Neural Radiance Fields with LiDAR Maps - Supplementary Material

Ming-Fang Chang[1] Akash Sharma[1] Michael Kaess[1] Simon Lucey[2]

[1]Carnegie Mellon University [2]The University of Adelaide

In this document, we provide additional information about the proposed work, including an additional study for the cGAN loss, details for experiments and datasets, an example showing the importance of introducing the depth loss, network architecture details, additional discussion, and more result visualizations.

1. Ablation Study for cGAN Loss

The proposed cGAN refinement module not only significantly improves the rendered image quality, but also provides the flexibility for users to select desired level of image details. In this experiment, we applied different weights ω_{cGAN} to $\mathcal{L}_{\text{cGAN}}$ in Eq. 7 in the main paper to observe the effect of cGAN loss to image quality. The quantitative results and visualizations are shown in Fig. 1 and 2. The results reveal interesting trade-off between PSNR and LPIPS: when increasing ω_{cGAN}, the PSNR decreases but LPIPS improves (decreases). This leads to several open research questions: what is the best image metric to use and how to find the optimal balance among multiple image metrics? How to design the training losses to obtain the most desirable balance of image metrics? The answer could be application-oriented and is worth more future research.

2. Additional Details for Applications

Object detection simulator. We visualized more side-by-side comparisons of the Detectron2 masks [22] from our rendered images and the ground truth in Fig. 3. We observed that besides the high Car IoU (0.74 for log 0a13, 0.75 for log 4d7b), the object detection masks are visually similar when generated from our images and the ground truth images.

Data augmentation. We compared three cases in this experiment: 1) 52 real images per sequence, 2) 52 real images + 96 synthetic images, 3) 52 real images + 160 synthetic images. For each case we trained MapNet [1] until the validation loss stopped improving for 8000 epochs. The detailed quantatitive results are shown in Tab. 1 and the corresponding mean validation loss curve is shown in Fig. 5. The validation loss in Fig. 5 is a combination of translation and rotation losses, as described in [1]. The results show that training with more synthetic images improves the results with less pose errors and validation loss. One exception we observed is that the log 2b04 got the best result in case 2 instead of case 3 unlike other sequences, which is caused by the training process of case 3 stuck in a local minimum. Overall, the results show significant benefit of this realistic and geometry-based data augmentation for pose regression.

Changing seasons. In the changing season experiment (Sec. 5 of the main paper), we used the 5Hz imagery from one front-facing camera in the NCLT dataset, and selected one in every three frames for validation. We extracted 3 sequences and the extracted sequences contain 137-167 training images and 67-78 validation images. The LiDAR scans were collected with a Velodyne HDL-32E LiDAR. The number of points in the LiDAR maps range from 7.2×10^5 to 1.4×10^6, and the camera trajectory lengths span from $17.8m$ to $23.1m$ (Tab. 3). For each training image, we extracted the LiDAR scan with closets timestamp and used the extracted LiDAR scans to build the LiDAR map (Fig. 8). Visualizations of the results from the three collected sequences are shown in Fig. 4.

3. Additional Dataset Details

Additional statistics of the collect Argoverse 2 sequences, including the number of train/val images, trajectory length, and number of LiDAR points in the map, are shown in Tab. 2. Visualizations of the LiDAR maps of both Argoverse 2 and NCLT datasets are shown in Fig. 7 and Fig. 8. One can observe that the noise level is lower in Argoverse 2 maps (Fig. 7) than NCLT dataset (Fig. 8). This was caused by more accurate LiDAR pose estimation and has motivated us for using the tight sampling strategy in the Argoverse 2 dataset (Sec. 3.1 in the main paper). On the other hand, the semantic labels in Argoverse 2 were used to remove dynamic objects (Fig. 6).

The LiDAR maps from NCLT consist of LiDAR points collected from different seasons. One can also observe the seasonal foliage shape change. For example, the green points in Fig. 8 (c) were collected in August and spread wider than the points collected in other seasons, reflecting the fact that the shape of seasonal foliage is larger in summer.

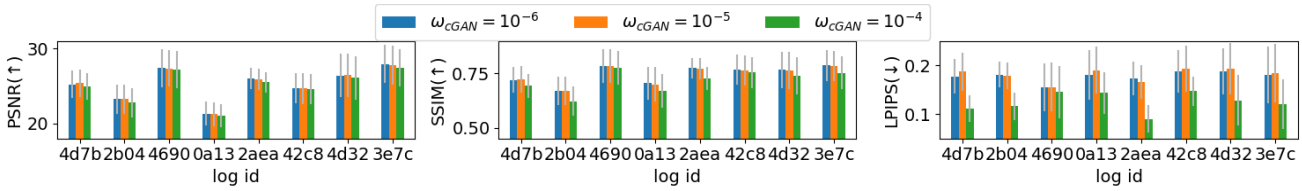


Figure 1: Ablation study for cGAN loss strength. We observed a trade-off between PSNR and LPIPS. Stronger cGAN loss adds more details to the output image, potentially making the image more perceptually pleasant with respect to LPIPS, but not necessarily faithful with respect to PSNR and SSIM.

Table 1: MapNet [1] results with different levels of data augmentation, where the median and mean for translation and rotation pose prediction errors are represented as $t_{med}(m)$, $t_{mean}(m)$, and $r_{med}(^\circ)$, $r_{mean}(^\circ)$. The cases with augmented images significantly outperformed the case trained with only real images.

	52 real				52 real + 96 syn.				52 real + 160 syn.			
log id	t_{med}	t_{mean}	r_{med}	r_{mean}	t_{med}	t_{mean}	r_{med}	r_{mean}	t_{med}	t_{mean}	r_{med}	r_{mean}
0a13	0.03	0.05	1.01	1.2	**0.02**	0.03	**0.17**	0.37	0.03	0.04	0.2	**0.2**
2aea	0.02	0.04	0.42	0.63	0.02	0.03	0.34	0.35	0.02	0.03	**0.15**	**0.2**
2b04	0.04	0.05	0.24	0.28	**0.01**	**0.03**	**0.21**	**0.21**	0.02	0.04	0.44	0.42
3e7c	0.04	0.07	1.82	2.2	0.04	0.04	**0.12**	0.24	**0.03**	**0.03**	0.18	**0.21**
4d7b	0.02	0.03	0.23	0.24	0.02	0.03	0.29	0.44	0.02	**0.02**	**0.16**	**0.21**
4d32	0.04	0.07	0.45	0.44	0.05	0.07	0.29	0.4	0.04	**0.06**	**0.2**	**0.22**
42c8	0.03	0.04	0.35	0.44	0.03	0.05	0.31	0.33	0.03	0.04	**0.24**	**0.29**
4690	**0.03**	0.04	0.36	0.34	0.04	0.04	0.3	0.29	0.04	0.04	**0.15**	**0.18**

Table 2: Dataset statistics for Argoverse 2 sequences

Argoverse 2 dataset				
log id	**# train**	**#val**	**traj.** (m)	**# LiDAR points**
0a13	155	22	22.9	502,567
2aea	155	22	26.3	517,136
2b04	155	22	6.5	338,572
3e7c	155	22	31.0	547,837
4d7b	155	22	19.9	298,121
4d32	155	22	28.4	415,469
42c8	155	22	31.6	580,788
4690	155	22	28.4	546,343

Table 3: Dataset statistics for NCLT sequences

NCLT dataset				
log id	**# train**	**#val**	**traj.** (m)	**# LiDAR points**
area 1	159	75	22.0	841,587
area 2	137	67	17.8	721,461
area 3	167	78	23.1	1,352,948

For the noisy LiDAR maps, we added LiDAR rain noise to individual scans according to [5] before building the map. We used $R = 8$ and $z_{max} = 200$ as the parameters. This noise model consists of two parts: 1) a threshold that removes faraway LiDAR measurements according to minimum detectable power and a LiDAR intensity decaying model in the rain. 2) a rain noise model for LiDAR range measurement. The resulting maps are shown in Fig. 9.

4. The Importance of LiDAR Supervision

Here we present an example to further demonstrate the importance of LiDAR depth loss, especially in the environment lacking photometric constraints. Given a camera frame t, we computed the photometric errors along the epipolar lines in its consecutive frames $t-1$ and $t+1$. In Fig. 10, we can observe the flat bottom in the photometric curves with respect to depth, showing there is no unique minimum to find the optimal depth in this condition.

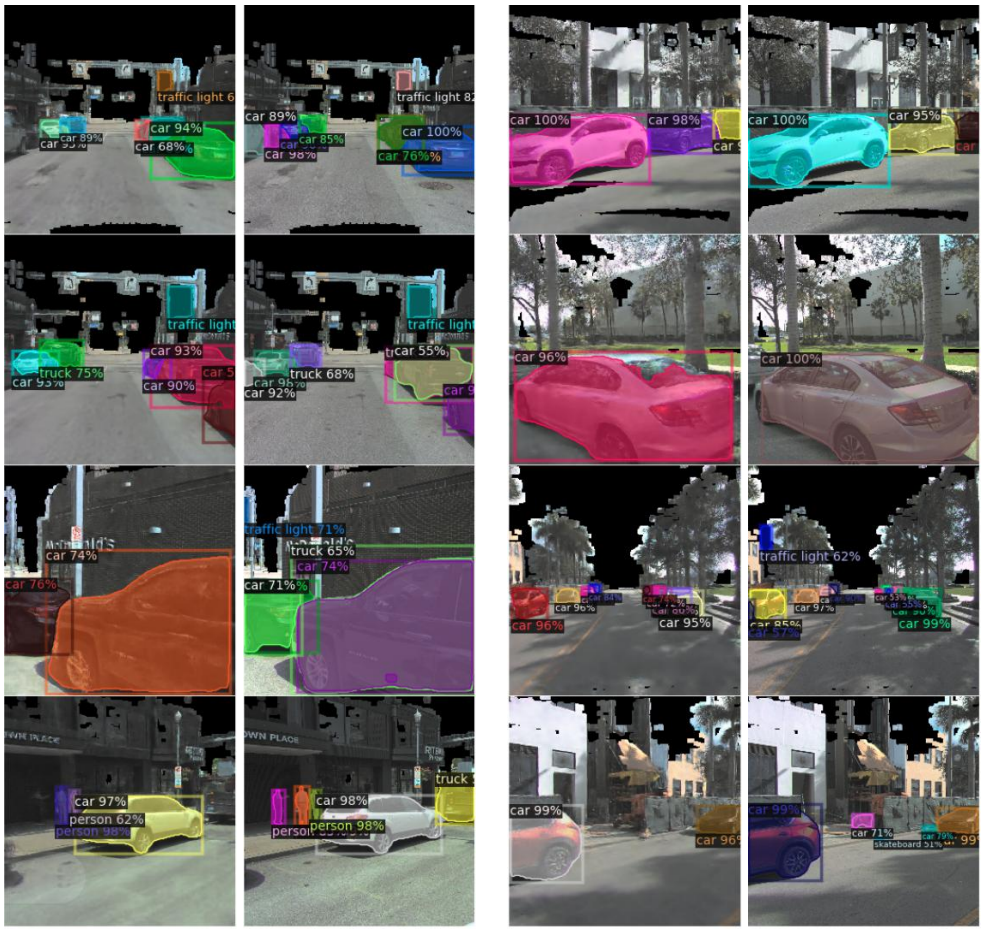
5. System Architecture Details

The architectures of our volume rendering networks and the refinement network $\mathcal{H}$ are shown in Fig. 2 in the main paper and Fig. 11. We used LeakyReLU in activation layers and did not use batchnorm layers.

In practice, real-world vision-LiDAR datasets do not always contain large number of training images and the network training can be prone to over-fitting. Empirically, we

(a) $\omega_{\text{cGAN}} = 10^{-6}$ (b) $\omega_{\text{cGAN}} = 10^{-5}$ (c) $\omega_{\text{cGAN}} = 10^{-4}$ (d) GT

Figure 2: The effect of cGAN loss strength. The results applying different cGAN loss strength are shown in (a)-(c). We can observe that the images with larger ω_{cGAN} contain more details but are not necessarily faithful to the ground truth.



(a) Left: ours/ right: GT (log 4d7b)

(b) Left: Ours/ right: GT (log 0a13)

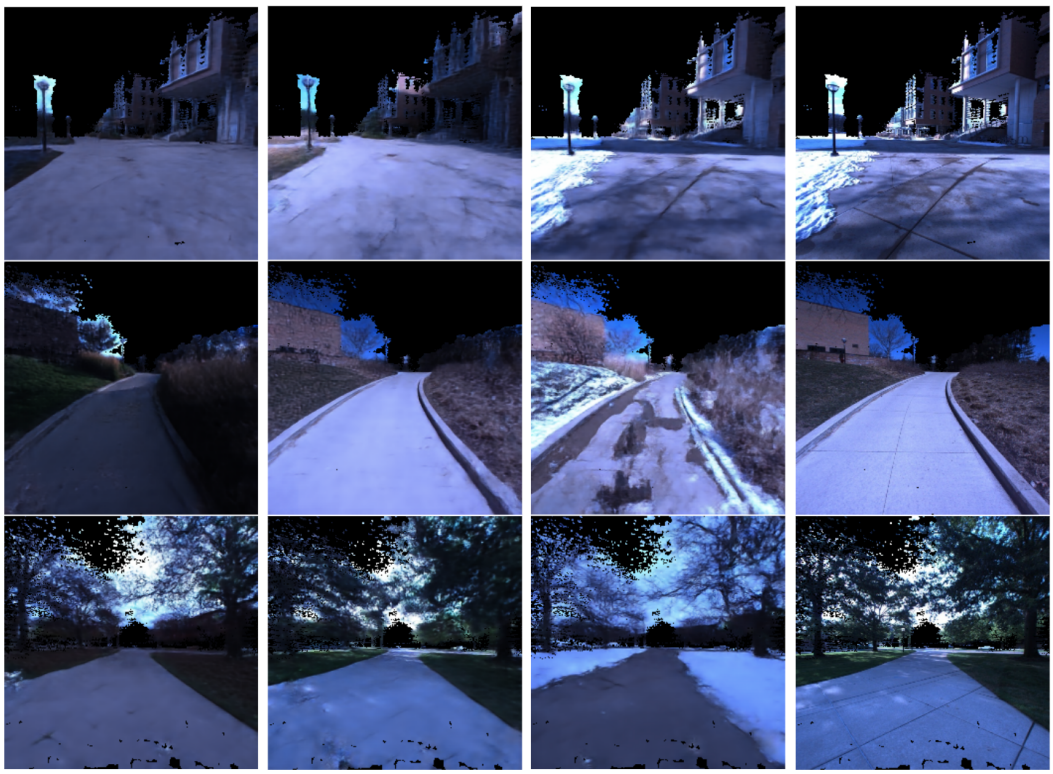
Figure 3: Side-by-side comparisons of the Detectron2 [22] results on our synthetic images and ground truth images from the validation set.

found the proposed ResNet-based encoder (6-layer autoencoder + 3-layer ResNet blocks) works well despite of its small size. A relevant observation of advocating small receptive fields for refinement CNNs is also mentioned by [6], indicating that small receptive fields are beneficial for generating view-consistent results.

6. Discussion

The advantage of using cGANs to generate realistic-looking images from real-world datasets has been well-proven in many previous works [6, 9, 10, 14, 19, 24]. On the other hand, our proposed framework supports other types of image refinement modules and we would like to point out other potential alternatives as additional discussion. Recently, although not as well-explored as cGANs yet, diffusion models also showed impressive performance in image synthesis tasks [4, 16, 17, 20]. In contrast to the one-step image refinement CNN used in our work, the application of iterative conditional diffusion models to real-world image quality refinement would be an interesting direction to explore. It is worth mentioning that recently, InfiniCity [11] demonstrated a 2D-3D hybrid approach that generates synthetic voxel grids to perform city-scale voxel-based neural rendering. It would be interesting to see how this 2D-3D hybrid approach can interact with our real-world LiDAR measurements in the future. Another interesting direction is to consider a joint refinement of the predicted depth and the rgb with the multi-view photometric consistency losses used commonly in deep multiview reconstruction works [2, 8, 12]. The photometric losses can be applied to a sequence of augmented virtual video poses and potentially refine the output depth quality in a self-supervised way.

A potential alternative system design for LiDAR-assisted NeRF is to follow NPLF [13]. NPLF aggregates point features into ray features with self-attention mechanism instead



(a) Season 1 (b) Season 2 (c) Season 3 (d) GT

Figure 4: Visualization of changing season results for the corresponding LiDAR maps in Fig. 8. From top to bottom are the results from area 1, 2, and 3.

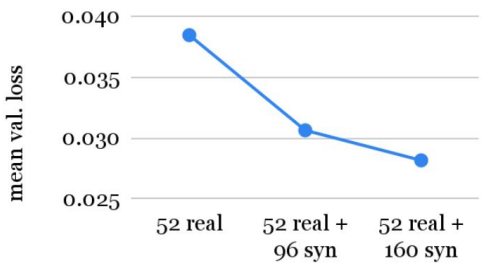
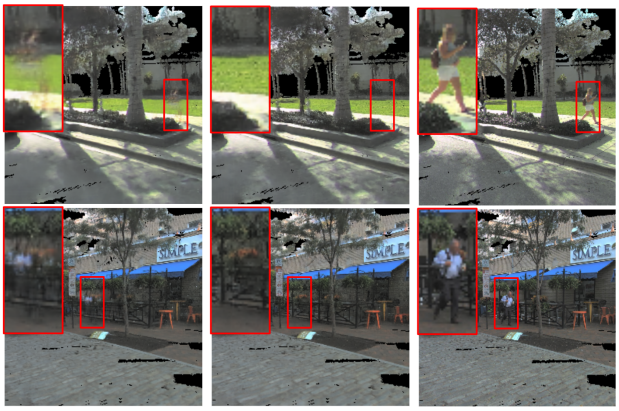


Figure 5: Mean validation loss for different data augmentation setups. Increasing the amount of augmented data significantly reduces the validation loss.

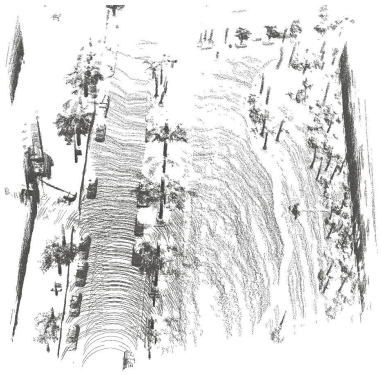


(a) W/o mask (b) W/ mask (c) GT

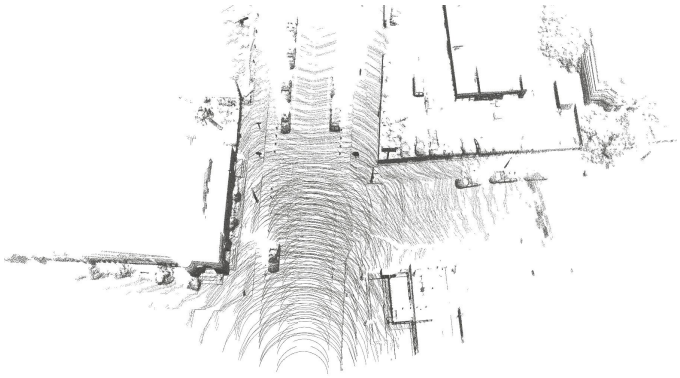
Figure 6: The results of w/o, w/ the moving object masks, and the ground truth. We observed the blurry shadows from moving objects in (a) and their removal in (b).

of explicit distance-based weights and volume rendering like ours and PointNeRF [23], and was trained without LiDAR depth loss. The attention mechanism provides more flexibility than the explicit method, and could potentially better overfit the training views. However, we expect the explicit method by PointNeRF to follow LiDAR geometry more faithfully.

Finally, we visualize some failure cases for future research reference (Fig. 12). Our system can produce unsatisfactory results when given inaccurate 3D labels and LiDAR depth. Also, thin objects are still challenging to render.

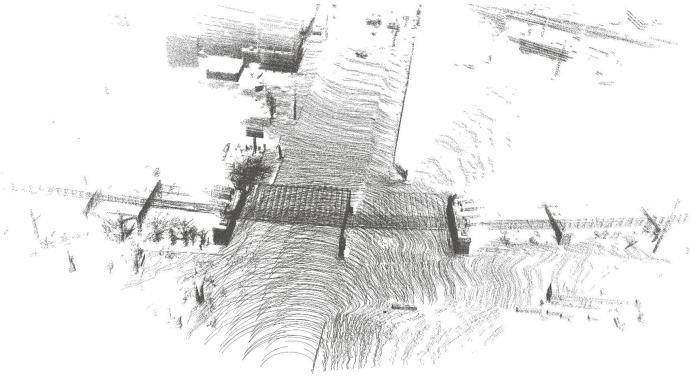


(a) 0a13

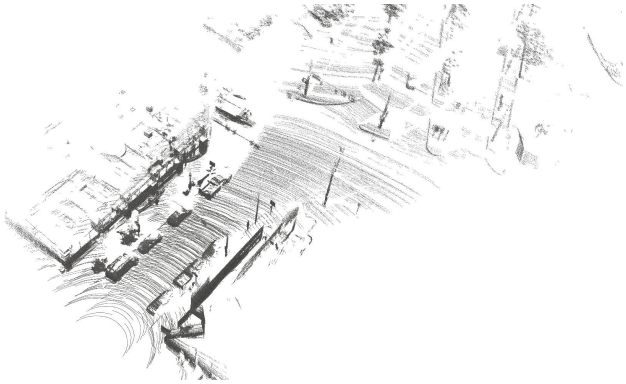


(b) 2aea

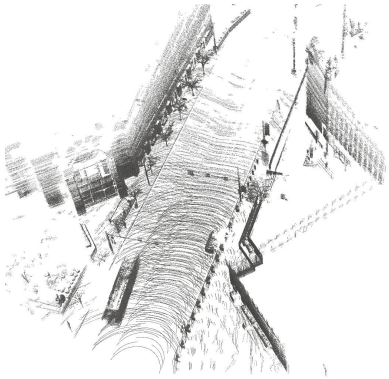
(c) 2b04



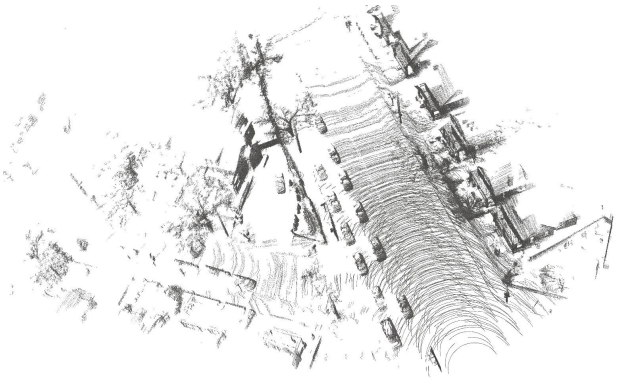
(d) 3e7c



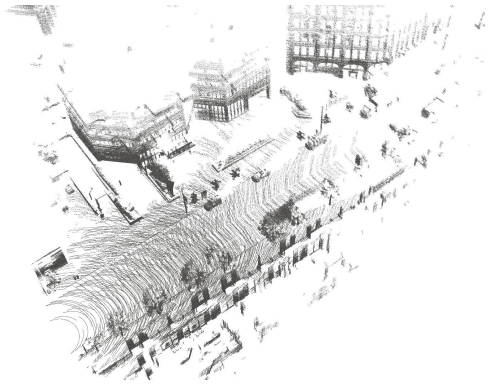
(e) 4d7b



(f) 4d32

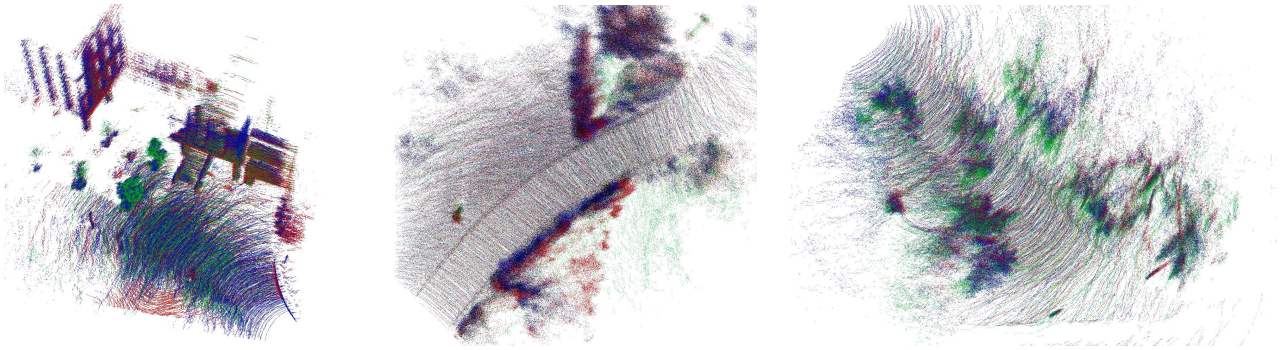


(g) 42c8



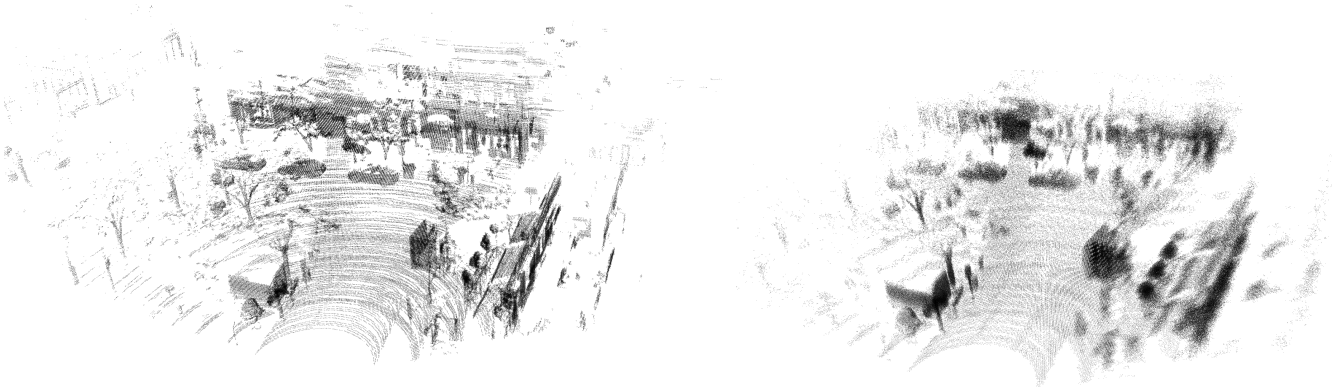
(h) 4690

Figure 7: Visualization of the collected LiDAR maps from Argoverse 2 dataset [21].



(a) Area 1 (a) Area 2 (c) Area 3

Figure 8: Visualization of the collected LiDAR maps from NCLT dataset [3]. LiDAR points collected in different seasons are shown by different colors.



(a) Clean (b) Noisy

Figure 9: The LiDAR maps before (a) and after (b) adding rain noise from [5]. The LiDAR measurements affected by the rain noise have shorter and noisier range.

7. More Results

In this section, we show more visualizations from our method and the baselines. The depth and RGB outputs from BlockNeRF with URF LiDAR depth loss, the point-based baseline, and our method are show in Fig. 13, 14, and 15. Additional visualizations for the noisy LiDAR map case are shown in Fig. 16. Resonating the findings of previous NeRF works, our study also showed that the positional encoding module is helpful (Tab. 4). The contribution of the proposed point-sampling strategy is shown quantitatively and qualitatively in Tab. 4 and Fig. 17.

References

[1] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-Aware Learning of Maps for Camera Localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2

[2] Ang Cao, Chris Rockwell, and Justin Johnson. FWD: Real-time Novel View Synthesis with Forward Warping and Depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 4

[3] Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. University of Michigan North Campus Long-Term Vision and Lidar Dataset. In *Int. J. of Robotics Res.*, 2016. 7

[4] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Conf. Neural Inform. Process. Syst.*, 2021. 4

[5] Christopher Goodin, Daniel Carruth, Matthew Doude, and Christopher Hudson. Predicting the Influence of Rain on LIDAR in ADAS. In *Electronics*, 2019. 2, 7

[6] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In *Int. Conf. Comput. Vis.*, 2021. 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 9

[8] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural Radiance Fields Approach to Deep Multi-View Photometric Stereo. In *IEEE Winter Conf. Applications of Comput. Vis.*, 2022. 4

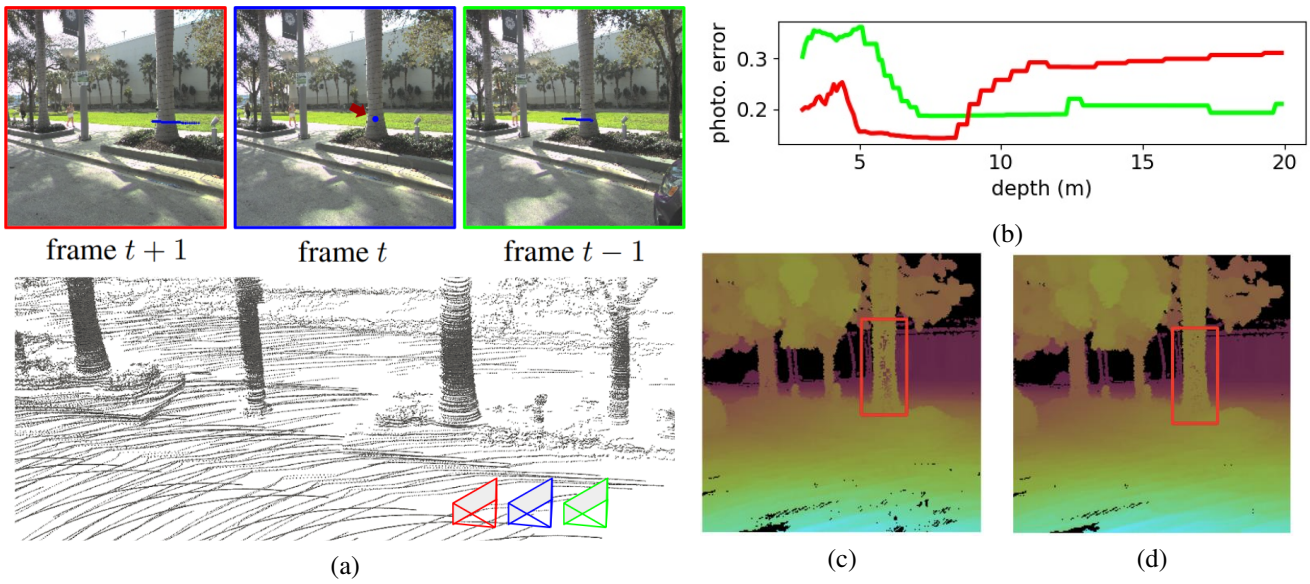


Figure 10: The effect of LiDAR depth loss. (a) The trunk region (blue dot on frame t) does not contain distinguished photometric information for determining depth. (b) The photometric error curves (within depth range $3-20m$) of the frame t blue dot in (a) on frame $t-1$ and frame $t+1$ are shown in green and red. The corresponding epipolar line segments are denoted by blue line segments on the frame $t-1$ and frame $t+1$ images in (a). Note that the photometric error curves in (b) contain large flat regions and no unique minimum for determining the optimal depth. The validation depths at this region w/o and w/ the LiDAR depth loss are shown in (c) and (d). We observed that the LiDAR depth provides guidance for correct depth on the trunk in (d).

Table 4: Ablation study. We show the contribution of each component quantitatively. The use of cGAN significantly improved results with all image metrics. The proposed tight point sampling strategy and the positional encoding module also helped. Note that the noisy and clean versions have different black regions and the numbers are not directly comparable.

map type	cGAN	tight sampling	$\gamma(.)$	ω_{cGAN}	PSNR (↑)	SSIM (↑)	LPIPS (↓)
clean		✓		$1e-5$	22.60	0.65	0.28
	✓	✓		$1e-5$	25.02	0.74	0.19
			✓	$1e-5$	23.02	0.67	0.25
	✓		✓	$1e-5$	25.19	0.74	0.18
		✓	✓	$1e-5$	23.14	0.67	0.25
	✓	✓	✓	$1e-5$	25.24	**0.75**	0.18
	✓	✓	✓	$1e-4$	24.94	0.72	**0.13**
	✓	✓	✓	$1e-6$	**25.28**	**0.75**	0.18
noisy		✓	✓	$1e-5$	23.79	0.70	0.24
	✓	✓	✓	$1e-5$	**25.93**	**0.76**	**0.17**

[9] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4

[10] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. BachGAN: High-Resolution Image Synthesis from Salient Object Layout. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[11] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. InfiniCity: Infinite-Scale City Synthesis. In *arXiv*, 2023. 4

[12] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF : Bundle-Adjusting Neural Radiance Fields. In *Int. Conf. Comput. Vis.*, 2021. 4

[13] Julian Ost, Issam Laradji, Alejandro Newell, Yuval Bahat, and Felix Heide. Neural Point Light Fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 4

Table 5: Quantitative number comparison for the Argoverse 2 sequences.

map type	method	metric	log ID							
			4d7b	2b04	4690	0a13	2aea	42c8	4d32	3e7c
clean	point-based [23]	PSNR	23.73	22.39	23.07	20.01	24.67	23.49	22.97	24.87
		SSIM	0.67	0.64	0.68	0.63	0.73	0.68	0.66	0.69
		LPIPS	0.25	0.23	0.24	0.28	0.21	0.26	0.28	0.27
	BlockNeRF [18]	PSNR	22.03	22.50	24.78	19.50	23.59	23.41	23.45	24.66
		SSIM	0.58	0.55	0.70	0.56	0.67	0.68	0.64	0.70
		LPIPS	0.44	0.44	0.26	0.39	0.34	0.38	0.37	0.35
	BlockNeRF+depth [15, 18]	PSNR	22.90	22.32	24.67	20.32	24.23	23.88	23.93	24.23
		SSIM	0.62	0.55	0.71	0.61	0.69	0.70	0.68	0.70
		LPIPS	0.44	0.48	0.31	0.41	0.36	0.39	0.39	0.39
	ours	PSNR	25.38	23.23	27.34	21.24	25.88	24.68	26.43	27.78
		SSIM	0.72	0.67	0.78	0.70	0.77	0.76	0.76	0.79
		LPIPS	0.19	0.18	0.16	0.19	0.17	0.19	0.19	0.18
noisy	point-based	PSNR	24.24	22.15	23.44	20.37	25.47	25.66	23.15	25.87
		SSIM	0.68	0.63	0.69	0.63	0.75	0.76	0.70	0.73
		LPIPS	0.24	0.25	0.23	0.30	0.19	0.21	0.26	0.24
	ours	PSNR	25.98	23.40	27.34	21.77	26.75	27.01	26.37	28.79
		SSIM	0.73	0.68	0.79	0.70	0.79	0.82	0.77	0.80
		LPIPS	0.17	0.18	0.16	0.20	0.16	0.17	0.19	0.16

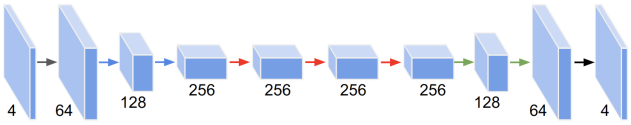
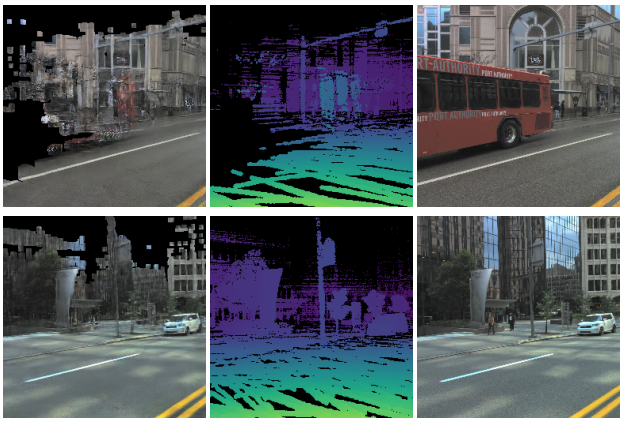


Figure 11: Our image refinement network $\mathcal{H}$. The downsampling 2D convolutional layers with stride 2 are represented by blue arrows, and the upsampling transposed 2D convolutional layers are represented by green arrows. The ResNet blocks [7] in the middle are represented by red arrows. Two 2D convolutional layers (black arrows) with stride 1 are used at input and output. LeakyReLU activations are appended to each layer except for the output layer.



(a) Ours (b) LiDAR depth (c) GT

Figure 12: Some failure cases. (Top) Incomplete dynamic object removal caused by inaccurate 3D labels. It leaves ghosts in the LiDAR map and affect our output. (Bottom) Thin objects with color similar to the background.

[14] Krishna Regmi and Ali Borji. Cross-View Image Synthesis Using Conditional GANs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4

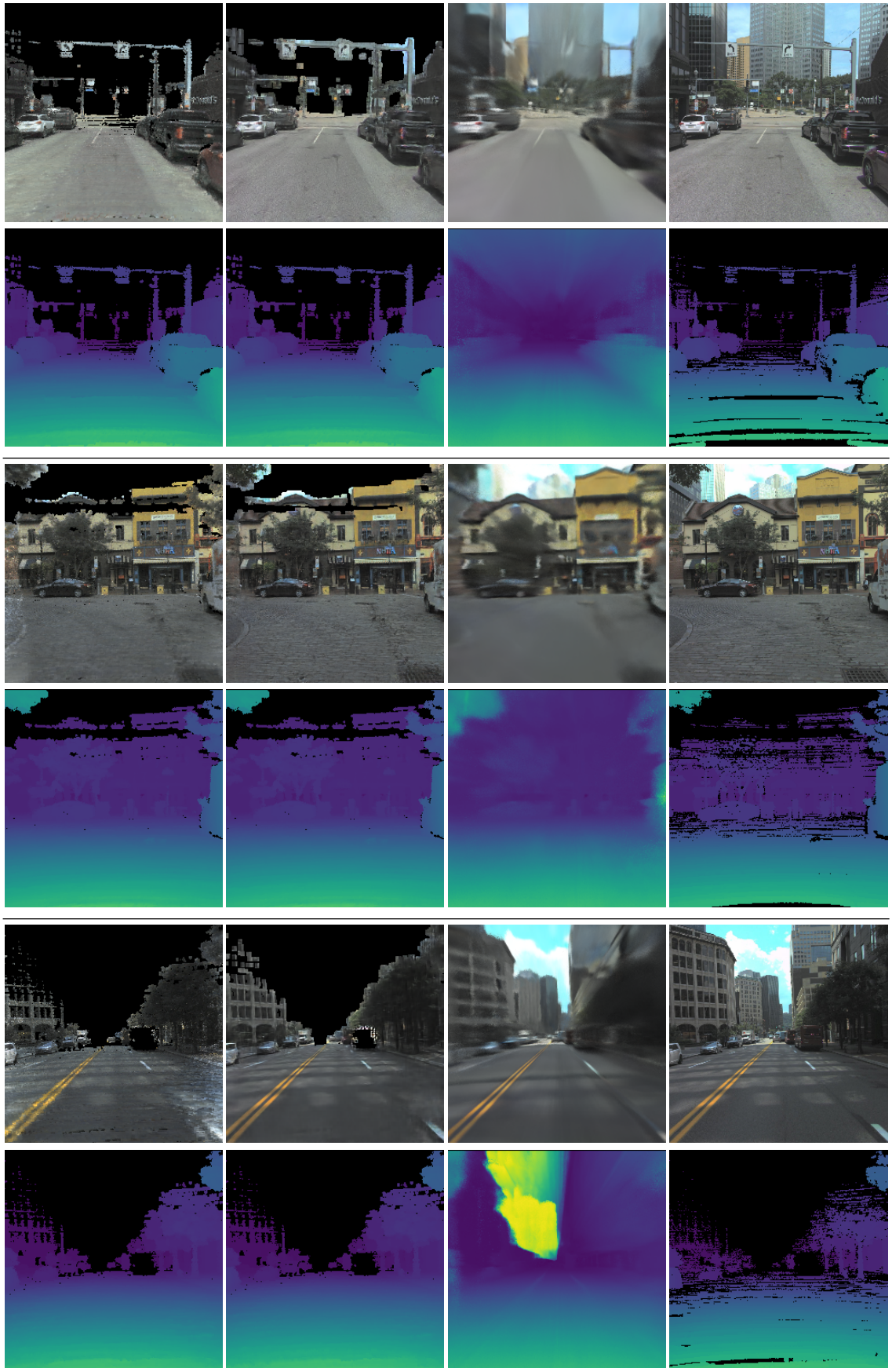
[15] Konstantinos Rematas, Andrew Liu, Pratul Srinivasan, Jonathan Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban Radiance Fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 9

[16] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *ACM Trans. Graph.*, 2022. 4

[17] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 4

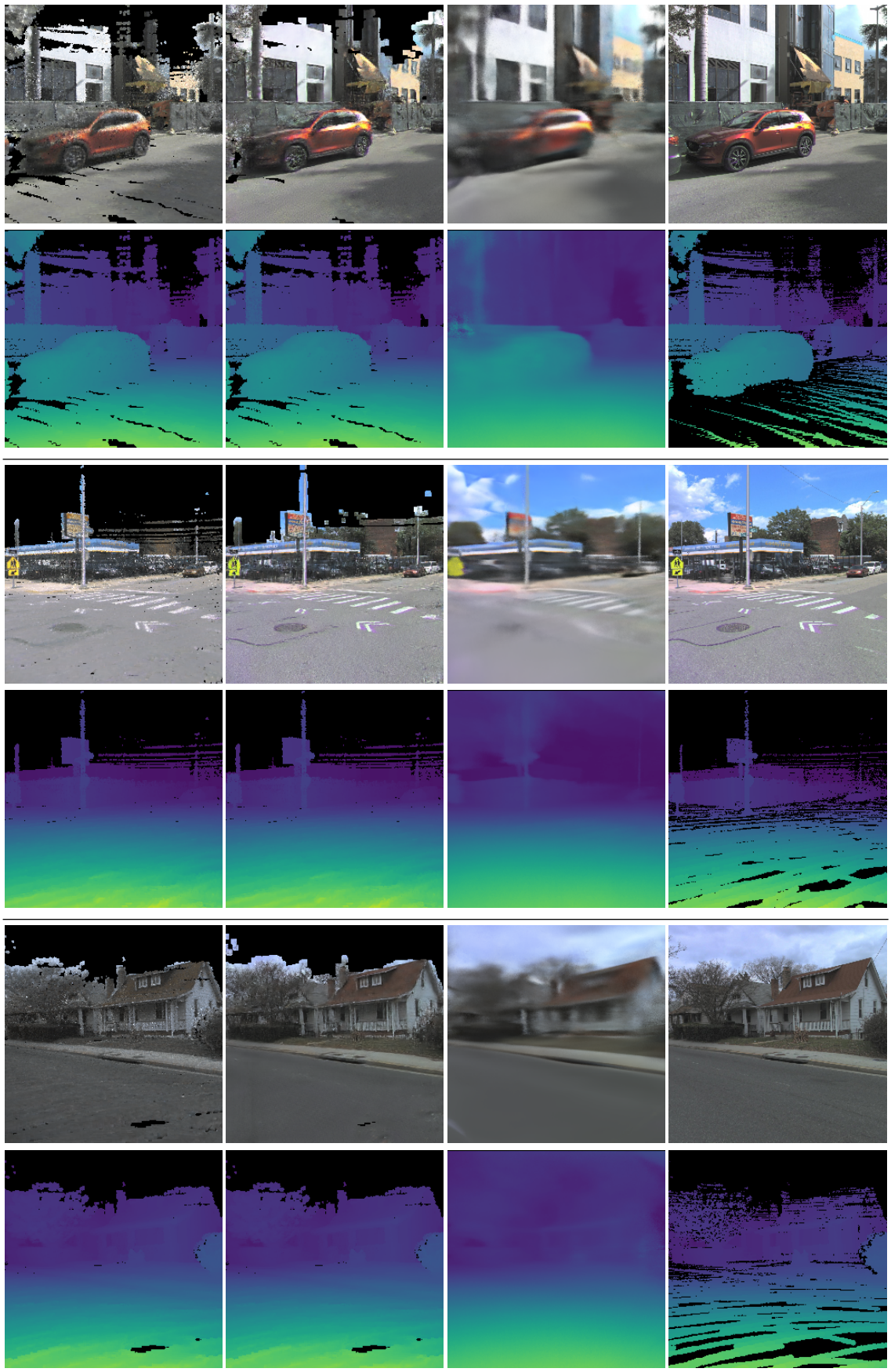
[18] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable Large Scene Neural View Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 9, 10, 11, 12

[19] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs.



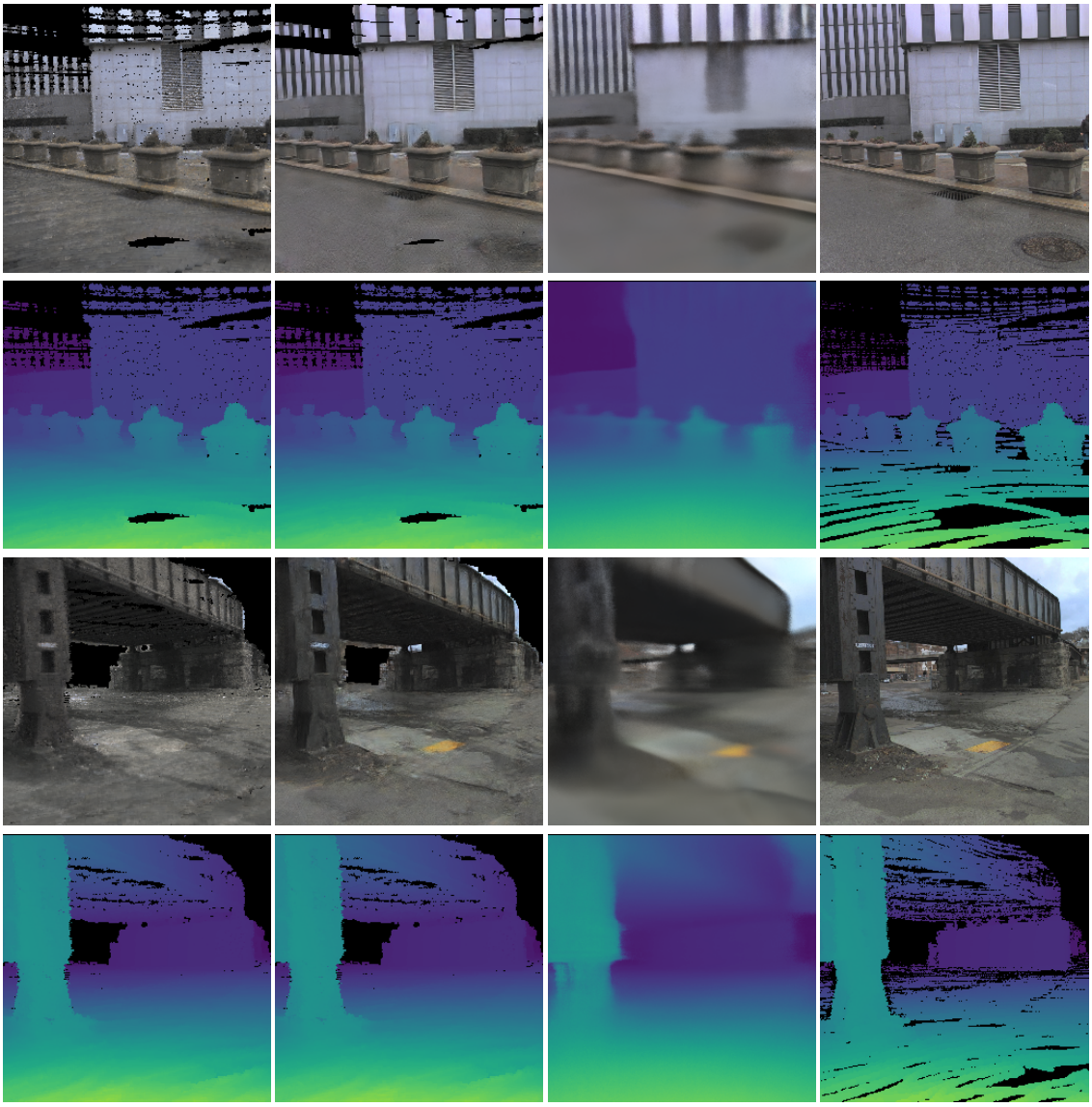
(a) Point-based (b) Ours (c) [18] + depth (d) GT

Figure 13: Additional visual comparison with baselines



(a) Point-based (b) Ours (c) [18] + depth (d) GT

Figure 14: Additional visual comparison with baselines



(a) Point-based (b) Ours (c) [18] + depth (d) GT

Figure 15: Additional visual comparison with baselines

In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4

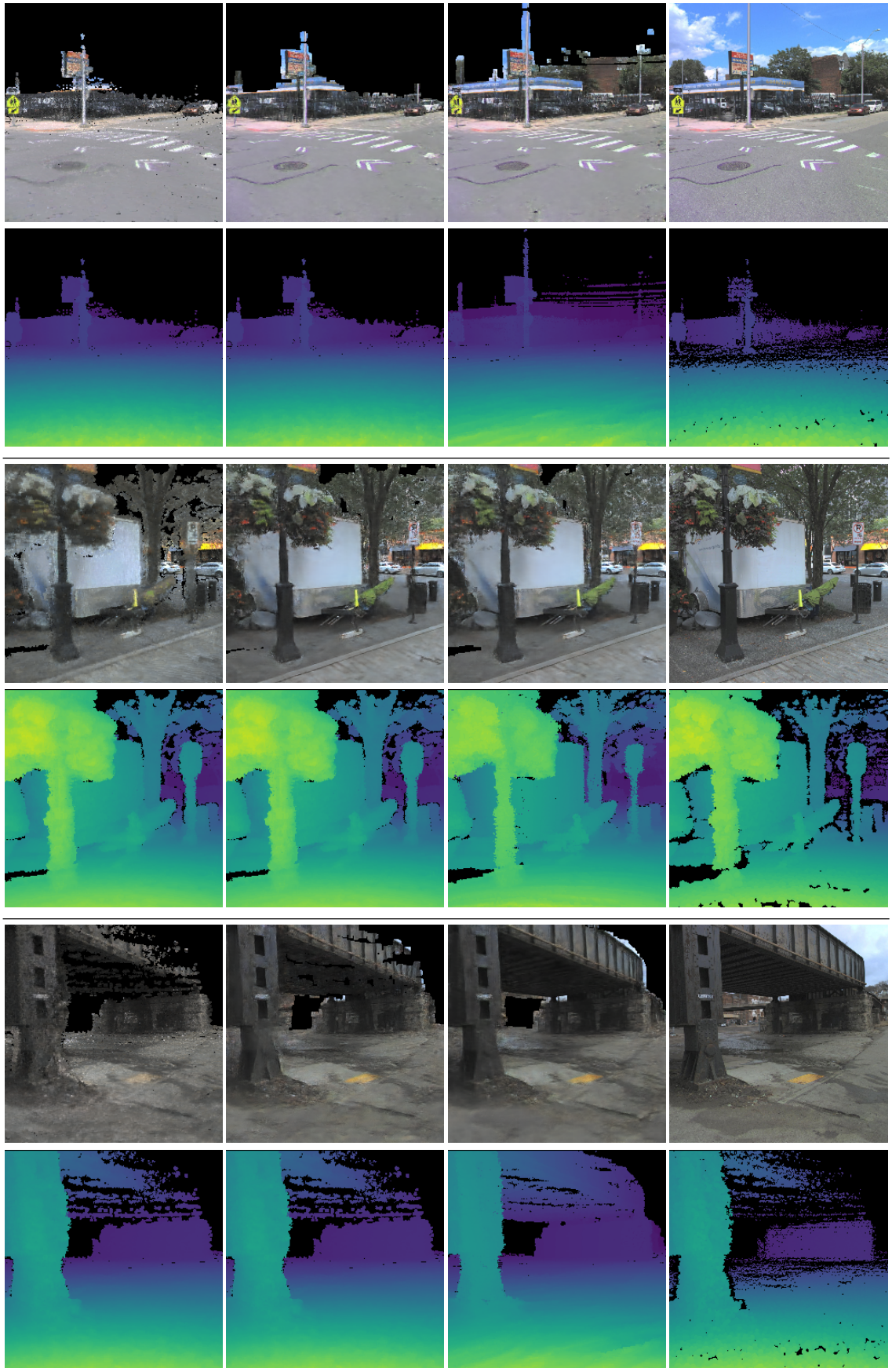
[20] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic Image Synthesis via Diffusion Models. In *arXiv*, 2022. 4

[21] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In *Conf. Neural Inform. Process. Syst.*, 2021. 6

[22] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1, 4

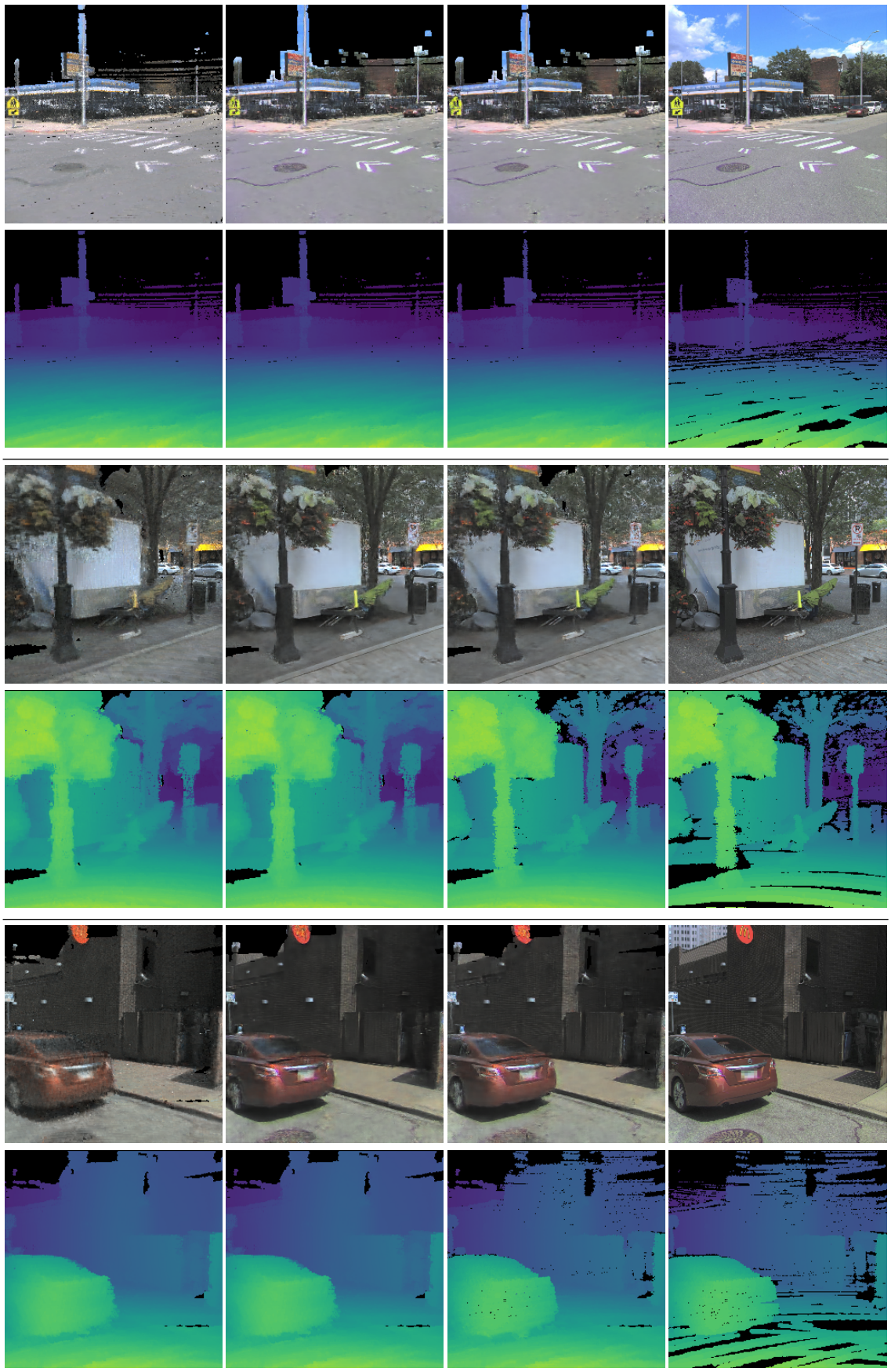
[23] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-NeRF: Point-based Neural Radiance Fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 5, 9

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Int. Conf. Comput. Vis.*, 2017. 4



(a) Point-based (noisy) (b) Ours (noisy) (c) Ours (clean) (d) GT (noisy)

Figure 16: Additional visual comparison with baselines on noisy LiDAR maps.



(a) Point-based (naive) (b) Naive (c) Ours (d) GT

Figure 17: Additional visual comparison for the point sampling strategy. The proposed point sampling strategy gives more accurate depth than the naive radius-based baseline.