Digging into Depth Priors for Outdoor Neural Radiance Fields Supplemental Material

ABSTRACT

In this supplementary material, we provide more implementation details and experimental results to support our findings. Firstly, more detailed experimental result and one more baseline model experiment is provided to sufficiently prove our claim. Then, we discuss the detailed ablation study and show a visual comparison between different depth priors.

A IMPLEMENTATION DETAILS

NeRF Method For both KITTI [1] and Argoverse [2] dataset, we train the mentioned methods on the selected sequences with a fixed number of steps and evaluate the testing viewpoints in terms of photorealistic metrics and depth accuracy metrics. Specifically, we train MipNeRF-360 for 75,000 iterations with a depth weight of 10 using the official codebase[1]. For Instant-NGP, we use the PyTorch re-implemented version[2]. The model is trained for 30 epochs with a depth weight of 0.5. All of the experiments are performed with Tesla V100 GPUs.

Depth Method For both KITTI [1] and Argoverse [2] dataset, we firstly re-split the training and testing dataset according to the selected sequence, i.e., using the selected sequence as the testing dataset and the rest as the training dataset. Then, we retrain the mentioned methods with their official implementation of BTS[3], CFNet[4], PCWNet[5], and re-implement MFFNet by ourselves. Note that for all generated depth maps, we crop the sky area which has an infinite distance and has no ground supervision. For binocular depth estimation, we select CFNet and PCWNet as the representative work in KITTI and Argoverse datasets, respectively. All of the experiments are performed with Tesla V100 GPUs.

B DETAILED EXPERINMENTAL RESULT

In this section, we will specifically introduce the used sequence in the two publicly available datasets and the corresponding result.

B.1 Dataset

KITTI: For the KITTI dataset, We use the following sequences:

(1) Seq00_2011_10_03_drive_0027_sync: frame 657 - 787
(2) Seq00_2011_10_03_drive_0027_sync: frame 890 - 1028
(3) Seq00_2011_10_03_drive_0027_sync: frame 2700 - 3000
(4) Seq02_2011_10_03_drive_0034_sync: frame 2749 - 2929
(5) Seq05_2011_09_30_drive_0018_sync: frame 400 - 725

Argoverse For the Argoverse dataset, We use all frames in the following sequence:

(1) Training set: 2c07fcda-6671-3ac0-ac23-4a232e0e031e
(2) Validation set: 70d2aea5-dbeb-333d-b21e-76a7f2f1ba1c
(3) Validation set: cb0cba51-dfaf-34e9-a0c2-d931404c3dd8

[1] https://github.com/google-research/multinerf
[2] https://github.com/kwea123/ngp_pl
[3] https://github.com/cleinc/bts
[4] https://github.com/gallenszl/CFNet
[5] https://github.com/gallenszl/PCWNet

B.2 Additional Results

The corresponding results in each sequence are shown in Tab. 2 and Tab.3. The conclusions in each sequence are consistent with the results reported in Tab.3 and Tab.5 of the main paper. Consequently, these results further support our finding1: Monocular depth is enough for sparse viewpoints (lines 735-743 of the main paper) and finding2: depth supervision is an option for dense viewpoints (lines 779-786 of the main paper).

C DETAILED ABLATION STUDY

In our *finding 4* of the main paper(lines 904-909), we claim that directly cropping the sky area with MSE supervision is enough. To further validate our claim, we investigate the influence of the cropping-based depth filtering strategy on all depth priors. The corresponding results are shown in Tab. 1. Note that because the raw LiDAR supervision originally does not have a valid value in the sky area, we exclude the corresponding results. It can be seen from the table that the cropping-based depth filtering strategy is beneficial for the performance of all depth priors, which verifies our finding. Moreover, as the estimation result of sky area is worse in monocular depth estimation and depth completion, the depth filtering strategy achieves a larger gain in these two depth priors.

D MORE VISUALIZATION

We visualize the point clouds for different depth supervision, which can be seen in Fig. 1. We can see that when using only RGB, the point clouds are extremely scattered and shows inaccurate geometry. Adding additional depth supervision will greatly alleviate this problem and helps NeRFs converge to a better geometry.

We also give a more qualitative comparison between different depth priors, which can be seen in Tab. 4 of the main paper. As shown in Fig. 2, depth completion achieves the best accuracy in GT valid area and then goes with binocular depth estimation and monocular depth estimation, which is consistent with the qualitative results. However, depth completion and monocular depth estimation cannot generate reasonable results in gt invalid area, i.e., the sky area. Hence, we need the cropping-based depth filtering strategy to filter out the unreasonable area. Tab. 1 shows the effectiveness of the proposed method.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012.
[2] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8748–8757.
[3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.
[4] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. on Graphics (TOG)*, 2022.

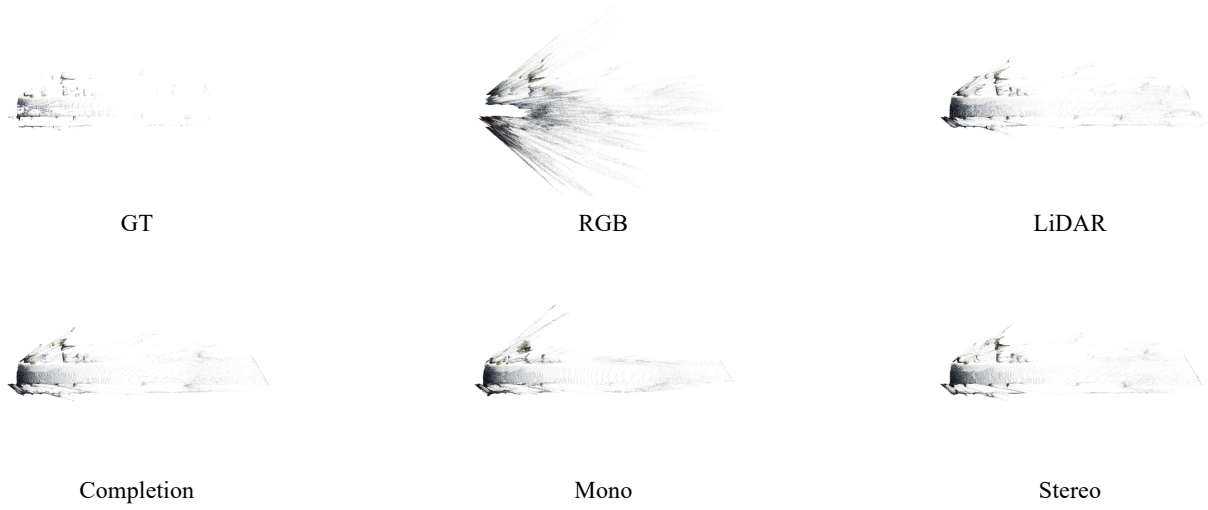

Figure 1: Point cloud visualization of MipNeRF-360 under different depth supervision.

Table 1: Detailed ablation study of the proposed cropping-based depth filtering strategy on all depth priors.

	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	Absrel↓
RGB	14.80	0.475	0.551	4.569	0.153
Completion	17.98	0.540	0.540	1.057	**0.038**
Completion Crop	**18.39**	**0.554**	**0.505**	**1.051**	**0.038**
Mono	17.12	0.510	0.543	2.492	0.075
Mono Crop	**17.97**	**0.542**	**0.510**	**2.383**	**0.073**
Stereo	18.80	**0.562**	0.508	**1.347**	0.042
Stereo Crop	**18.87**	**0.562**	**0.501**	1.349	**0.040**

[5] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.

Table 2: Quantitative comparison with selected methods on each sequence of KITTI dataset. The best results are bolded.

Method	Sequence	Depth Supervision	Dense PSNR↑	Dense SSIM↑	Dense LPIPS↓	Dense RMSE↓	Dense Absrel↓	Sparse PSNR↑	Sparse SSIM↑	Sparse LPIPS↓	Sparse RMSE↓	Sparse Absrel↓
MipNeRF-360 [3]	(1)	RGB-Only	**22.66**	**0.755**	**0.403**	3.719	0.112	16.54	0.640	0.480	6.284	0.174
		GT Depth	22.30	0.744	0.420	0.644	0.022	20.08	**0.696**	**0.446**	0.714	0.026
		Depth Completion	21.71	0.723	0.453	**0.627**	**0.018**	**20.38**	0.694	0.467	**0.651**	**0.020**
		Stereo Depth	21.71	0.720	0.457	1.219	0.026	20.40	0.691	0.469	1.264	0.028
		Mono Depth	21.73	0.723	0.449	2.362	0.056	19.79	0.681	0.468	2.418	0.061
	(2)	RGB-Only	**21.57**	**0.683**	**0.406**	2.656	0.058	16.86	0.579	0.467	3.704	0.106
		GT Depth	21.48	0.673	0.421	0.821	0.020	**19.18**	**0.621**	**0.441**	0.876	0.025
		Depth Completion	21.01	0.650	0.452	**0.711**	**0.017**	19.16	0.612	0.465	**0.811**	**0.019**
		Stereo Depth	20.93	0.648	0.455	1.224	0.022	19.17	0.612	0.466	1.256	0.025
		Mono Depth	21.14	0.651	0.451	2.096	0.053	18.60	0.600	0.470	2.197	0.057
	(3)	RGB-Only	**21.83**	**0.641**	**0.460**	2.482	0.071	14.80	0.475	0.551	4.569	0.153
		GT Depth	21.56	0.629	0.473	1.015	0.036	17.47	0.542	0.507	1.173	0.045
		Depth Completion	21.26	0.621	0.482	**0.894**	**0.030**	18.39	0.554	0.505	**1.051**	**0.038**
		Stereo Depth	21.40	0.621	0.482	1.249	0.033	**18.87**	**0.562**	**0.501**	1.349	0.040
		Mono Depth	21.15	0.615	0.486	2.287	0.062	17.97	0.542	0.510	2.383	0.073
	(4)	RGB-Only	21.61	**0.678**	**0.466**	4.050	0.115	17.81	0.609	0.502	5.386	0.170
		GT Depth	**21.67**	0.672	0.474	1.009	0.040	19.29	0.631	0.491	1.267	0.055
		Depth Completion	21.60	0.669	0.479	**0.946**	**0.031**	**19.74**	**0.636**	**0.489**	**1.049**	**0.036**
		Stereo Depth	21.59	0.669	0.480	1.181	0.035	19.69	**0.636**	0.491	1.315	0.038
		Mono Depth	21.59	0.667	0.481	1.953	0.057	19.61	0.632	0.492	2.026	0.063
	(5)	RGB-Only	**22.30**	**0.702**	**0.453**	2.543	0.084	18.65	0.643	0.491	3.370	0.116
		GT Depth	22.19	0.693	0.465	1.099	0.043	19.71	0.657	0.485	1.193	0.050
		Depth Completion	21.99	0.689	0.470	**0.912**	**0.033**	20.57	0.663	0.484	**0.975**	**0.037**
		Stereo Depth	22.03	0.688	0.470	1.090	**0.033**	**20.62**	**0.665**	**0.481**	1.148	0.038
		Mono Depth	21.81	0.685	0.473	2.106	0.068	20.20	0.655	0.488	2.194	0.076
InstantNGP [4]	(1)	RGB-Only	21.59	0.701	0.433	9.352	0.534	14.74	0.543	0.534	14.020	0.727
		GT Depth	**21.84**	**0.705**	**0.429**	**0.985**	**0.032**	**19.35**	**0.652**	**0.454**	**1.083**	0.036
		Depth Completion	21.31	0.684	0.462	1.041	**0.032**	19.33	0.638	0.487	1.106	**0.035**
		Stereo Depth	21.19	0.677	0.467	1.311	0.044	19.13	0.629	0.492	1.384	0.046
		Mono Depth	20.59	0.658	0.478	2.457	0.062	18.62	0.606	0.500	2.571	0.067
	(2)	RGB-Only	20.45	0.621	0.426	9.141	0.479	13.28	0.418	0.555	15.783	0.832
		GT Depth	**21.02**	**0.641**	**0.408**	**0.913**	0.026	18.06	**0.570**	**0.436**	**1.117**	0.035
		Depth Completion	20.77	0.624	0.441	0.986	**0.024**	**18.23**	0.560	0.464	1.204	**0.032**
		Stereo Depth	20.94	0.620	0.440	1.253	0.032	18.11	0.558	0.467	1.471	0.040
		Mono Depth	20.53	0.607	0.456	2.108	0.064	17.77	0.535	0.480	2.371	0.077
	(3)	RGB-Only	20.70	0.644	0.471	9.338	0.475	16.47	0.554	0.510	14.919	0.810
		GT Depth	**21.53**	**0.658**	**0.457**	**1.365**	0.055	18.63	0.596	**0.481**	1.715	0.070
		Depth Completion	21.47	0.652	0.468	1.398	**0.047**	**19.06**	**0.603**	0.483	**1.593**	**0.056**
		Stereo Depth	21.36	0.649	0.472	1.398	0.050	**19.06**	0.599	0.484	1.698	0.062
		Mono Depth	21.42	0.647	0.475	2.175	0.072	18.85	0.593	0.489	2.451	0.085
	(4)	RGB-Only	17.65	0.502	0.526	10.861	0.494	13.47	0.360	0.600	16.730	0.806
		GT Depth	**19.65**	**0.548**	**0.496**	**3.475**	**0.101**	**16.63**	**0.464**	**0.518**	**3.589**	**0.108**
		Depth Completion	19.14	0.531	0.515	3.613	0.102	16.59	0.454	0.537	3.866	0.116
		Stereo Depth	19.39	0.530	0.514	3.760	0.111	16.87	0.455	0.536	3.774	0.117
		Mono Depth	18.95	0.517	0.525	4.262	0.135	16.40	0.442	0.540	4.405	0.147
	(5)	RGB-Only	22.17	0.681	0.445	9.180	0.555	19.23	0.620	0.481	13.604	0.792
		GT Depth	**22.47**	**0.697**	**0.431**	**1.115**	0.047	**19.96**	**0.646**	**0.458**	**1.250**	**0.053**
		Depth Completion	21.78	0.671	0.464	1.268	**0.045**	19.88	0.625	0.487	1.393	0.055
		Stereo Depth	21.80	0.671	0.465	1.429	0.048	19.81	0.626	0.488	1.596	0.056
		Mono Depth	21.46	0.657	0.479	2.395	0.095	19.21	0.610	0.499	2.540	0.104

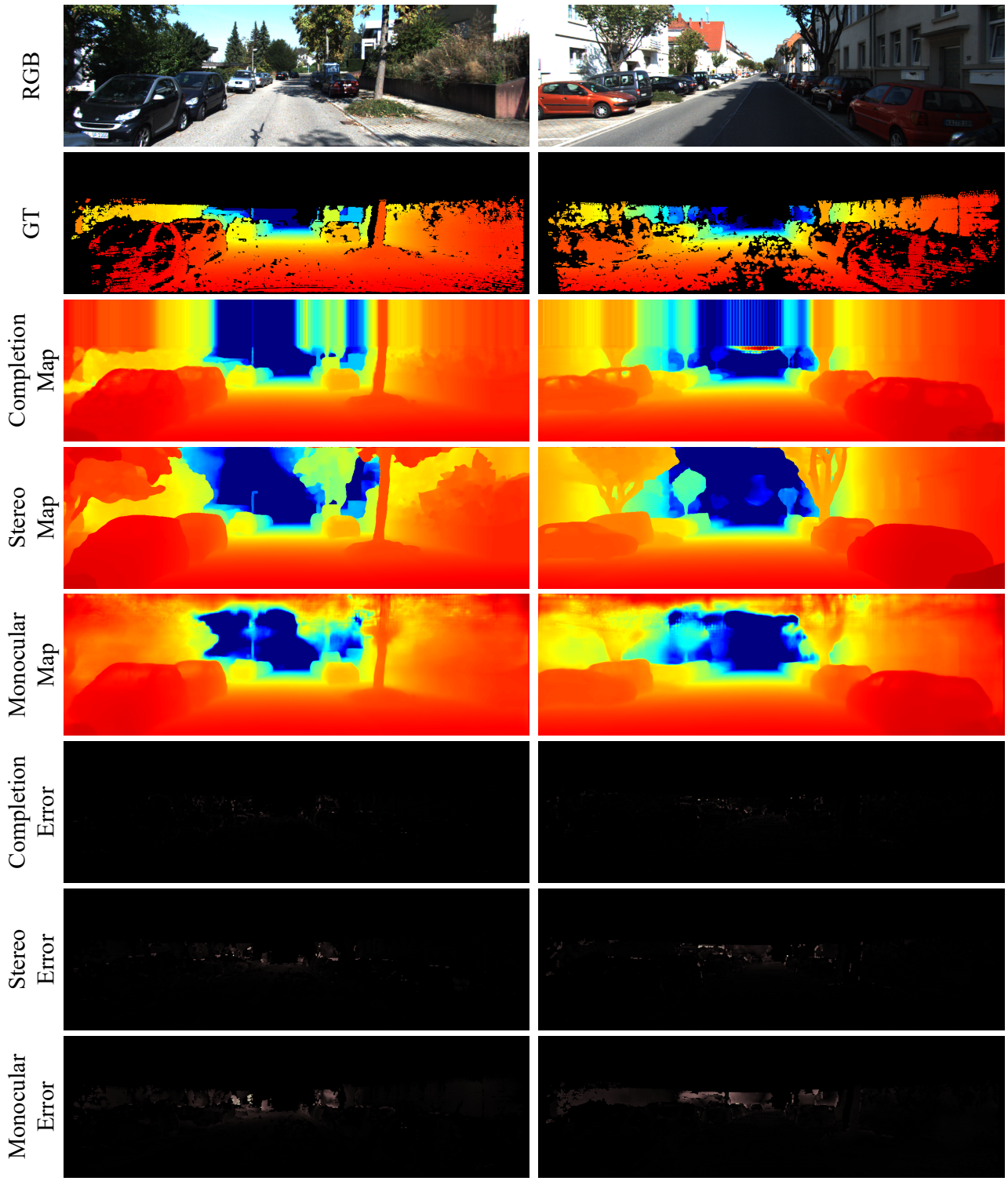

Figure 2: Qualitative results on the KITTI dataset with different depth recovery methods. White and blue denotes larger error and further distance in the error map and colorized depth map, respectively. For depth recovery accuracy, depth completion has the best accuracy in gt valid area then goes with binocular depth estimation and monocular depth estimation

Table 3: Quantitative comparison with selected methods on each sequence of Argoverse dataset. The best results are bolded.

Method	Sequence	Depth Supervision	Dense PSNR↑	Dense SSIM↑	Dense LPIPS↓	Dense RMSE↓	Dense Absrel↓	Sparse PSNR↑	Sparse SSIM↑	Sparse LPIPS↓	Sparse RMSE↓	Sparse Absrel↓
MipNeRF-360 [3]	(1)	RGB-Only	**27.81**	**0.839**	**0.459**	5.253	0.114	24.60	0.819	0.480	6.158	0.136
		GT Depth	27.39	0.830	0.469	**1.735**	**0.041**	**26.35**	**0.824**	**0.474**	**1.812**	**0.043**
		Stereo Depth	26.81	0.820	0.482	3.173	0.056	26.31	0.817	0.483	3.194	0.058
		Mono Depth	27.15	0.825	0.481	3.866	0.078	26.59	0.821	0.481	4.085	0.077
	(2)	RGB-Only	**29.42**	**0.865**	**0.430**	6.884	0.129	24.44	0.822	0.458	8.185	0.144
		GT Depth	28.91	0.856	0.443	**3.073**	**0.056**	**28.03**	**0.849**	**0.446**	**3.571**	**0.062**
		Stereo Depth	28.52	0.849	0.452	4.670	0.061	27.46	0.841	0.452	4.841	0.068
		Mono Depth	28.76	0.852	0.446	5.004	0.096	27.94	0.844	0.450	5.338	0.096
	(3)	RGB-Only	**30.81**	**0.863**	**0.448**	6.204	0.118	28.40	0.845	**0.466**	6.569	0.137
		GT Depth	30.04	0.852	0.464	**1.944**	**0.037**	**29.64**	**0.848**	0.467	**1.947**	**0.039**
		Stereo Depth	29.63	0.843	0.476	4.970	0.076	29.40	0.841	0.477	4.895	0.074
		Mono Depth	29.84	0.846	0.472	4.957	0.106	29.59	0.844	0.472	5.181	0.106
InstantNGP [4]	(1)	RGB-Only	25.85	**0.823**	0.482	18.462	0.774	22.03	0.804	0.505	19.026	0.751
		GT Depth	**27.83**	**0.823**	**0.473**	**1.947**	**0.061**	**26.30**	0.811	0.482	**2.018**	**0.066**
		Stereo Depth	27.21	0.817	0.478	5.365	0.091	26.04	**0.814**	**0.481**	5.535	0.100
		Mono Depth	27.03	0.816	0.486	5.653	0.113	25.71	0.808	0.488	5.812	0.112
	(2)	RGB-Only	28.40	**0.867**	**0.424**	8.804	0.263	24.38	0.844	0.449	11.884	0.378
		GT Depth	**28.56**	**0.867**	0.426	**1.993**	**0.043**	**26.97**	**0.849**	**0.440**	**1.934**	**0.043**
		Stereo Depth	27.83	0.857	0.441	4.977	0.075	26.28	0.839	0.452	5.386	0.082
		Mono Depth	28.27	0.859	0.442	5.612	0.111	26.48	0.845	0.451	6.469	0.129
	(3)	RGB-Only	29.96	**0.851**	**0.445**	13.167	0.443	20.12	0.800	0.528	21.407	0.650
		GT Depth	**30.38**	0.850	0.449	**1.471**	**0.031**	28.88	**0.842**	**0.457**	**1.691**	**0.035**
		Stereo Depth	29.93	0.843	0.460	6.497	0.104	**28.98**	0.831	0.470	6.610	0.109
		Mono Depth	29.63	0.838	0.471	6.984	0.142	28.71	0.834	0.475	7.649	0.153

Table 4: Quantitative comparison of NeRF++ on each sequence of KITTI dataset. The best results are bolded.

Method	Sequence	Depth Supervision	Dense					Sparse				
			PSNR↑	SSIM↑	LPIPS↓	RMSE↓	Absrel↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	Absrel↓
NeRF++ [5]	(1)	RGB-Only	**20.64**	**0.657**	**0.512**	51.983	4.214	17.88	0.598	0.534	53.995	4.372
		GT Depth	19.98	0.621	0.556	**1.490**	**0.061**	19.39	0.622	0.543	**1.529**	**0.058**
		Mono Depth	20.41	0.639	0.534	2.569	0.077	19.04	0.600	**0.560**	2.731	0.085
		Depth Completion	20.03	0.623	0.550	1.581	0.066	**19.37**	0.616	0.544	1.594	0.066
		Stereo Depth	18.92	0.581	0.587	3.235	0.084	19.34	**0.617**	0.544	1.658	0.066
	(2)	RGB-Only	19.83	0.234	0.539	42.549	2.938	16.50	0.503	0.553	57.436	4.998
		GT Depth	20.09	0.567	0.539	**1.942**	**0.047**	18.84	0.540	0.549	**1.208**	**0.050**
		Mono Depth	**20.27**	0.570	0.537	2.107	0.074	18.63	0.540	0.553	2.147	0.078
		Depth Completion	20.19	**0.571**	0.534	1.282	0.051	**18.90**	**0.543**	**0.548**	1.370	0.057
		Stereo Depth	20.18	**0.571**	**0.532**	1.356	0.052	18.58	0.525	0.574	1.530	0.068
	(3)	RGB-Only	**19.83**	**0.517**	**0.566**	42.549	2.938	16.91	0.452	**0.588**	55.046	4.735
		GT Depth	19.59	0.503	0.583	**2.110**	**0.087**	17.99	0.472	0.594	**2.292**	**0.111**
		Mono Depth	19.55	0.501	0.582	2.847	0.118	17.77	**0.471**	0.595	2.953	0.127
		Depth Completion	19.64	0.506	0.580	2.296	0.100	18.03	0.476	0.592	2.398	0.110
		Stereo Depth	19.67	0.507	0.578	2.233	0.108	**18.16**	0.476	0.590	2.352	0.117
	(4)	RGB-Only	**20.26**	**0.585**	**0.559**	46.588	3.704	17.78	0.544	0.577	53.181	4.544
		GT Depth	20.08	0.576	0.577	**2.128**	**0.096**	18.66	**0.554**	0.585	**2.300**	**0.110**
		Mono Depth	20.00	0.574	0.579	2.913	0.117	18.63	0.551	0.586	2.985	0.133
		Depth Completion	20.05	0.574	0.577	2.406	0.108	**18.70**	0.553	**0.587**	2.503	0.116
		Stereo Depth	20.01	0.573	0.577	2.319	0.109	18.65	**0.554**	0.586	2.426	0.117
	(5)	RGB-Only	**20.91**	**0.608**	**0.549**	59.522	5.792	18.94	0.578	0.558	61.611	6.148
		GT Depth	20.66	0.602	0.560	**1.900**	**0.098**	19.60	**0.583**	0.572	**2.083**	**0.114**
		Mono Depth	20.51	0.596	0.566	2.655	0.122	**19.62**	0.580	0.576	2.667	0.130
		Depth Completion	20.60	0.601	0.562	2.104	0.111	19.51	0.580	**0.577**	2.245	0.122
		Stereo Depth	20.58	0.600	0.563	2.138	0.110	19.52	0.577	0.575	2.343	0.123