Fine-Tuning Image-Conditional Diffusion Models is Easier than You Think

Supplementary Material

A. DDIM Inference

During training, the highest noise level corresponds to the last timestep $t = T$, and $t = 1$ corresponds to a very small noise level. The DDIM inference scheduler iterates over a series of k timesteps $\tau_1 > \tau_2 > \ldots > \tau_k > 0$ and iteratively denoises the initial noise input $\mathbf{z}_{\tau_1}$. We consider the `leading` and `trailing` schedules that are also discussed by Lin *et al*. [12] and show the selected timesteps for different k in Tab. A-1. The original `leading` timestep selection strategy of the DDIM scheduler excludes the final timestep T. This leads to a mismatch between training and inference; using the `leading` schedule, the model receives noise as input, even though the timestep embedding indicates a partially denoised input. In contrast, the fixed `trailing` strategy always starts with $t = T$ for the first denoising step, properly aligning training and inference. In the limit of $k \rightarrow T$ inference steps, both strategies converge to the same behavior.

In Fig. A-1, we illustrate the difference between single-step predictions using the broken `leading` and the fixed `trailing` DDIM scheduler for Marigold [10] and Stable Diffusion [14]. Both models output noise when using the broken scheduler. With the fixed implementation, both models predict the mean of their respective conditional distribution. For single-step Marigold this results in a well-defined depth map, whereas for single-step Stable Diffusion, it produces a blurry image with coarse structures that roughly align with the input prompt.

Fig. A-3 further demonstrates the scheduler's impact when multiple steps are considered. It clearly shows that the effect of the broken scheduler becomes less noticeable as the number of inference steps increases. Additionally, the weak text conditioning in Stable Diffusion leads to blurry images, which gradually sharpen as more inference steps are taken. In contrast, the strong image conditioning in Marigold allows the model to predict reasonably accurate depth maps already in the first step. As shown by the heatmap in Fig. 2b in the main text, subsequent steps only lead to small changes in the predicted distances, and most of the scene remains unchanged.

B. Detailed Experimental Setup

Training Datasets. For a direct comparison with Marigold [10], we use the same synthetic training datasets offering high quality ground-truth annotations, *i.e.*, Hypersim [13] and Virtual KITTI 2 [3].

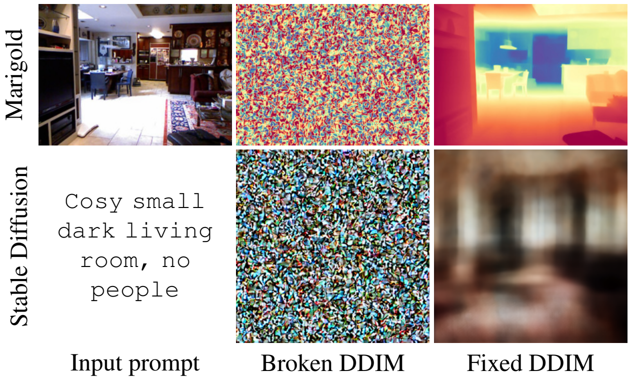

Figure A-1. **Single-step outputs of Marigold and Stable Diffusion.** With a single step, Stable Diffusion produces a blurry image at best, while Marigold outputs a sensible depth map. Note that the input prompt is text for Stable Diffusion, but an RGB image for Marigold.

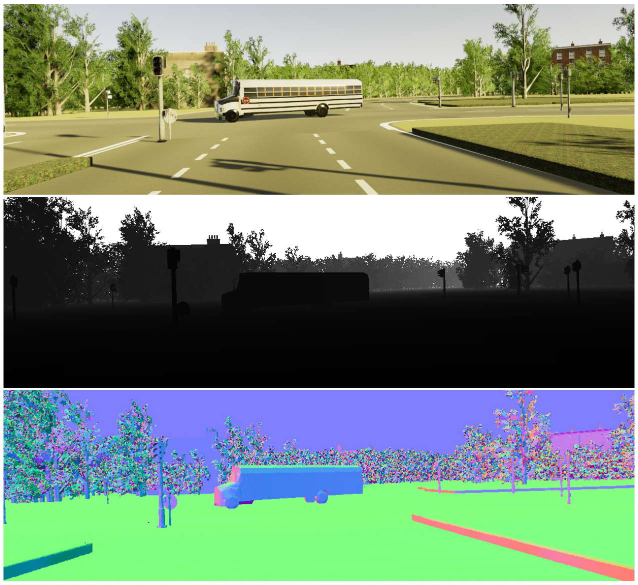

Figure A-2. **Virtual KITTI 2 example.** Top: Synthetic RGB image. Middle: Ground-truth depth map. Bottom: Ground-truth surface normals, generated using discontinuity-aware gradient filters [6].

Hypersim consists of 54K photorealistic images from 365 indoor scenes, which we resize to a resolution of 480 × 640 with a far plane at 65 meters. Virtual KITTI 2 contains approximately 20K samples from four synthetic driving scenarios under various weather conditions. These

Table A-1. **Comparison of `leading` vs. `trailing` timestep selection.** The timesteps selected by two DDIM scheduler timestep selection strategies for $T = 1000$ timesteps and varying numbers of inference steps.

Inference Steps	`leading` timestep selection	`trailing` timestep selection
1	[1]	[1000]
2	[501, 1]	[1000, 500]
4	[751, 501, 251, 1]	[1000, 750, 500, 250]
10	[901, 801, 701, 601, 501, 401, 301, 201, 101, 1]	[1000, 900, 800, 700, 600, 500, 400, 300, 200, 100]

images are cropped to 352 × 1216 pixels, and the far plane is set to 80 meters.

Since Virtual KITTI 2 does not provide annotations for surface normals, we compute them ourselves with the ground-truth depth maps, employing discontinuity-aware gradient filters from [6]. A qualitative example of the resulting normals can be seen in Fig. A-2.

Data Preprocessing. Following Marigold's approach for depth estimation, we remove outliers, *i.e.*, values below the 2nd percentile and above the 98th percentile, and normalize the depth map to the range $[-1, 1]$. Then, we repeat the normalized depth map 3 times along the color channel to match the VAE encoder's expected input shape. Normals, on the other hand, can be encoded directly since they are already in the desired range of $[-1, 1]$ and match the number of channels. The only data augmentation we utilize is random horizontal flipping.

Training Details. We mask out undefined depth values in the Hypersim dataset, and pixels surpassing the far plane for Virtual KITTI 2. When training Marigold for normal prediction as a diffusion estimator, the mask is downsampled by a factor of 8 to match the latent resolution. Thus, we neither enforce nor supervise undefined regions. For the end-to-end fine-tuning of GeoWizard, both the scale and shift invariant depth loss and the angular loss are optimized jointly. Scaling the depth loss by a factor of 0.5 roughly ensures equal magnitude.

Evaluation Datasets. For monocular depth estimation, we follow the evaluation strategy of Marigold and evaluate on commonly used benchmarks. NYUv2 [16] and ScanNet [4] provide RGB-D data of indoor environments captured with Kinect cameras. We use the official NYUv2 test split, consisting of 654 instances, while for ScanNet, Marigold's set of 800 randomly sampled images from the 312 validation scenes [10] is employed. ETH3D [15] and DIODE [17] offer high-resolution depth data for both indoor and outdoor scenes, derived from LiDAR sensors. We evaluate on all 454 samples in ETH3D and on DIODE's validation set, comprising 325 indoor and 446 outdoor examples. For KITTI [8], consisting of outdoor driving scenes captured by vehicle-mounted cameras and LiDAR sensors, the Eigen test split [5] is used, containing 652 images.

Table A-2. **Frozen vs. fine-tuned VAE decoder.** We conduct end-to-end fine-tuning of Marigold [10] for depth estimation, and assess the effect of freezing or fine-tuning the weights of the pre-trained VAE decoder.

Decoder	NYUv2 [16]		KITTI [8]		ETH3D [15]		ScanNet [4]		DIODE [17]	
	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
Frozen	**5.2**	**96.6**	**9.6**	**91.9**	**6.2**	95.9	**5.8**	**96.2**	**30.2**	**77.9**
Fine-tuned	5.3	96.5	**9.6**	**91.9**	**6.2**	**96.0**	**5.8**	96.1	**30.2**	77.7

Regarding surface normal estimation we utilize the official DSINE [1] evaluation pipeline and data, comprised of the NYUv2 test split, 300 ScanNet [16] samples, the full iBims-1 [11] dataset, which is a small high-quality RGB-D dataset of 100 samples, and Sintel [2], made up of 1064 synthetic outdoor examples derived from an open-source 3D animated short film.

Evaluation Details. For most existing methods in Tab. 5 and Tab. 6 we obtain the performance metrics either from the papers introducing these methods or from the Marigold and DSINE papers. The missing scores, like those of the newer GeoWizard [7] and DepthFM [9] models, are obtained by reevaluating the respective models with their official inference code and released checkpoints. In the case of DepthFM, the prediction alignment with respect to the ground-truth metric depth happens in the log metric space.

C. Additional Results

GeoWizard for Depth Estimation. GeoWizard [7] jointly predicts depth and surface normals, using a similar training and evaluation setup as Marigold. We find that GeoWizard suffers from the same flaw in the DDIM implementation as Marigold, and end-to-end fine-tuning the model for depth and normal estimation significantly boosts the performance (see Tab. A-3 and Tab. 3 in the main text). In particular, the fine-tuned model performs better than both the fixed single-step model and the previously best reported results with 50 steps and ensembling of 10 predictions.

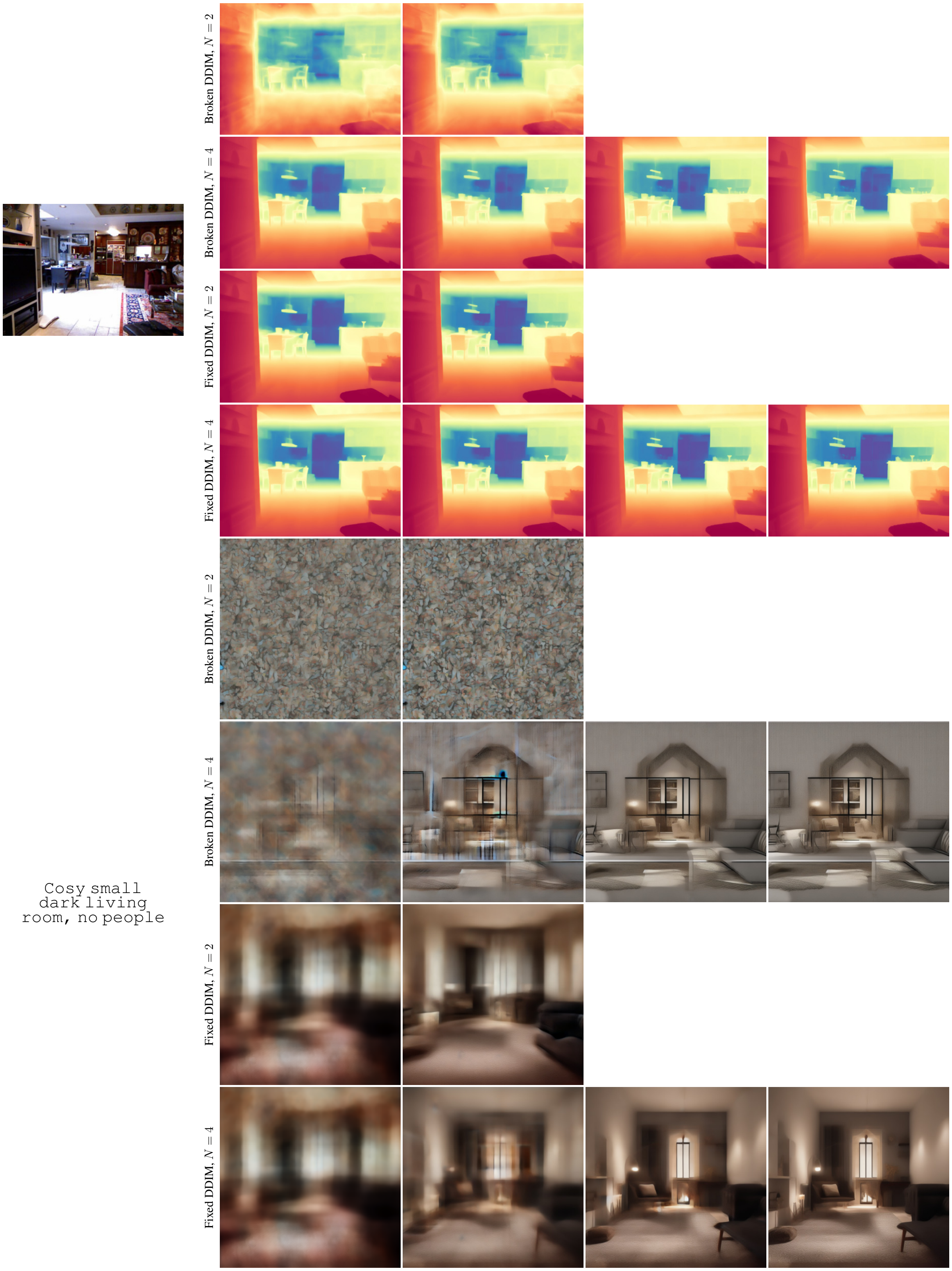

Figure A-3. **Few-step inference of Marigold and Stable Diffusion.** With more steps, the adverse effects of the broken DDIM scheduler get less pronounced. Both Marigold and Stable Diffusion produce sharper outputs with more steps, but the difference is much greater for Stable Diffusion.

Table A-3. **Fixed DDIM scheduler and end-to-end fine-tuning (E2E FT) for GeoWizard's [7] depth estimation.** We use the official code and model weights to re-evaluate the method on all datasets. Inference time is for a single 576×768-pixel image, evaluated on an NVIDIA RTX 4090 GPU. We obtain significant speed-ups, improving results.

Method	Steps	Ensemble	Inference time	NYUv2 [16]		KITTI [8]		ETH3D [15]		ScanNet [4]		DIODE [17]	
				AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
GeoWizard [7]	50	10	72 s	5.2	96.6	9.7	92.1	6.4	96.1	6.1	95.3	29.7	79.2
↳ reproduced by us	50	10	72 s	5.7	**96.2**	14.4	82.0	7.5	94.3	6.1	95.8	31.4	77.1
GeoWizard + DDIM fix	1	1	**254** ms	5.8	96.1	13.3	84.7	7.8	94.3	6.2	95.7	32.0	76.0
GeoWizard + E2E FT	1	1	**254** ms	**5.6**	96.1	**9.8**	**91.4**	**6.3**	**95.7**	**5.9**	**96.2**	**30.6**	**77.9**

Table A-4. **Comparison of DepthFM [9] with the DDIM-fixed and end-to-end fine-tuned (E2E FT) Marigold and Stable Diffusion models.** We re-evaluated DepthFM [9] on all datasets using the official code and model weights, with 4 inference steps and an ensemble size of 6. Inference time is for a single 576×768-pixel image, evaluated on an NVIDIA RTX 4090 GPU.

Method	Steps	Ensemble	Inference time	NYUv2 [16]		KITTI [8]		ETH3D [15]		ScanNet [4]		DIODE [17]	
				AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑	AbsRel↓	$\delta1$↑
DepthFM [9]	4	6	1.67 s	6.5	95.6	8.3	93.4	—	—	—	—	22.5	80.0
↳ reproduced by us	4	6	1.67 s	6.9	95.4	11.4	88.1	6.5	**96.2**	8.1	92.5	**25.0**	78.3
DepthFM	1	1	132 ms	7.5	95.0	11.6	87.5	6.7	96.0	8.3	92.3	25.3	**77.9**
Marigold [10] + E2E FT	1	1	**121** ms	**5.2**	**96.6**	**9.6**	91.9	**6.2**	95.9	**5.8**	96.2	30.2	**77.9**
Stable Diffusion [14] + E2E FT	1	1	**121** ms	5.4	96.5	**9.6**	**92.1**	6.4	95.9	**5.8**	**96.5**	30.3	77.6

Further Comparisons to DepthFM. DepthFM [9] proposes a direct mapping from input images to depth maps through flow matching, leveraging Stable Diffusion v2 [14] as a prior. We observe that, apart from the ETH3D $\delta1$ and DIODE [17] metrics, a simpler approach like E2E FT achieves better performance with a more than 10× speedup as seen in Tab. A-4.

Fine-Tuning the VAE Decoder. By default, we keep the pretrained VAE decoder frozen while conducting end-to-end fine-tuning. Tab. A-2 shows that fine-tuning the weights of this decoder does not improve performance.

Further Qualitative Samples. Fig. A-4 and Fig. A-5 show qualitative results for depth and normals estimation, respectively, comparing Marigold [10] and the end-to-end fine-tuned models. The fixed single-step model fails to produce sharp results, while the multi-step model exhibits noticeable over-sharpening and high-frequency noise artifacts (even after ensembling), particularly in the normals estimations. In contrast, the end-to-end fine-tuned models do not exhibit these issues.

Addendum

We were made aware of recent work by Xu *et al.* [18]. Similar to us, they directly fine-tune Stable Diffusion in an end-to-end fashion, however, we arrive to this point in a very different way. We initially discovered the issue with the DDIM scheduler, fixed this in Marigold, and in turn arrived to an end-to-end fine-tuning scheme that works for Marigold. Surprisingly, our ablations showed that this also works well for direct fine-tuning of Stable Diffusion. The main contribution of Xu *et al.* is an approach to fine-tune Stable Diffusion (for a broader spectrum of tasks). However, even with additional modules on top, their method achieves lower scores than some of the baselines. As such, these results might lead one to conclude that end-to-end fine-tuning is not a suitable alternative to multi-step, diffusion-based depth and normal estimation. In contrast, our simple end-to-end fine-tuning setup *does* outperform diffusion baselines, demonstrating that it is an effective and efficient alternative.

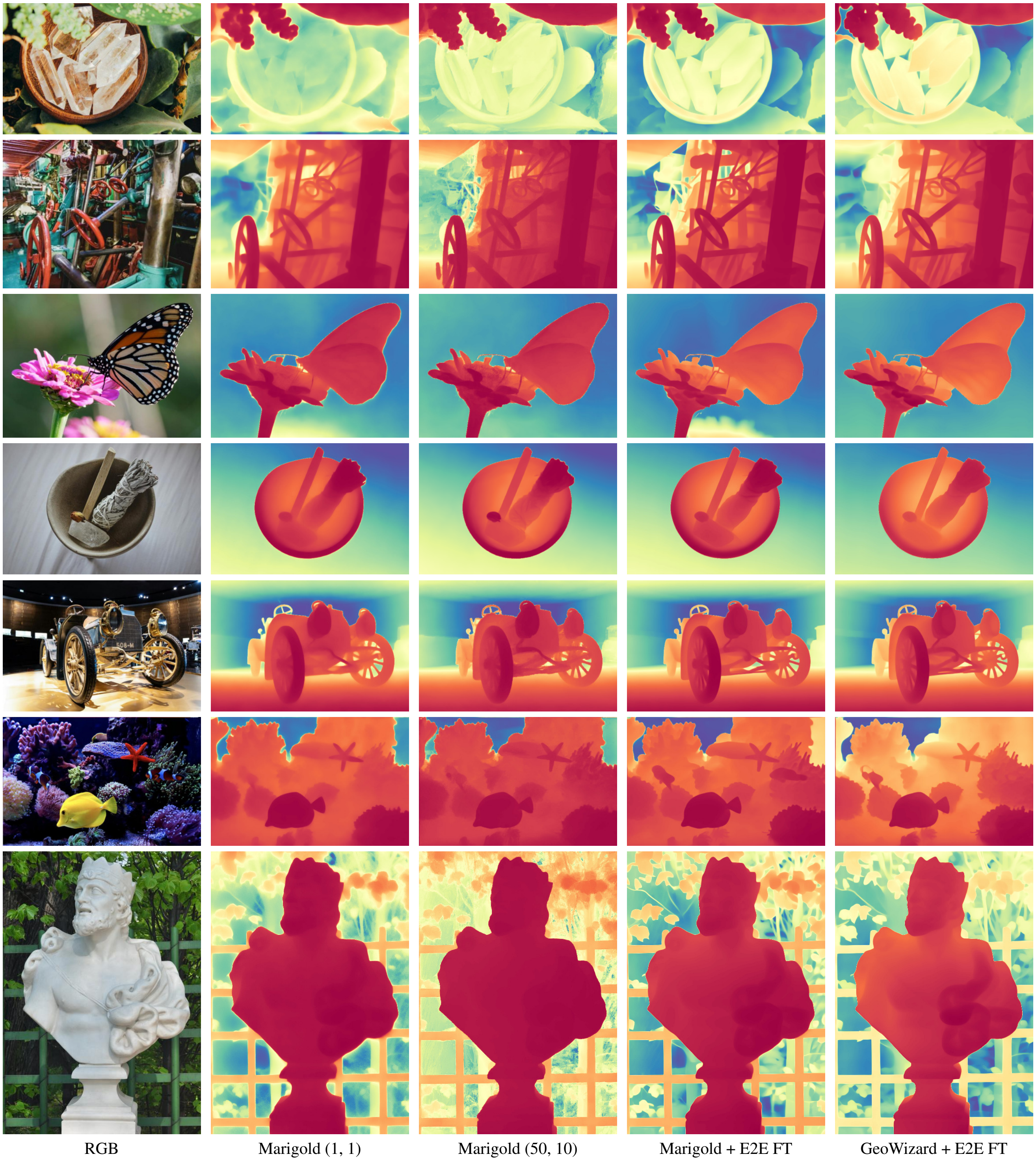

Figure A-4. **Additional qualitative samples for depth estimation.** "Marigold (X, Y)" denotes Marigold using X inference steps with an ensemble of size Y.

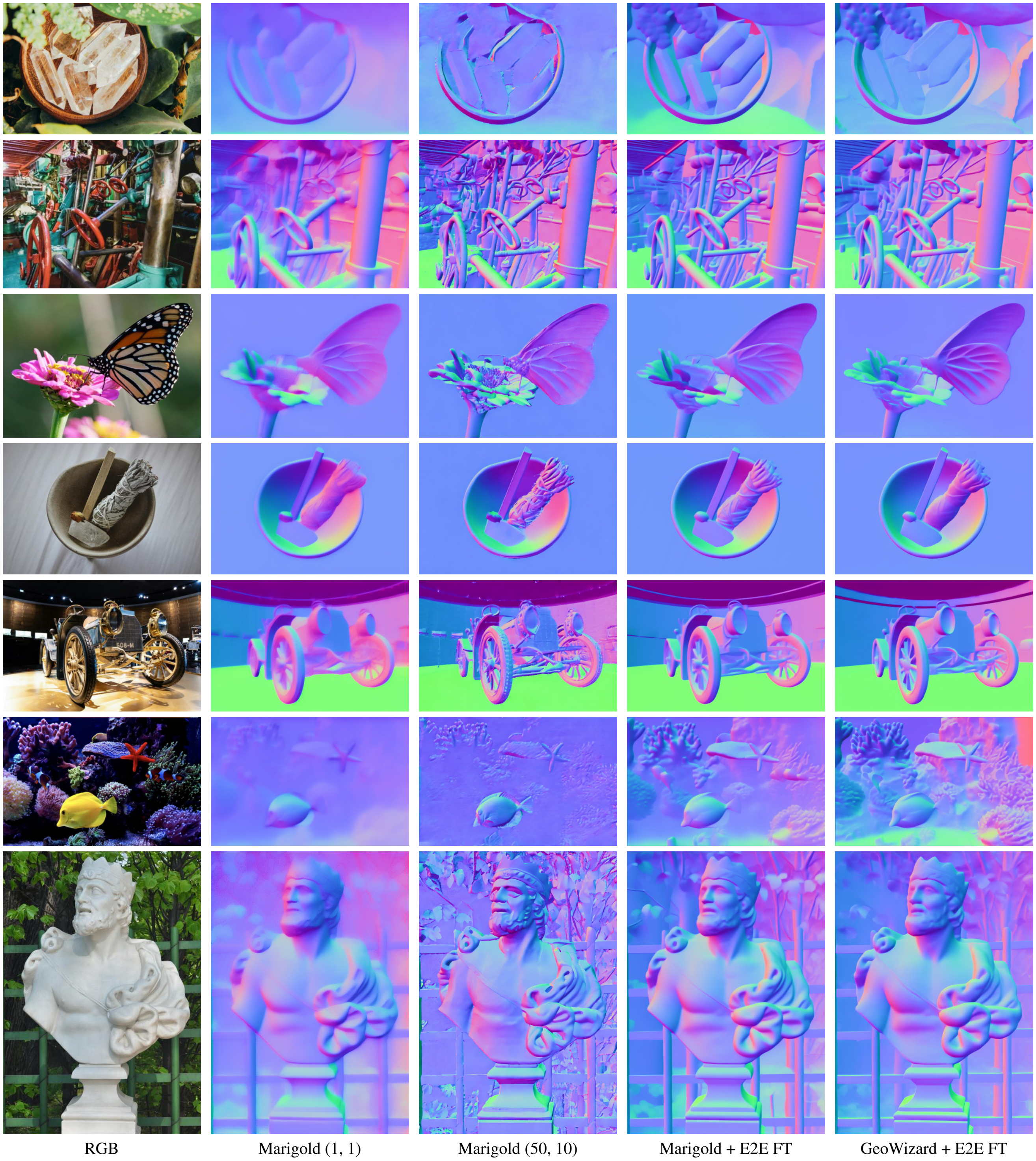

Figure A-5. **Additional qualitative samples for normal estimation.** “Marigold (X, Y)” denotes Marigold using X inference steps with an ensemble of size Y.

References

[1] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *CVPR*, 2024. 2

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 2

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

[4] 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. 2, 4

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

[6] Yi Feng, Bohuan Xue, Ming Liu, Qijun Chen, and Rui Fan. D2NT: A high-performing depth-to-normal translator. In *ICRA*, 2023. 1, 2

[7] 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. *ECCV*, 2024. 2, 4

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4

[9] Ming Gui, Johannes S. Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 2, 4

[10] 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*, 2024. 1, 2, 4

[11] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *ECCVW*, 2018. 2

[12] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *WACV*, 2023. 1

[13] 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

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 4

[15] 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*, 2017. 2, 4

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 4

[17] 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. 2, 4

[18] 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:2403.06090v2*, 2024. 4