SharpDepth: Sharpening Metric Depth Predictions Using Diffusion Distillation

Supplementary Material

In this supplementary material, we provide additional datasets details in Sec. 1. We then provide additional results in Sec. 2. Finally, we demonstrate the effectiveness of our estimated depths in a downstream application implementing metric SLAM in Sec. 3.

Limitations: SharpDepth relies on the pre-trained metric depth estimator for both conditioning input and supervision, so its accuracy is tied to the metric estimator's performance. As our training pipeline distills details from a diffusion-based estimator, the sharpness of SharpDepth is limited by the diffusion model's quality. We anticipate that advancements in these areas will enhance our approach.

1. Dataset Details

As described in the main paper, we train SharpDepth using approximately 1% of the data from six real datasets and evaluate it on seven real datasets (for metric depth accuracy) and three synthetic datasets (for metric depth boundary detail). This approach guarantees a diverse training dataset capable of encompassing various camera configurations. Details of the training and test sets are provided in Tab. 1.

	Dataset	Images	Scene	Acquisition
Training Set	Argoverse2 [21]	15k	Outdoor	LiDAR
	Waymo [18]	12k	Outdoor	LiDAR
	PandaSet [23]	12k	Outdoor	LiDAR
	ARKit [1]	16k	Indoor	RGB-D
	ScanNet [4]	11k	Indoor	RGB-D
	Taskonomy [25]	30k	Indoor	RGB-D
Test Set	KITTI [5]	652	Outdoor	LiDAR
	NYU [16]	654	Indoor	RGB-D
	ETH3D []	454	Outdoor	RGB-D
	Diode [20]	325	Indoor	LiDAR
	Booster [15]	456	Indoor	RGB-D
	NuScenes [3]	1000	Outdoor	LiDAR
	IBims-1 [10]	100	Indoor	RGB-D
	Sintel [22]	1065	Synthetic	-
	UnrealStereo4K [19]	200	Synthetic	-
	Spring [13]	1016	Synthetic	-

Table 1. **Datasets.** List of the training and test datasets along with their number of images, scene type, and acquisition method.

2. Additional Results

In-the-wild image samples. We evaluate the robustness of our method on a diverse set of "in-the-wild" images. Qualitative results for Internet-sourced images are shown in Fig. 1, while results from handheld mobile device captures are presented in Fig. 2. Our approach consistently generates accurate metric depth maps, exhibiting improved depth discontinuities and overall structural coherence. Notably, it excels at capturing thin structures, comparable to affine-invariant diffusion depth models [7, 9], while preserving the precision of metric depth.

Method	PSNR↑	SSIM↑	LPIPS↓
MonoGS + UniDepth	18.472	0.718	0.305
MonoGS + Ours	**18.857**	**0.735**	**0.289**

Table 2. Performance of SharpDepth on the fr1/desk sequence of TUM RGB-D dataset [17].

Generalization to another metric depth estimator. To show our method's generalization capabilities, we evaluate SharpDepth on Metric3Dv2 [8], a recent versatile metric depth estimation model. In this experiment, we leverage our previously trained model on UniDepth, and no further retraining is performed. We directly apply our model trained on UniDepth to Metric3Dv2 depths during test time. We provide the qualitative results in Fig. 3. As can be seen, our method generalizes well to metric depths produced by Metric3Dv2.

More depth metrics on test datasets. We provide an extended version of zero-shot metric accuracy on 6 zero-shot datasets in Tab. 4. We report absolute mean relative error (A.Rel), root mean square error (RMSE), scale-invariant error in log scale ($\mathrm{SI}_{\log}$) and the percentage of inlier pixel (δ_1). As shown in Tab. 4, SharpDepth achieves competitive metric accuracy compared to UniDepth and other metric depth models. Moreover, our method consistently outperforms other metric refinement techniques, such as PatchRefiner. This highlights the effectiveness of our approach in enhancing high-frequency details in depth maps while maintaining robust zero-shot performance. We further report the Pseudo Depth Boundary Error (PDBE), including the accuracy $\epsilon^{\mathrm{acc}}_{\mathrm{PDBE}}$ and completion $\epsilon^{\mathrm{compl}}_{\mathrm{PDBE}}$, along with visual samples in Fig. 4 and Fig. 5. The results demonstrate that both the accuracy and completion rates effectively capture the boundary details of the depth map.

Training and Inference costs. We show the train and inference costs compared to DepthAnythingV2-Metric [24], PatchRefiner [11], and BetterDepth [27] in Tab. 3 (using an A100 40GB GPU on the KITTI dataset). Compared with DepthAnythingV2, our training cost is 427 times lower. Compared with other refiners, we have better inference time compared to PatchRefiner and have a similar training cost

Methods	DAv2	PatchRefiner	BetterDepth	SharpDepth
Train samples	62M	21K	74K	93K
GPU hours	30720	N/A	36	72
Inf. time (ms)	334.9	1642.9	N/A	551.1
FLOPs (G)	4056.5	12534.0	N/A	7989.5

Table 3. Training and inference cost analysis.

to BetterDepth, while not needing ground-truth depth for supervision.

More visual results. We present additional qualitative results on our test datasets in Fig. 6, Fig. 7, and Fig. 8. These figures show predictions from UniDepth, UniDepth-aligned Lotus, and our method. As observed, our approach generates depth maps with more detailed representations of fine structures.

3. Applications

In this section, we demonstrate that our predicted sharper depth maps can significantly benefit downstream 3D reconstruction tasks, such as Visual SLAM [6] and Volumetric TSDF Fusion [26]. By providing more detailed and accurate depth information, our method enhances the quality and reliability of these reconstruction pipelines.

3.1. Visual SLAM

Dense visual SLAM focuses on reconstructing detailed 3D maps, which are crucial for applications in AR and robotics. In this work, we demonstrate that high-frequency depth maps can significantly improve the performance of SLAM methods in reconstructing the scene. We conduct experiments using a Gaussian Splatting-based SLAM method, i.e., MonoGS [12], on the fr1/desk sequence of TUM RGBD dataset [17], using the depth maps from UniDepth and SharpDepth as inputs to the system. Quantitative results are provided in Tab. 2, where our method consistently outperforms UniDepth in terms of photometric errors, showcasing its potential to enhance SLAM performance. Additionally, we present qualitative results in Fig. 9. As shown, our method better captures the underlying geometry of the scene, leading to improved novel view renderings.

3.2. Volumetric TSDF Fusion

Existing 3D reconstruction pipelines rely on multiple pairs of RGB-D inputs that are multi-view consistent. To achieve high-quality point clouds, it is crucial to have accurate metric depth predictions with sharp details. In this section, we demonstrate that our predicted depth maps can be used with TSDF Fusion [26], to further enhance their reconstruction quality.

As can be seen in Fig. 10, SharpDepth can render less distorted point clouds compared to those produced by the UniDepth [14] approach.

Dataset	**Method**	A.Rel ↓	RMSE ↓	$\text{SI}_{\log}$ ↓	δ_1 ↑
KITTI	Marigold [9]	0.095	3.221	13.240	92.284
	Lotus [7]	0.113	3.538	18.383	87.703
	UniDepth [14]	0.051	2.236	7.078	97.921
	ZoeDepth [2]	0.057	2.390	7.470	96.500
	Metric3Dv2 [8]	0.053	2.481	7.449	97.589
	PatchRefiner [11]	0.158	6.043	13.061	79.245
	DAv2-Metric (Indoor) [24]	0.552	12.783	14.649	1.000
	DAv2-Metric (Outdoor) [24]	0.124	3.924	13.843	85.312
	UniDepth-aligned Lotus	0.130	3.935	16.077	83.633
	SharpDepth (Ours)	0.059	2.374	8.100	97.315
NYUv2	Marigold [9]	0.055	0.224	8.114	96.384
	Lotus [7]	0.054	0.222	7.993	96.612
	UniDepth [14]	0.055	0.200	5.367	98.417
	ZoeDepth [2]	0.077	0.278	7.190	95.200
	Metric3Dv2 [8]	0.066	0.254	7.498	97.391
	PatchRefiner [11]	2.482	5.900	19.089	1.000
	DAv2-Metric (Indoor) [24]	0.205	0.594	8.229	69.613
	DAv2-Metric (Outdoor) [24]	2.798	6.328	22.333	1.000
	UniDepth-aligned Lotus	0.087	0.281	8.916	93.921
	SharpDepth (Ours)	0.064	0.228	6.179	96.949
ETH3D	Marigold [9]	0.064	0.616	9.217	95.956
	Lotus [7]	0.062	0.581	9.266	96.001
	UniDepth [14]	0.456	3.008	7.728	25.308
	ZoeDepth [2]	0.567	3.272	13.015	34.210
	Metric3Dv2 [8]	0.138	0.903	6.081	82.420
	PatchRefiner [11]	1.781	8.830	11.715	4.974
	DAv2-Metric (Indoor) [24]	0.346	2.230	8.064	39.336
	DAv2-Metric (Outdoor) [24]	2.089	9.473	11.269	3.564
	UniDepth-aligned Lotus	0.493	3.267	13.092	20.347
	SharpDepth (Ours)	0.474	3.092	12.119	22.606
Diode	Marigold [9]	0.307	3.755	29.230	76.685
	Lotus [7]	0.330	3.877	30.999	73.751
	UniDepth [14]	0.265	4.216	23.370	66.031
	ZoeDepth [2]	0.484	6.637	29.374	30.195
	Metric3Dv2 [8]	0.158	2.552	19.455	88.765
	PatchRefiner [11]	1.264	7.064	29.563	25.031
	DAv2-Metric (Indoor) [24]	0.432	7.691	27.315	24.845
	DAv2-Metric (Outdoor) [24]	1.502	7.543	30.240	22.045
	UniDepth-aligned Lotus	0.357	5.321	30.671	55.876
	SharpDepth (Ours)	0.297	4.644	25.340	61.486
Booster	Marigold [9]	0.049	0.074	6.392	97.384
	Lotus [7]	0.041	0.063	5.333	98.779
	UniDepth [14]	0.492	0.532	7.686	28.041
	ZoeDepth [2]	0.642	0.674	10.563	20.855
	Metric3Dv2 [8]	0.668	0.720	5.795	15.490
	PatchRefiner [11]	5.551	5.994	18.136	1.000
	DAv2-Metric (Indoor)	0.311	0.352	7.327	57.284
	DAv2-Metric (Outdoor)	7.075	7.646	15.536	1.000
	UniDepth-aligned Lotus	0.494	0.519	6.382	26.429
	SharpDepth (Ours)	0.491	0.528	7.089	27.717
nuScenes	Marigold [9]	0.267	6.158	35.628	65.881
	Lotus [7]	0.363	7.263	49.047	50.911
	UniDepth [14]	0.144	4.771	21.959	83.861
	ZoeDepth [2]	0.587	8.155	33.076	21.838
	Metric3Dv2 [8]	0.199	7.371	28.267	84.215
	PatchRefiner [11]	0.582	10.589	30.193	31.726
	DAv2-Metric (Indoor) [24]	0.411	11.489	30.750	14.828
	DAv2-Metric (Outdoor) [24]	0.588	8.573	30.260	19.100
	UniDepth-aligned Lotus	0.432	7.850	49.524	41.243
	SharpDepth (Ours)	0.184	5.208	25.584	78.479

Table 4. **Detailed results on different datasets.** We ranked methods that do not require GT alignment as best, second-best, and third-best. Gray indicates the method that has been trained on the training set.

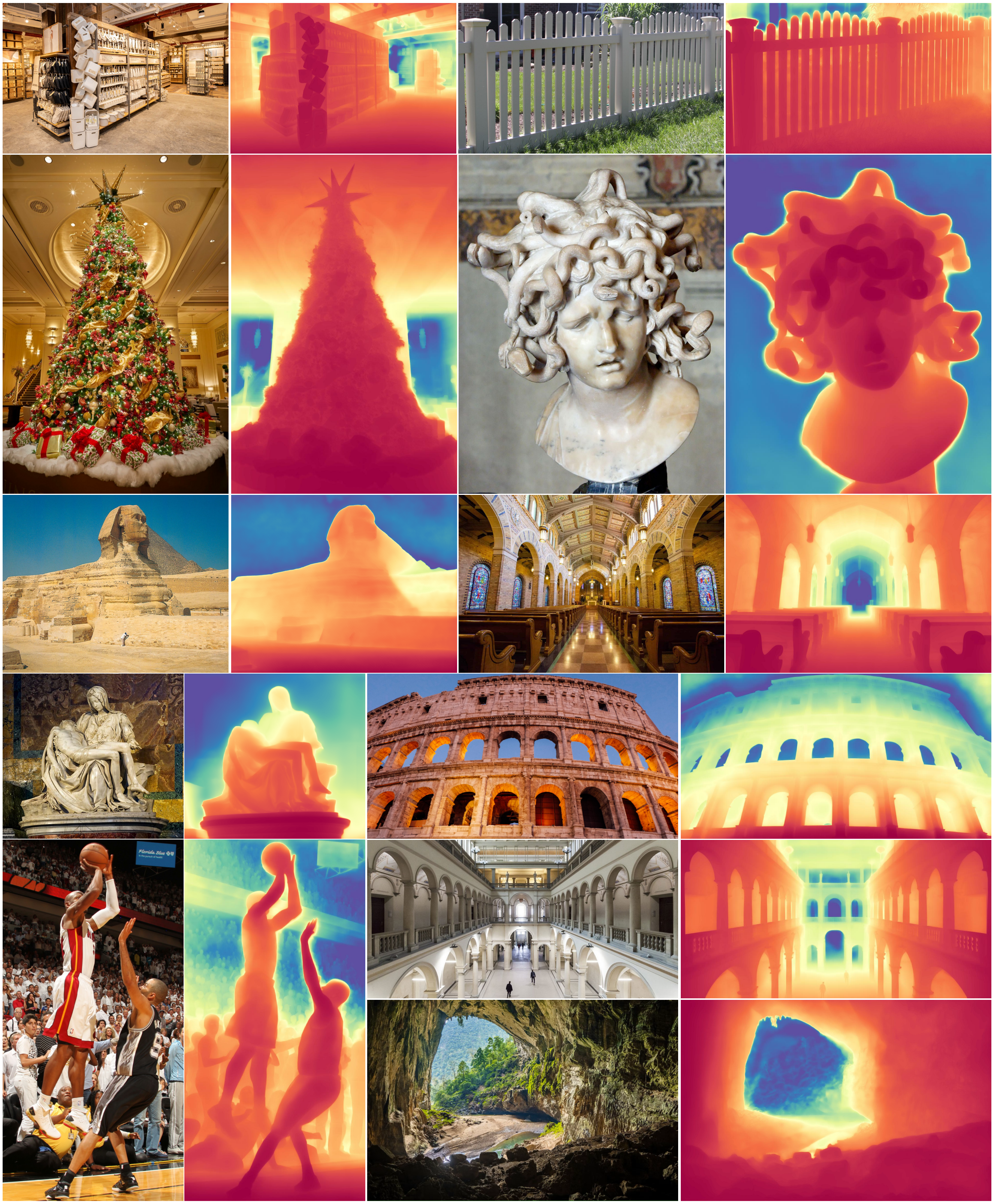


Figure 1. **In-the-wild depth estimation from Internet images**. Red indicates the close plane and blue means the far plane.

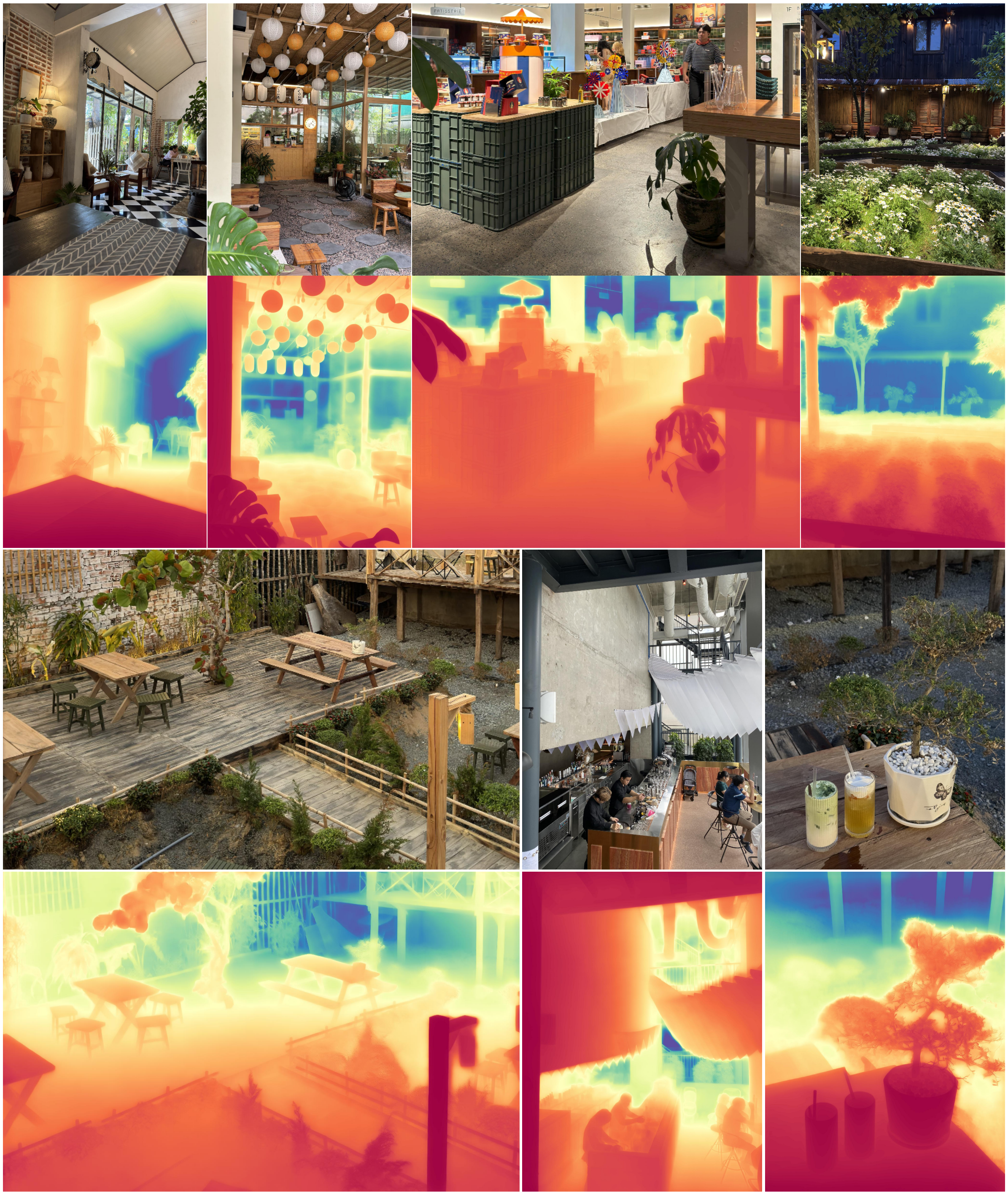
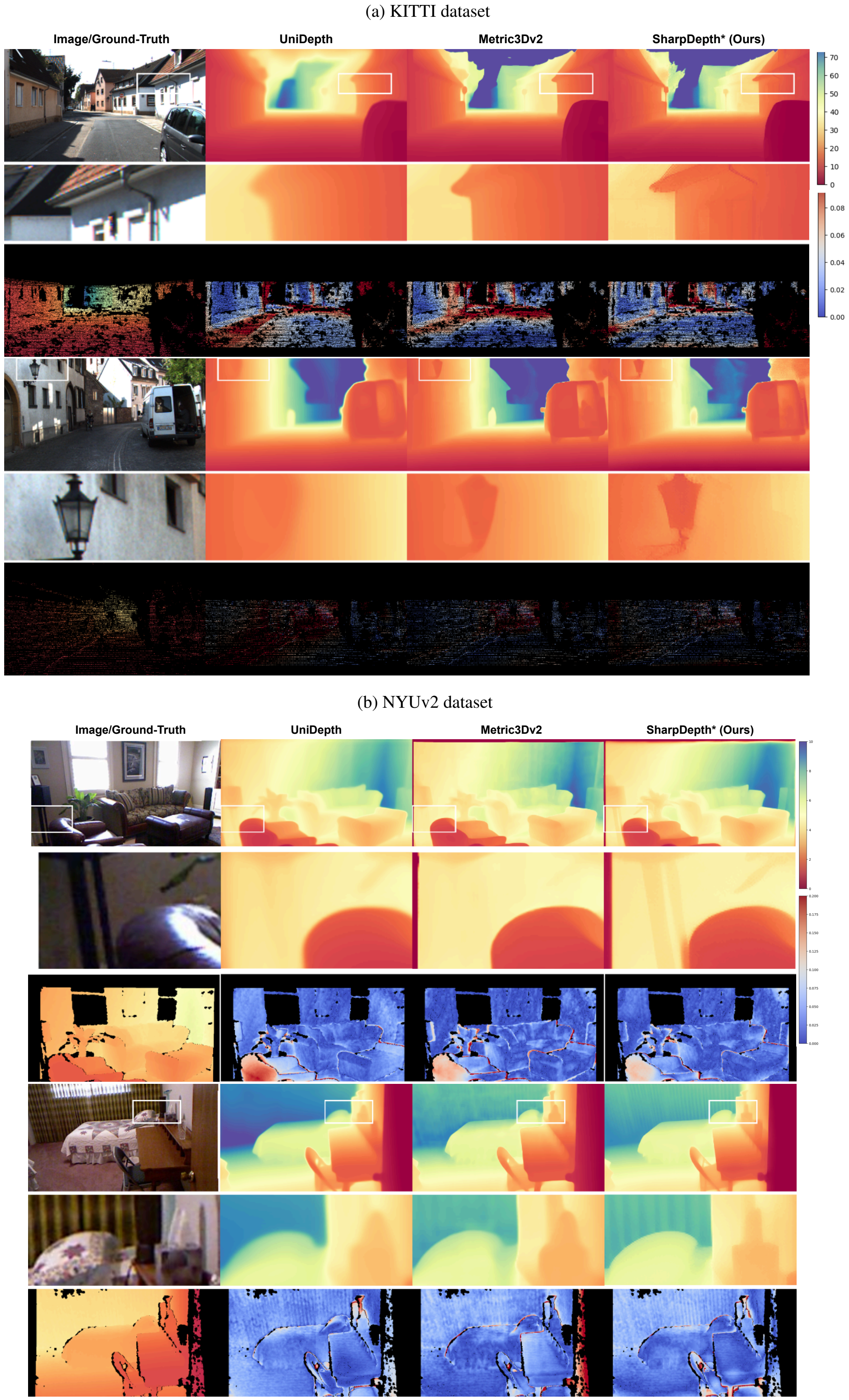


Figure 2. **In-the-wild depth estimation from images captured by a mobile phone**. Red indicates the close plane and blue means the far plane.


(a) KITTI dataset

(b) NYUv2 dataset

Figure 3. **Qualitative results on KITTI and NYUv2**. SharpDepth* denotes our method when using depth by Metric3Dv2 as input. The last column represents the colormap ranges for depth (*spectral*) and AbsRel error (*coolwarm*).

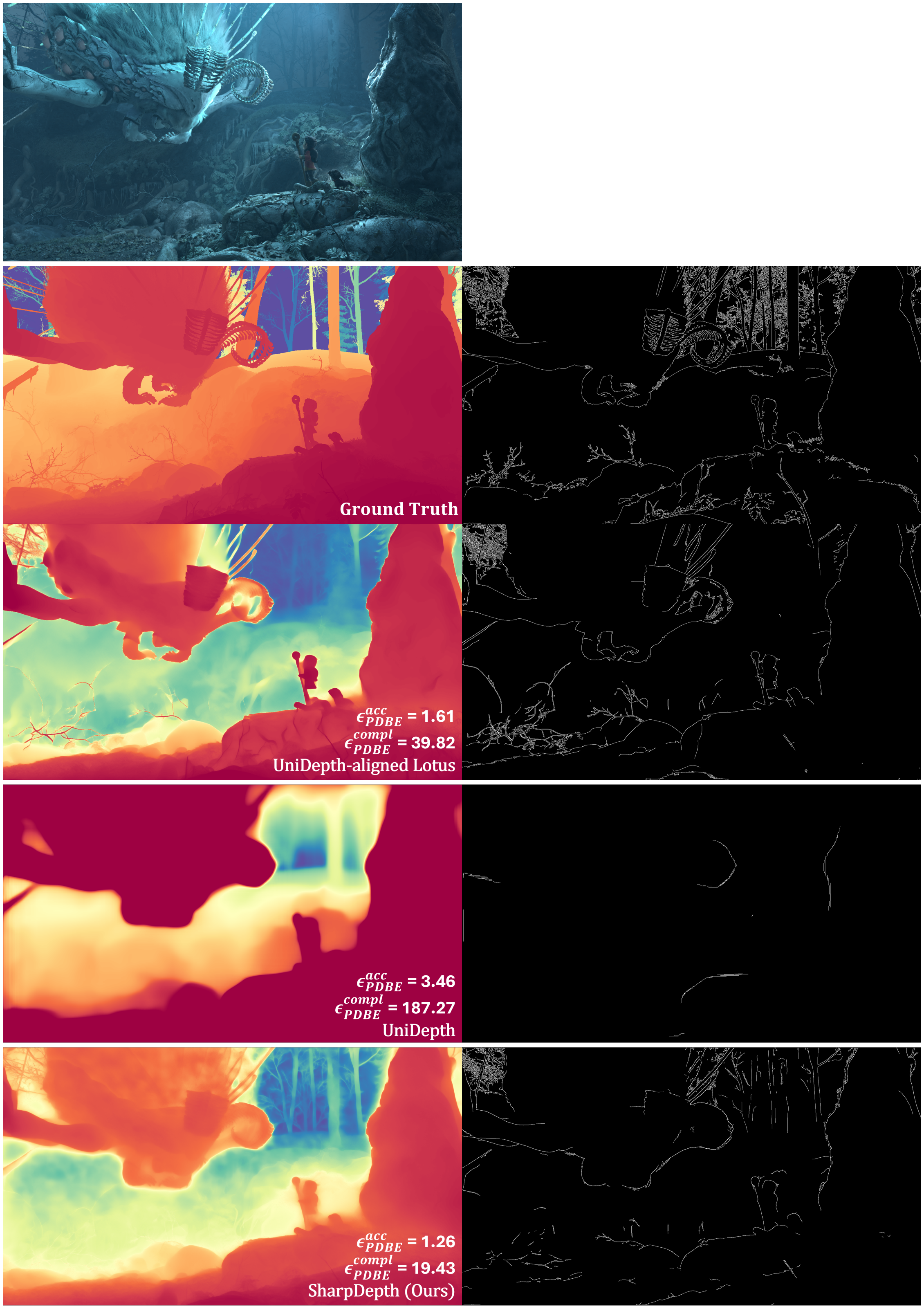


Figure 4. **Illustration of the depth boundary metrics on the Spring dataset**. We show the depth maps and extracted boundaries for each prediction. Compared to UniDepth, our method extracts more edges due to better depth discontinuities. Compared to Lotus, our method can capture more precise edges, due to the global prior from pre-trained UniDepth.

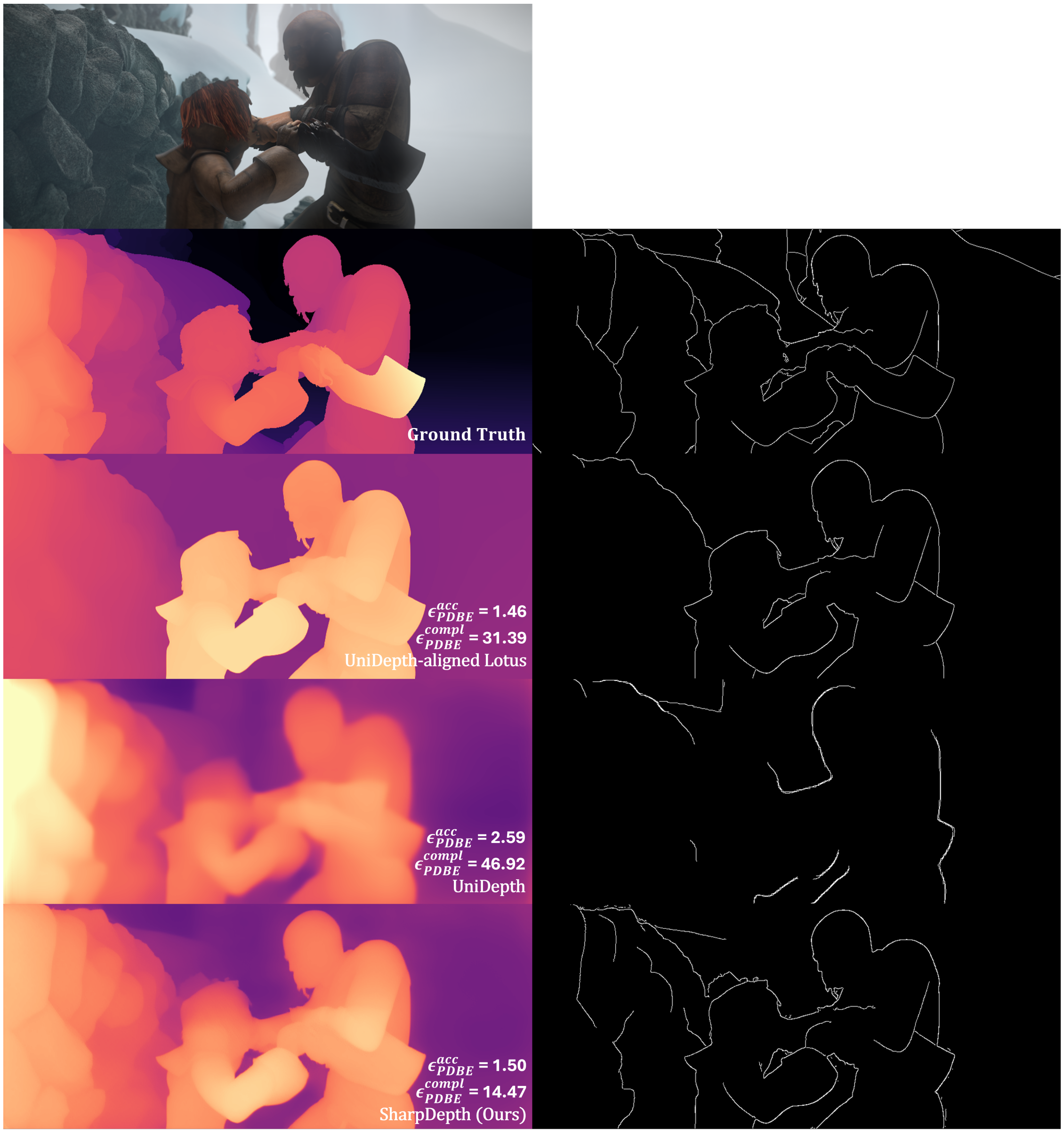
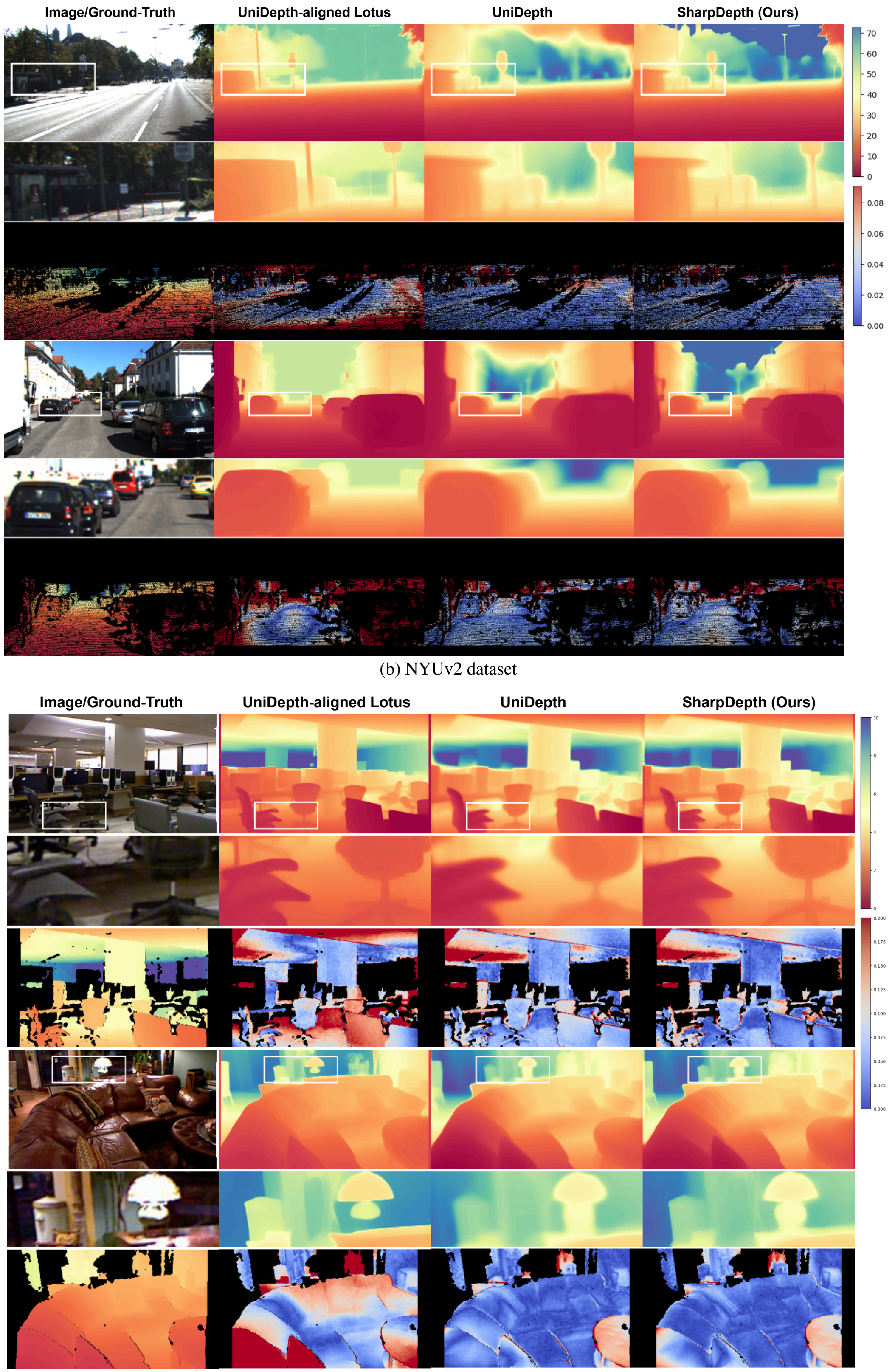


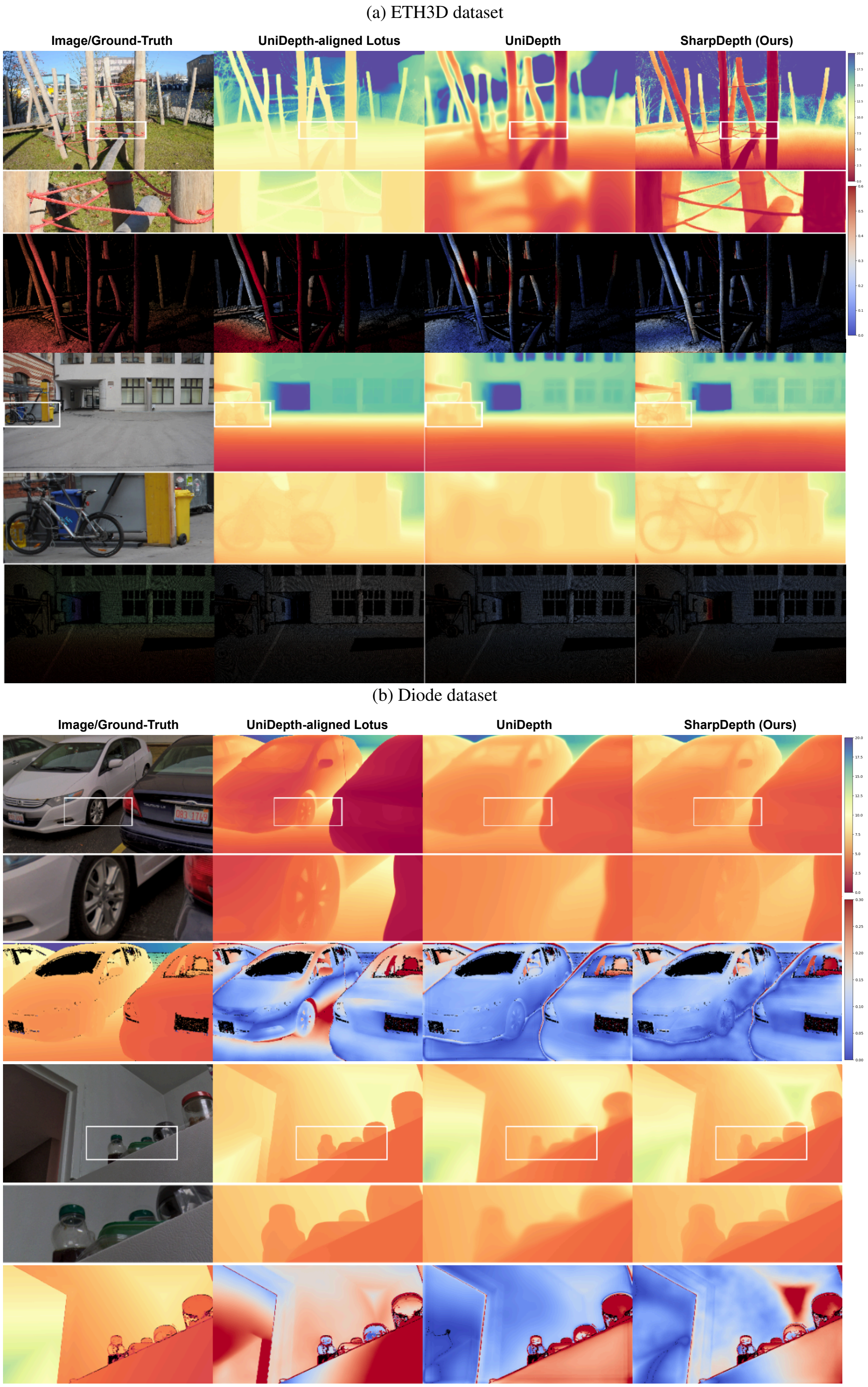
Figure 5. **Illustration of the depth boundary on the Sintel dataset**. We show the depth maps and extracted boundaries for each prediction.



(a) KITTI dataset

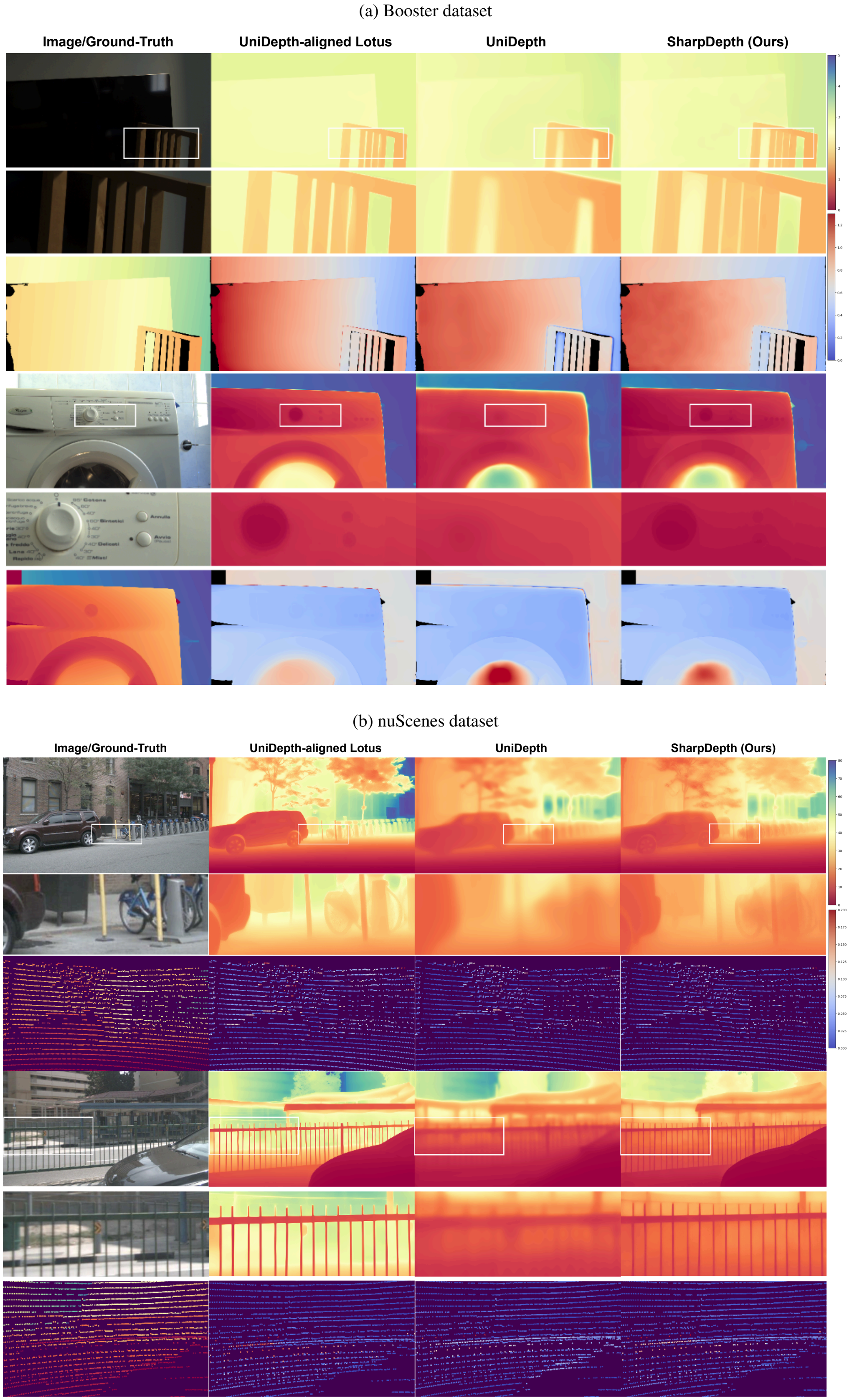
(b) NYUv2 dataset

Figure 6. **Qualitative comparisons on different datasets** (1/3). The last column represents the colormap ranges for depth (*spectral*) and AbsRel error (*coolwarm*).


(a) ETH3D dataset

(b) Diode dataset

Figure 7. **Qualitative comparisons on different datasets** (2/3). The last column represents the colormap ranges for depth (*spectral*) and AbsRel error (*coolwarm*).


(a) Booster dataset

(b) nuScenes dataset

Figure 8. **Qualitative comparisons on different datasets** (3/3). The last column represents the colormap ranges for depth (*spectral*) and AbsRel error (*coolwarm*).

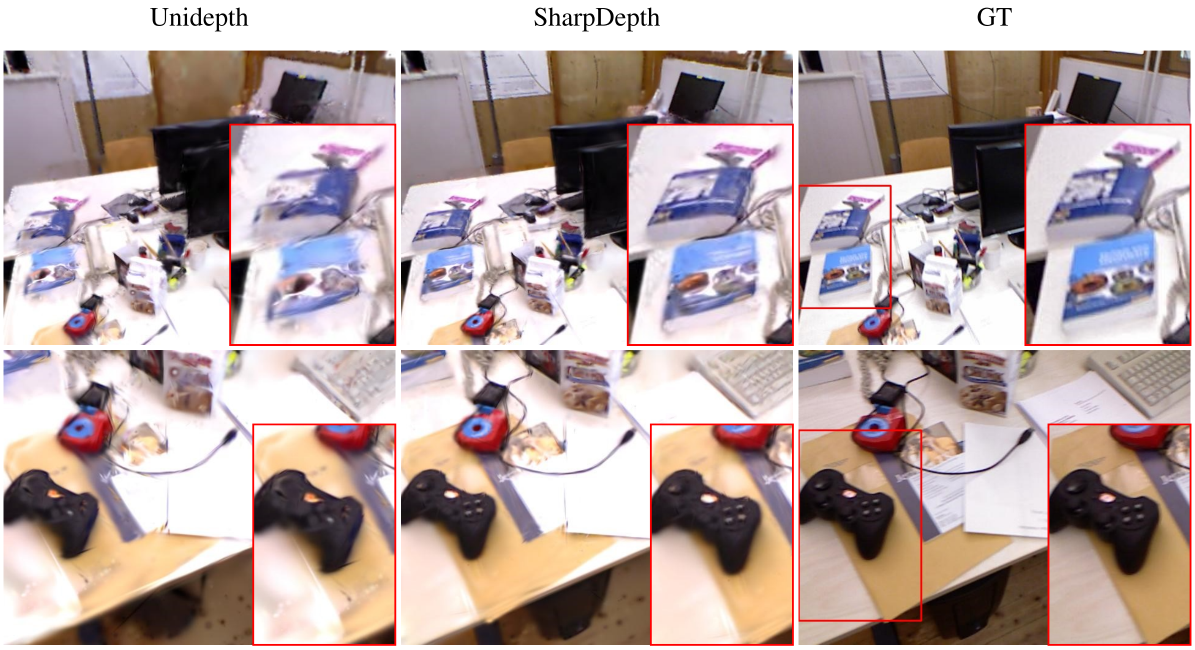


Figure 9. **Rendering comparison on TUM fr1/desk sequence**. For each method, we show the novel view rendering. Compared to UniDepth (leftmost column), using SharpDepth (middle column) can result in finer details of objects, such as the books in the first row and the game console in the second row.

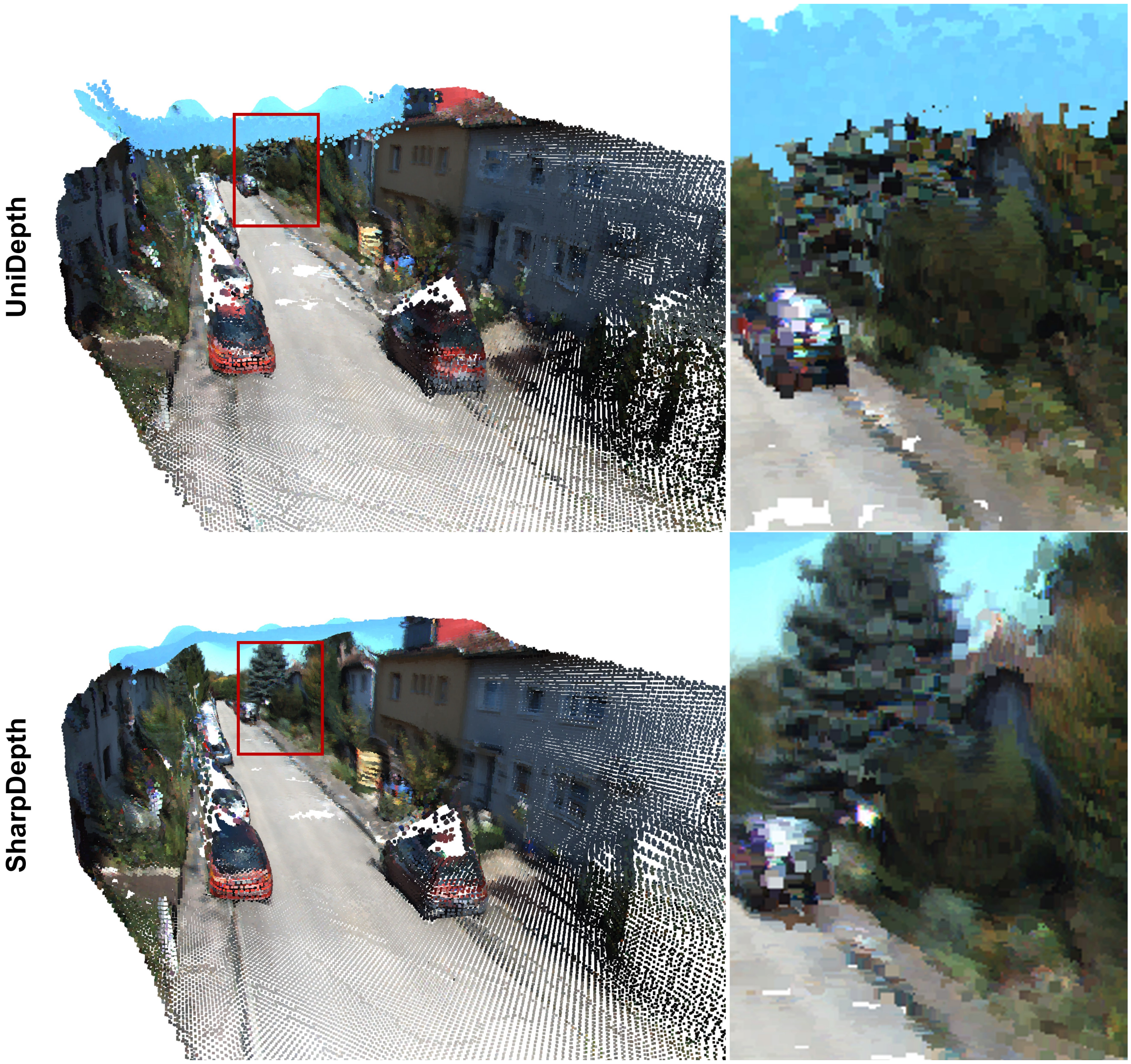


Figure 10. **Multi-view scene reconstruction on KITTI dataset**. We predict depth maps using UniDepth and SharpDepth for each frame and use TSDF-Fusion to generate the point cloud. SharpDepth's point cloud achieves less shape distortion in vehicles.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 1

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32 (11), 2013. 1

[6] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2), 2022. 2

[7] 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, 3

[8] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 1, 3

[9] 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, 3

[10] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *ECCV Workshops*, 2018. 1

[11] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In *ECCV*. Springer, 2024. 1, 3

[12] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[13] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, 2023. 1

[14] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 2, 3

[15] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: a benchmark for depth from images of specular and transparent surfaces. *PAMI*, 2023. 1

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. Springer, 2012. 1

[17] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 1, 2

[18] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1

[19] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *CVPR*, 2021. 1

[20] 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

[21] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1

[22] Jonas Wulff, Daniel J Butler, Garrett B Stanley, and Michael J Black. Lessons and insights from creating a synthetic optical flow benchmark. In *ECCV*. Springer, 2012. 1

[23] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021. 1

[24] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 3

[25] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1

[26] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017. 2

[27] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffu-

sion refiner for zero-shot monocular depth estimation. *arXiv preprint arXiv:2407.17952*, 2024. 1