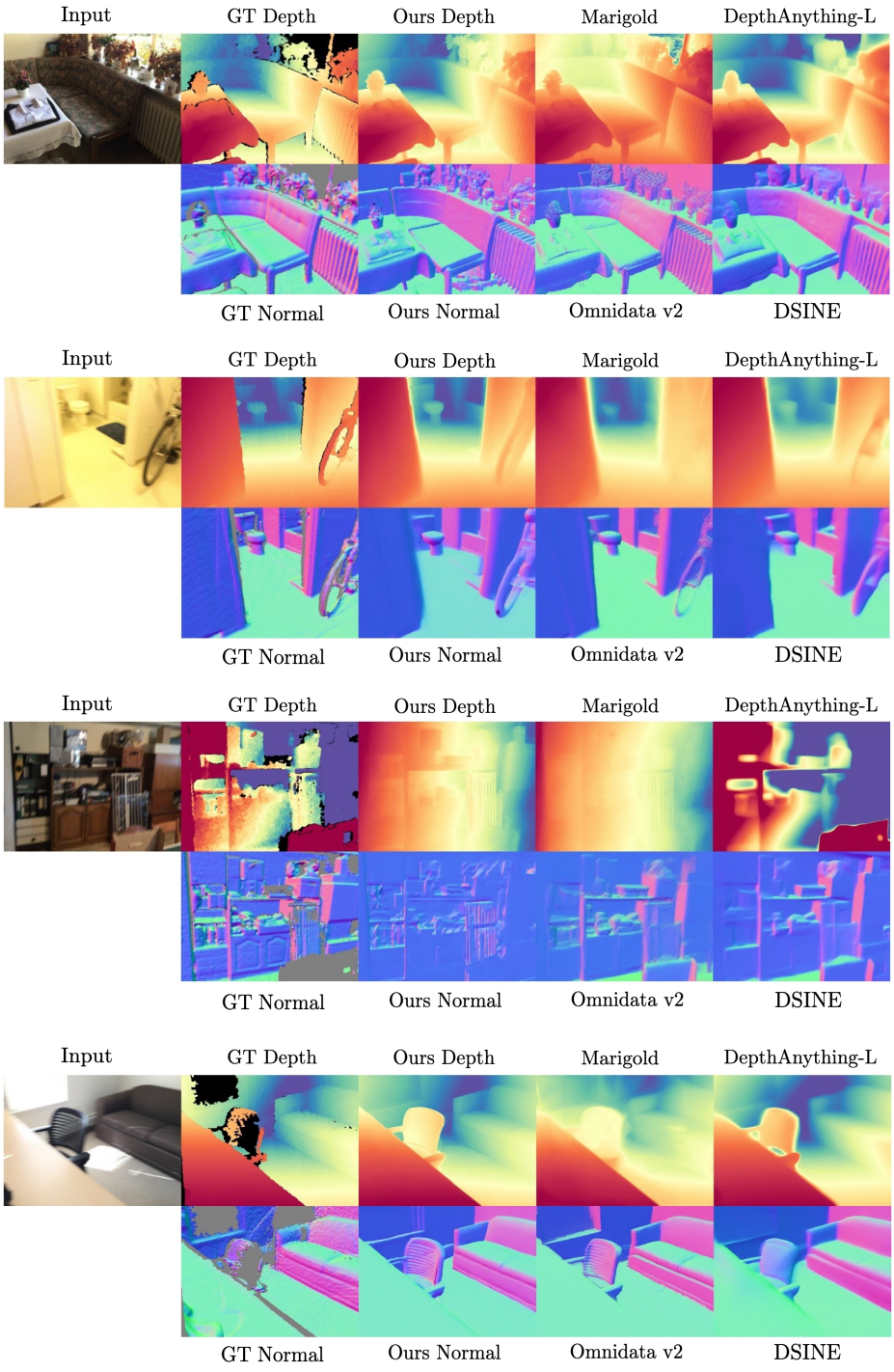

Fig. S8: Qualitative comparison on ScanNet [4].

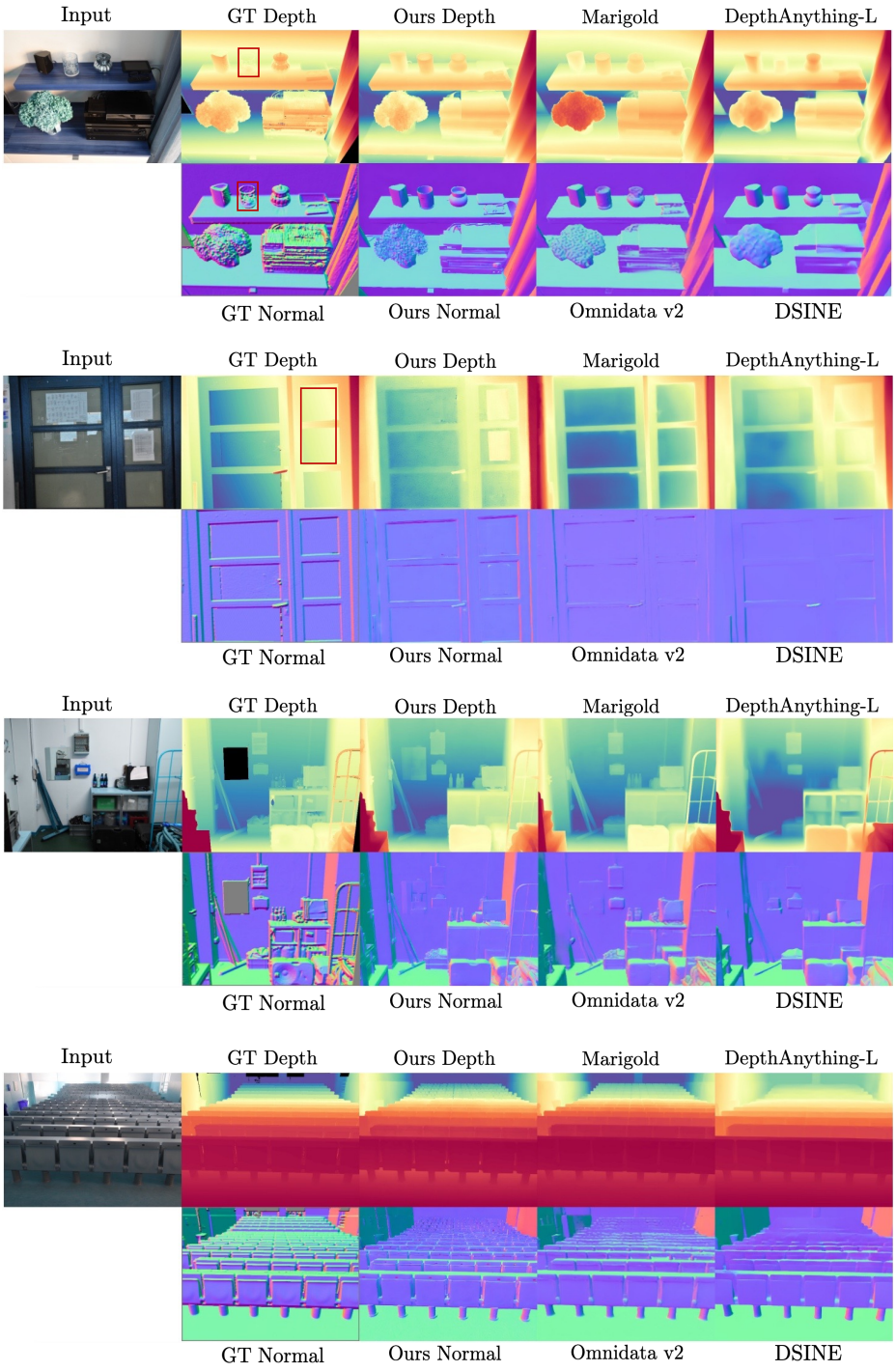

Fig. S9: Qualitative comparison on iBims-1 [10]. The red box marks the part where GT is erroneous.

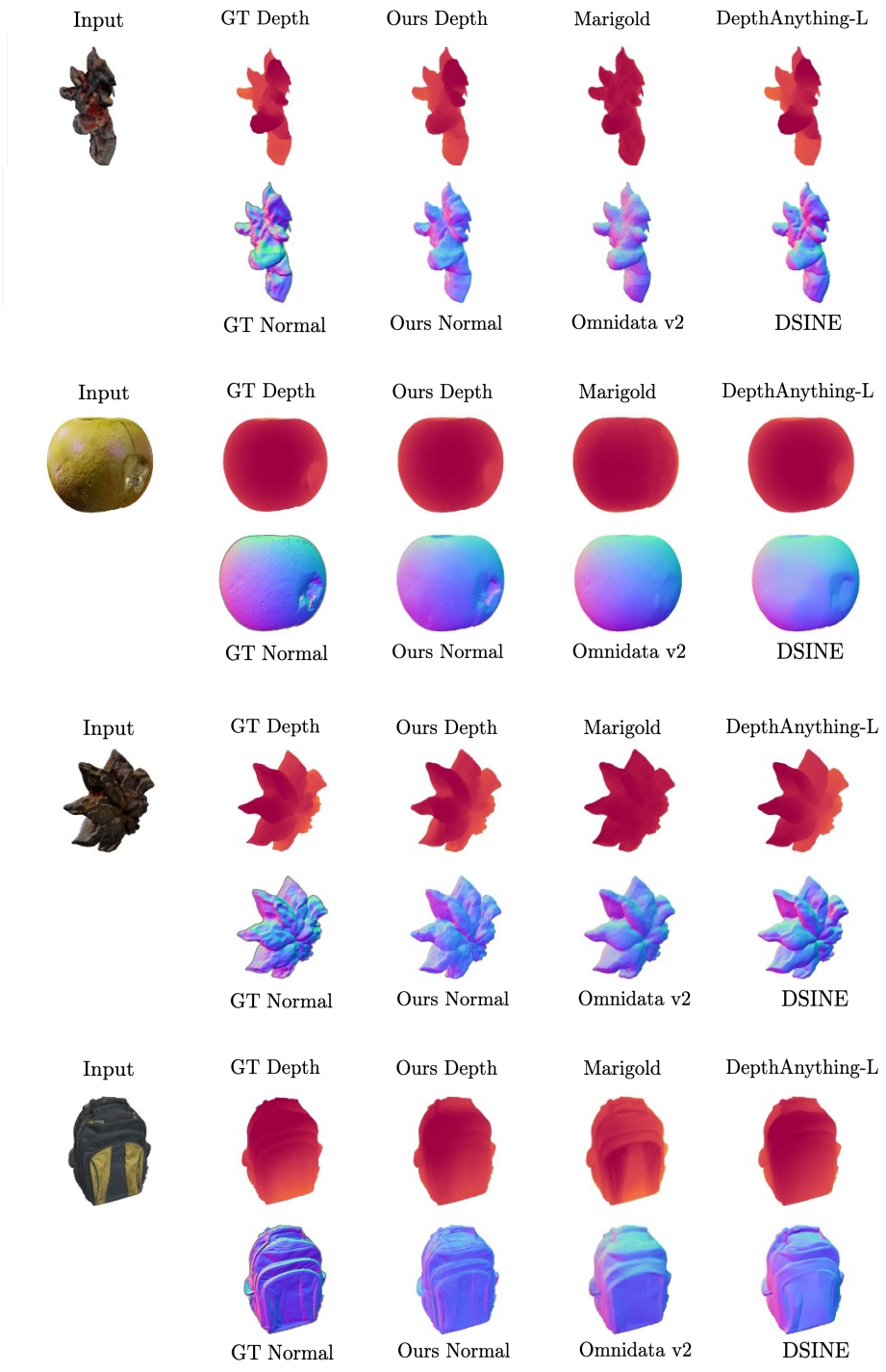

Fig. S10: Qualitative comparison on OmniObject3D [23].

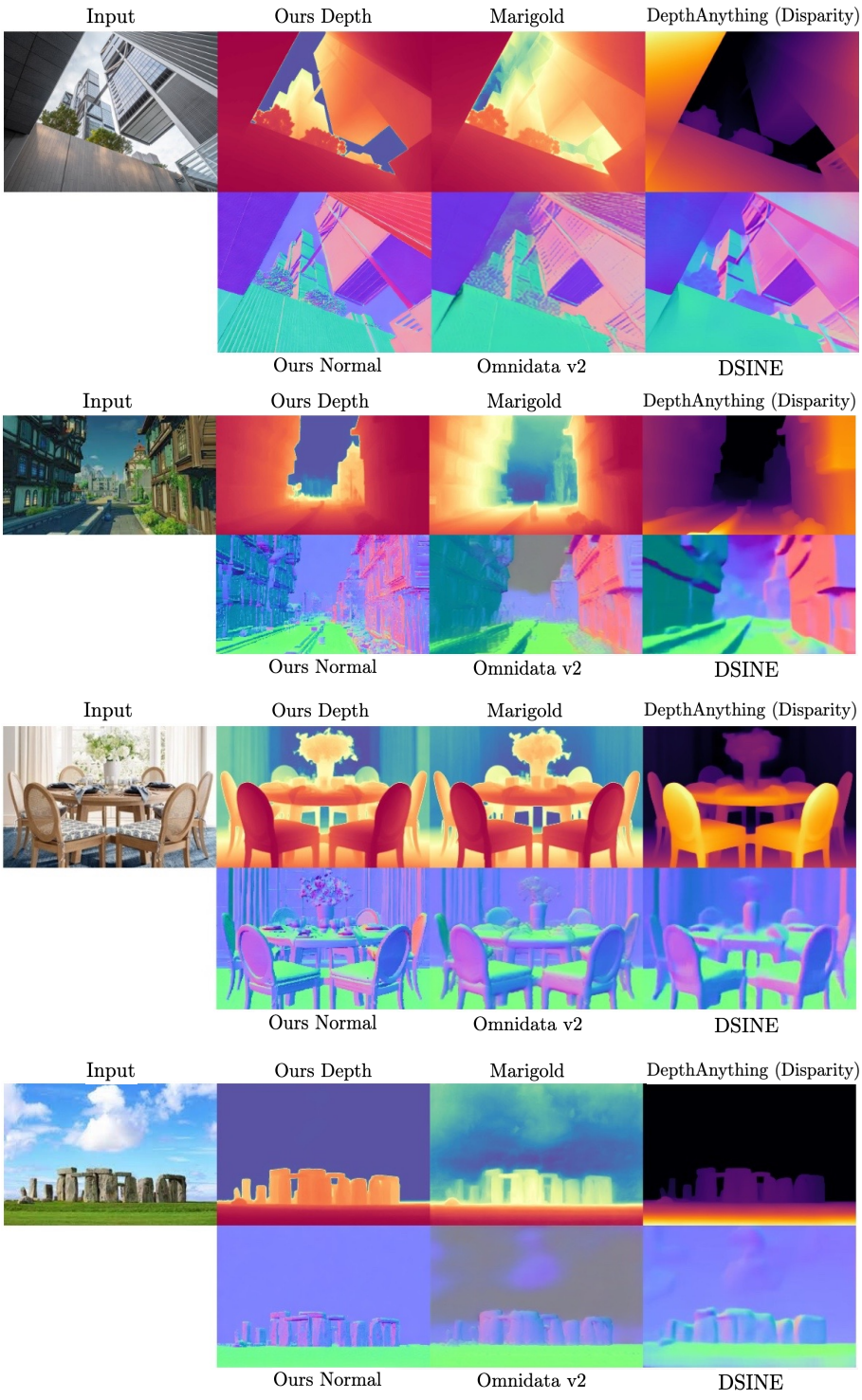

Fig. S11: Qualitative geometry comparison on in-the-wild images (1/11).

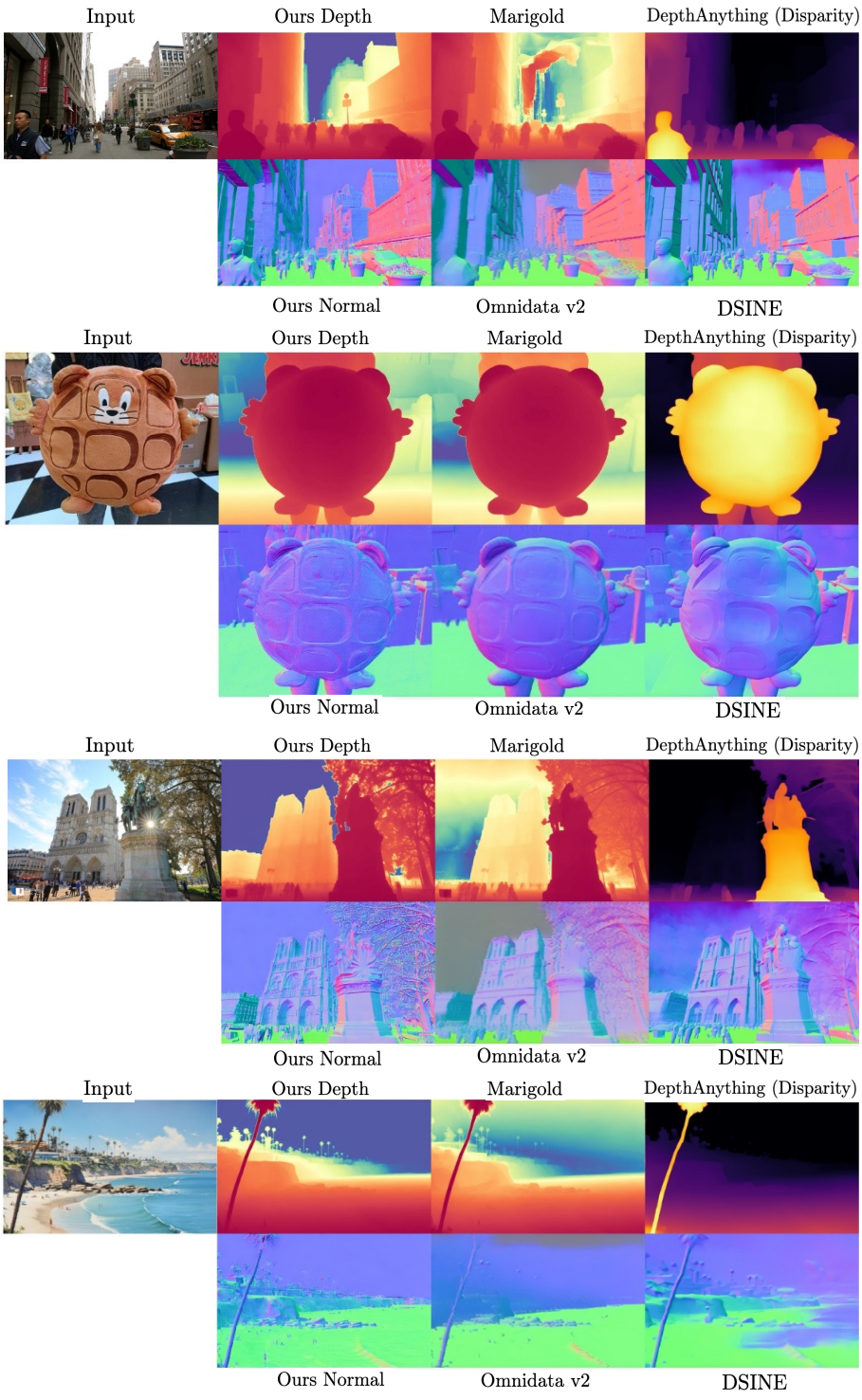

Fig. S12: Qualitative geometry comparison on in-the-wild images (2/11).

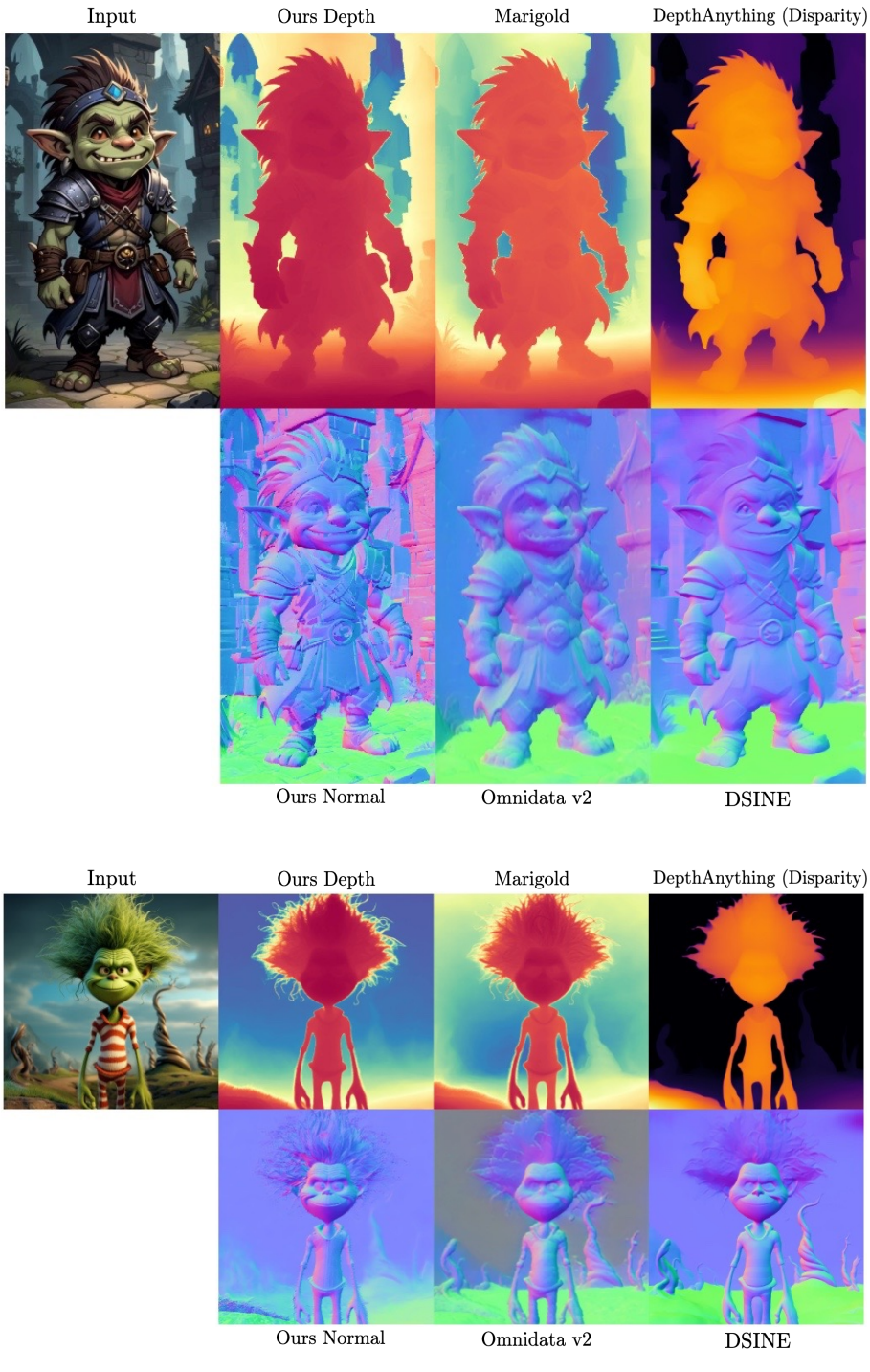

Fig. S13: Qualitative geometry comparison on in-the-wild images (3/11).

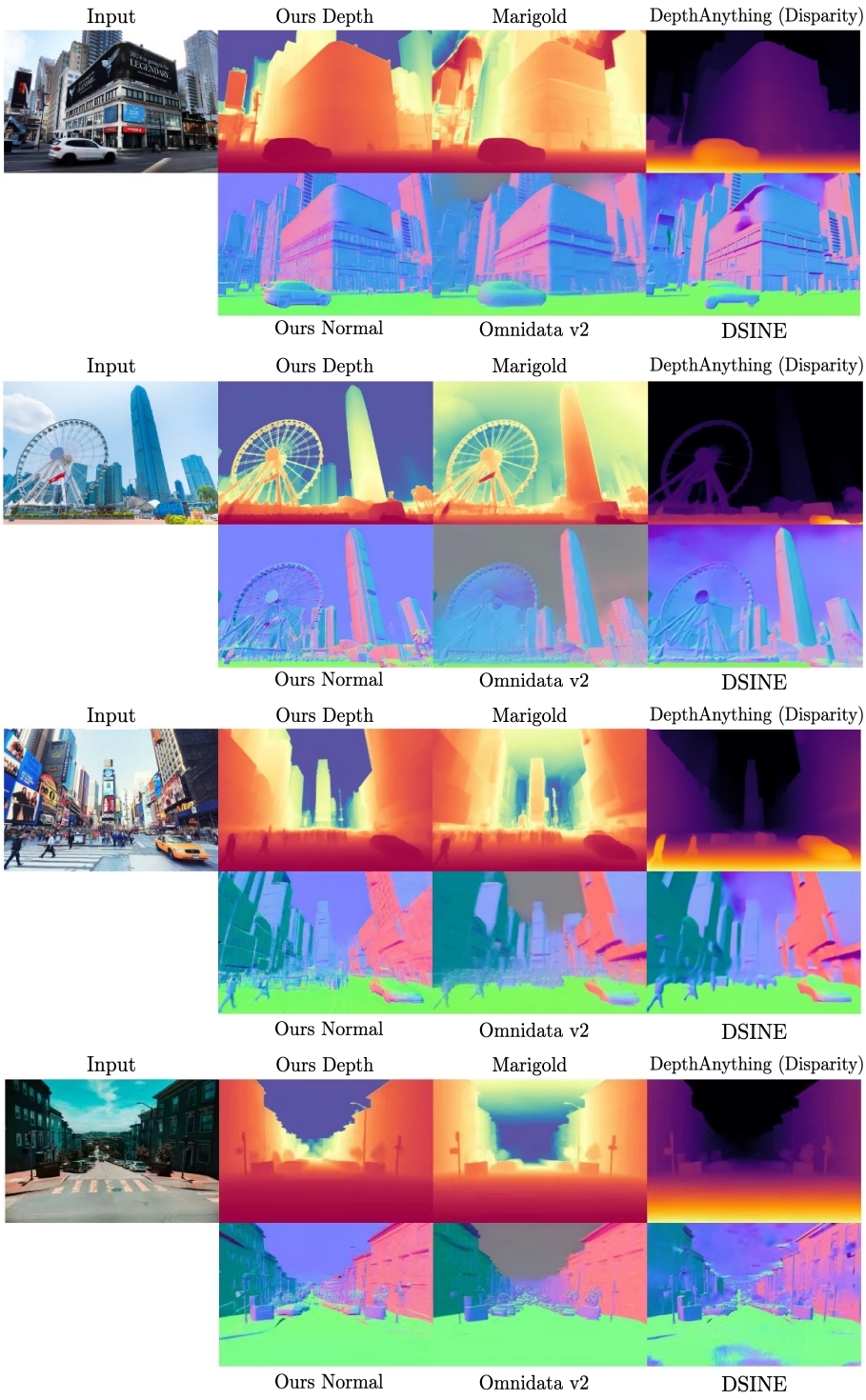

Fig. S14: Qualitative geometry comparison on in-the-wild images (4/11).

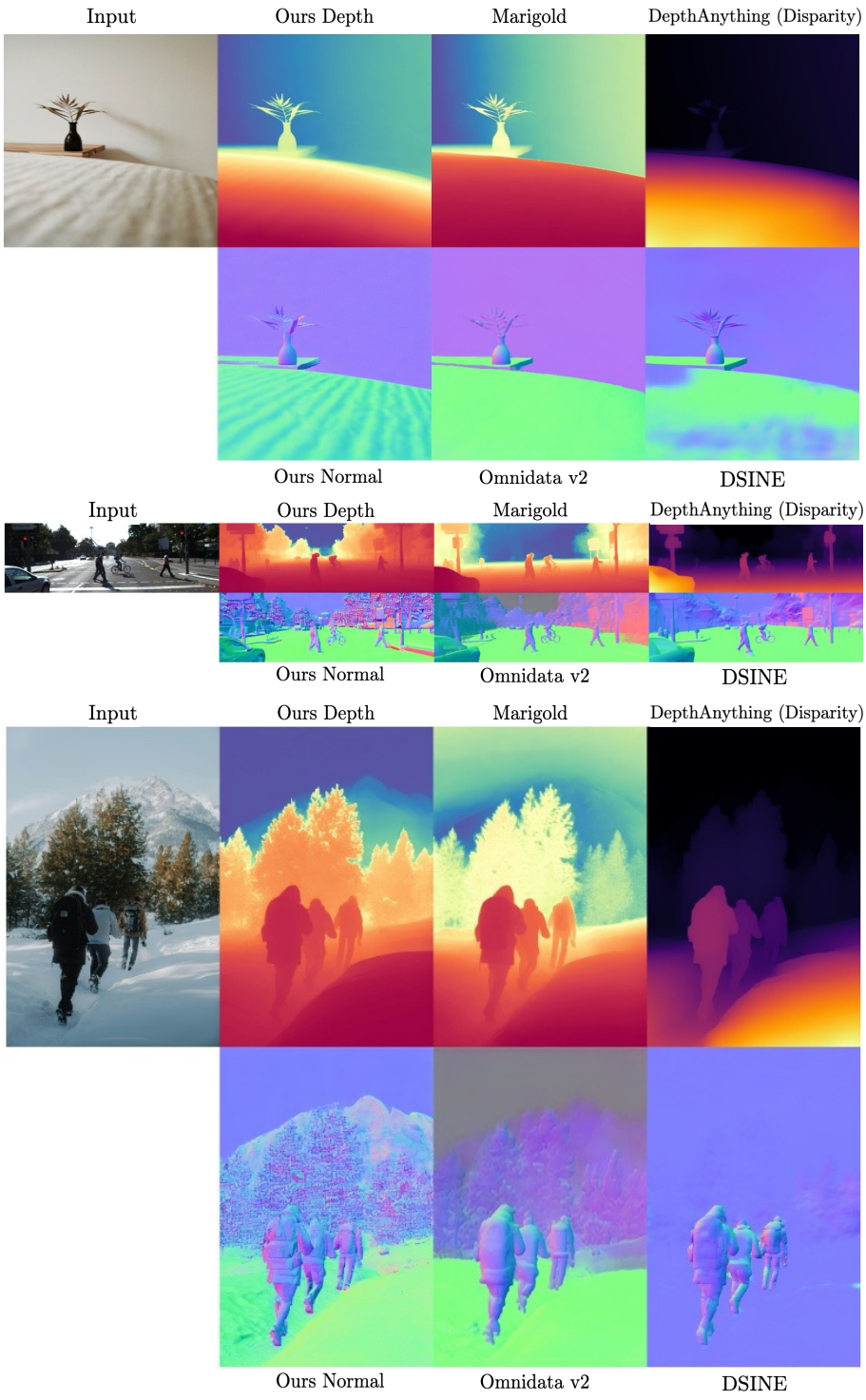

Fig. S15: Qualitative geometry comparison on in-the-wild images (5/11).

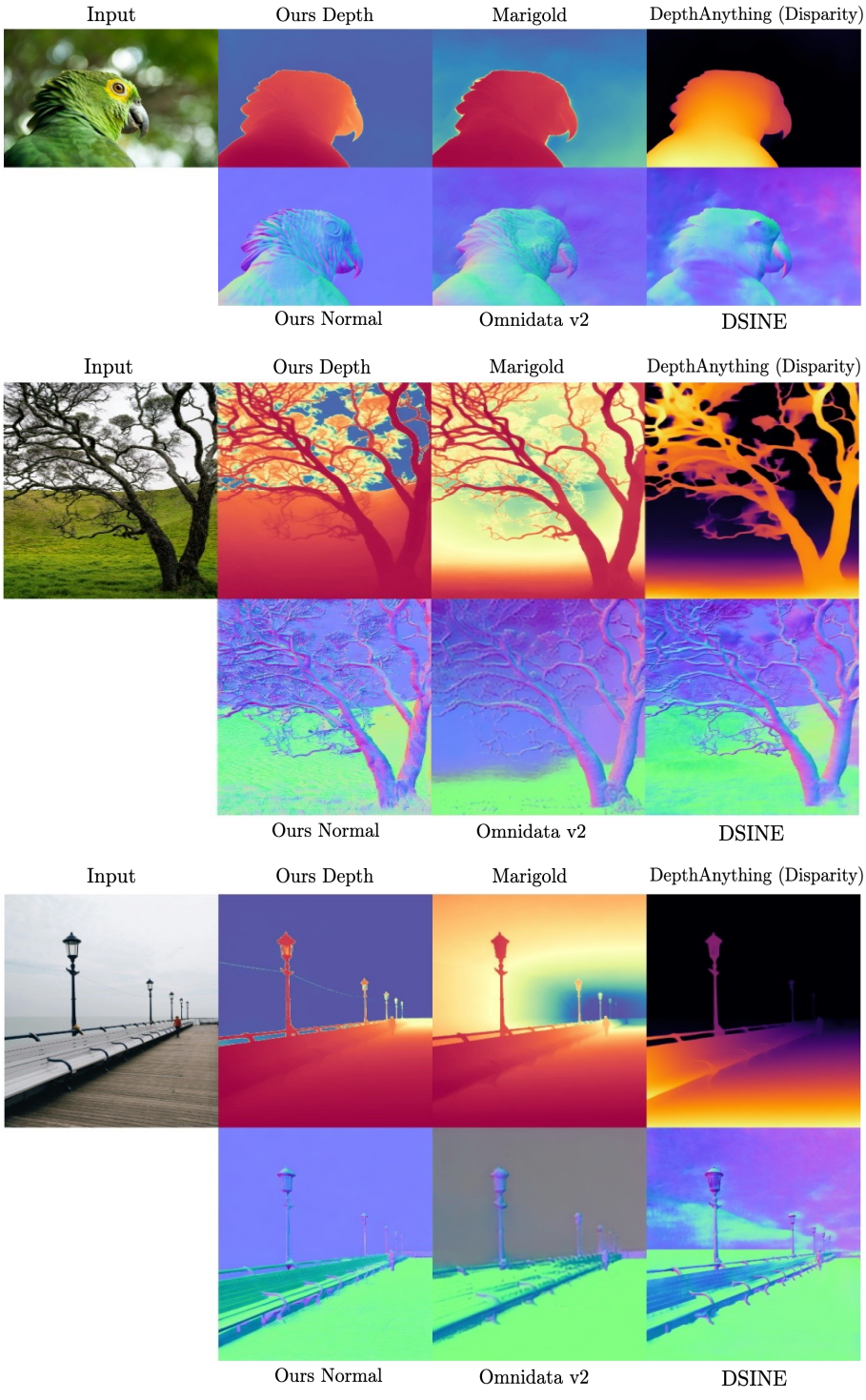

Fig. S16: Qualitative geometry comparison on in-the-wild images (6/11).

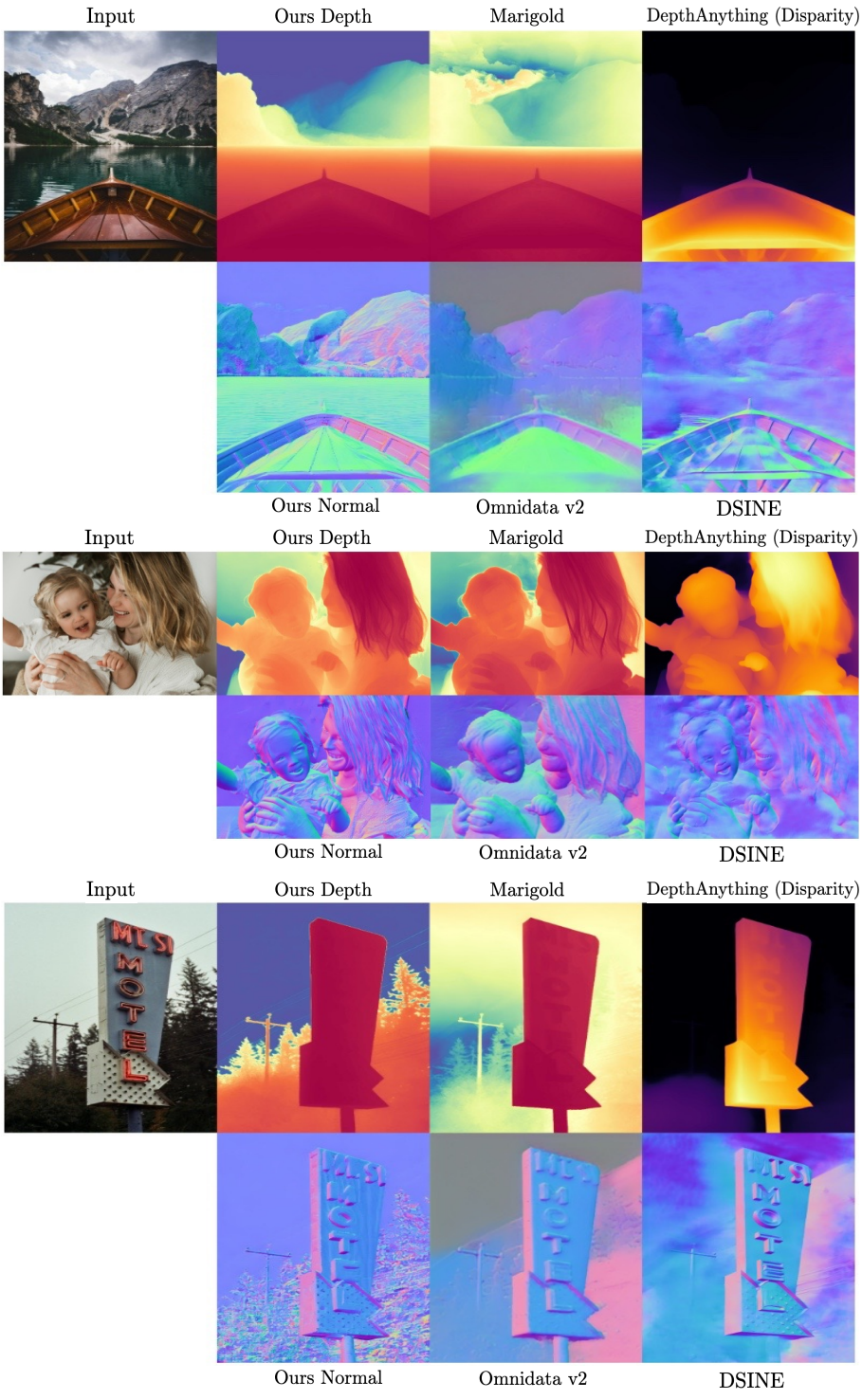

Fig. S17: Qualitative geometry comparison on in-the-wild images (7/11).

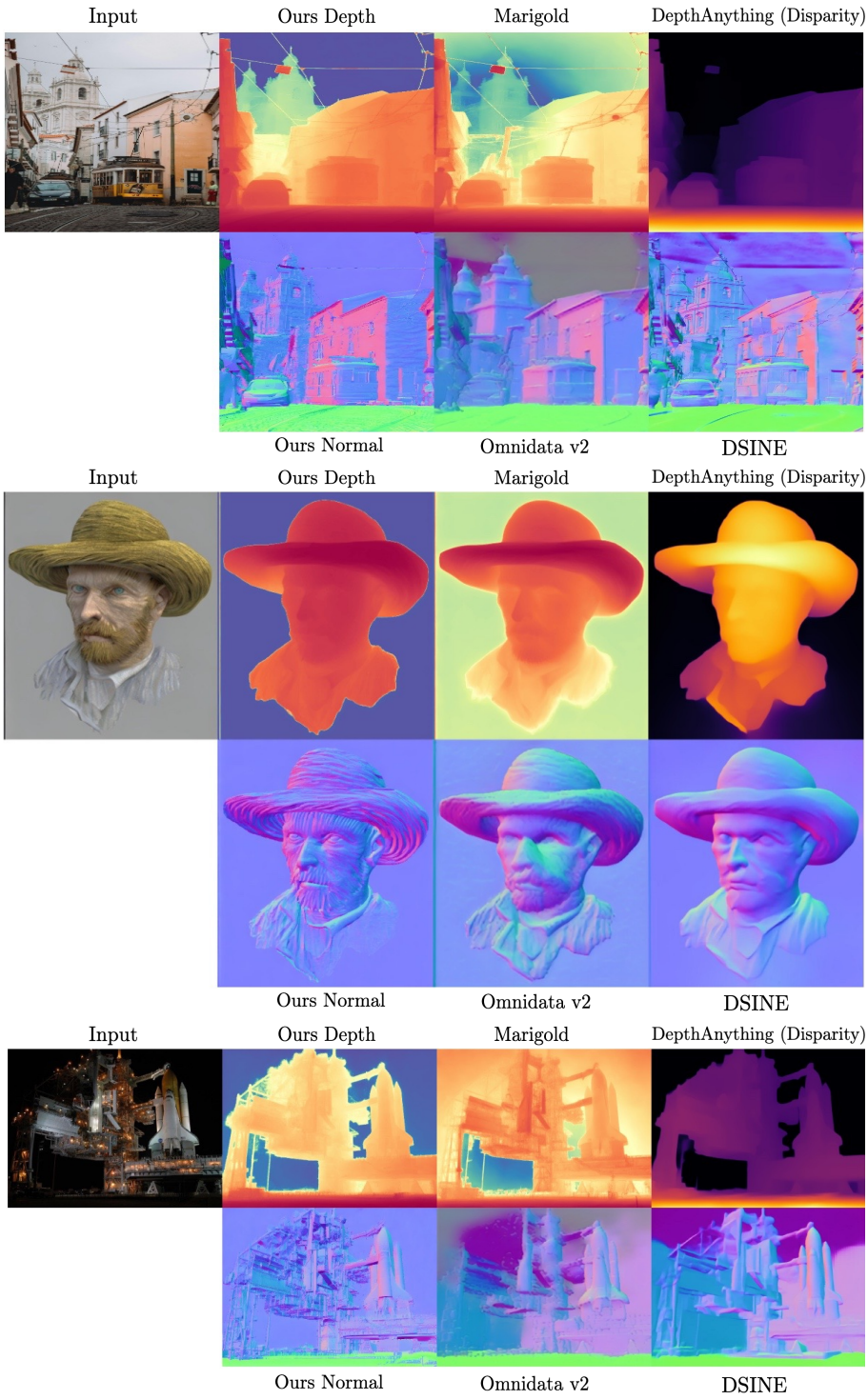

Fig. S18: Qualitative geometry comparison on in-the-wild images (8/11).

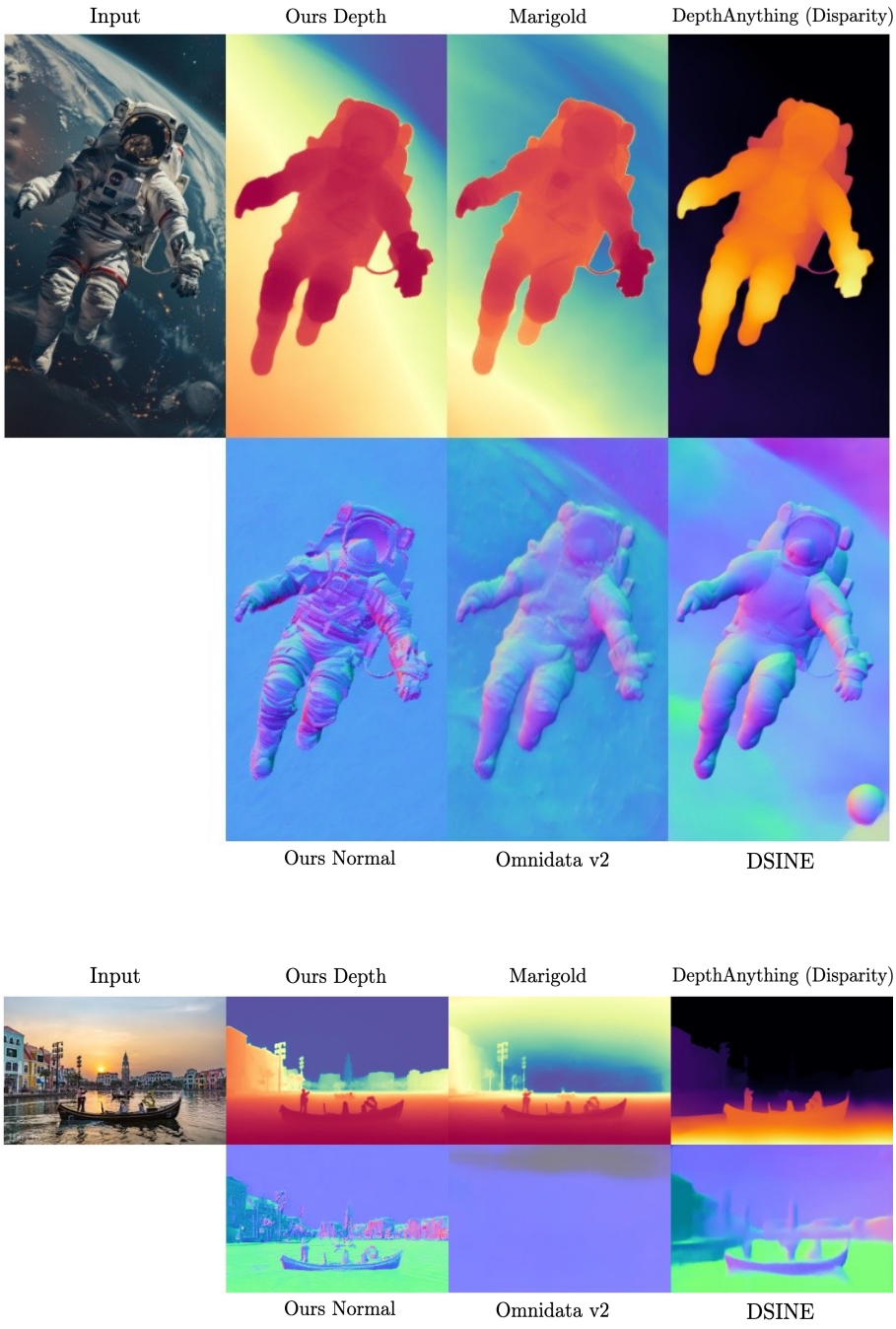

Fig. S19: Qualitative geometry comparison on in-the-wild images (9/11).

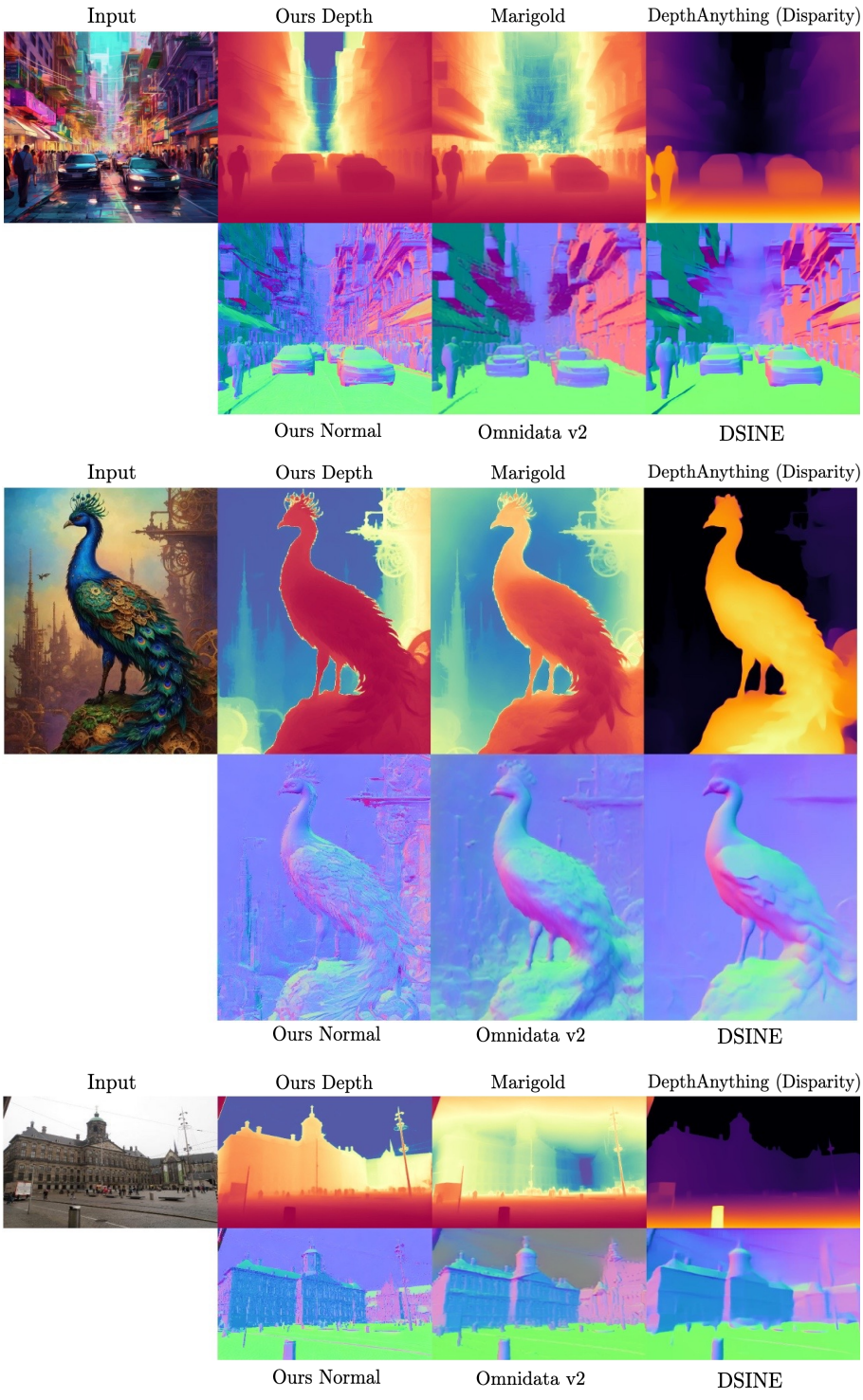

Fig. S20: Qualitative geometry comparison on in-the-wild images (10/11).

Fig. S21: Qualitative geometry comparison on in-the-wild images (11/11).

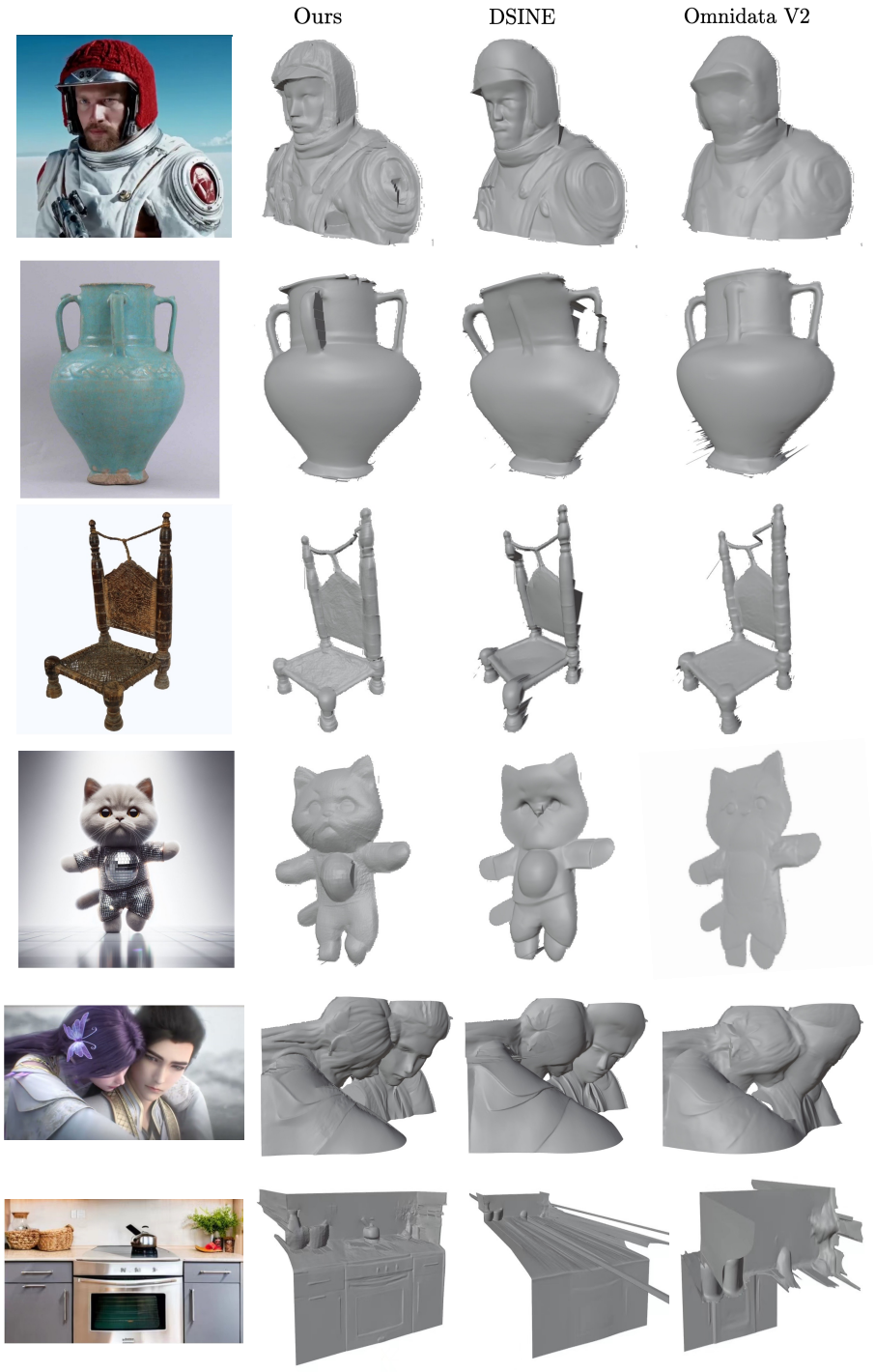

Fig. S22: Qualitative comparison on 3D reconstruction. We segment out the foreground objects using SAM [9]. The meshes rotate left and right along the z-axis.

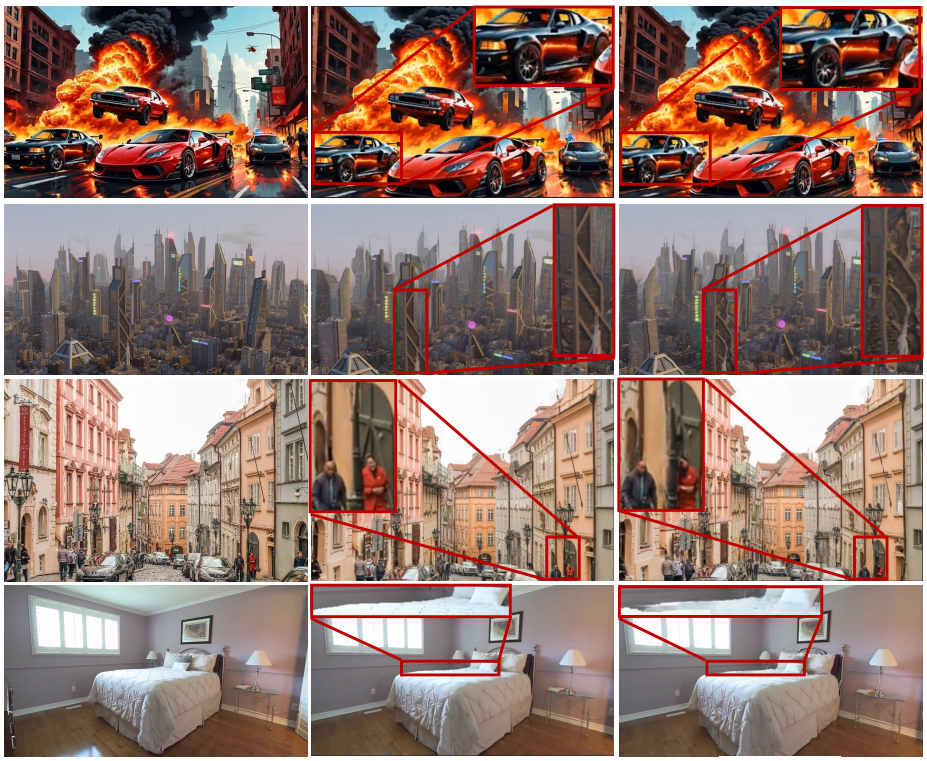

Fig. S23: Novel view synthesis comparison on more scenes.

References

1. Bae, G., Davison, A.J.: Rethinking inductive biases for surface normal estimation. In: CVPR (2024)
2. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv.org (2020)
3. Cao, X., Santo, H., Shi, B., Okura, F., Matsushita, Y.: Bilateral normal integration. In: ECCV (2022)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
5. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR (2023)
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013)
7. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3d common corruptions and data augmentation. In: CVPR (2022)
8. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
10. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: ECCVW (2018)
11. Liu, X., Ren, J., Siarohin, A., Skorokhodov, I., Li, Y., Lin, D., Liu, X., Liu, Z., Tulyakov, S.: Hyperhuman: Hyper-realistic human generation with latent structural diffusion. In: ICLR (2024)
12. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. In: ACM TOG (2019)
13. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv.org (2023)
14. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In: IEEE TPAMI (2022)
15. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
16. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: CVPR (2020)
17. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
18. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv.org (2019)
19. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv.org (2023)
20. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. CoRR (2019)

21. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. arXiv.org (2024)
22. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv.org (2021)
23. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR (2023)
24. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)