Dense Metric Depth Completion from Sparse Direct Time-of-Flight Sensors

Supplementary Material

Table I. List of train datasets with domain, frame count, and sampling weight.

Name	Domain	#Frames	Weight
ApolloSynthetic	Outdoor/driving	194k	3.8
EDEN	Outdoor/garden	369k	1.2
GTA-SfM	Outdoor/in-the-wild	19k	2.7
Hypersim	Indoor	75k	4.8
IRS	Indoor	91k	5.4
KenBurns	In-the-wild	76k	1.5
MatrixCity	Outdoor/driving	391k	1.3
MidAir	Outdoor/in-the-wild	424k	3.8
MVS-Synth	Outdoor/driving	12k	1.2
ObjaverseV1	Object	168k	4.6
Structured3D	Indoor	77k	4.6
Synscapes	Outdoor/driving	25k	1.9
Synthia	Outdoor/driving	96k	1.1
UrbanSyn	Outdoor/driving	7.5k	2.0
UnrealStereo4K	In-the-wild	8k	1.6
TartanAir	In-the-wild	293k	4.8

1. Additional Results

Affine depth alignment comparison. We compare our method against state-of-the-art RGB-only MDE models with post-hoc sensor depth alignment to assess whether sparse dToF can be replaced by scale alignment alone. We align Depth Anything v2 and MoGe-1 predictions to sparse sensor depth using least-squares fitting, and additionally report robust ROE alignment for MoGe. As shown in Table IV, our method consistently outperforms these aligned baselines, demonstrating that encoder-stage fusion provides substantially richer geometric guidance than post-hoc alignment, even with strong monocular priors.

Full table of performance on varying input depth sparsity. For clarity, we report the full quantitative results in Tab. VII on varying input depth sparsity, which corresponds to the plot shown in Figure 6 of the main paper. The result shows that our method achieves best score in extreme sparse inputs and exhibits a smaller performance drop as sparsity increases compared to all competing methods, except for PriorDA, which has lower accuracy. Both the visualization and the table indicate that our model achieves superior relative stability under variations in input depth sparsity.

Depth completion qualitative results. We present qualitative comparisons for depth completion across all evaluation benchmarks, using both metric depth maps and point cloud visualizations. Figure I shows examples from ETH3D and iBims-1, where our method, especially in the Electro scene (row 1), accurately recovers thin structures in areas with sparse depth measurements. Results on DDAD in Figure II further demonstrate that our model correctly reconstructs trees (red box, row 1) and a telegraph pole (red arrow, row 2). Figure III presents qualitative results on DIODE. While existing methods fail to preserve local geometry, such as the metal bottle (row 1) or distort the lecture-room floor (row 2), our method maintains these structures and their spatial consistency. Finally, Figure IV shows results on real sensor datasets, including KITTI-DC and ZJUL5. Extremely sparse depth and textureless regions cause state-of-the-art methods fail to distinguish between a thin ornament with the bookshelf (row 1) or collapse desk legs into the foreground (row 2), whereas our method avoids such artifacts. In outdoor scenes such as KITTI-DC, our estimation prevents background regions from leaking into foreground objects.

Table II. Ablation study on flood-fill preprocessing and loss combinations. We report the average relative depth error (Rel) and threshold accuracy (δ) over evaluation benchmarks: KITTI-DC, ZJUL5, DDAD, DIODE, ETH3D, and iBims-1.

Methods	Rel	δ
Ours, w/o flood-fill	3.54	76.7
$\mathcal{L}_1 + \mathcal{L}_m$	3.53	77.4
$\mathcal{L}_1 + \mathcal{L}_g + \mathcal{L}_m$	3.47	77.8
$\mathcal{L}_1 + \mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_m$ (Ours)	3.46	77.9

Table III. Ablation study on number of the depth tokens N_d. We report the average relative depth error (Rel) and threshold accuracy (δ) over evaluation benchmarks: KITTI-DC, ZJUL5, DDAD, DIODE, ETH3D, and iBims-1. We measure inference speed under the same evaluation settings as the main paper.

Method		Inference time	Rel	δ
Ours	$N_d = 400$	32 ms	3.87	71.9
	$N_d = 800$	32 ms	3.55	76.5
	$N_d = 1000$	34 ms	3.47	77.4
	$N_d = 1200$	34 ms	3.47	77.6
	$N_d = 2000$	36 ms	3.43	78.1

2. Dataset details

Evaluation datasets. We utilize all images in the ETH3D (454 frames), iBims-1 (100 frames), and ZJUL5 datasets (1021 frames) for evaluation. For DDAD, we randomly select 1000 frames from the validation set. For KITTI-DC and DIODE, we use official validation splits. To ensure reliable evaluation, we further remove boundary artifacts in DIODE by detecting and masking out edge regions. For clarity, we additionally provide the input resolution used when reporting inference time. The resolutions are: KITTI-DC (1216×352), ZJUL5 (640×480), DDAD (1936×973), DIODE (1024×768), ETH3D (2048×1365), and iBims-1 (2048×1365).

Training dataset. Table I lists the synthetic datasets used to train our model. Our synthetic dataset covers diverse domains, including both indoor and outdoor scenes, as well as a wide variety of objects. The sampling ratio for training

Table IV. Comparison of metric depth accuracy after affine alignment of RGB-only monocular depth estimation models to sensor depth. Each dataset block reports the relative depth error (Rel) and threshold accuracy (δ). Best and second-best results are **bold** and underlined.

Method	KITTI-DC		ZJUL5		DDAD		DIODE		ETH3D		iBims-1		Average	
	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$
DAv2+LS	9.87	27.7	13.19	25.3	60.76	26.1	16.39	49.8	6.32	47.9	3.96	59.2	18.41	39.4
MoGe+LS	9.19	19.5	11.39	26.4	12.93	20.0	6.02	58.3	3.49	60.8	3.07	70.2	7.68	42.5
MoGe+ROE	4.61	48.7	**10.30**	**34.2**	8.46	37.0	3.05	75.7	2.67	71.6	2.50	77.4	5.27	57.5
Ours (ViT-S)	**2.00**	**84.0**	11.13	32.3	**4.61**	**68.8**	**0.89**	**96.7**	**0.84**	**94.8**	**1.29**	**90.9**	**3.46**	**77.9**

Table V. Full table of ablation study on the depth fusion architecture. We compared the depth-prompting variant decoder, the joint attention encoder, and the masked joint attention encoder architecture on KITTI-DC, ZJUL5, DDAD, DIODE, ETH3D, and iBims-1 datasets. Each dataset block reports the relative depth error (Rel) and threshold accuracy (δ). Best results in each group are **bolded**

Method	KITTI-DC		ZJUL5		DDAD		DIODE		ETH3D		iBims-1		Average	
	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ
Depth prompting	2.14	82.1	12.16	31.2	5.93	58.2	1.23	94.1	1.39	88.2	1.60	87.6	4.07	73.6
Joint attention	2.14	81.6	12.42	30.8	5.02	64.2	1.04	95.7	1.07	92.6	1.55	88.6	3.87	75.6
Masked joint attention (Ours)	**2.00**	**84.0**	**11.13**	**32.3**	**4.61**	**68.8**	**0.89**	**96.7**	**0.84**	**94.8**	**1.29**	**90.9**	**3.46**	**77.9**

Table VI. Full table of ablation study on the ViT backbone scale. We evaluated our method with ViT-S, ViT-B, and ViT-L model as the encoder backbone on KITTI-DC, ZJUL5, DDAD, DIODE, ETH3D, and iBims-1 datasets. Each dataset block reports the relative depth error (Rel) and threshold accuracy (δ). Best results in each group are **bolded**

Method	KITTI-DC		ZJUL5		DDAD		DIODE		ETH3D		iBims-1		Average	
	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ	Rel	δ
Ours, ViT-S	2.00	84.0	11.13	32.3	4.61	68.8	0.89	96.7	0.84	94.8	1.29	90.9	3.46	77.9
Ours, ViT-B	1.93	84.7	10.56	33.5	4.53	70.6	0.80	97.1	0.72	95.9	1.16	92.0	3.28	79.0
Ours, ViT-L	**1.90**	**85.0**	**9.95**	**34.7**	**4.27**	**72.7**	**0.68**	**97.7**	**0.58**	**97.0**	**1.04**	**93.2**	**3.07**	**80.1**

Table VII. Full table of quantitative results showing performance with varying numbers of depth input points. Each block reports the average relative depth error (%) over the evaluation benchmarks. Best and second-best results are **bold** and underlined.

	# Depth Points				
Method	100	250	500	1000	2500
OMNI-DC	5.43	3.05	2.10	**1.53**	**1.04**
Prior DA	3.71	3.07	2.69	2.40	2.10
Prompt DA	7.51	4.86	3.64	2.96	2.45
Marigold DC	6.18	5.11	4.56	4.18	3.84
Ours	**3.28**	**2.44**	**2.00**	1.67	1.41

follows the weighting strategy of MoGe [1].

3. More ablation results

Flood-fill. We use flood-fill to convert sparse depth into a 2D representation for ViT tokenization, while a validity mask preserves sparsity and avoids discontinuity bias. Flood-fill introduces no artifacts and stabilizes learning under extreme sparsity, with density-invariant efficiency fixed by the 2D token budget. An ablation study in Table II confirms its effectiveness, as removing flood-fill degrades performance.

Loss. We conduct an ablation study on different loss combinations and evaluate the performance on the same datasets used in the depth completion experiments. The average relative depth error and threshold accuracy are used as evaluation metrics. As shown in Table II, incorporating the global scale-invariant loss $\mathcal{L}_g$ yields a slight performance gain, and further adding the local scale-invariant loss $\mathcal{L}_l$ provides an additional improvement.

Number of depth tokens. We also study the effect of varying the number of depth tokens at inference time. We evaluate models with 400–2000 depth tokens (trained with 800–1200). Table III shows that reducing depth tokens can decrease inference time by $\sim$4 ms with only a 0.4% increase in error, demonstrating a flexible trade-off between efficiency and accuracy.

Full table of depth fusion architecture ablation. We report detailed ablation results of the network architecture on each benchmark in Table V. The table shows that an encoder with masked joint attention consistently outperforms both the encoder with joint attention and the decoder based on depth prompting.

Full table of ViT backbone ablation. We include detailed results of ViT backbone ablation in Table VI. All variants are trained under identical settings for 100K iterations. The results show that increasing network scale consistently improves performance across all benchmarks, demonstrating that our proposed fusion architecture scales well with backbone capacity.

References

[1] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 2

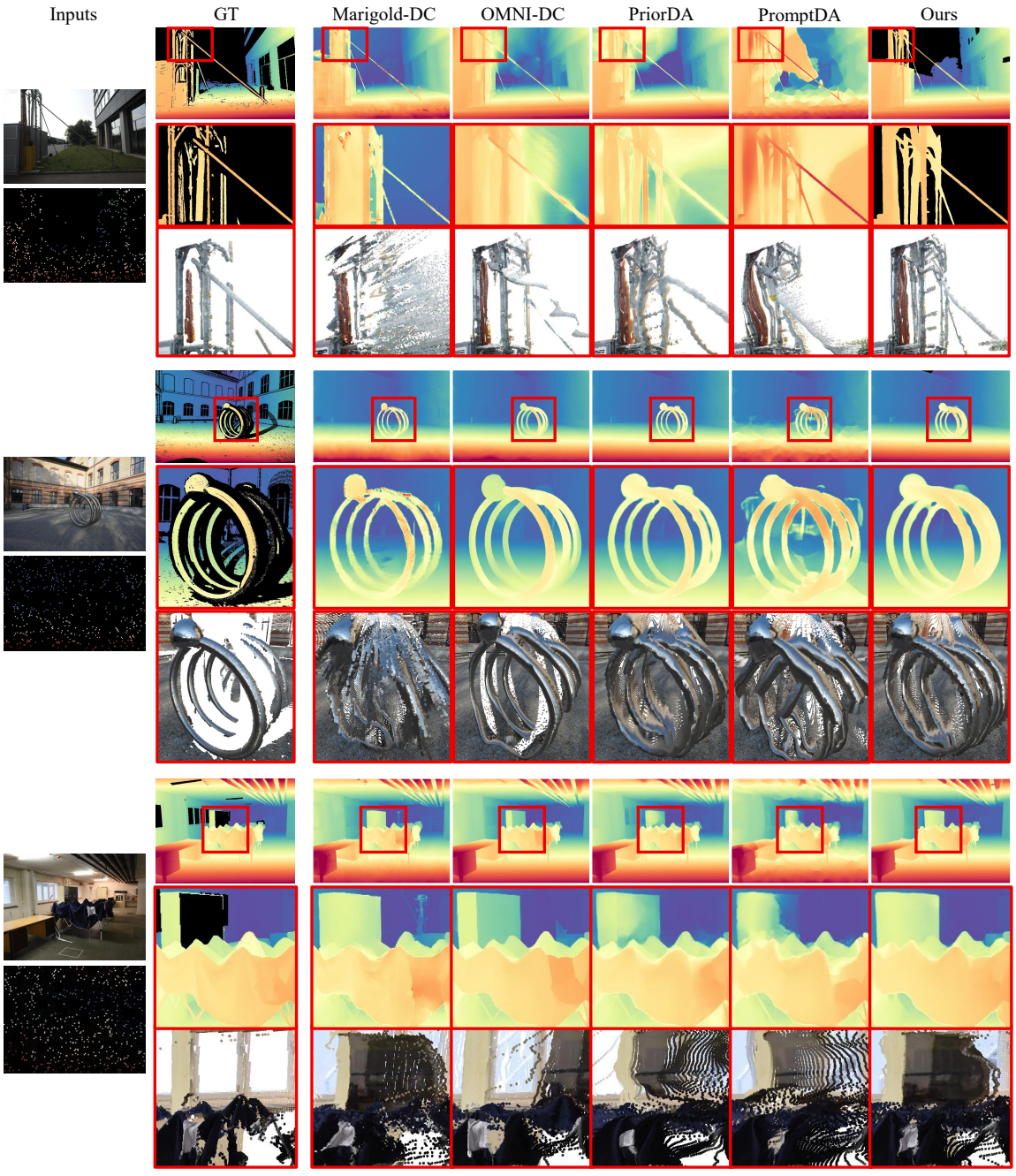

Figure I. Qualitative comparison of metric depth maps, and reconstructed point clouds. Examples are from the ETH3D dataset's electro (top), facade (middle) scenes, and iBims-1's factory (bottom) scene.

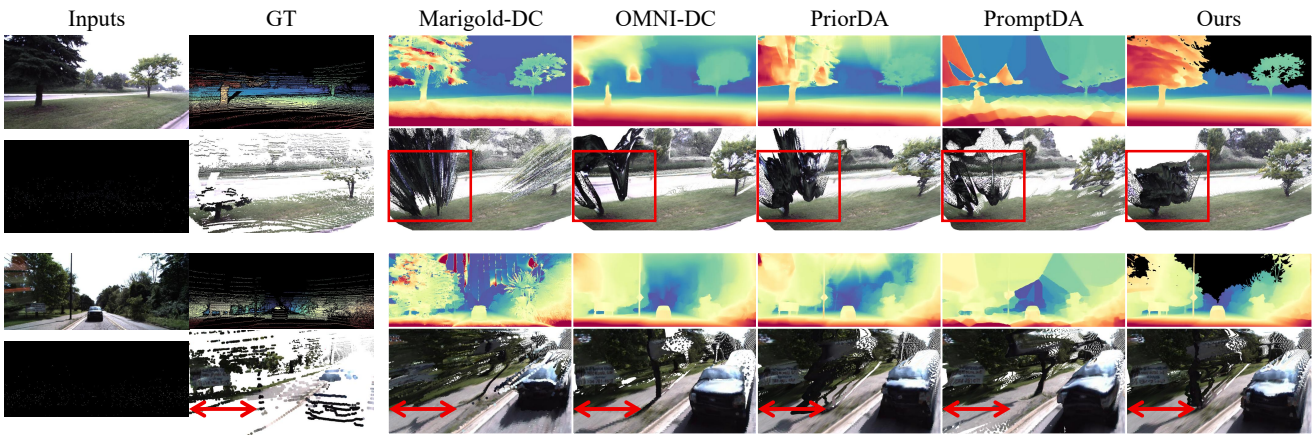

Figure II. Qualitative comparison of metric depth maps, and reconstructed point clouds. Examples are from DDAD dataset

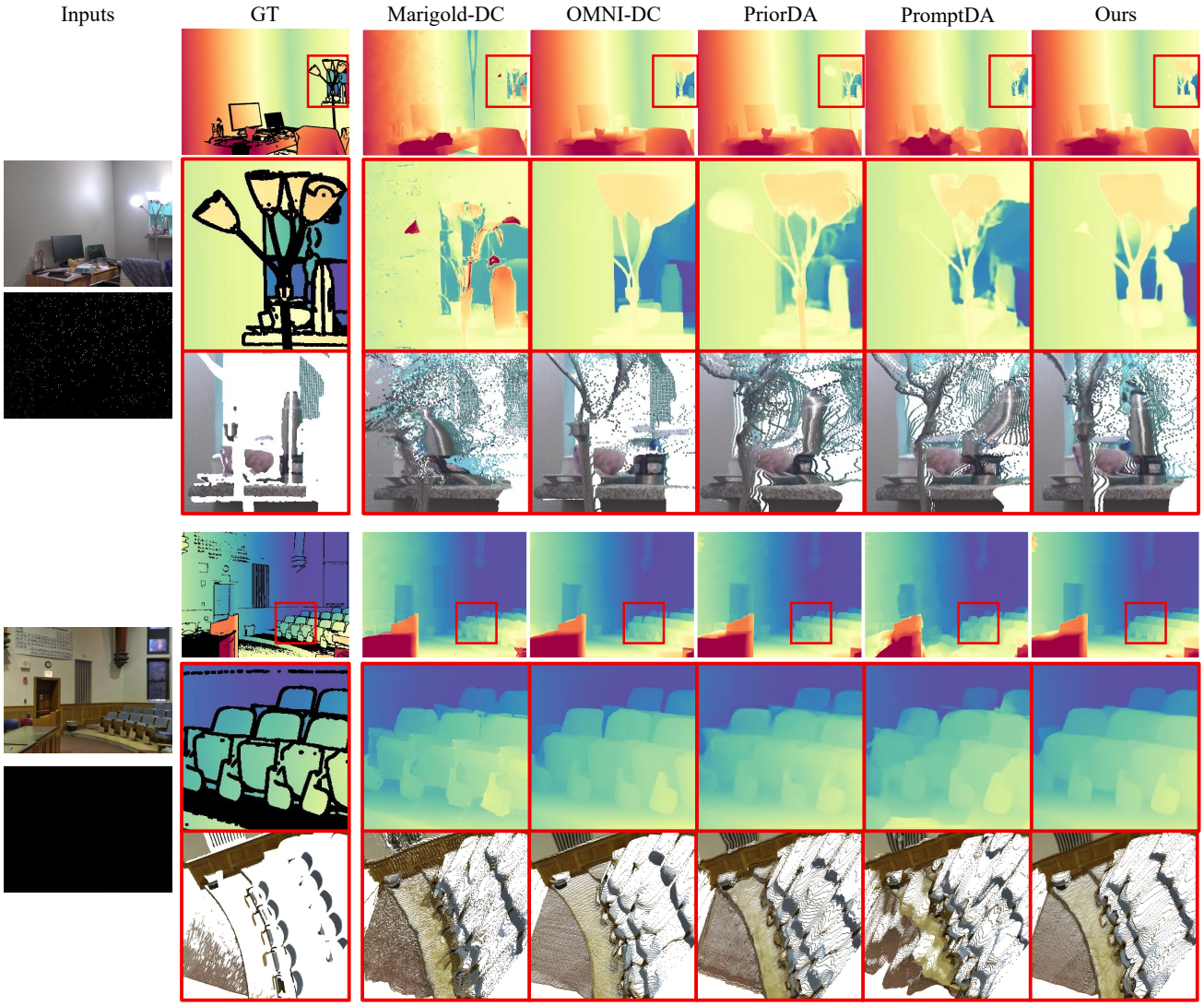

Figure III. Qualitative comparison of metric depth maps, and reconstructed point clouds. Examples are from the DIODE dataset

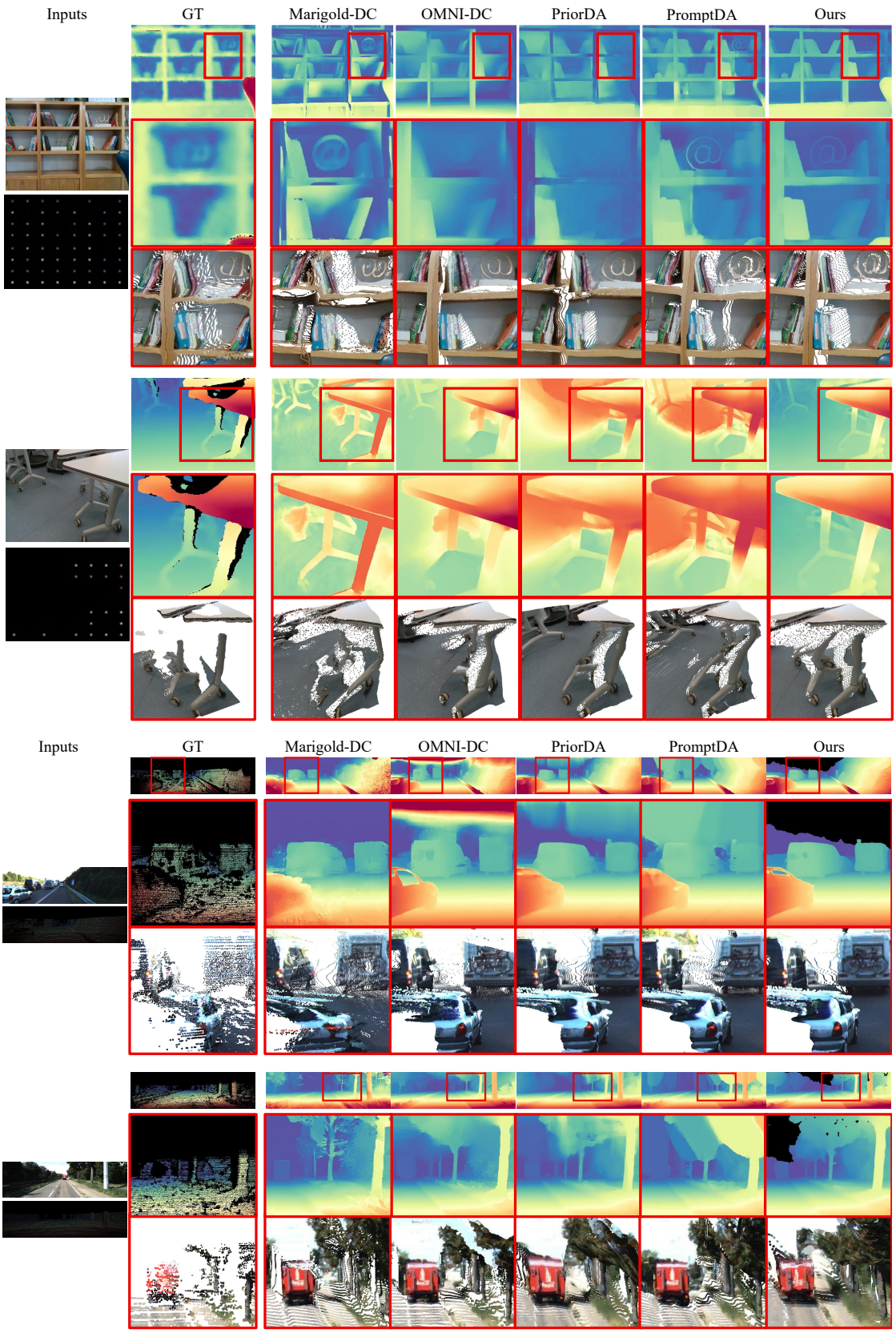

Figure IV. Qualitative comparison of metric depth maps, and reconstructed point clouds. The first row shows examples from the cafe1 scene and the second from the teaching region scene of the ZJUL5 dataset, while the third and fourth rows present examples from the KITTI-DC dataset.