FiffDepth: Feed-forward Transformation of Diffusion-Based Generators for Detailed Depth Estimation

Supplementary Material

1. Implementation Details

We implement FiffDepth using PyTorch, employing Stable Diffusion v2 as the backbone and adhering to the original pre-training setup with a v-objective. Text conditioning is disabled. To maximize the benefits of pre-trained models, we adopt the Depth Anything V2-Large model as the DINO supervision model, as it was the most advanced version available at the time. To enable training on a single GPU, gradients are accumulated over 16 steps. The Adam optimizer is employed with a learning rate of $3 \cdot 10^{-5}$. Random horizontal flipping is applied as a data augmentation technique. The training process used a batch size of 32, and the model converged after approximately $10,000$ iterations. Training to convergence takes approximately 1.5 days on a single Nvidia Tesla V100 GPU.

Training datasets. Hypersim [5] is a photorealistic indoor dataset with 461 scenes. From the official split, we use 54k samples from 365 scenes, filtering out incomplete ones. RGB images and depth maps are resized to 480×640, and the original distances relative to the focal point are converted into depth values relative to the focal plane. The second dataset, Virtual KITTI [1], is a synthetic street scene data set that features five scenes under varying conditions, such as weather and camera perspectives. We use four scenes, totaling around 20K samples, cropping the images to the KITTI benchmark resolution [3] and setting the far plane at 80 meters.

Evaluation datasets. For the evaluation of affine-invariant depth, we use the same datasets and evaluation protocol as Marigold. These datasets include NYUv2 [7], ScanNet [2], KITTI [3], ETH3D [6], and DIODE [8]. NYUv2 [7] and ScanNet [2] are indoor datasets collected using RGB-D Kinect sensors. For NYUv2, we use the test split containing 654 images. From ScanNet, 800 images are randomly sampled from the 312 validation scenes. KITTI [3], a dataset of street scenes with sparse metric depth captured by LiDAR, is evaluated using the Eigen test split [14], which consists of 652 images. ETH3D [6] and DIODE [8] are high-resolution datasets with depth maps generated from LiDAR measurements. ETH3D contains 454 samples with ground truth depth maps, while DIODE's validation set includes 325 indoor and 446 outdoor samples.

2. Detailed Ablation Studies

Studies about L_k. When keeping the original trajectories, if we only predict image features, the results are distorted because image features are very different from depth features (Figs. 2 and 3). Please refer to the w/o blend depth results in Table 1.

If we only predict depth features, the U-Net fails to improve the results because the previous trajectories are fully adapted to depth and cannot be applied at $t = 0$ when the input is an image latent. This effectively cuts off the connection between the input at $t = 0$ and the earlier stages, leading to results similar to those without keeping trajectories. For without keeping trajectories results, please refer to the w/o L_k results in Table 1. Without L_k, the details are also reduced accordingly (Figs. 2 and 3), as reflected in the results shown in Table 2. When predicting the blended latent, we can achieve a smooth transition from image to depth.

Regarding the effects of different blend ratio weights, please refer to Figure 1. It can be observed that as the proportion of depth increases, the accuracy of the results improves accordingly. For simplicity, we set $\gamma = 0.5$, although this is not optimal for all datasets. The γ value remains constant throughout the entire training process.

Usage of DINO supervision. As for the use of DINO, we found that the performance drops significantly without DINO supervision (Table 1). If we use depth at the d_0 stage, there is no noticeable impact on metrics such as AbsRel, but it has a substantial effect on details, as shown by the Zero-shot Boundary metrics in Table 2. These phenomena can also be observed in Figs. 4 and 5.

Different size of DINOv2. Since DINOv2 has different size versions, we use various sizes of Depth Anything models—small, base, and large—as our DINOv2 model. The results, shown in Table 1, indicate that while there are some differences in test metrics across these versions, all results demonstrate that our method effectively transfers the generalization capabilities of DINOv2, regardless of its size, to the diffusion model. In all experiments except for those mentioned here, our method uses the Depth Anything Large version.

3. More Comparison with Other Methods

We have added more comparisons with GenPercept [9] and Lotus [4] here. Qualitative comparisons are shown in Figs. 6 to 11. Although their feed-forward approach has a runtime almost identical to ours, their method falls short in terms of accuracy, generalization ability, and detail preservation compared to ours.

We also add more comparison results with Depth Any-

thing v2 and Depth Pro in Figs. 12 to 15, and with GeoWizard and DepthFM in Figs. 16 to 20.

4. More visual results

We have also included the results of depth to point cloud conversion in Figs. 21 to 27.

References

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

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[4] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 1

[5] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 1

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 1

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1

[8] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 1

[9] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024. 1

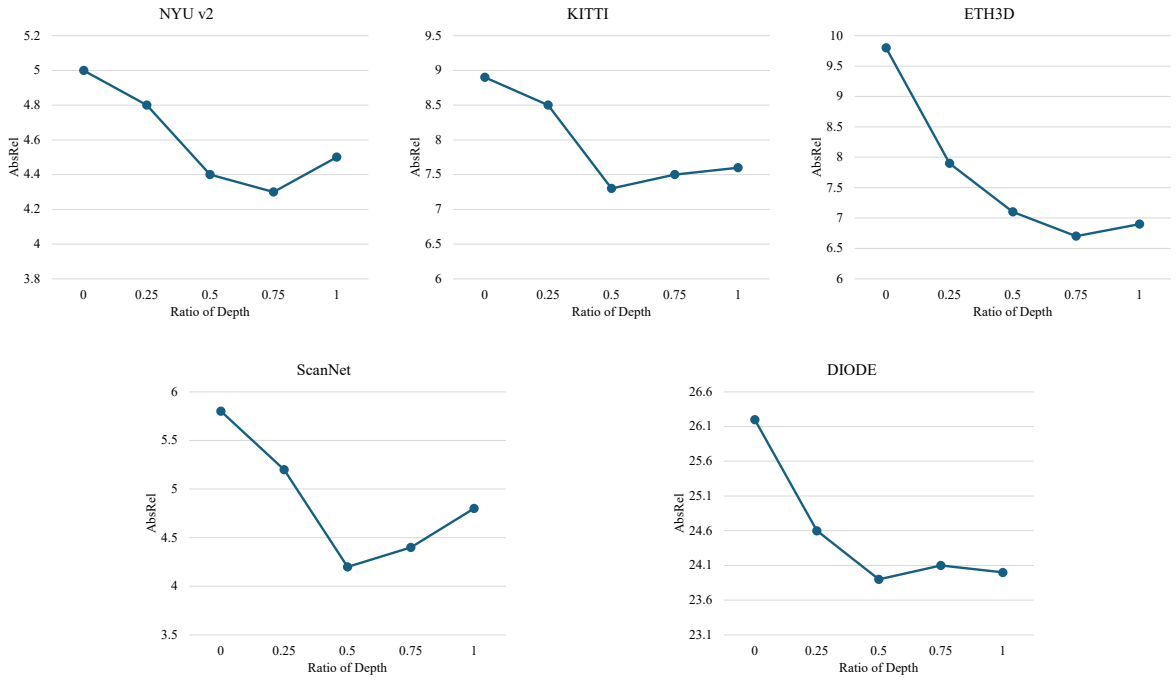


Figure 1. **Effects of different blend weight ratios.**

Method	Training Data	NYUv2		KITTI		ETH3D		ScanNet		DIODE-Full		DA-2K
		AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	Acc (%)
w/o blend depth	274K	5.0	96.9	8.9	93.1	9.8	95.7	5.8	96.5	26.2	79.0	91.2
w/o L_k	274K	4.8	96.1	7.9	91.4	7.3	95.2	5.4	96.1	25.3	76.4	95.2
w/o DINO	74K	5.5	96.5	9.8	91.3	6.3	95.9	6.0	95.8	28.5	77.8	87.9
DINO-L on d_0	274K	4.6	96.4	7.8	92.9	7.6	96.7	4.9	96.6	25.1	77.6	95.7
FiffDepth-S (Ours)	274K	5.2	97.1	7.9	93.4	7.5	96.8	4.6	97.3	24.2	77.5	94.8
FiffDepth-B (Ours)	274K	4.7	97.6	7.8	93.7	7.2	97.0	4.3	97.8	24.1	77.9	96.9
FiffDepth-L (Ours)	274K	4.4	97.8	7.3	93.5	7.1	97.2	4.2	97.9	23.9	78.1	97.1

Table 1. Quantitative comparison. We use AbsRel (absolute relative error: $|d^* - d| / d$) and δ_1 (percentage of $\max(d^*/d, d/d^*) < 1.25$). All metrics are reported as percentages.

Method	Sintel F1↑	Spring F1↑	iBims F1↑	AM R↑	P3M R↑	DIS R↑
GenPercept	0.080	0.040	0.126	0.074	0.115	0.049
e2e-ft	0.088	0.049	0.136	0.088	0.107	0.051
Lotus-D	0.081	0.062	0.141	0.065	0.109	0.047
Lotus-G	0.072	0.054	0.130	0.067	0.112	0.043
Marigold	0.068	0.032	0.149	0.064	0.101	0.049
GeoWizard	0.087	0.038	0.137	0.070	0.104	0.052
DepthFM	0.064	0.030	0.145	0.058	0.97	0.039
DepthAnything v1	0.261	0.045	0.127	0.058	0.094	0.023
w/o blend depth	0.113	0.047	0.148	0.072	0.114	0.063
w/o L_k	0.283	0.066	0.157	0.157	0.154	0.074
w/o DINO	0.394	0.073	0.161	0.168	0.157	0.087
DINO-L on d_0	0.312	0.061	0.154	0.146	0.152	0.069
FiffDepth (Ours)	**0.423**	**0.086**	**0.189**	**0.176**	**0.179**	**0.091**

Table 2. **Zero-shot boundary accuracy.** We provide the F1 score for datasets containing ground-truth depth and boundary recall (R) for those with matting or segmentation labels.

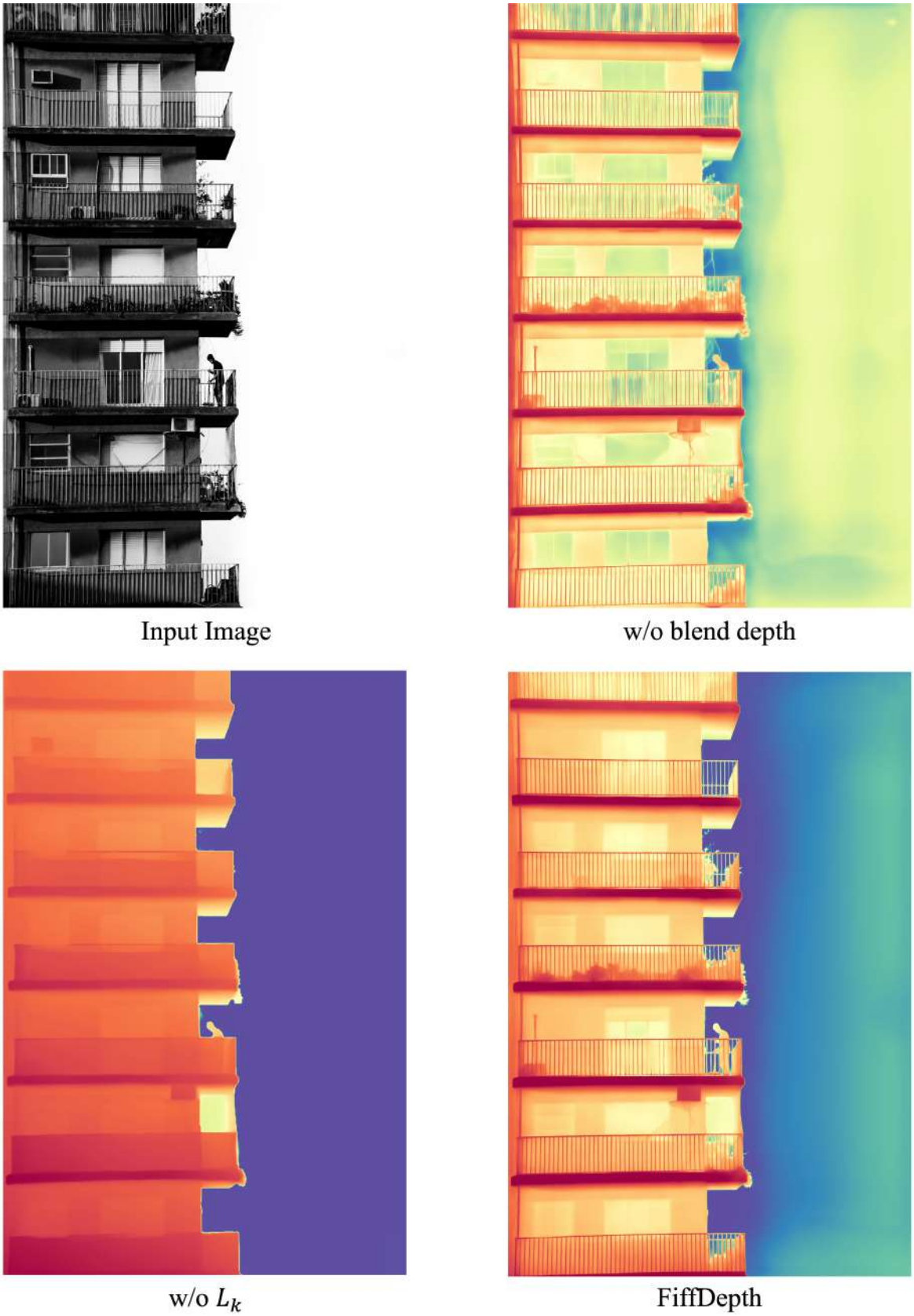


Figure 2. **Ablation studies.** The generalization capability and depth details of the method are affected when some essential components are missing.

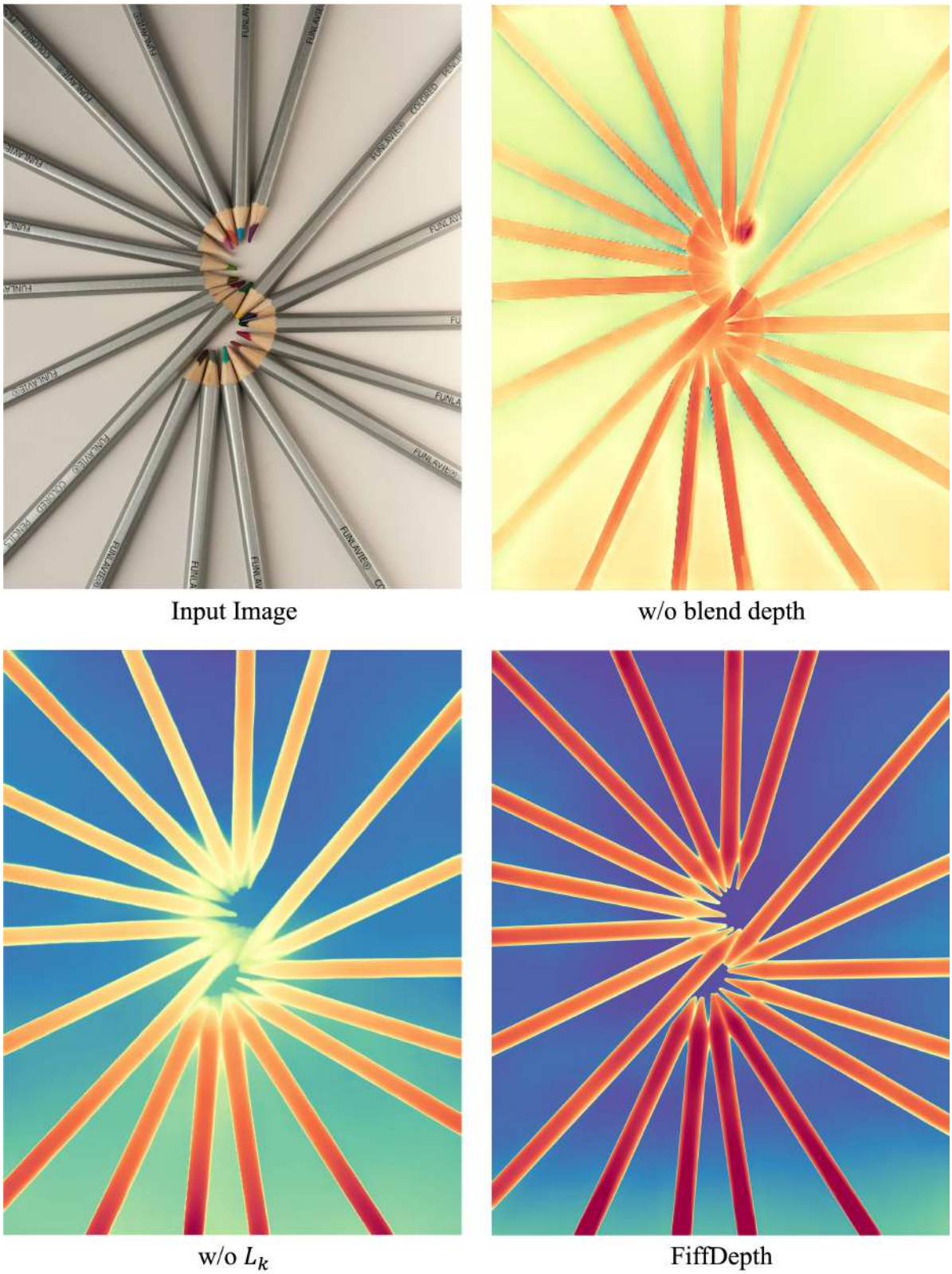


Figure 3. **Ablation studies.** The generalization capability and depth details of the method are affected when some essential components are missing.

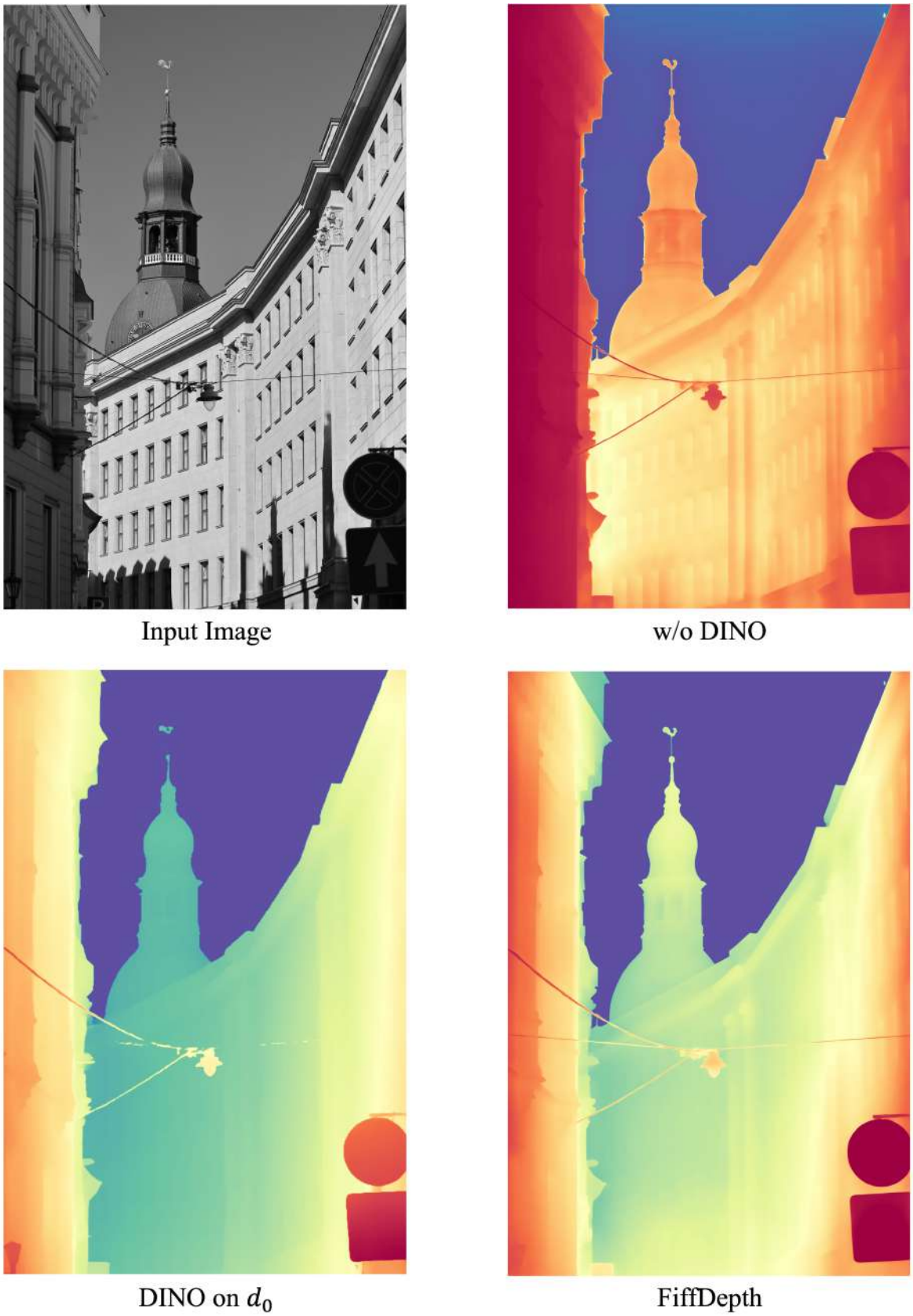


Figure 4. **Ablation studies.** The generalization capability and depth details of the method are affected when some essential components are missing.

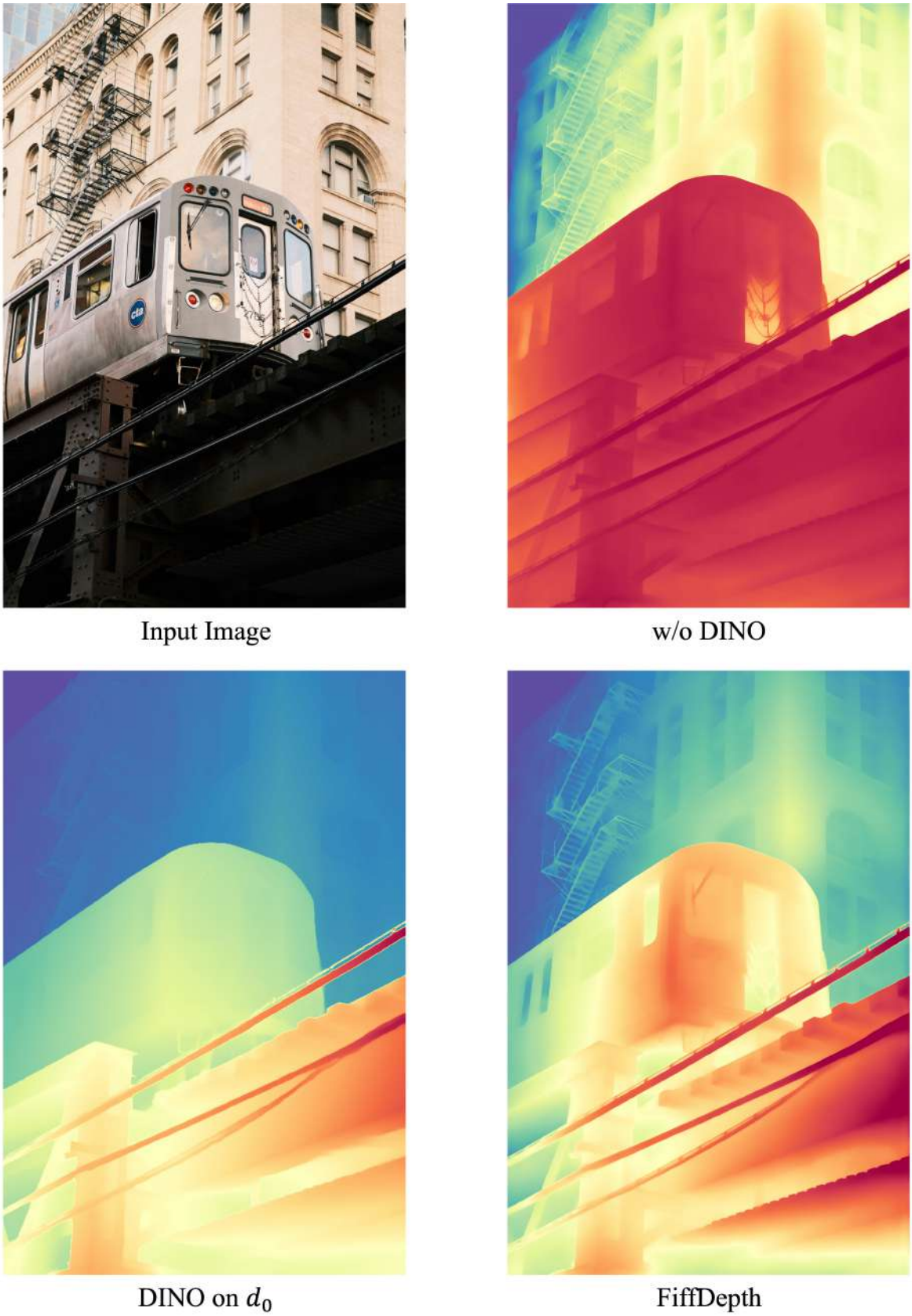


Figure 5. **Ablation studies.** The generalization capability and depth details of the method are affected when some essential components are missing.

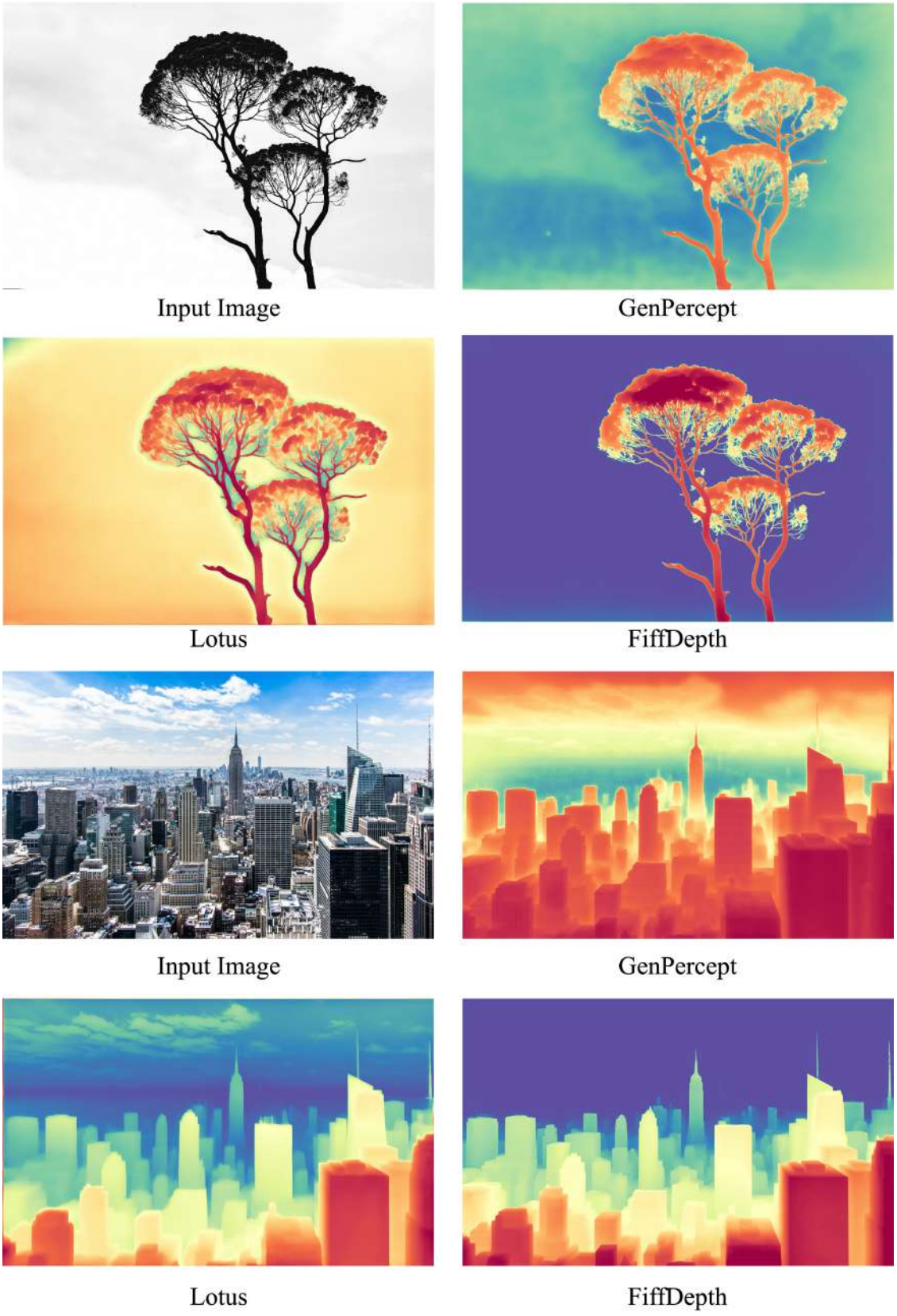


Figure 6. **Qualitative comparison with GenPercept and Lotus.**

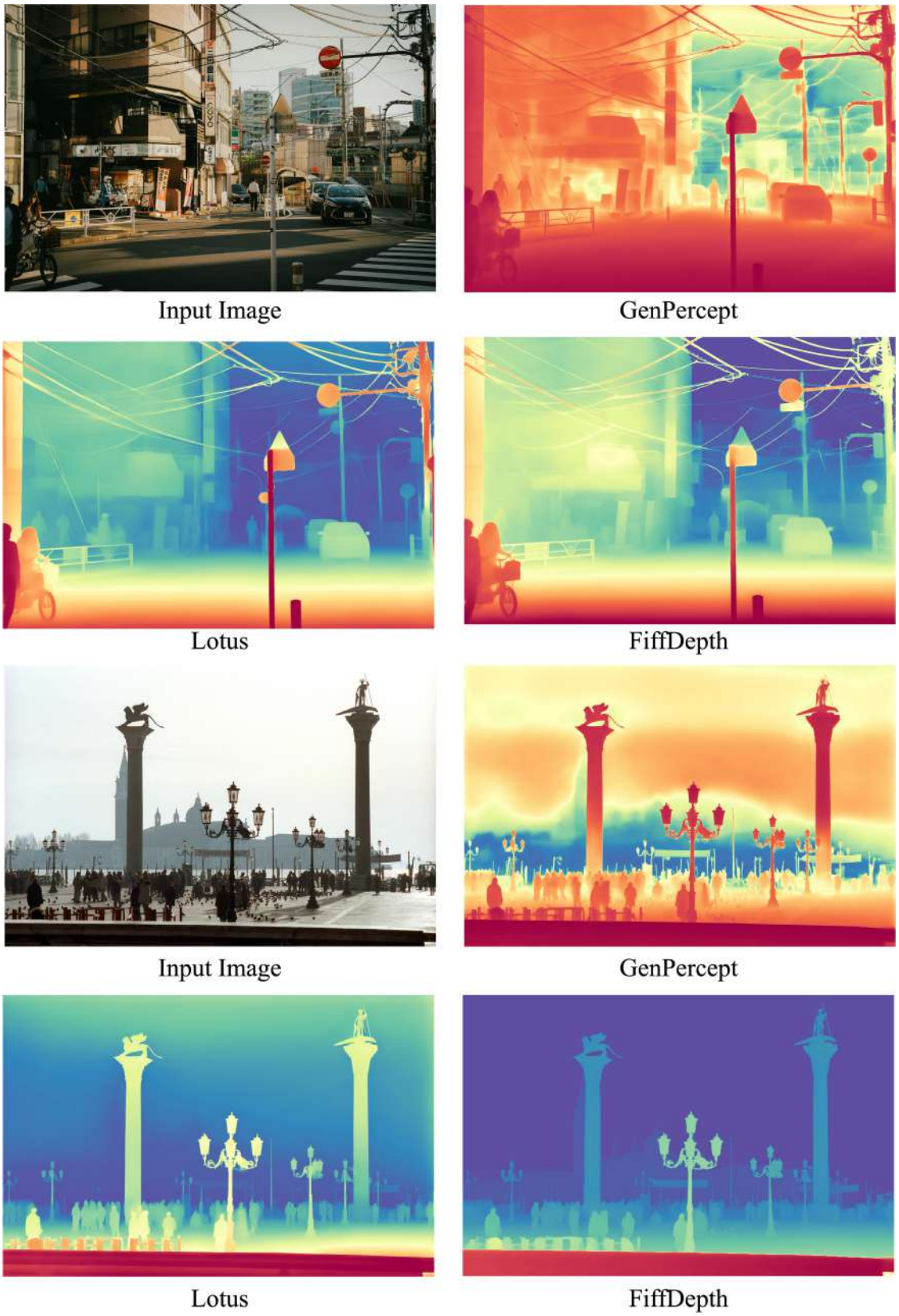


Figure 7. **Qualitative comparison with GenPercept and Lotus.**

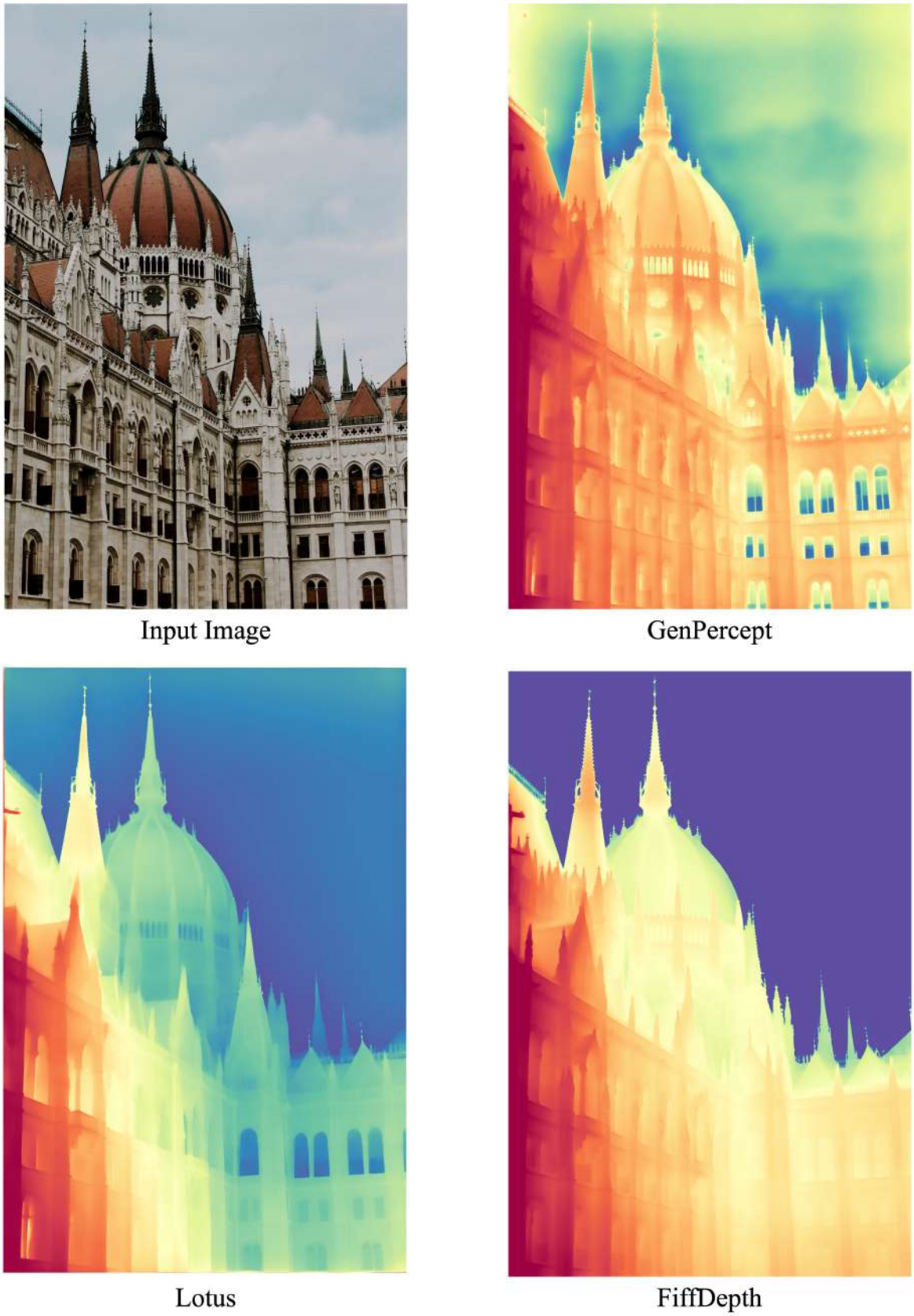


Figure 8. **Qualitative comparison with GenPercept and Lotus.**

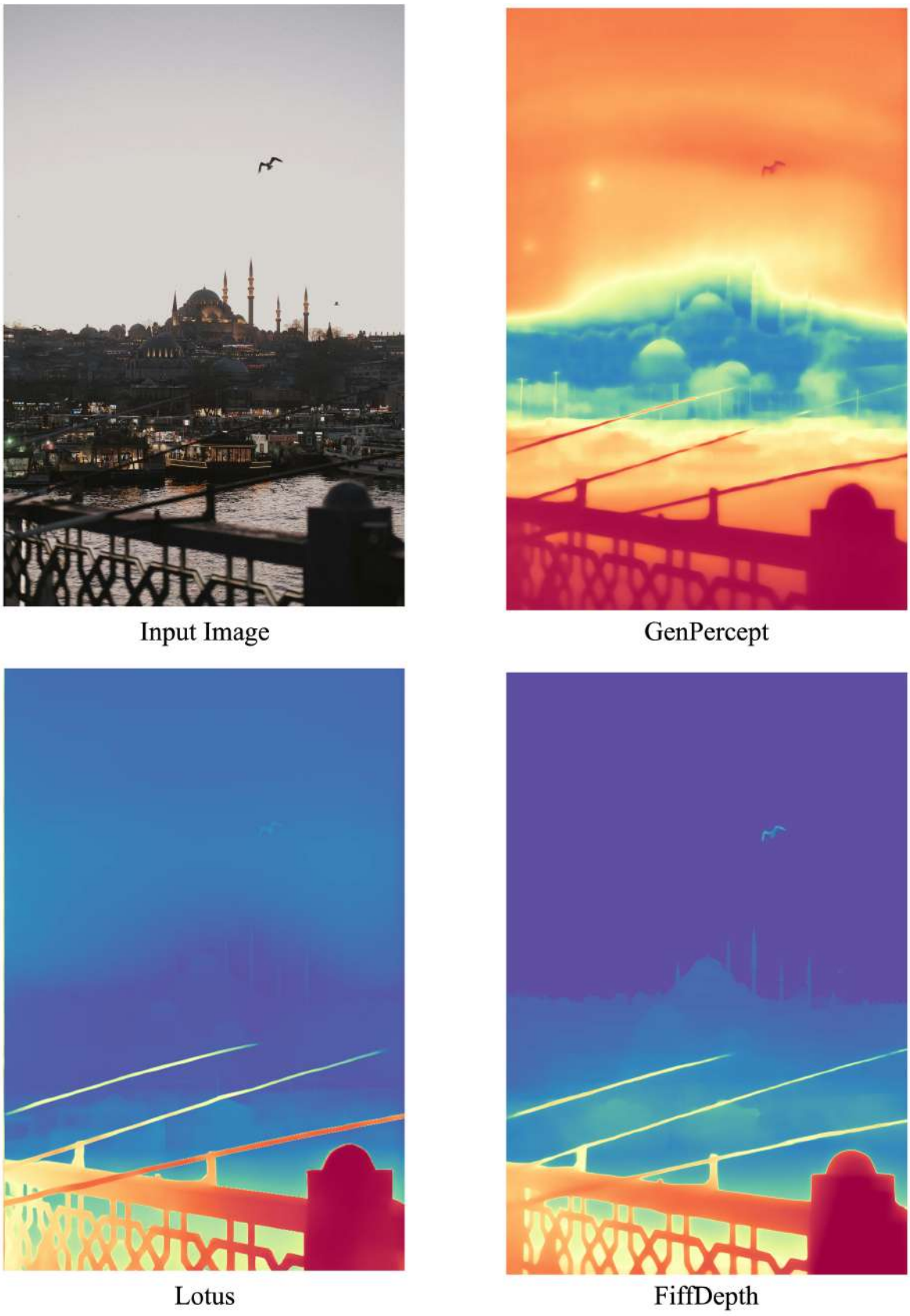


Figure 9. **Qualitative comparison with GenPercept and Lotus.**

Figure 10. **Qualitative comparison with GenPercept and Lotus.**

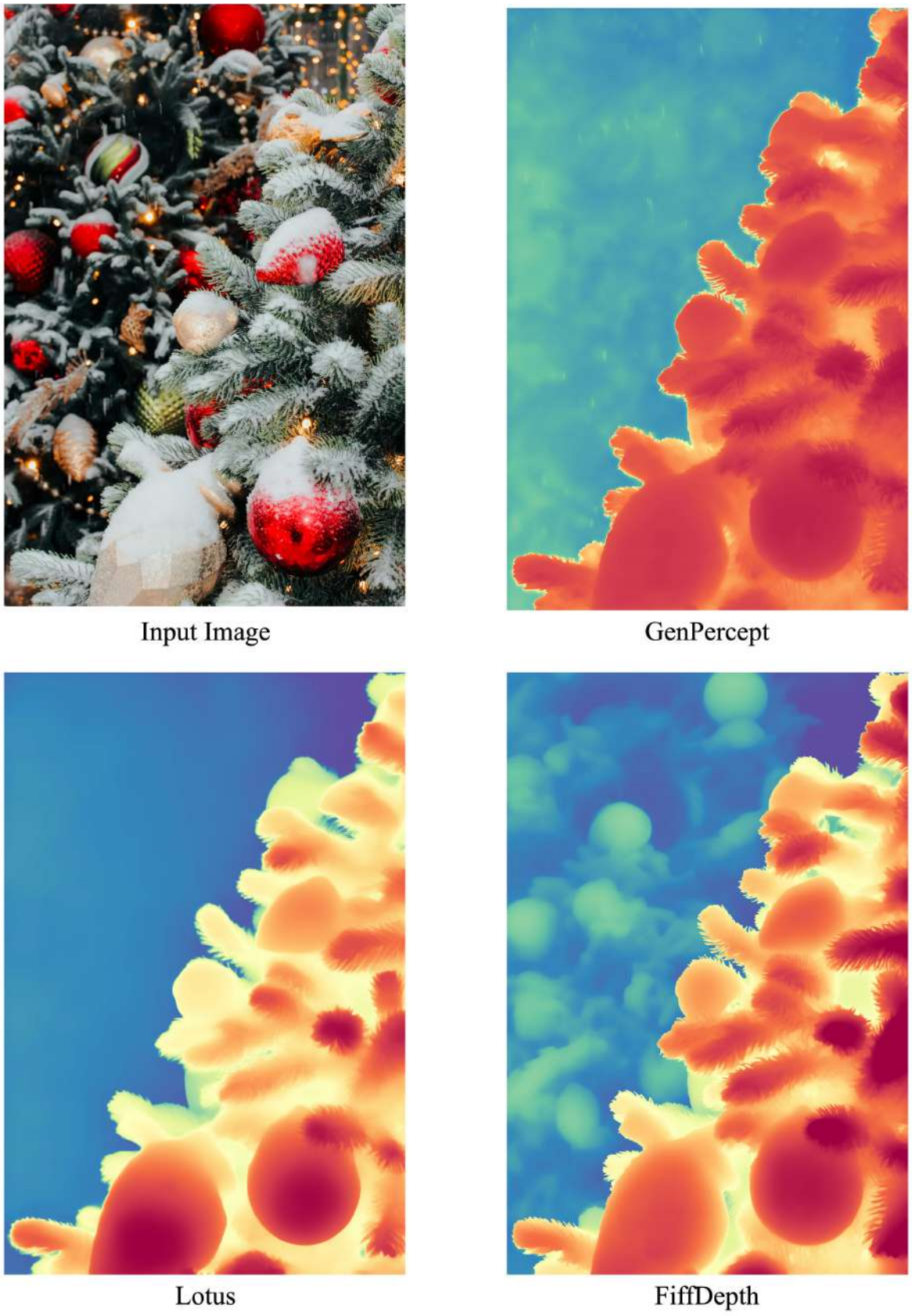


Figure 11. **Qualitative comparison with GenPercept and Lotus.**

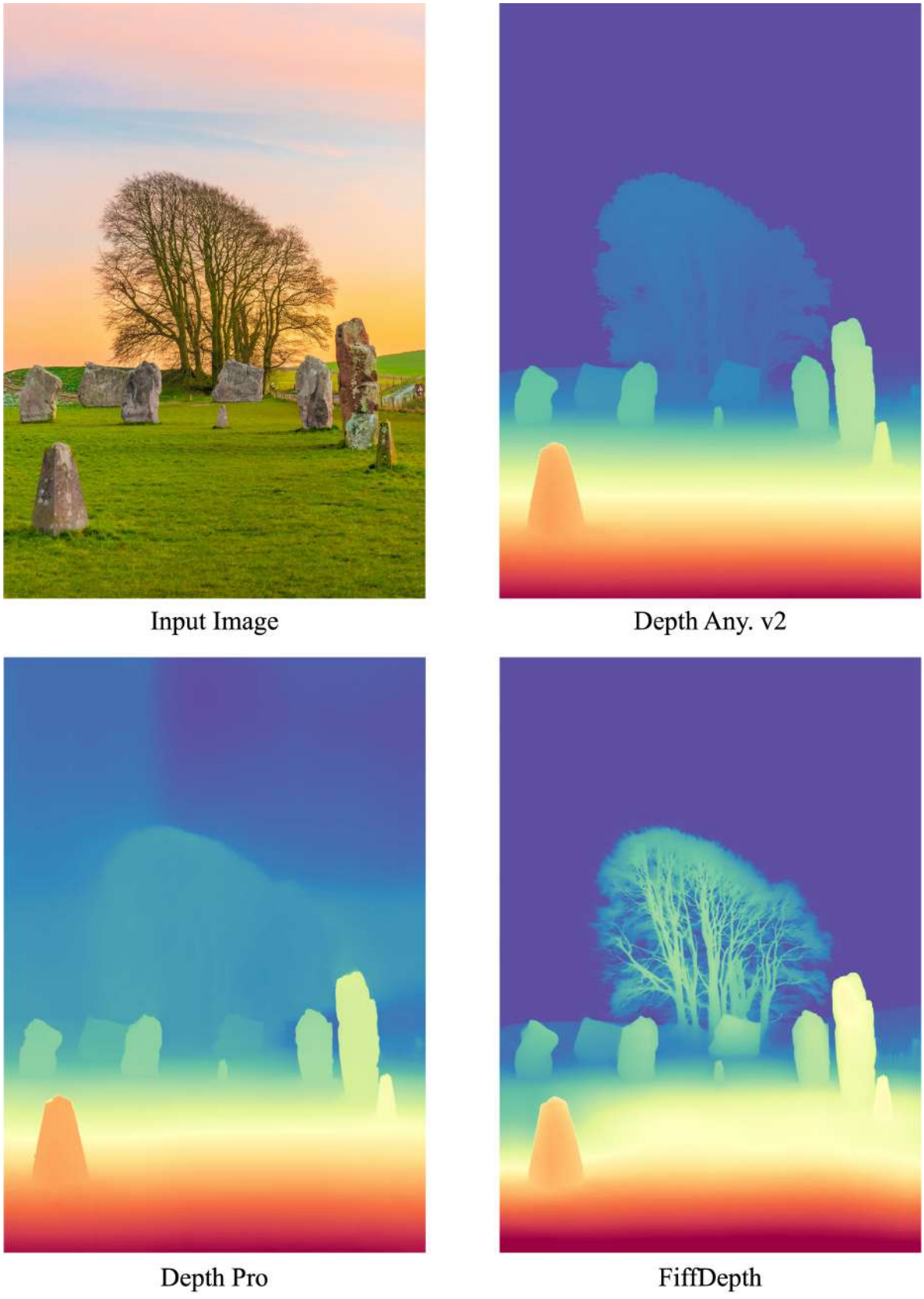


Figure 12. **Qualitative comparison with Depth Pro and Depth Anything v2.**

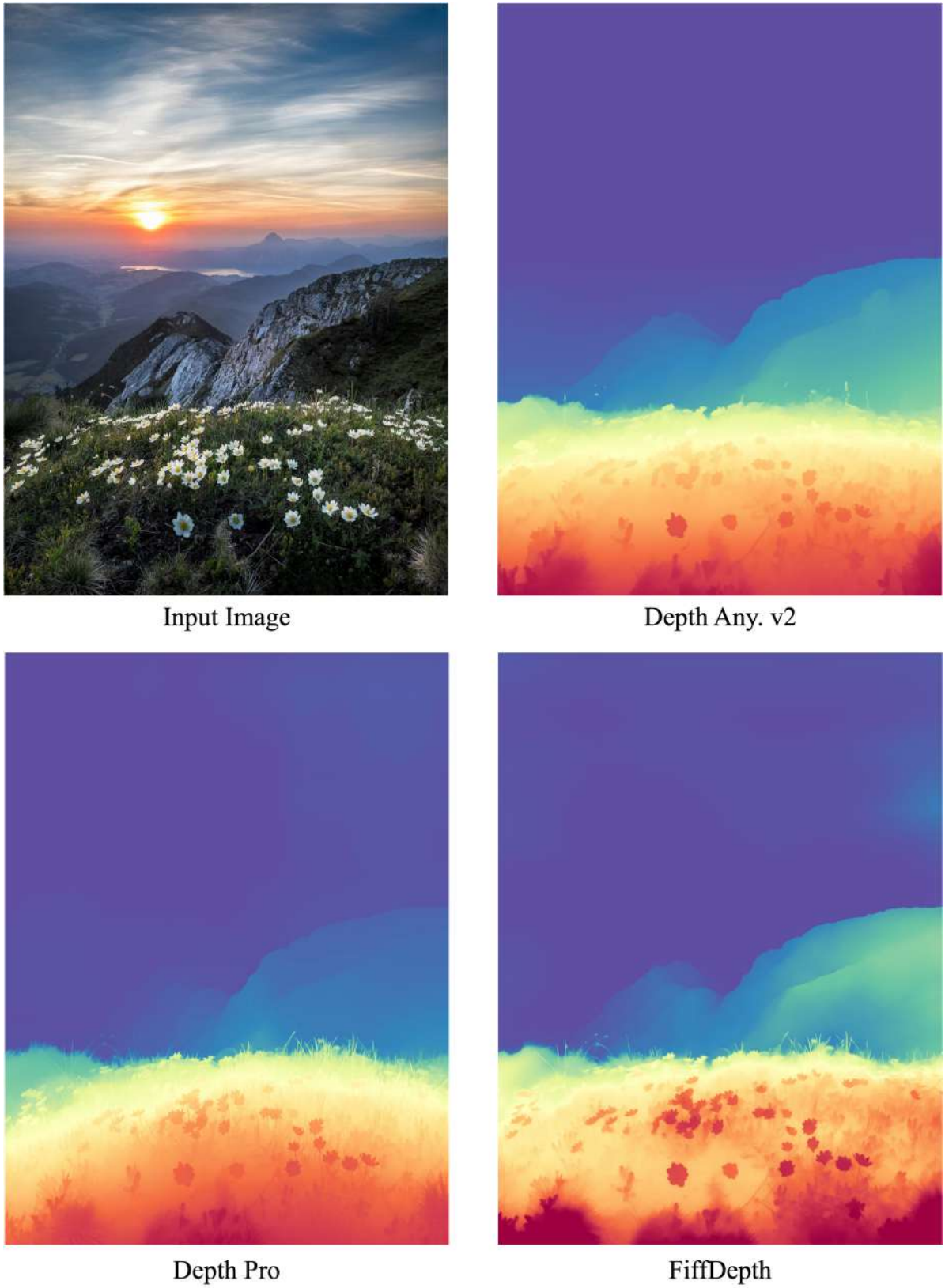


Figure 13. **Qualitative comparison with Depth Pro and Depth Anything v2.**

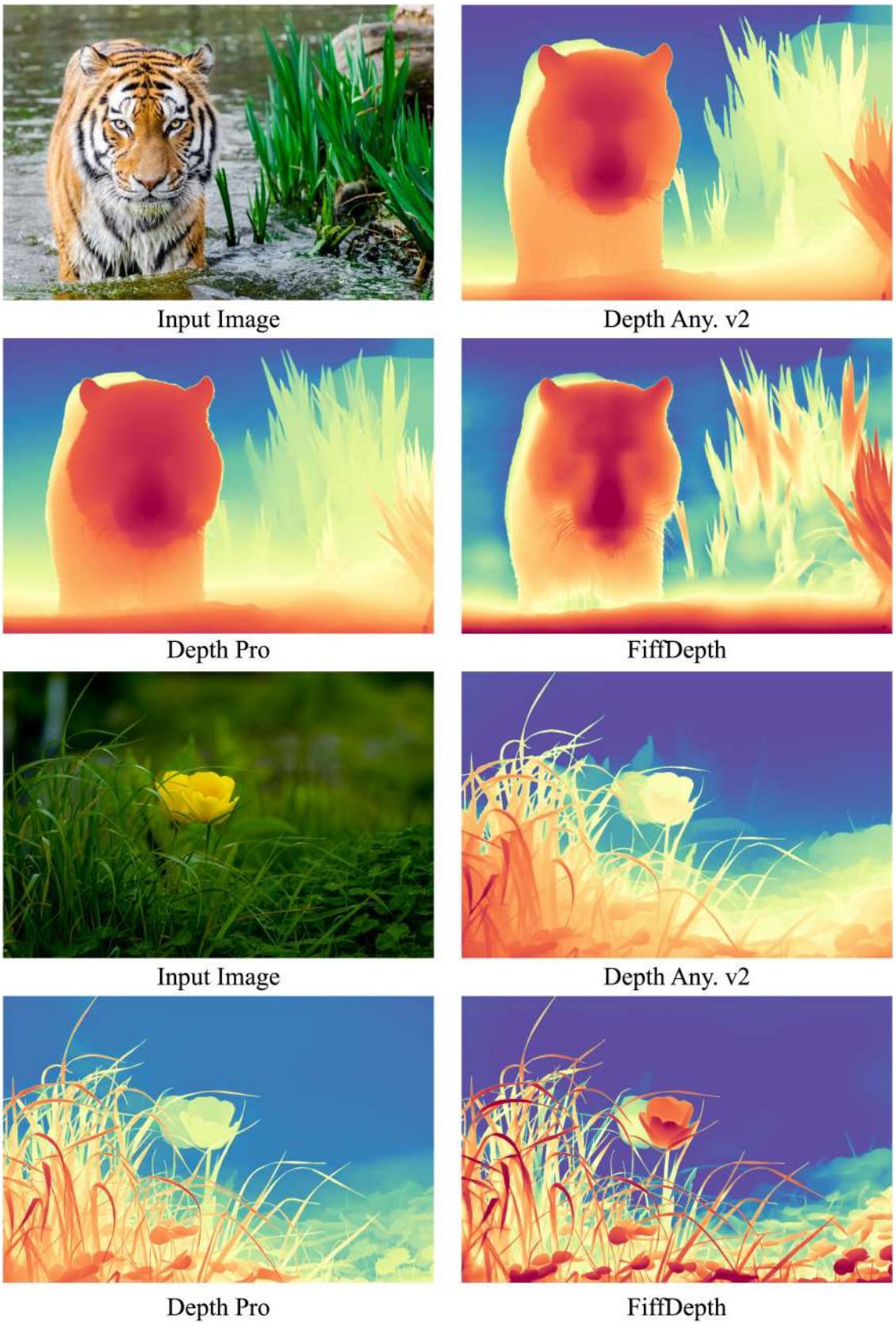


Figure 14. **Qualitative comparison with Depth Pro and Depth Anything v2.**

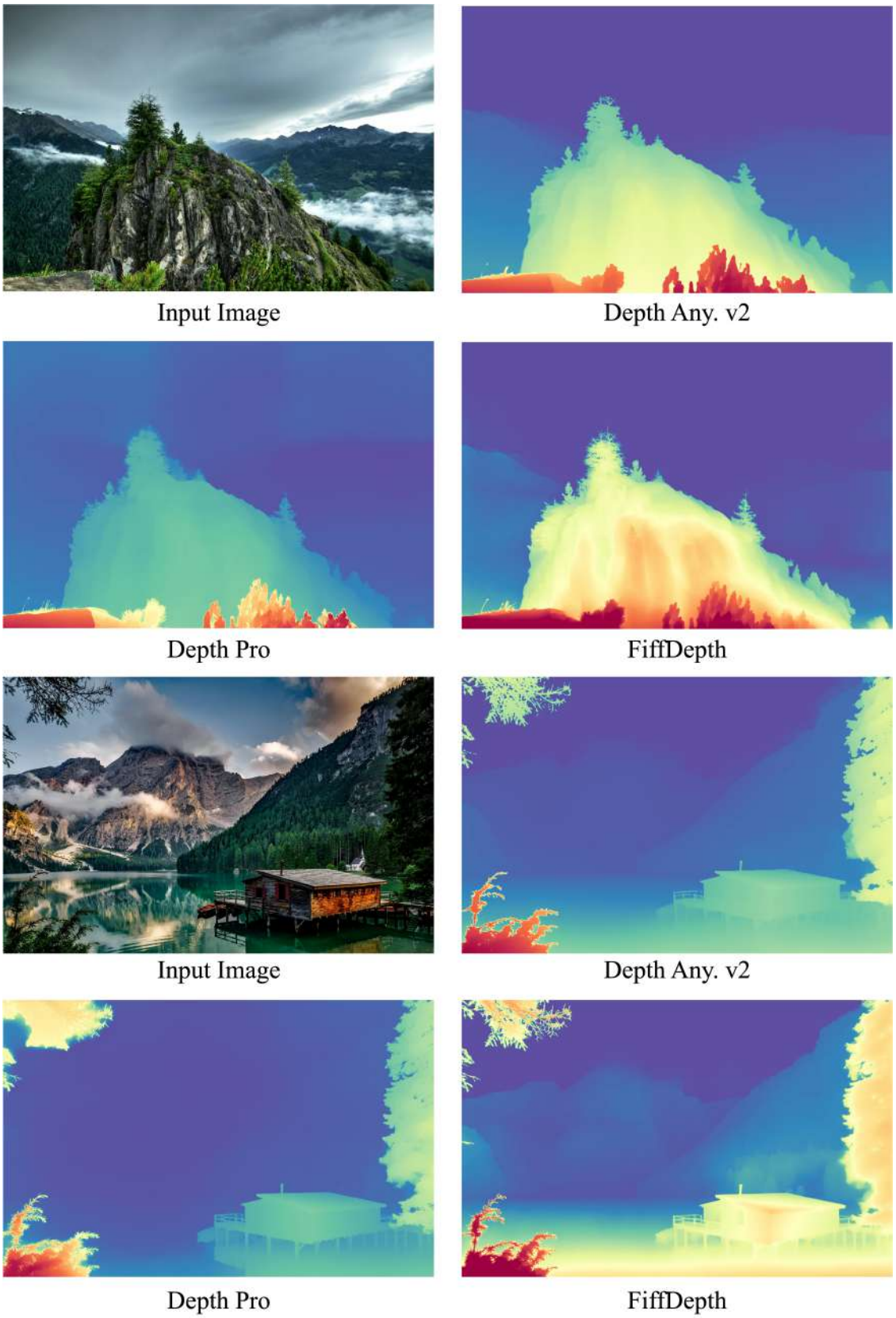


Figure 15. **Qualitative comparison with Depth Pro and Depth Anything v2.**

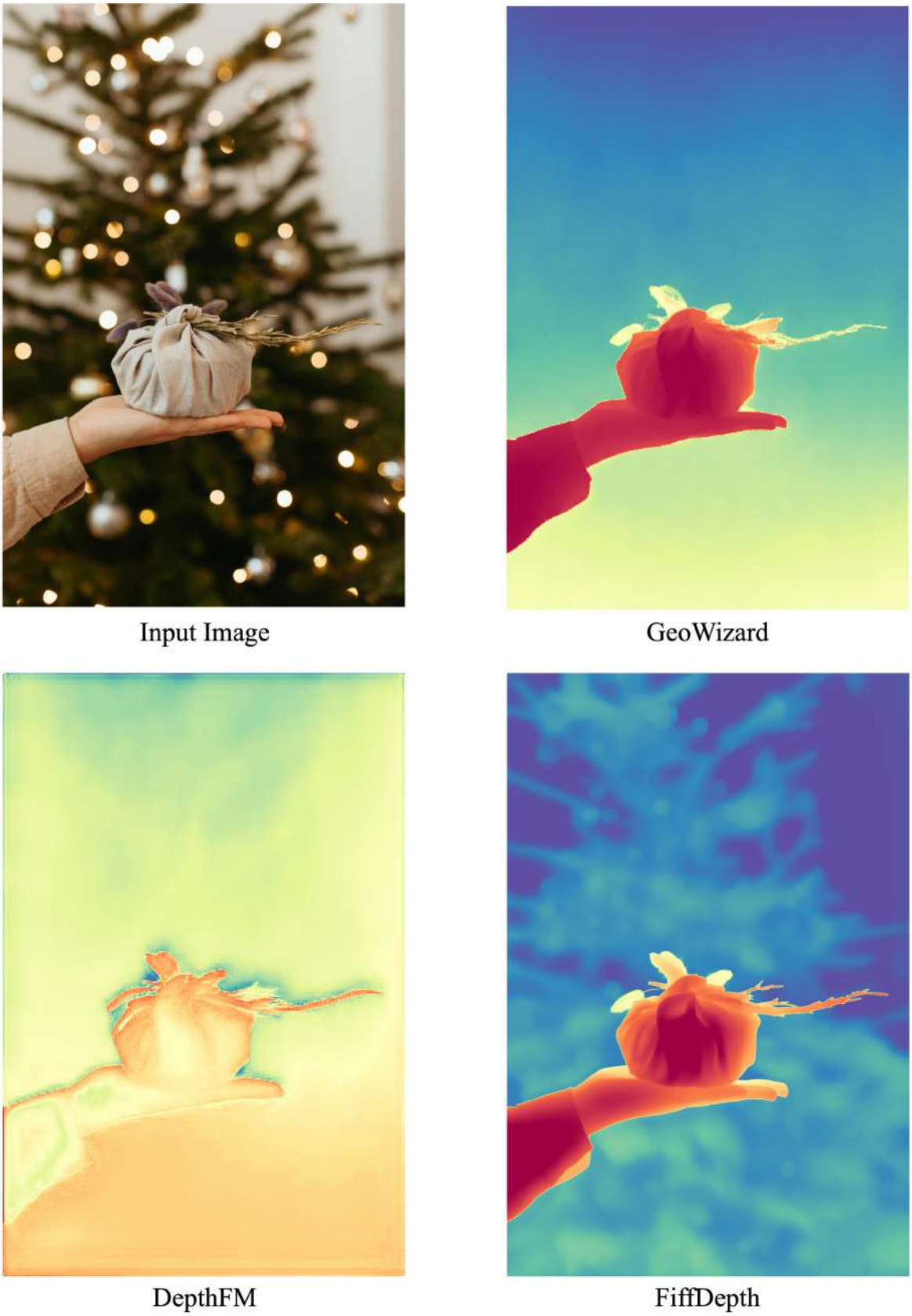


Figure 16. **Qualitative comparison with GeoWizard and DepthFM.**

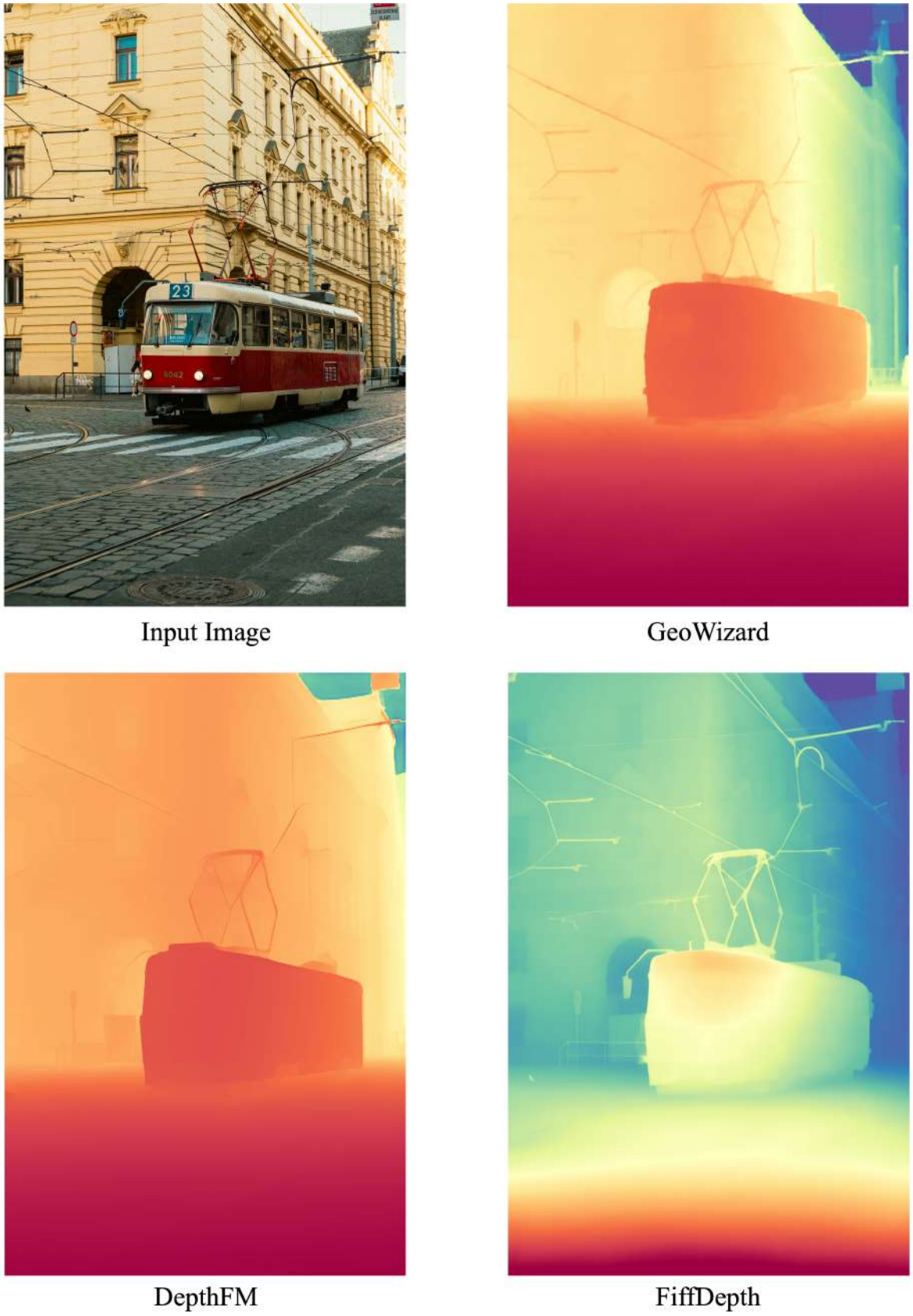


Figure 17. **Qualitative comparison with GeoWizard and DepthFM.**

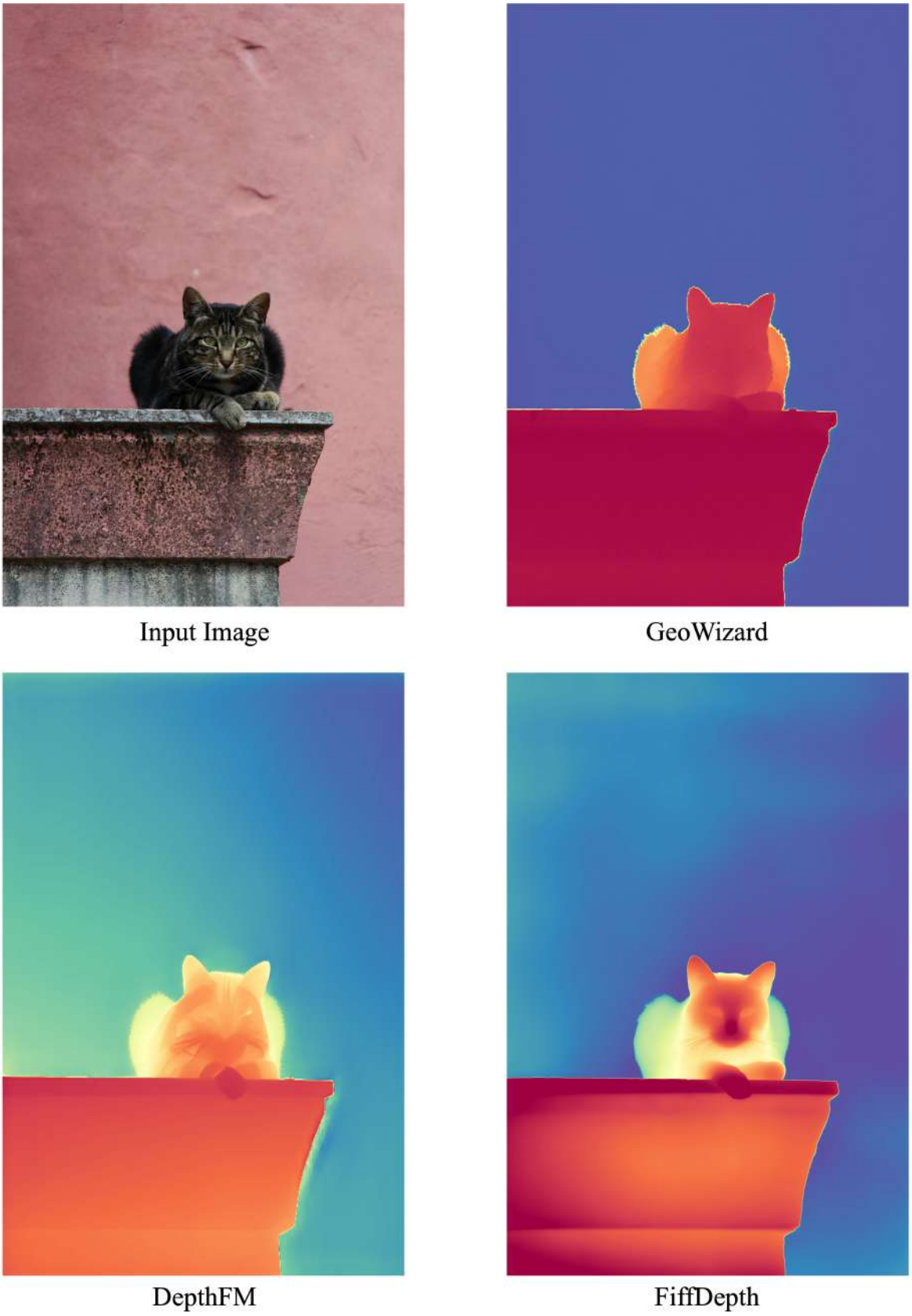


Figure 18. **Qualitative comparison with GeoWizard and DepthFM.**

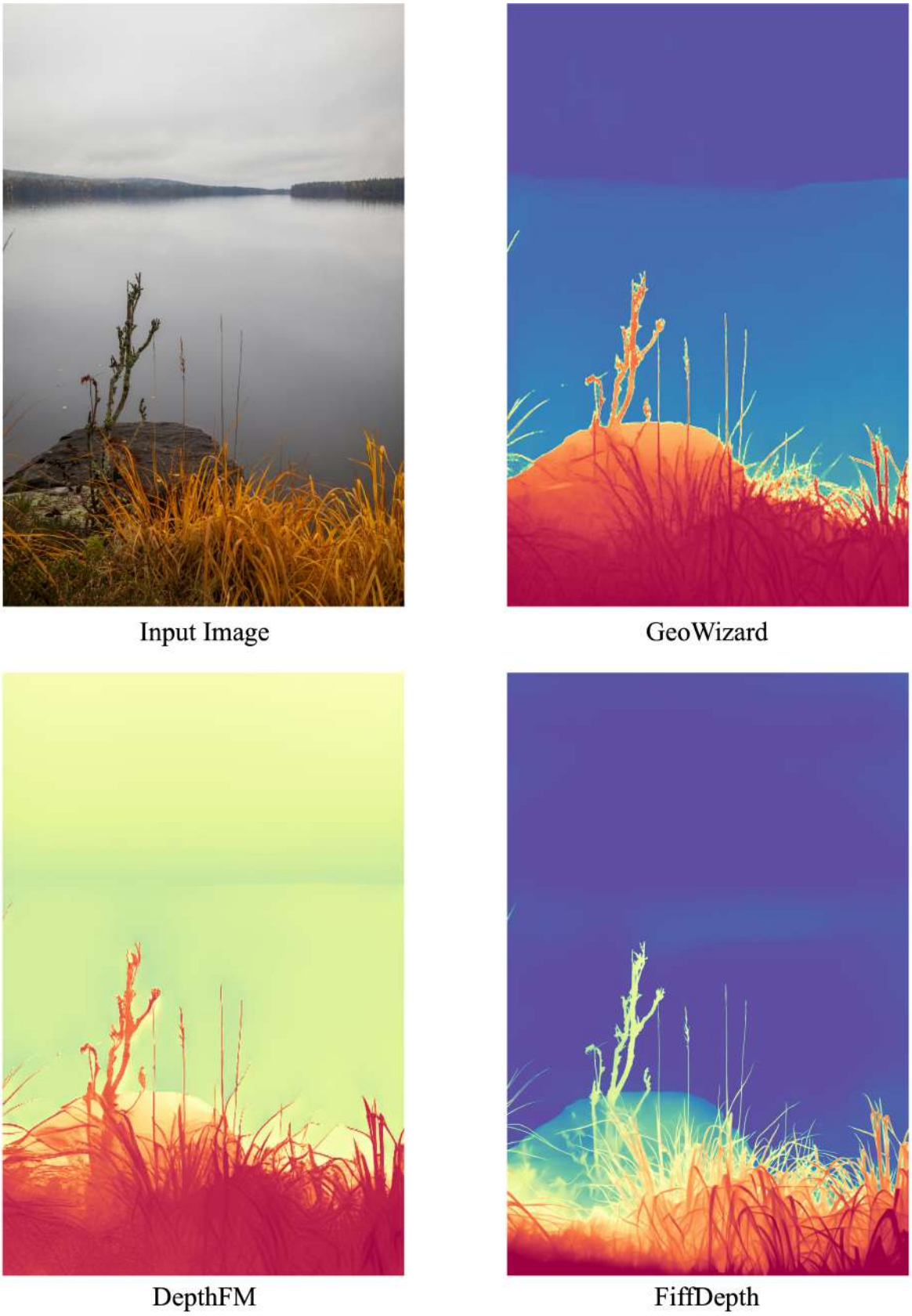
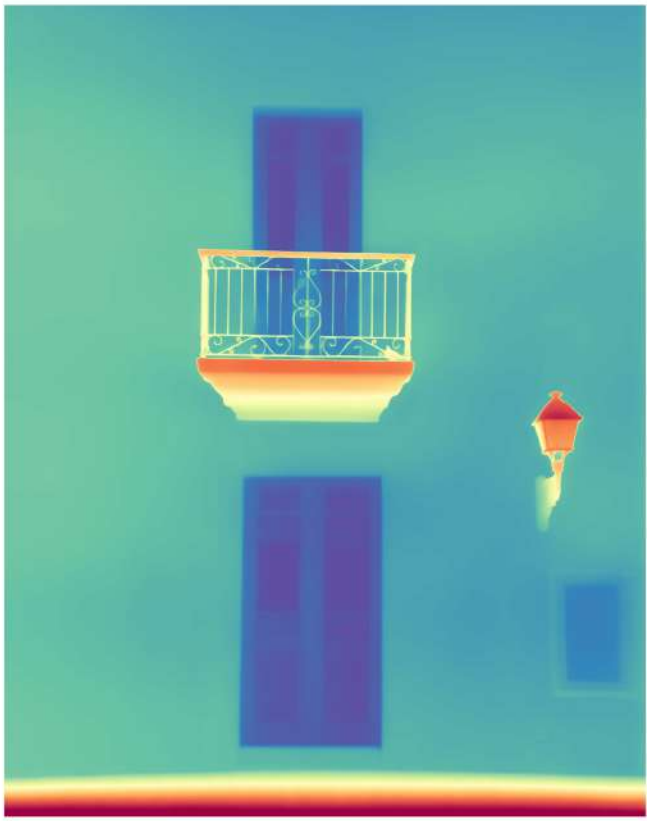


Figure 19. **Qualitative comparison with GeoWizard and DepthFM.**

Input Image

GeoWizard

DepthFM



FiffDepth

Figure 20. **Qualitative comparison with GeoWizard and DepthFM.**

Figure 21. **Depth to 3D Point Clouds.**

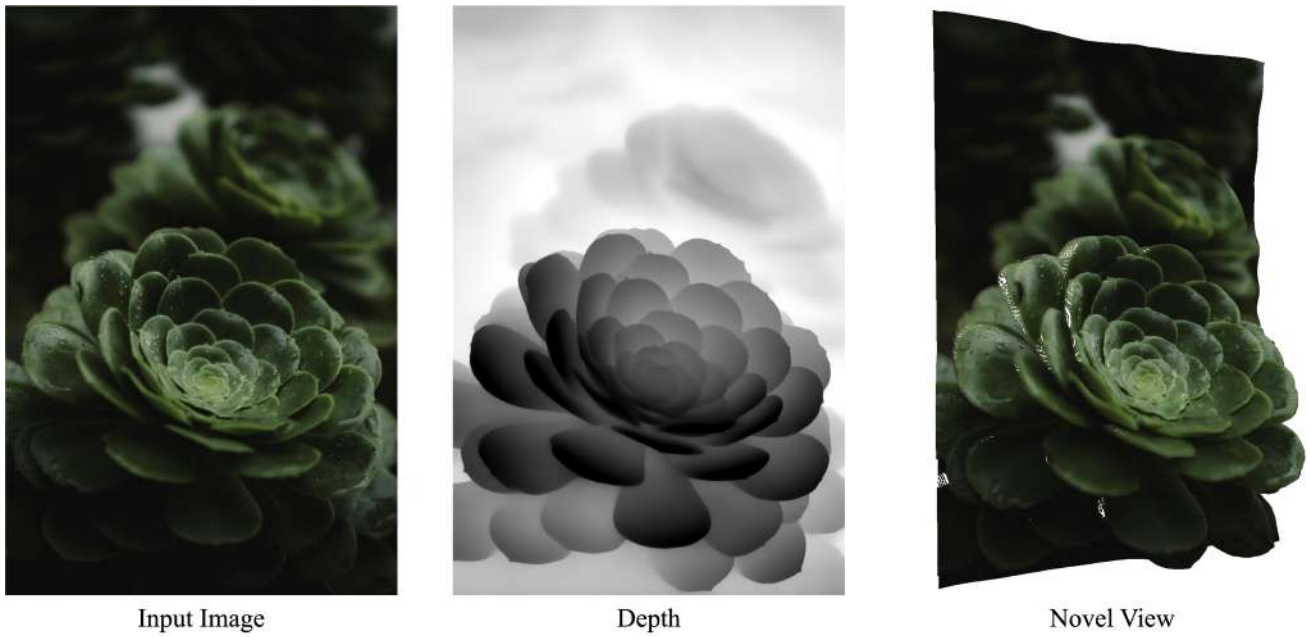


Figure 22. **Depth to 3D Point Clouds.**

Figure 23. **Depth to 3D Point Clouds.**

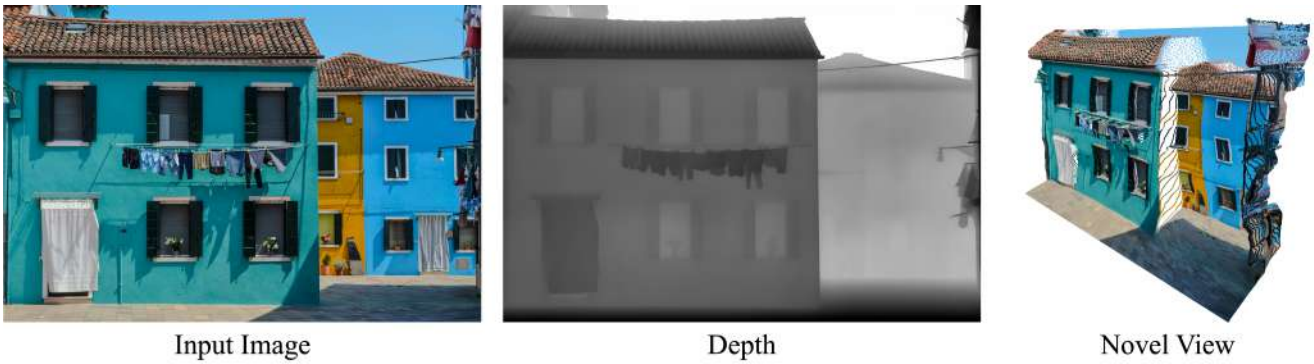


Figure 24. **Depth to 3D Point Clouds.**

Figure 25. **Depth to 3D Point Clouds.**

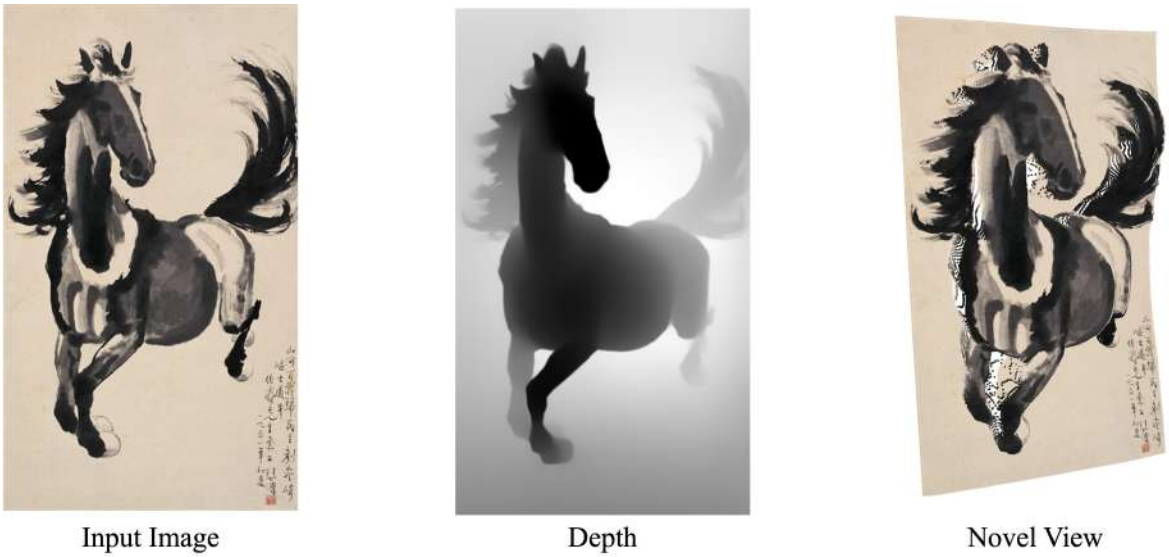


Figure 26. **Depth to 3D Point Clouds.**

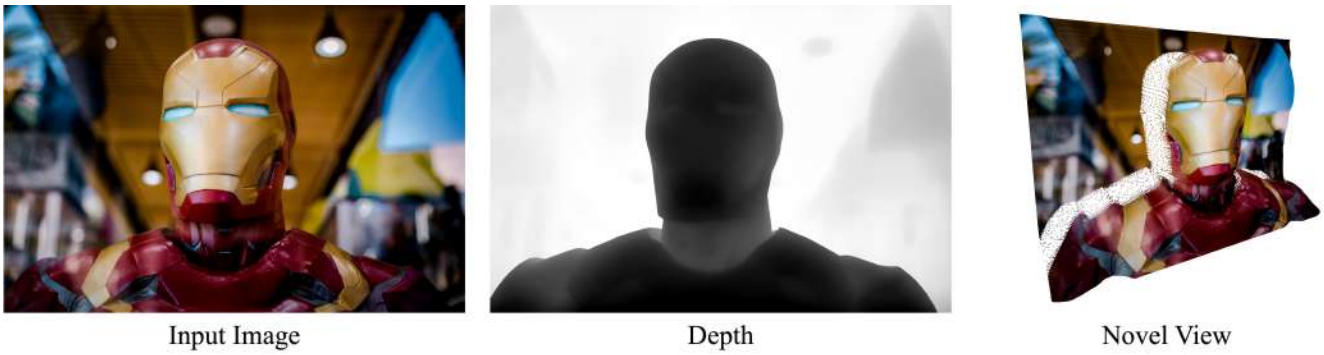


Figure 27. **Depth to 3D Point Clouds.**