Supplementary Material: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

Bingxin Ke Anton Obukhov Shengyu Huang
Nando Metzger Rodrigo Caye Daudt Konrad Schindler
Photogrammetry and Remote Sensing, ETH Zürich

In this supplementary material, we provide additional implementation details in Sec. A and present additional quantitative and qualitative results in Sec. B and Sec. C, respectively.

A. Implementation Details

A.1. Mixed Dataset Training

We train on two synthetic datasets, Hypersim [9] and Virtual KITTI [1], whose images have different resolutions and aspect ratios. For each batch, we probabilistically choose the dataset and then draw samples from it. We ablate the Bernoulli parameter of dataset sampling in Sec. B.4.

A.2. Annealed Multi-Resolution Noise

In the standard multi-resolution noise, multiple Gaussian noise images are sampled to form a pyramid of resolutions and then subsequently combined by upsampling, weighted averaging, and renormalization. The weight for the i-th pyramid level is computed as s^i, where $0 < s < 1$ is a strength of influence of lower-resolution noise. To bring such noise closer to the Gaussian used in the original DDPM formulation, we propose to anneal the weight of levels $i > 0$ based on the diffusion schedule. Specifically, we assign the i-th level at timestep t the weight $(st/T)^i$, where T is the total number of diffusion steps. Thus, a smaller weight is given to lower-resolution levels at timesteps closer to the noise-free end of the schedule. In addition to the ablation study in the main paper, we further demonstrate the effectiveness of annealing and other noise settings in Sec. B.3.

A.3. Alignment with Ground Truth Depth

Following the established evaluation protocol [7], we use least squares fitting over pixels with valid ground truth values to compute the scale and shift factors of the affine-invariant predictions. Note that, while some methods predict affine-invariant disparities [3, 7, 8], others (including ours) predict affine-invariant depth values [13–15]. We apply least squares fitting accordingly, *i.e*. the disparities are aligned to the inverse ground truth depth.

A.4. Visualization in 3D

We compute the scale and shift scalars between the prediction and ground truth. Subsequently, we unproject pixels into the metric 3D space using the camera intrinsics. We manually estimate the scale, shift, and intrinsics of "in-the-wild" samples, where ground truth and camera intrinsics are unavailable. For some samples, camera intrinsics can also be extracted from the EXIF metadata. To visualize normals, we perform least squares plane fitting at each position, considering a neighborhood area of 3×3 pixels around it.

B. Experimental Results

B.1. Stable Diffusion VAE with Depth

To assess how well the pre-trained image variational autoencoder of Stable Diffusion [10] works with depth maps, we tested it with 800 samples from the Hypersim [9] training set. To this end, each sample is normalized to the operational range of VAE as explained in the main paper, and replicated three times to accommodate the RGB interface. Upon decoding the latents, the reconstructed depth map is derived by averaging the three RGB channels. Over the chosen set of depth maps, the Mean Absolute Error (MAE) of reconstructions is 0.0095 ± 0.0091, which is safely below the current state-of-the-art depth estimation errors.

B.2. Consistency of Channels After VAE Decoder

To further understand the suitability of the Stable Diffusion latent space for depth representation, we evaluate the agreement of depth channels obtained from the VAE decoder during inference. We validate with the training split of NYUv2 [6] and a subsampled Eigen training split [4] of the KITTI dataset [5]. As shown in Tab. S1, the channel-wise discrepancy resulting from decoding depth from the latent space is small relative to the value range of the decoder output, *i.e*., $[-1, 1]$. This could be related to the ability of VAE to represent gray-scale RGB images.

Table S1. **Consistency of channels after VAE decoder.** The reported numbers are averaged over the respective datasets.

	std	max − min
NYU	0.0027	0.0062
KITTI	0.0022	0.0052

B.3. Prediction Variance and Training Noise

Since Marigold is a generative model, the predictions vary depending on the initial noise starting the diffusion process. We evaluate the consistency of predictions of three models, trained differently, *i.e.*, with Gaussian noise, multi-resolution noise, and annealed multi-resolution noise. We train with two synthetic datasets and validate with the training split of NYUv2 [6] and a subsampled Eigen training split [4] of the KITTI dataset [5]. Specifically, we perform inference 10 times for each sample and compute pixel-wise statistics over the resulting depth predictions. Subsequently, we aggregate these statistics across entire datasets and report them in Tab. S2. As seen from the values, training with the multi-resolution noises increases the prediction consistency at inference, and the annealed version brings further improvement. Fig. S1 demonstrates predictions for a single sample with three models and varying starting noise.

Table S2. **Pixel-wise consistency of depth predictions made by models trained with three different noise types.** The reported numbers are averaged over entire datasets, wherein each sample was processed 10 times, starting from a new noise sample.

Multi-res. noise	Annealed	NYUv2		KITTI	
		std	max − min	std	max − min
✗	✗	0.086	0.260	0.050	0.152
✓	✗	0.037	0.117	0.030	0.094
✓	✓	0.033	0.106	0.025	0.079

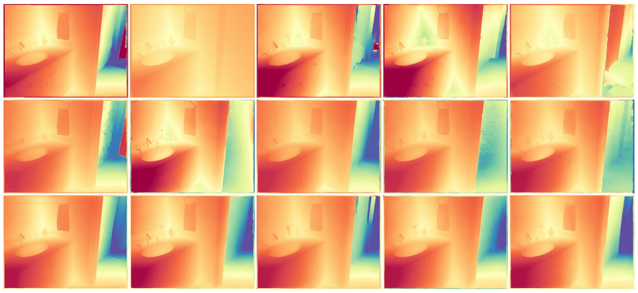


Figure S1. **Example of predictions on the same input** by models trained with (top-down) Gaussian, multi-resolution, and annealed multi-resolution noise. The last row exhibits the least variance.

B.4. Ratio of Mixed Training Datasets

To further investigate the impact of the synthetic datasets used in our fine-tuning protocol, we ablate the mixing ratio of the datasets, discussed in Sec. A.1. We train with two synthetic datasets, Hypersim [9] and Virtual KITTI [1], and validate with the training split of NYUv2 [6] and a subsampled Eigen training split [4] of the KITTI dataset [5]. As shown in Tab. S3, training with a mixture of these two synthetic datasets yields better results on both indoor and outdoor real datasets, than training with a single synthetic dataset. Interestingly, based on the higher-quality indoor dataset, Hypersim [9], adding a small portion (5%) of Virtual KITTI [1], a street-view dataset, can already increase the performance on the outdoor dataset. We find a sweet spot at around 10% where the performance is improved on both indoor and outdoor scenes. When the ratio of Virtual KITTI keeps increasing, the overall performance is impaired. This is likely caused by the varying scene diversity and rendering quality of these two datasets.

Table S3. **Ablation study of the training dataset mixing strategy.** Our method trained with only Hypersim delivers strong results. Outdoor performance is further enhanced with a small portion of Virtual KITTI. The zero-shot transfer is attained at 10% ratio.

Hypersim	Virtual KITTI	NYUv2		KITTI	
		AbsRel↓	$\delta 1$↑	AbsRel↓	$\delta 1$↑
100%	0%	5.7	96.3	13.7	82.5
95%	5%	5.8	96.2	11.1	88.8
90%	10%	5.6	96.5	11.3	88.7
50%	50%	6.0	96.0	12.8	85.5
0%	100%	13.9	83.4	15.4	79.3

B.5. Inference Speed

In Fig. S2, we report inference runtime, aligned with the settings from Figs. 6, 7. We acknowledge the slower speed *vs.* higher quality trade-off compared to feed-forward methods. Speed can be enhanced in future research, *e.g.* distillation for 2- or 4-step denoising schedules, and reducing prediction variance for smaller ensemble sizes.

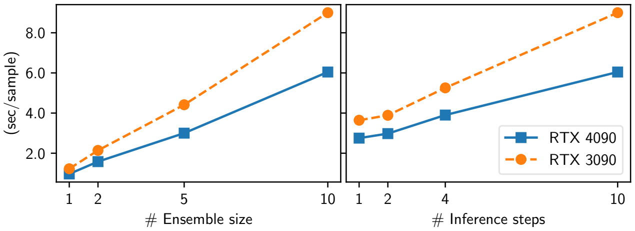


Figure S2. Inference speed on a single GPU with NYUv2 dataset.

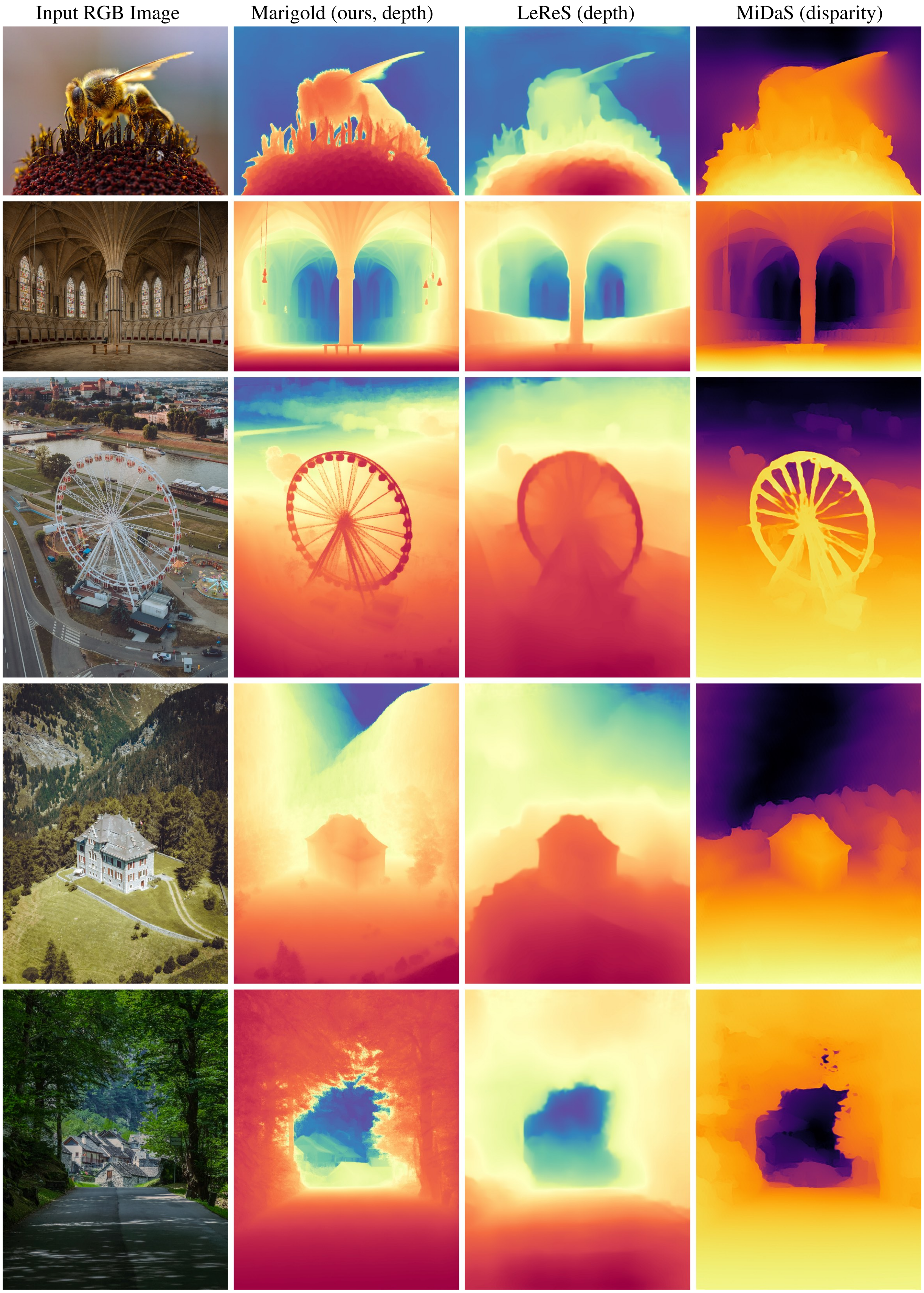
C. Qualitative Comparisons

C.1. In-the-Wild

We present the gallery of "in-the-wild" images and corresponding predictions in Fig. S3. The input images are taken in daily life or downloaded from the internet. Our method, Marigold, predicts accurate depth maps, exhibiting better overall layout and fine details. We show the final predictions for each method, that is, depth for Marigold and LeReS, and disparity for MiDaS.

C.2. Test Datasets

We show additional qualitative comparisons with our competitors [3, 7, 8, 13, 14], on 5 test datasets [2, 5, 6, 11, 12]. The depth maps are visualized in Fig. S4, and the normal maps can be found in Fig. S5. Marigold excels at capturing fine scene details and reflecting the global scene layout.

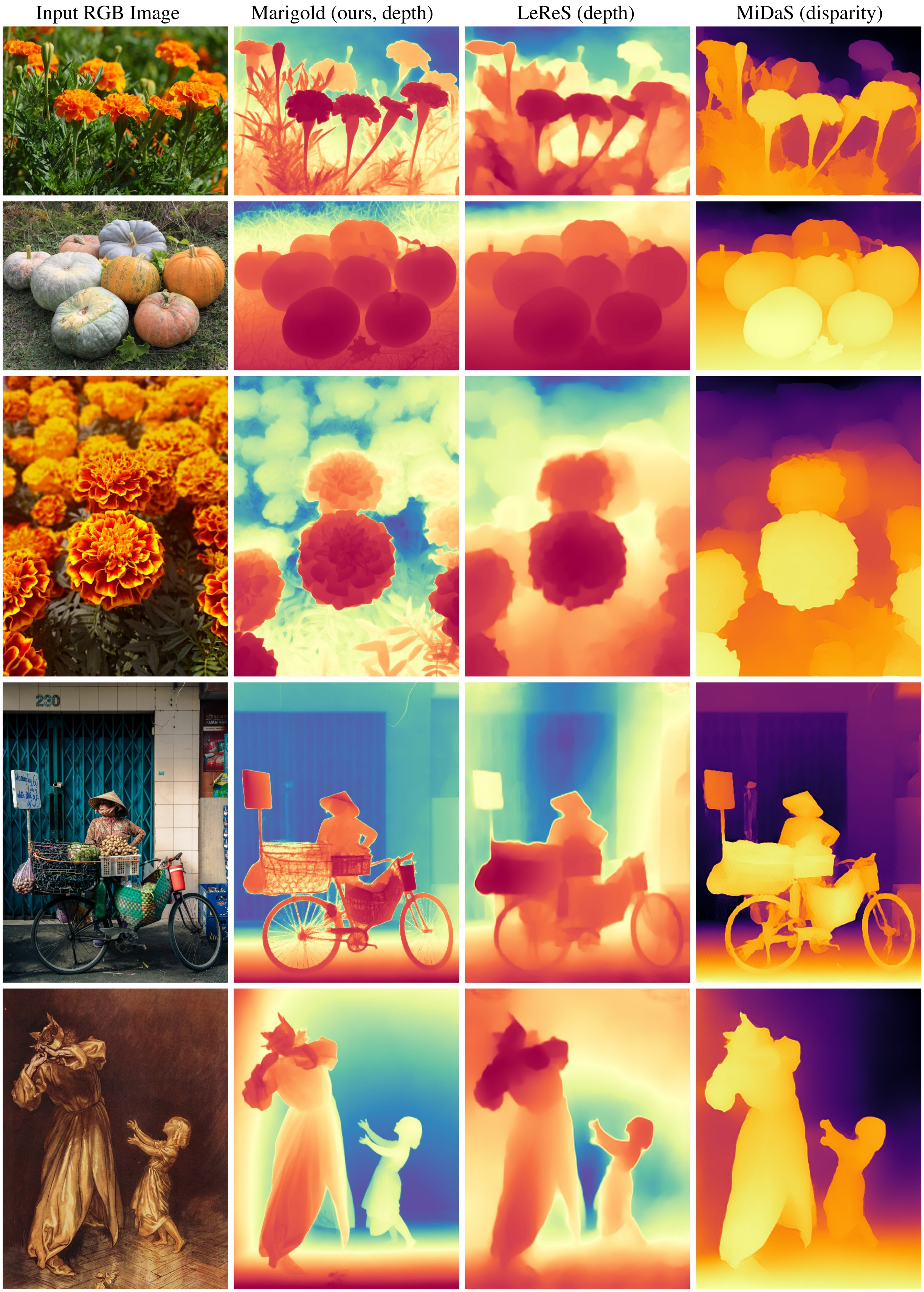


Input RGB Image

Marigold (ours, depth)

LeReS (depth)

MiDaS (disparity)

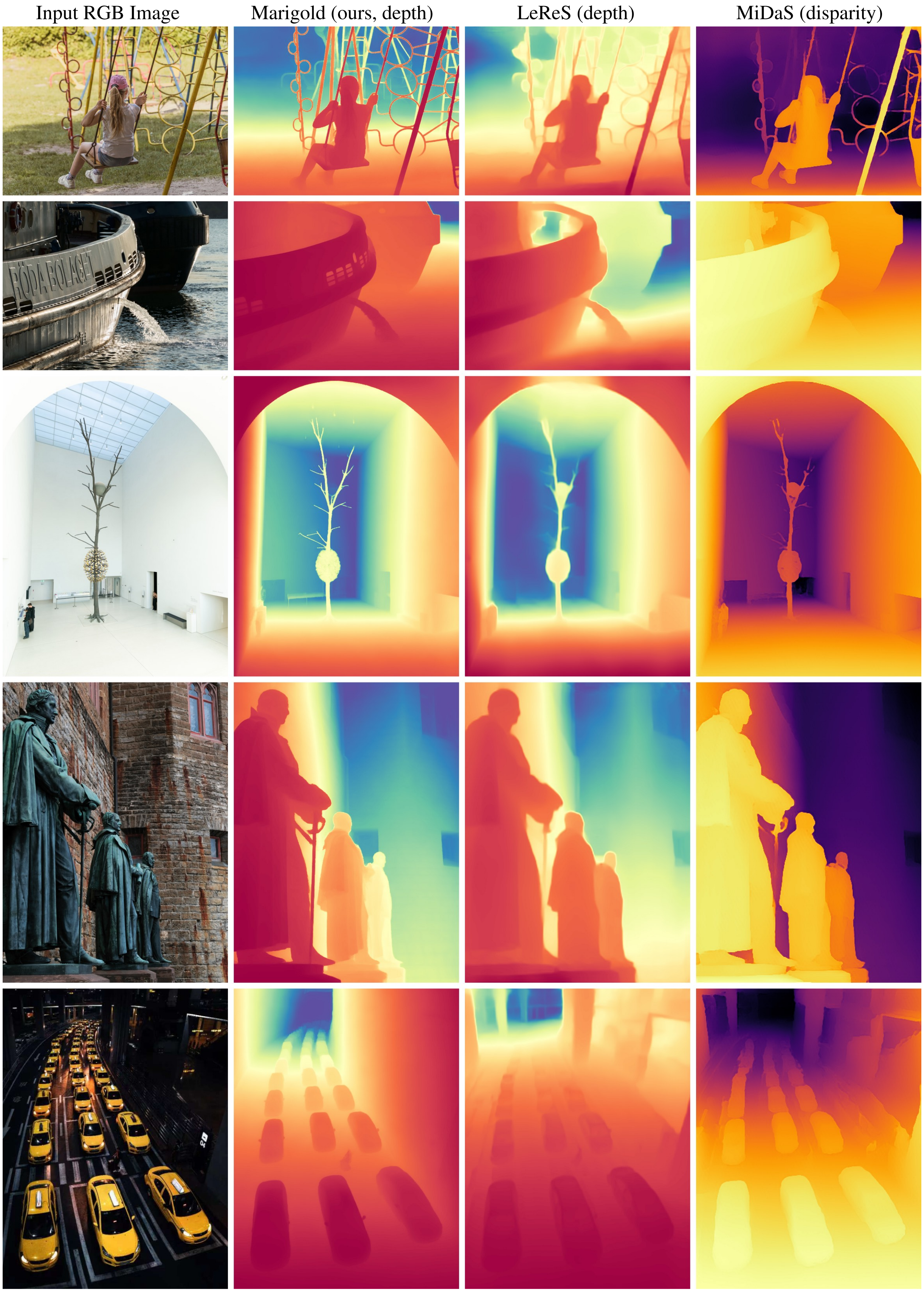


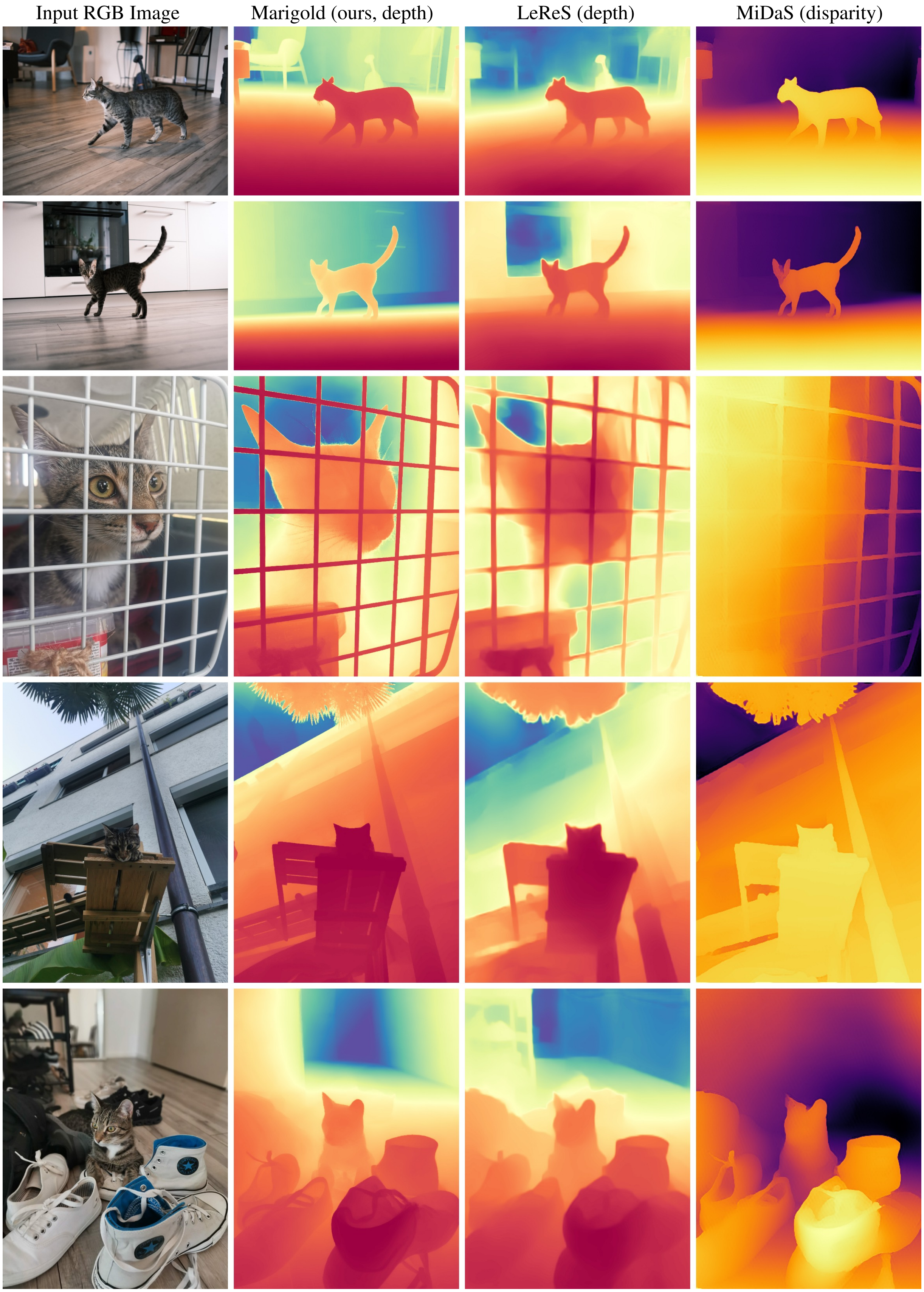
Input RGB Image

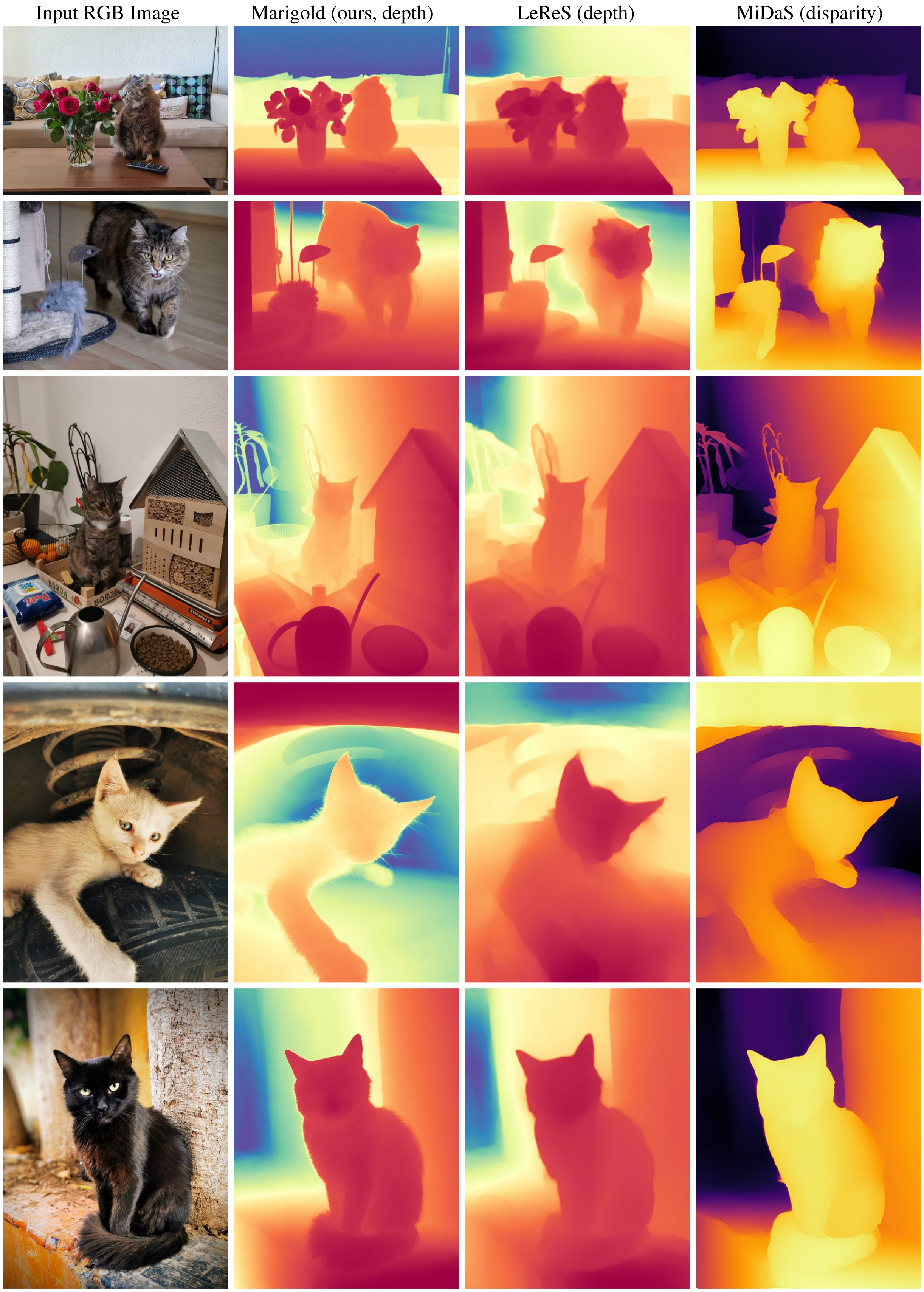
Marigold (ours, depth)

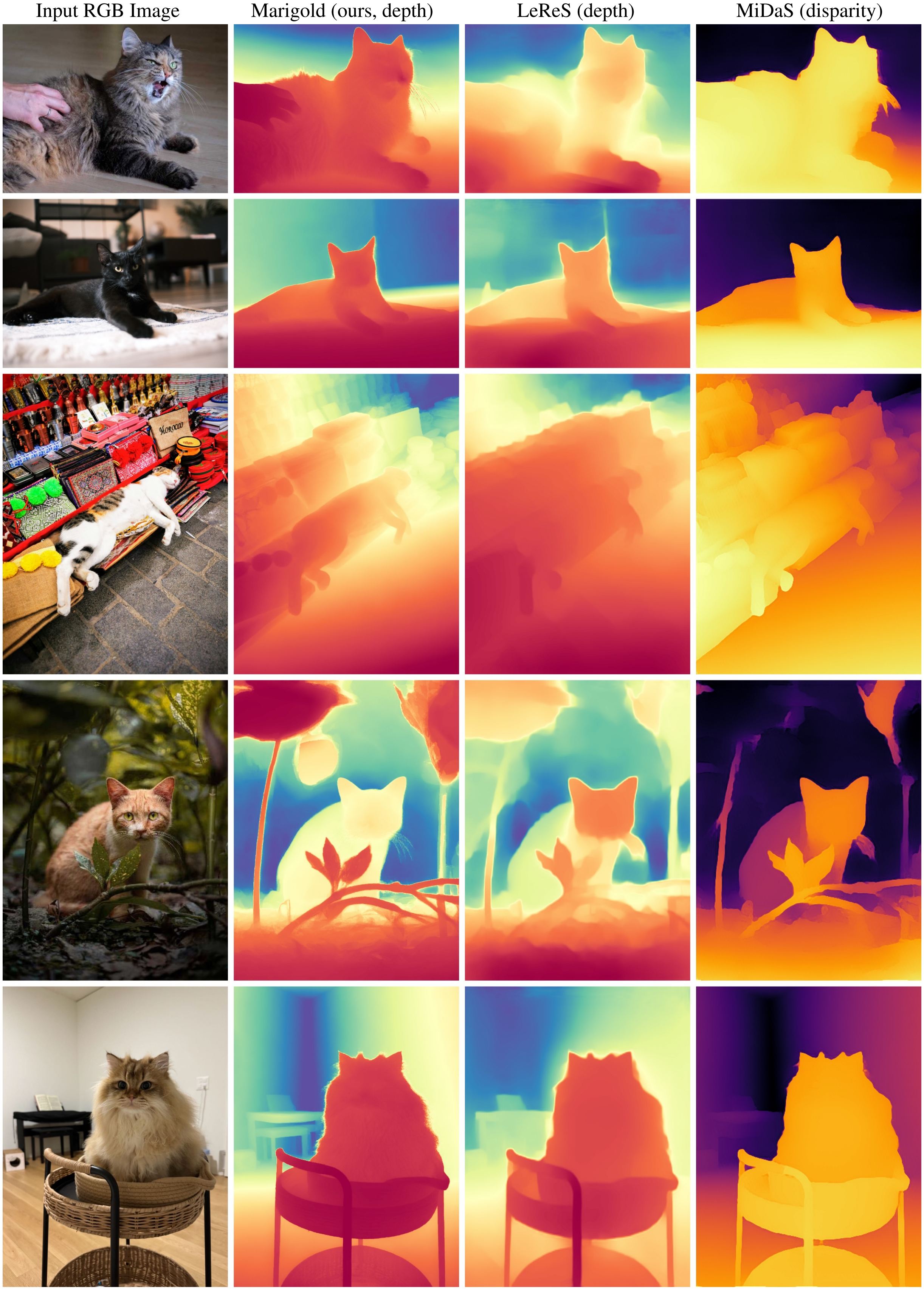
LeReS (depth)

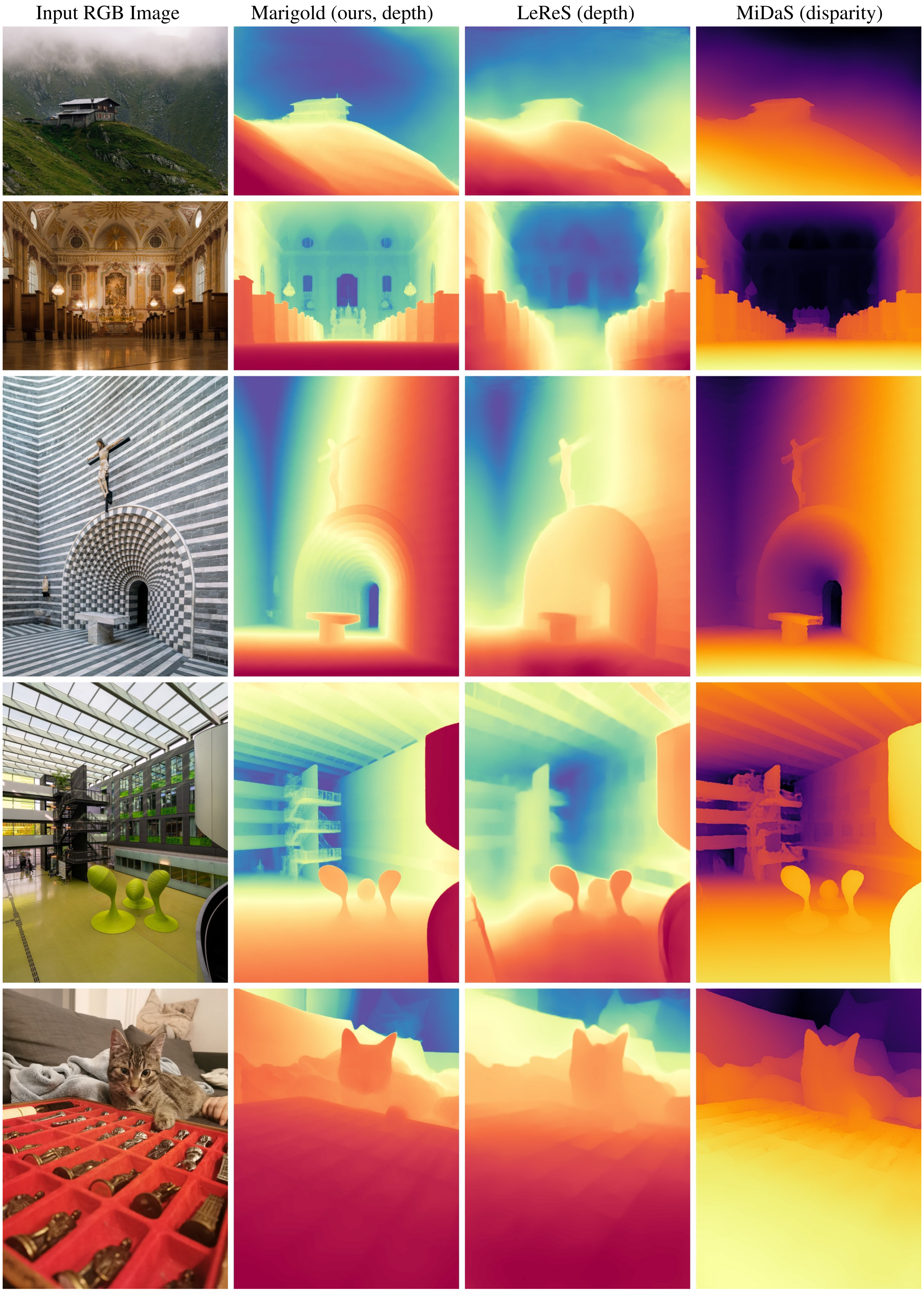
MiDaS (disparity)


Input RGB Image
Marigold (ours, depth)
LeReS (depth)
MiDaS (disparity)


Input RGB Image
Marigold (ours, depth)
LeReS (depth)
MiDaS (disparity)


Input RGB Image
Marigold (ours, depth)
LeReS (depth)
MiDaS (disparity)


Input RGB Image
Marigold (ours, depth)
LeReS (depth)
MiDaS (disparity)

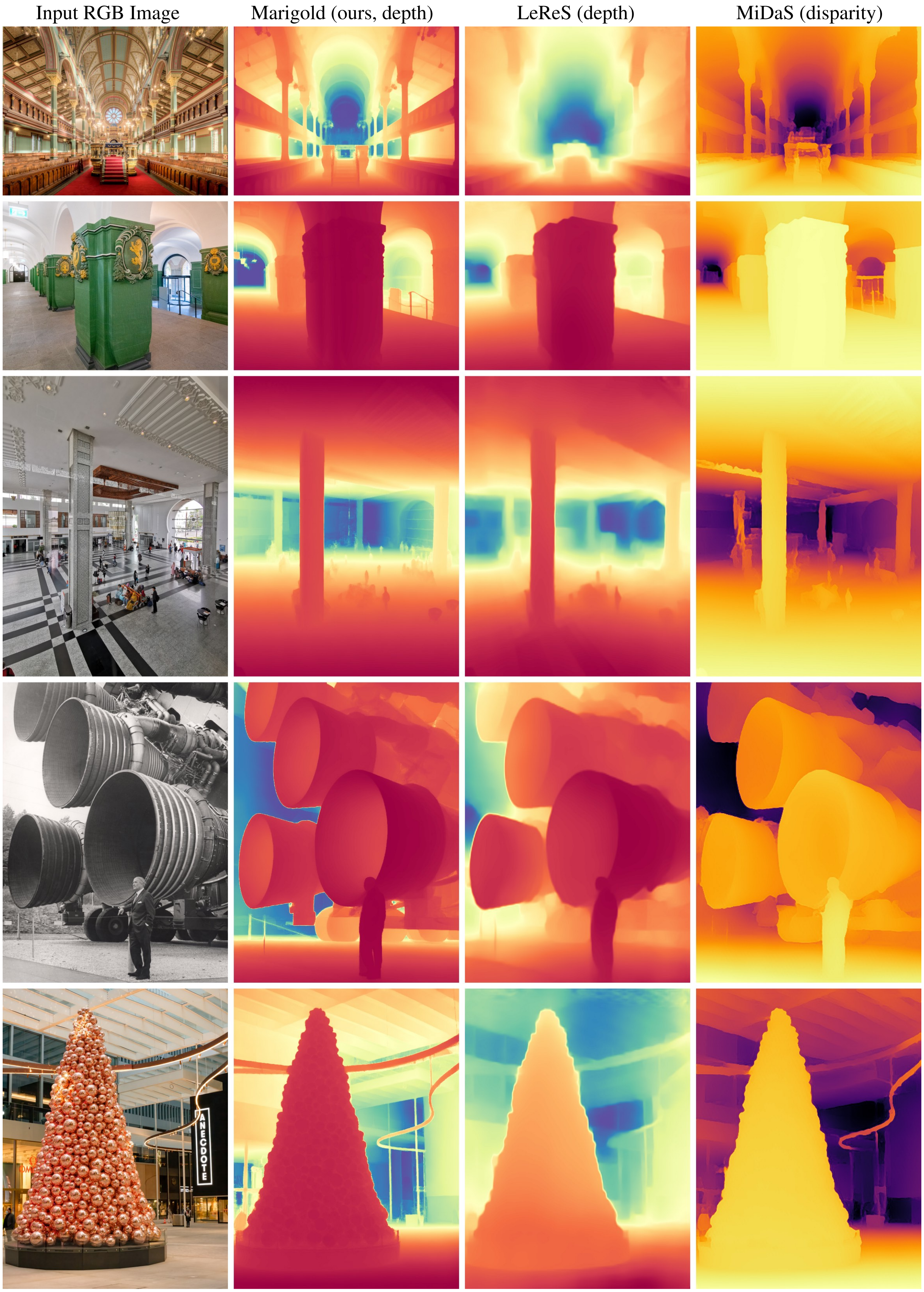


Input RGB Image

Marigold (ours, depth)

LeReS (depth)

MiDaS (disparity)

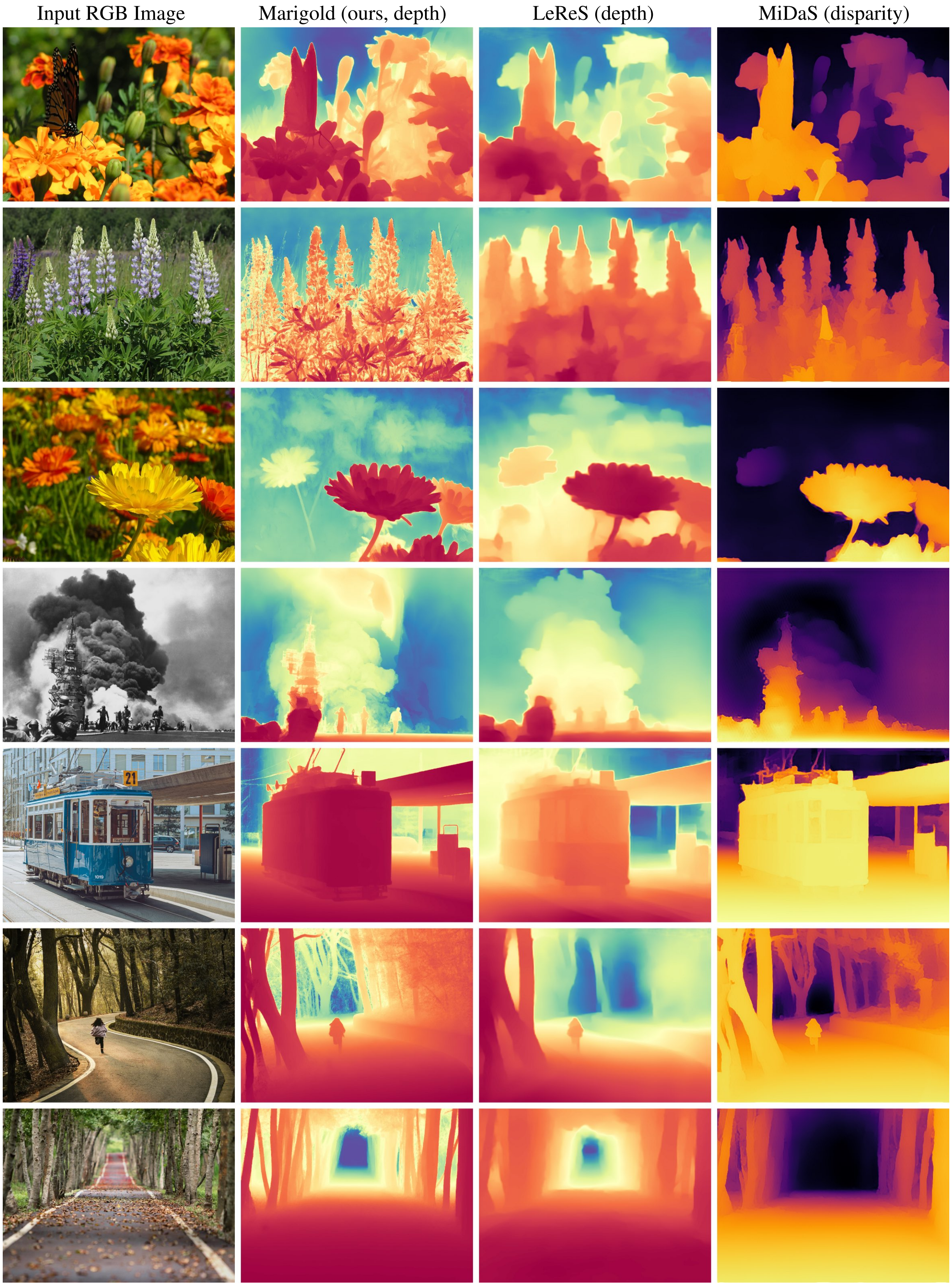


Input RGB Image

Marigold (ours, depth)

LeReS (depth)

MiDaS (disparity)



Input RGB Image

Marigold (ours, depth)

LeReS (depth)

MiDaS (disparity)

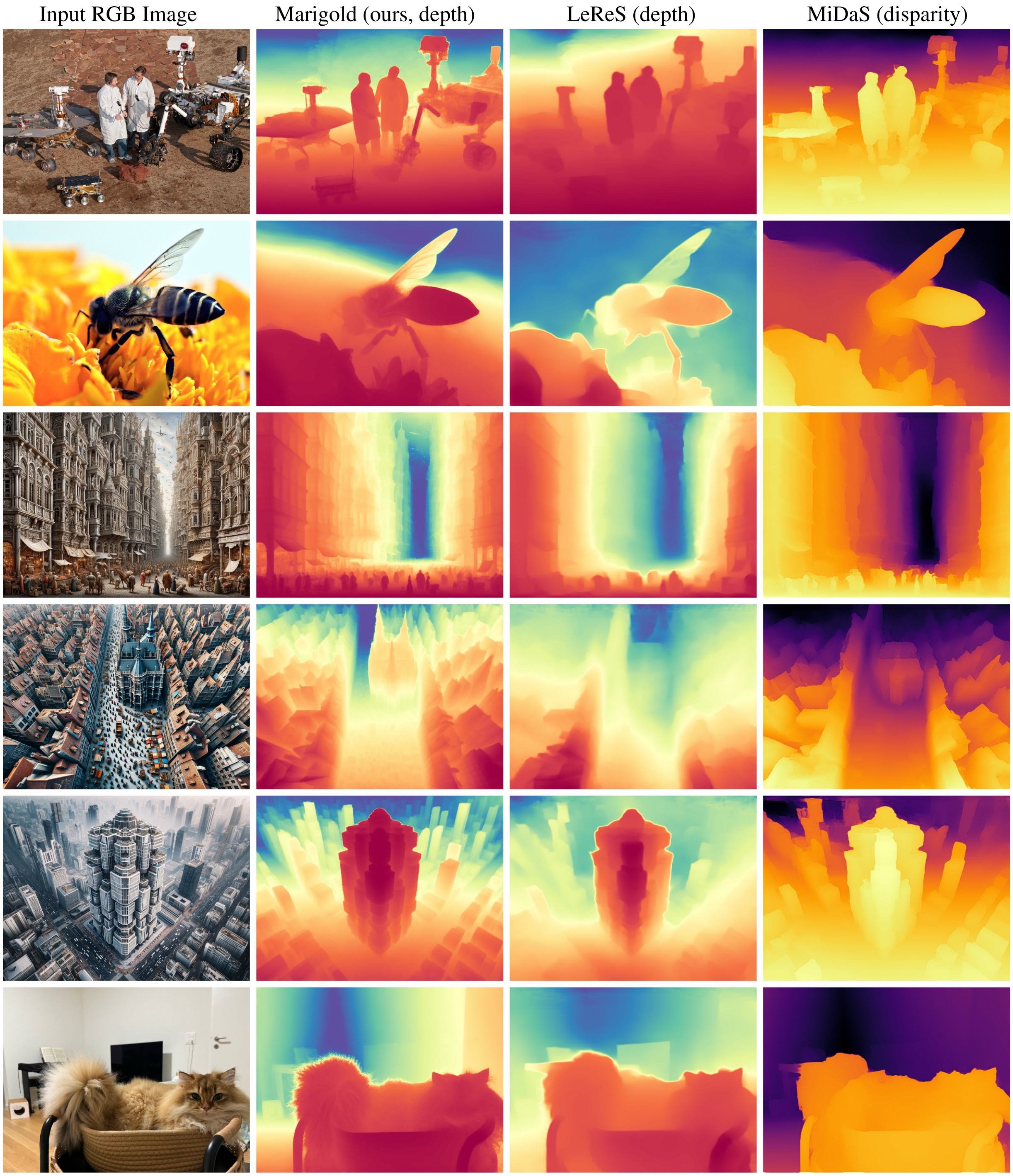
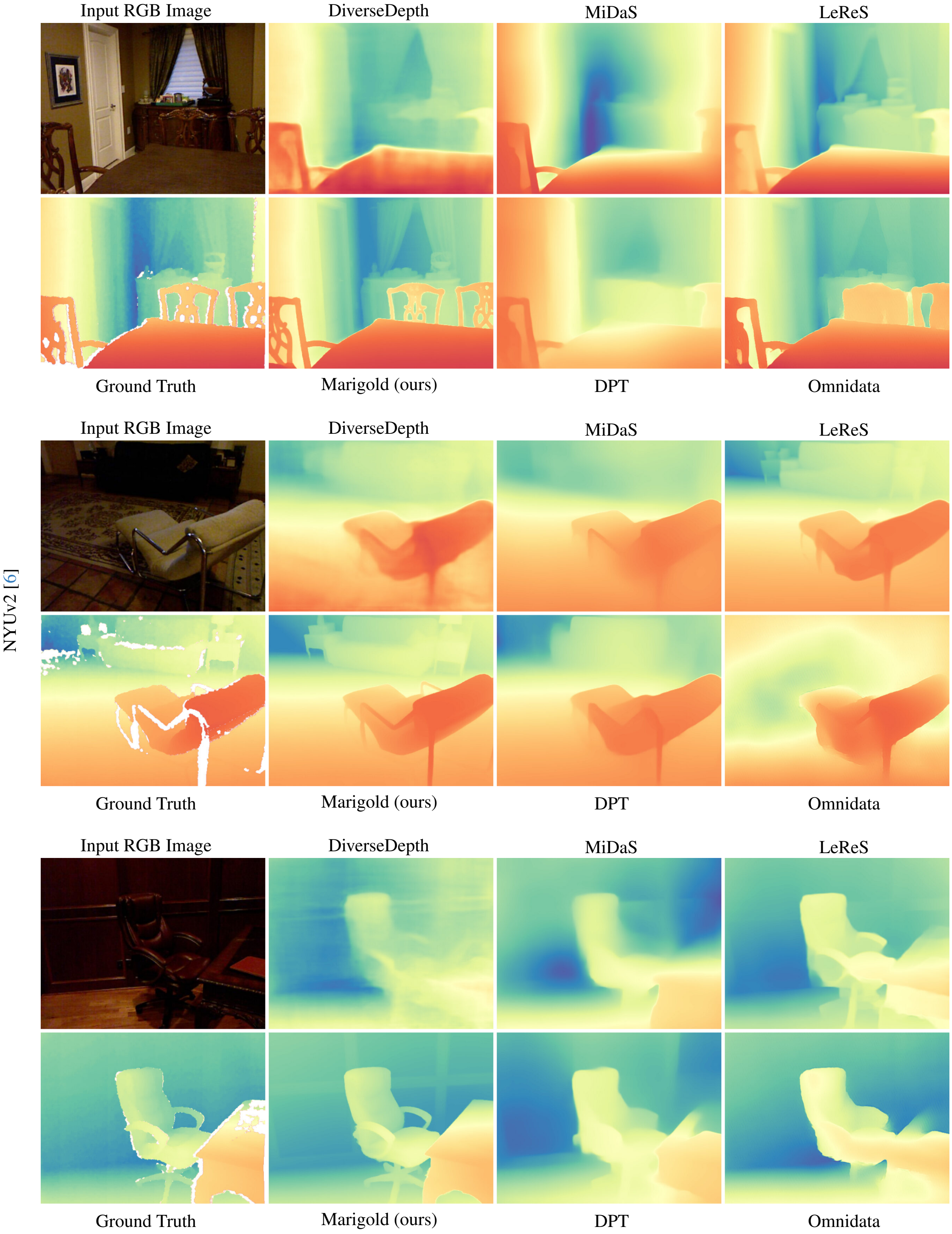
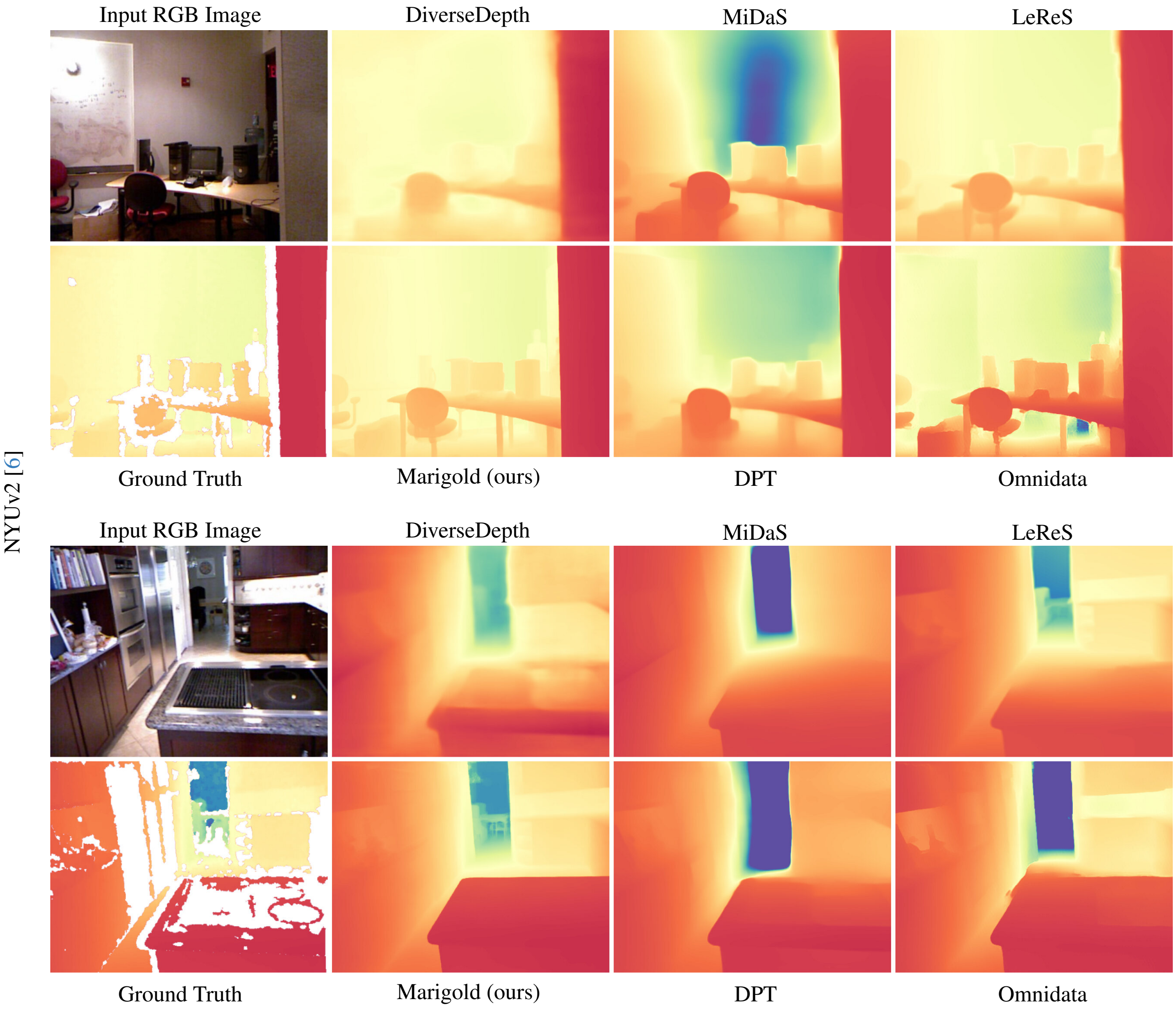
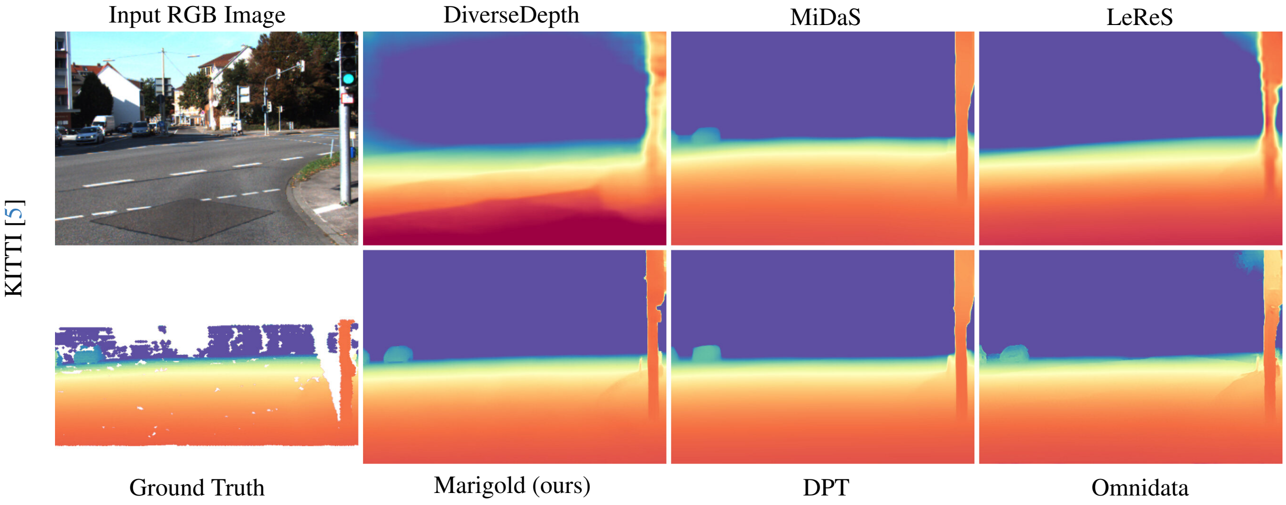
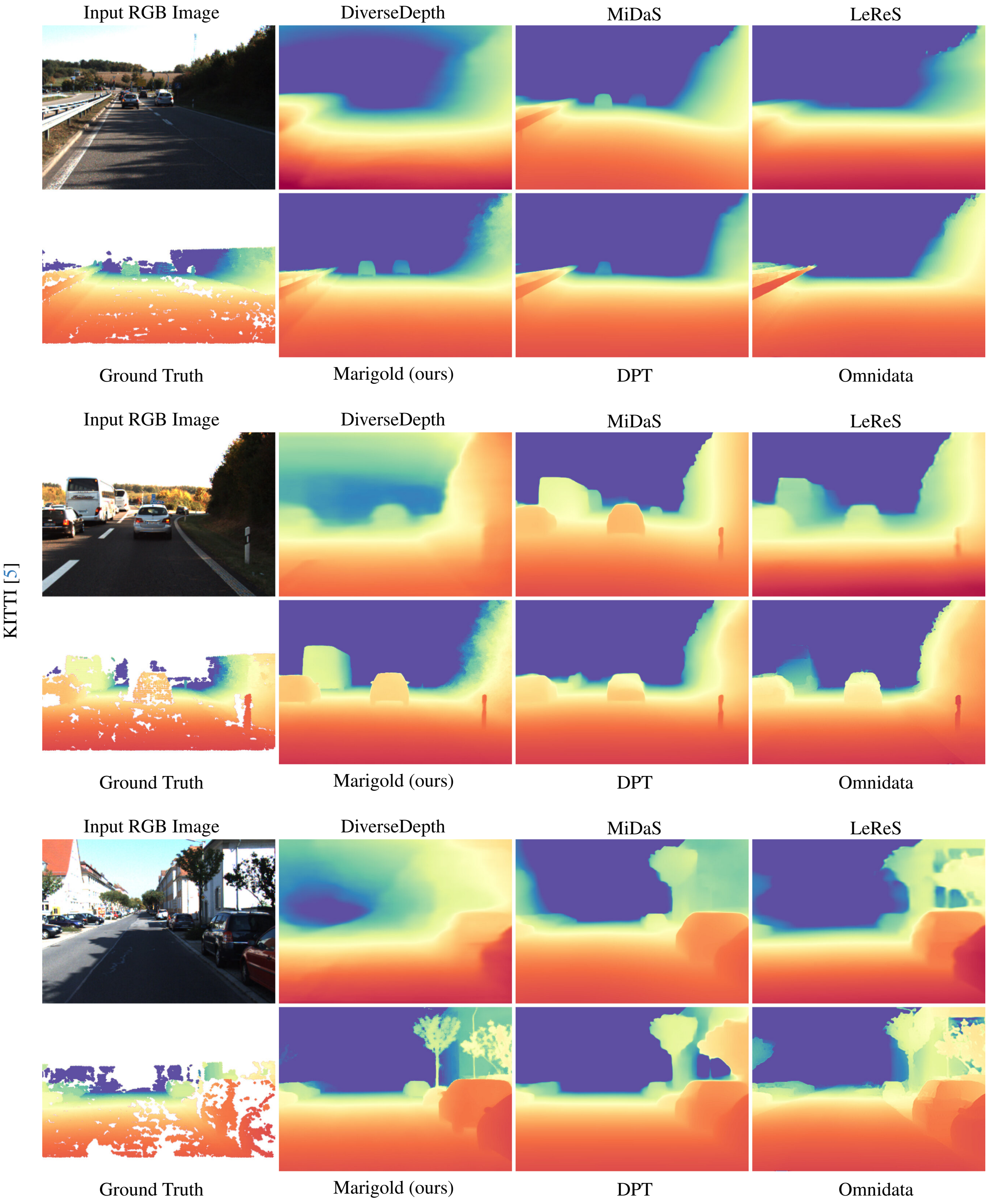


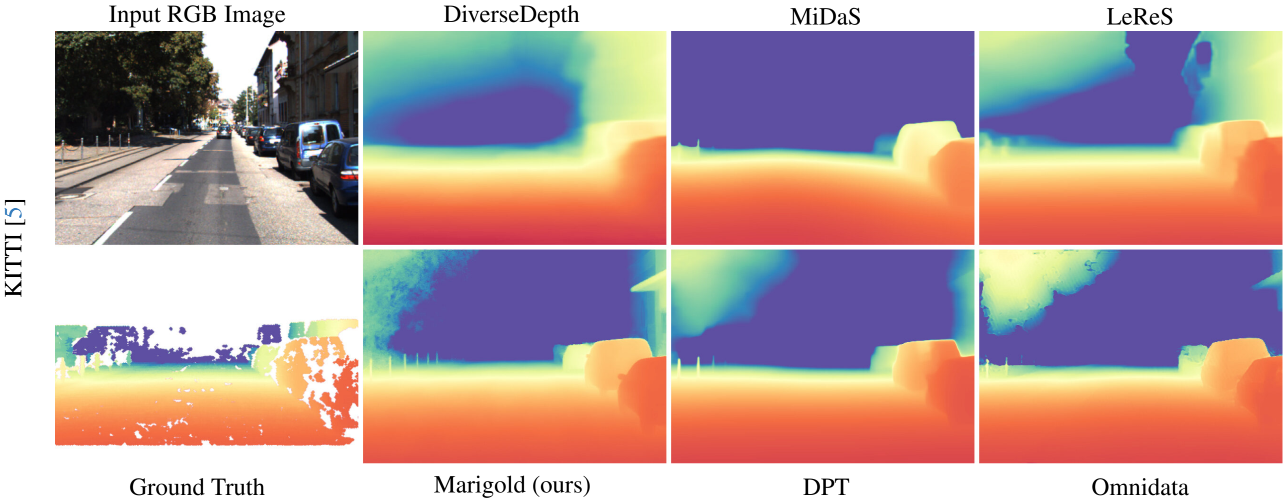
Figure S3. **Qualitative comparison on in-the-wild scenes.** Marigold and LeReS predict depth (with red indicating closer and blue indicating farther distances), while MiDaS predicts disparity (with yellow signifying closer and purple signifying farther distances).

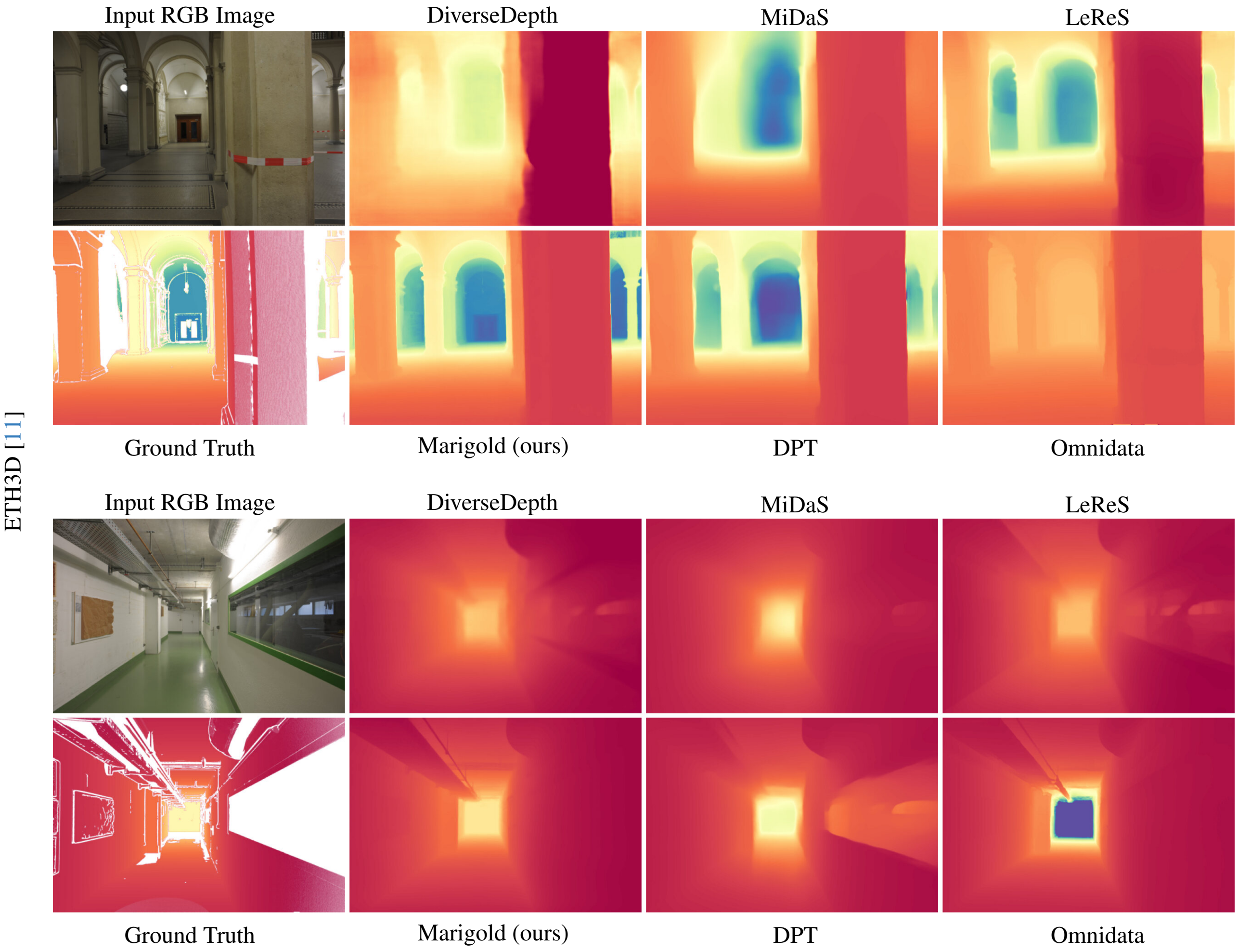

Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
NYUv2 [6]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

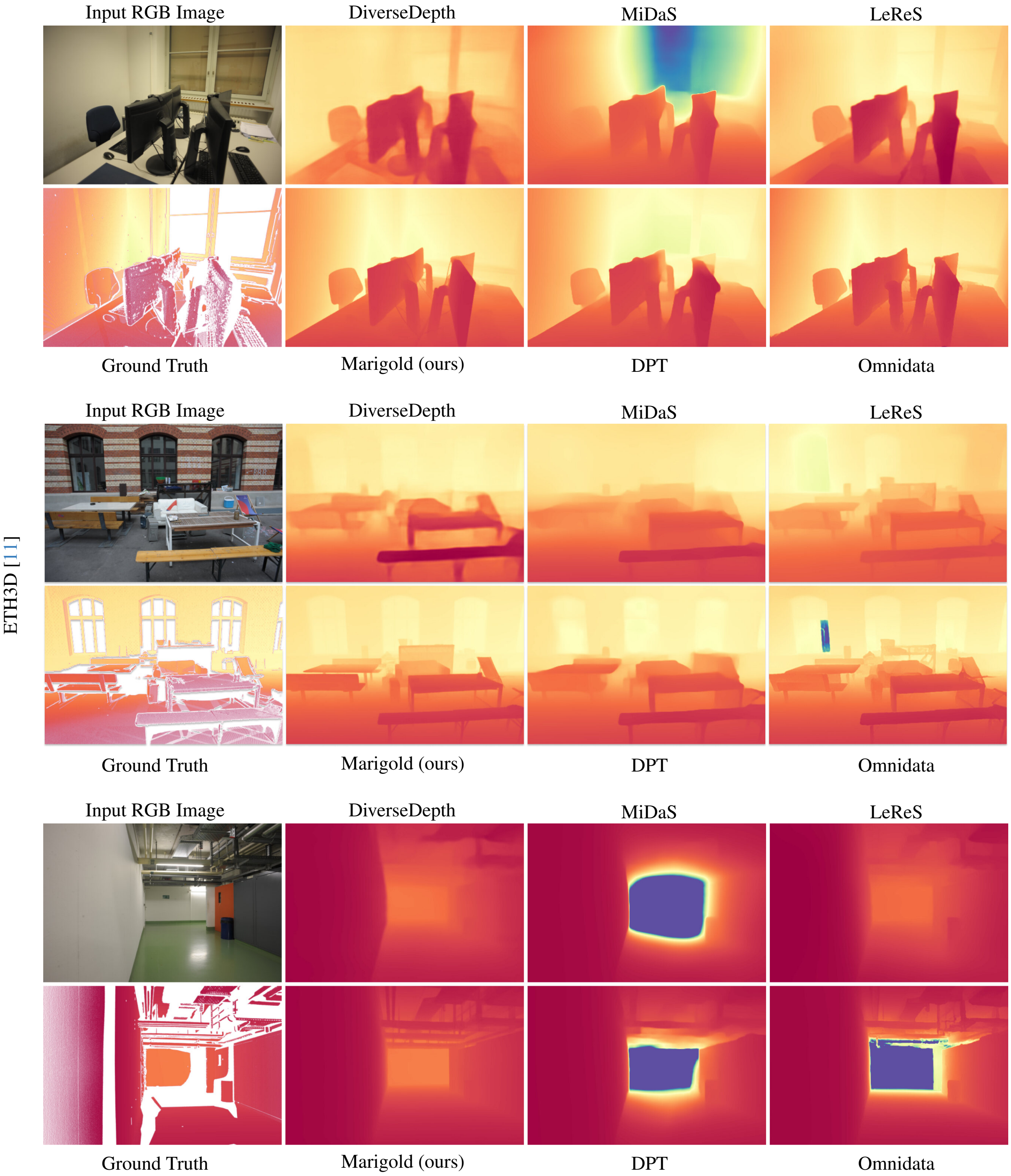

NYUv2 [6]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

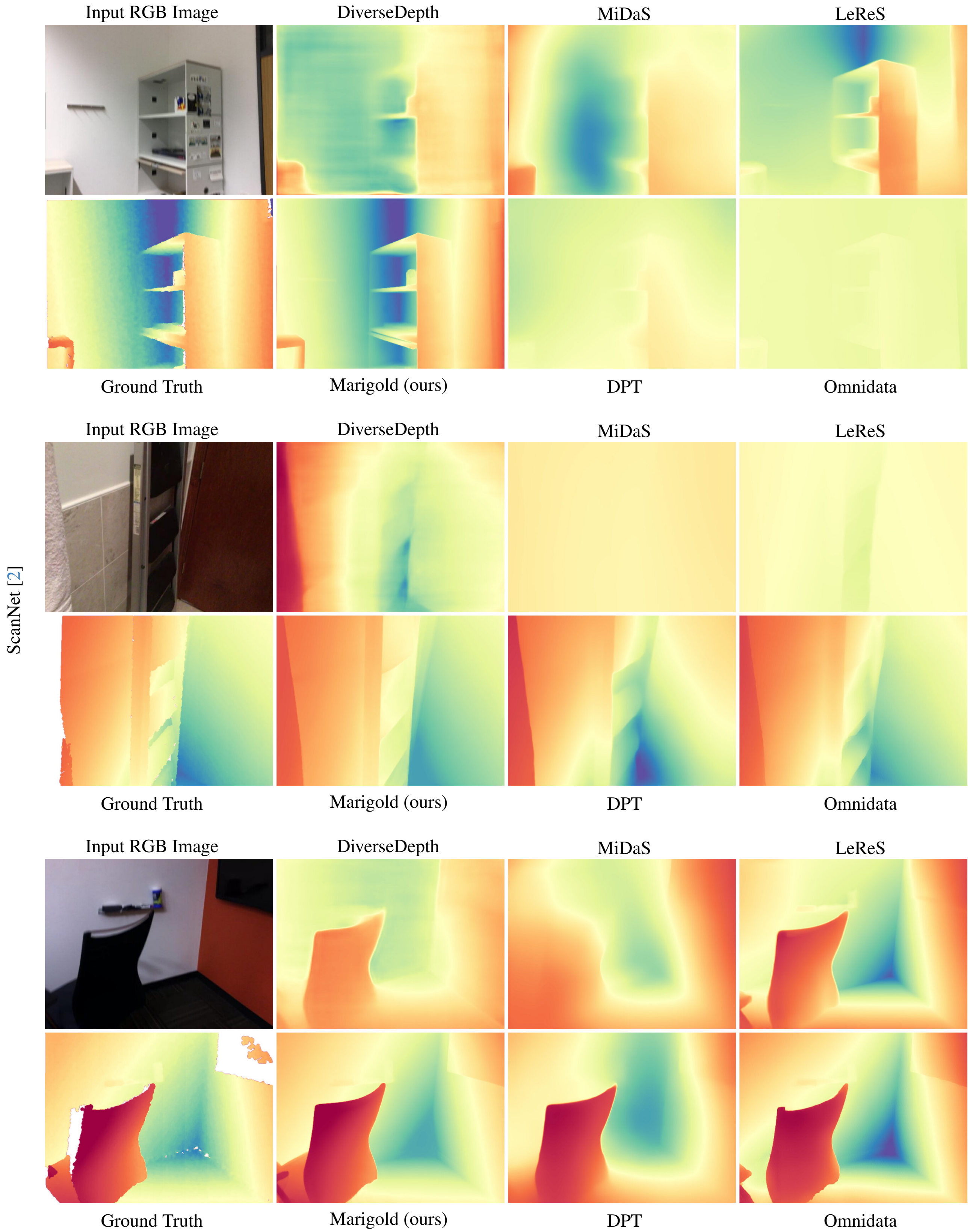

KITTI [5]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

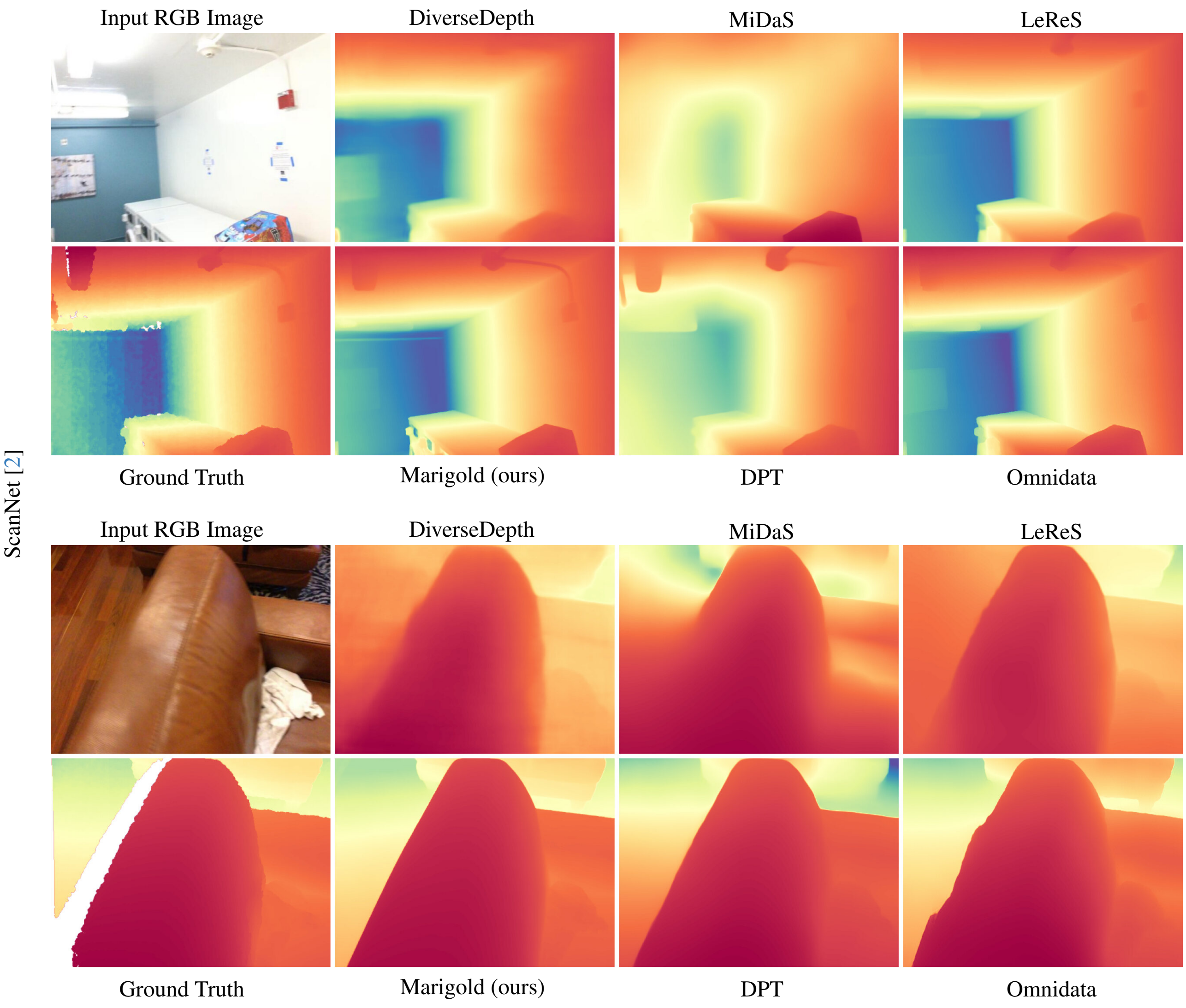

Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
KITTI [5]
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

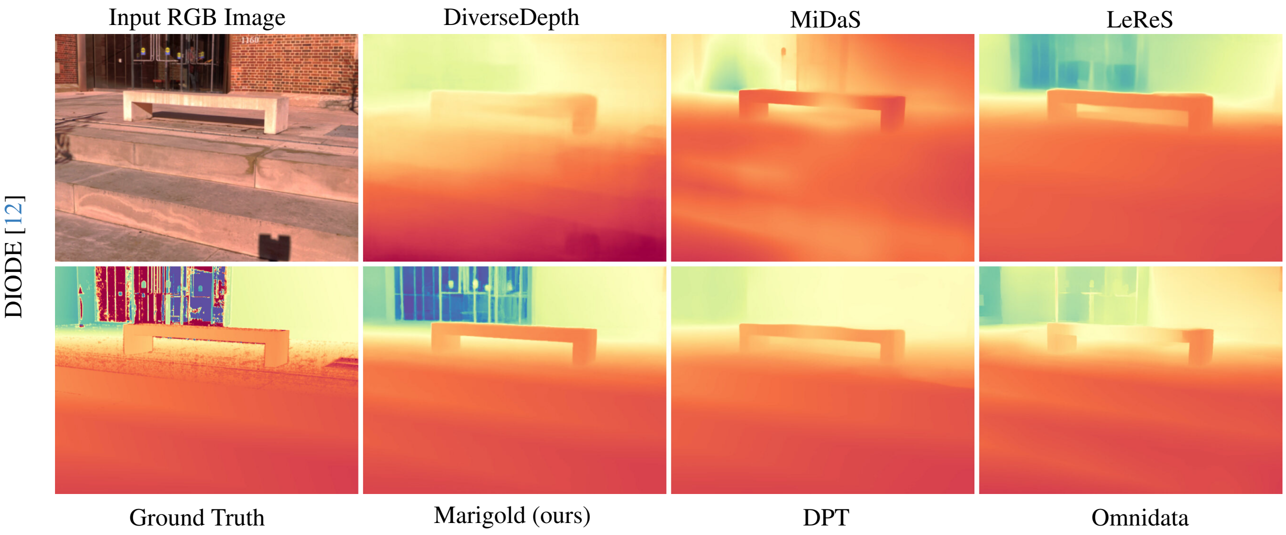

KITTI [5]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


ETH3D [11]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
ETH3D [11]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


ScanNet [2]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


ScanNet [2]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


DIODE [12]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

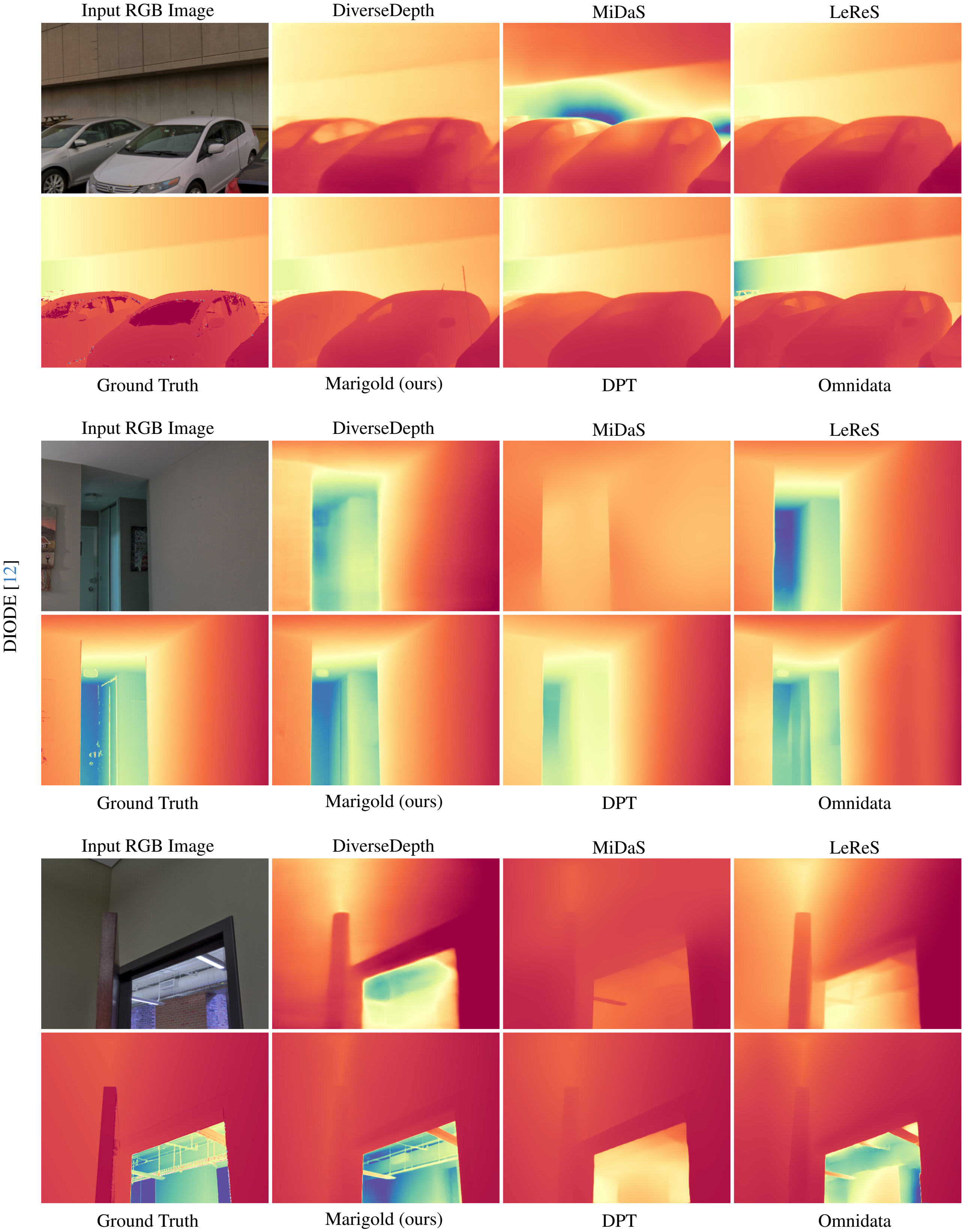
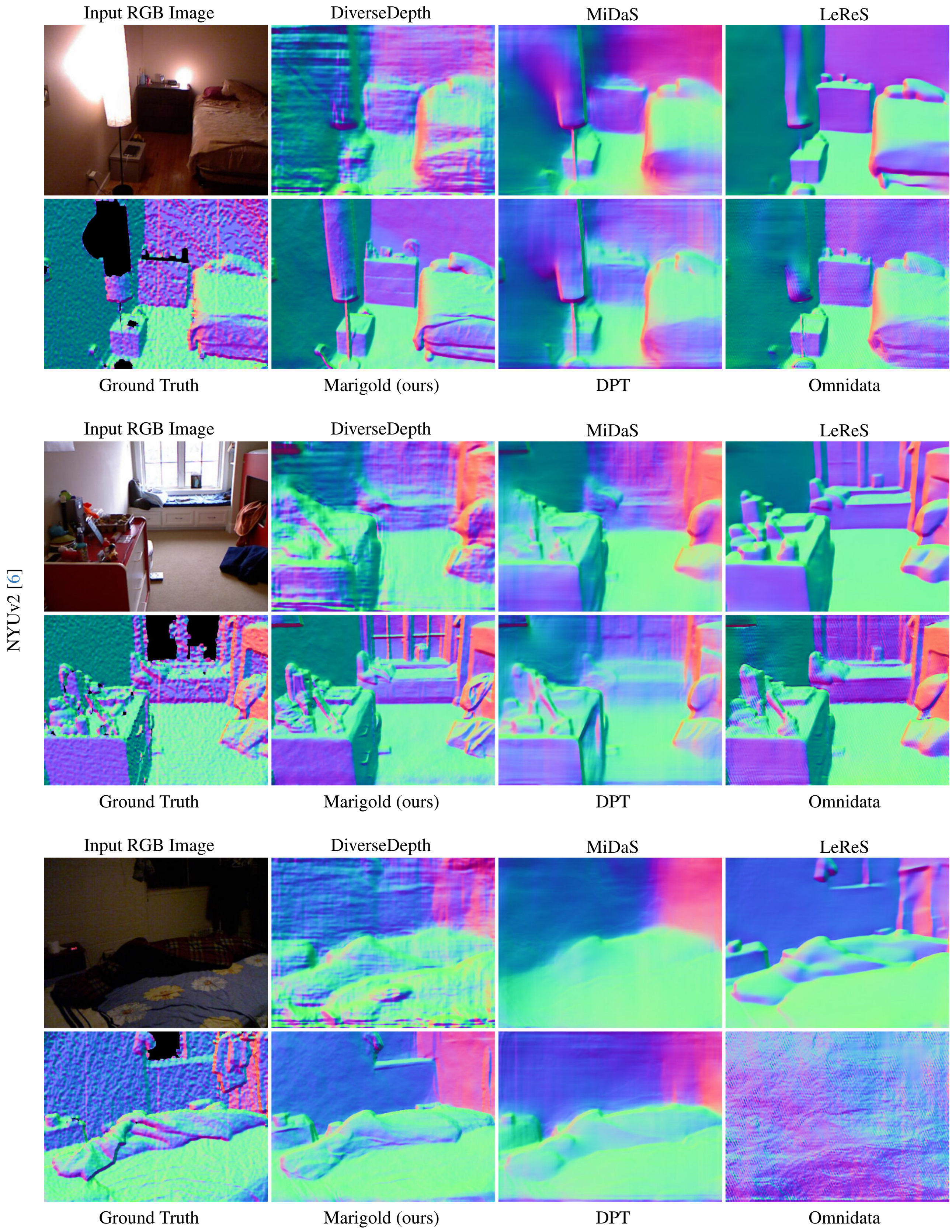
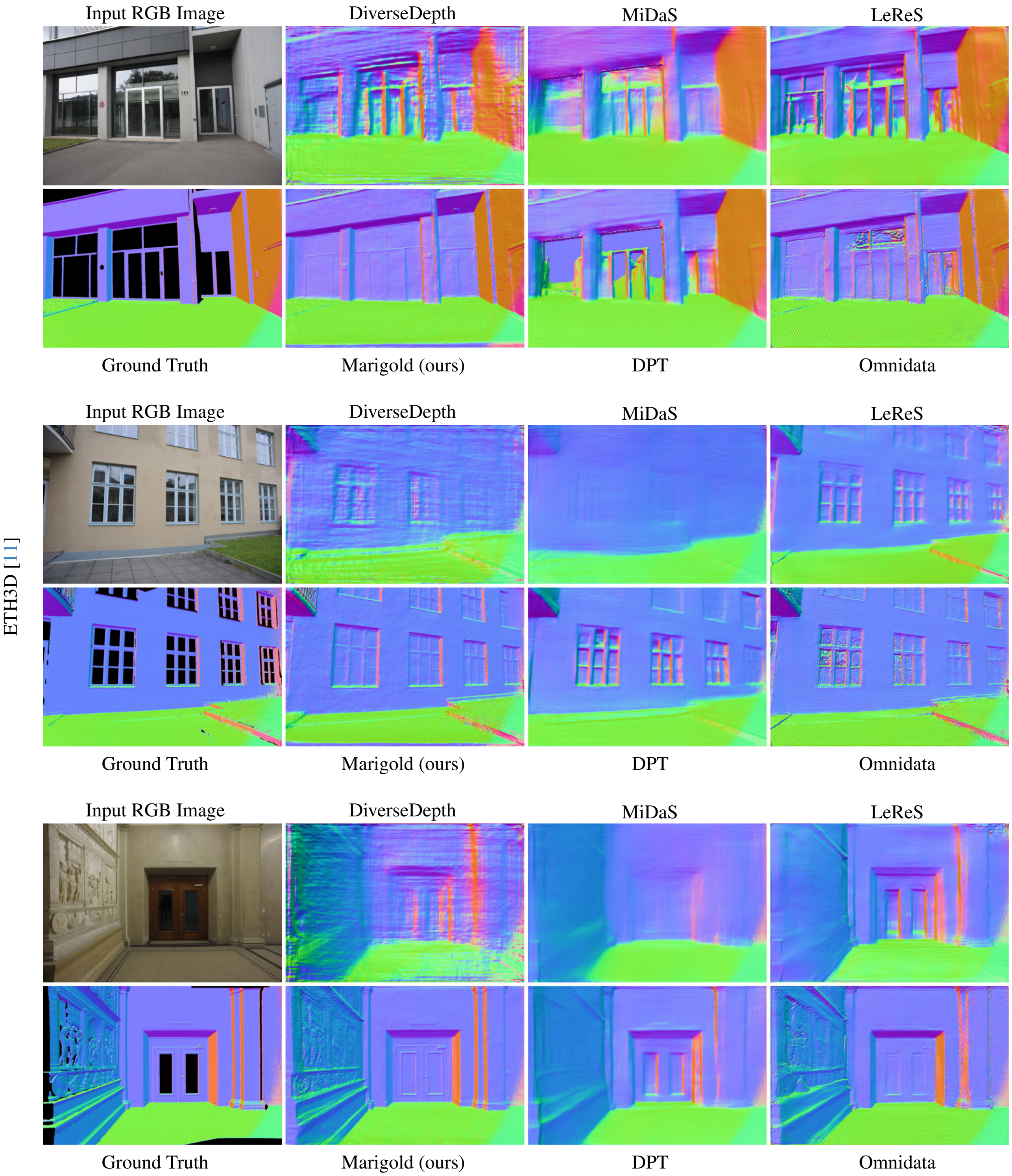


Figure S4. **Qualitative comparison (depth)** of monocular depth estimation methods across different datasets. Predictions are aligned to ground truth. For every sample, the color coding is consistent across all depth maps.


NYUv2 [6]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata


Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
ETH3D [11]
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata
Input RGB Image
DiverseDepth
MiDaS
LeReS
Ground Truth
Marigold (ours)
DPT
Omnidata

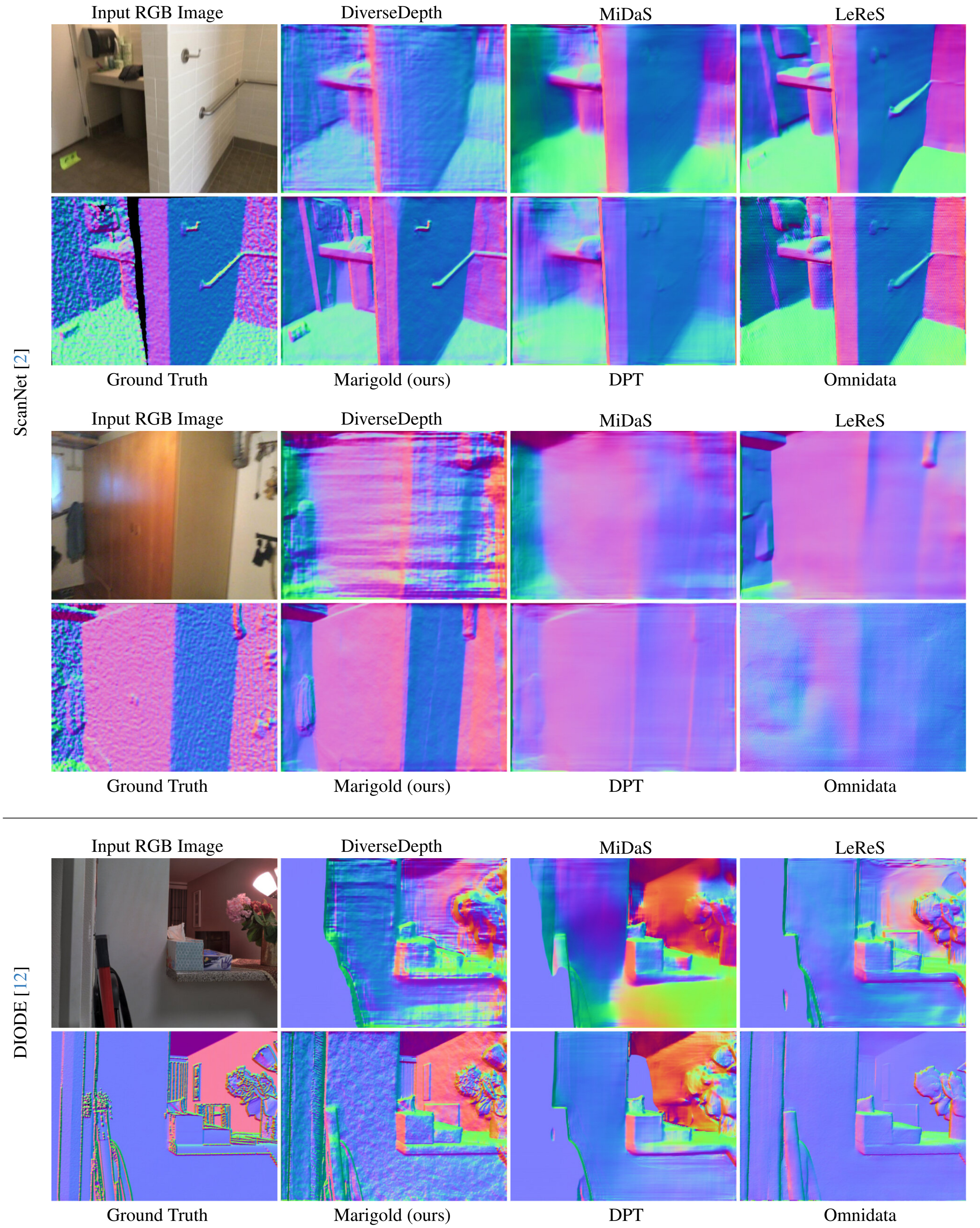


Figure S5. **Qualitative comparison (unprojected, colored as normals)** of monocular depth estimation methods across different datasets. Ground truth normals are derived from the ground truth depth maps.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. 1, 2

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 3, 19, 20

[3] 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. 1, 3

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 1, 2, 3, 15, 16, 17

[6] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 1, 2, 3, 14, 15, 22

[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*, 2020. 1, 3

[8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1, 3

[9] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 1, 2

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[11] 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. 3, 17, 18, 23

[12] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv preprint arXiv:1908.00463*, 2019. 3, 20

[13] 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. 1, 3

[14] 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*, 2021. 3

[15] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *NeurIPS*, 35, 2022. 1