Transfer Your Perspective: Controllable 3D Generation from Any Viewpoint in a Driving Scene

Supplementary Material

In this supplemental material, we provide more details and experiment results in addition to the main paper:

- Suppl. A: provides more related works, including 3D scene generation, diffusion models, domain adaptation, and NeRF.
- Suppl. B: summarizes existing datasets for CAV.
- Suppl. C: conducts experiments on generated point clouds with additional ground-truths.
- Suppl. D: shows the statistical results of the experiments.
- Suppl. E: further demonstrates the generation quality of two-stage training with experiments on CAV setting.
- Suppl. F: shows more qualitative results.
- Suppl. G: concludes future directions.

A. More Related Work

3D Scene Generation. As generative models have gained traction, recent research has increasingly focused on applying these methods to 3D point cloud data. Initially, the synthesis of point clouds was primarily limited to fixed-size data, such as single objects [1, 10, 34, 41]. However, recent advancements have extended beyond single-object generation to encompass entire 3D scenes. Early pioneering works in this domain employed generative adversarial networks (GANs) [15], demonstrating the feasibility of 3D scene generation, albeit with significant challenges in quality and realism.

More recent efforts have aimed at improving the quality and realism of 3D scene generation. For instance, LiDARGen [51] and UltraLiDAR [44] leverage diffusion models to enhance scene quality, incorporating realistic effects like ray drop. However, these methods struggle to generate scenes based on user-defined conditions, such as specific locations or diverse traffic scenarios.

To address these limitations, works like LidarDM [52] have introduced more controllable scene generation using consecutive video frames and user-defined conditions. Similarly, Text2LiDAR [43] employs text prompts for conditioning, enabling diverse scene generation tailored to user inputs.

Other advancements prioritize flexibility, efficiency, and quality. R2DM [27] proposes efficient training pipelines and a LiDAR completion framework that enhances scene quality. Meanwhile, RangeLDM [18] combines latent diffusion models with improved speed and quality for scene generation. LiDM [30] synthesizes recent advancements to achieve state-of-the-art results in realistic 3D scene generation, balancing quality, realism, and user control with multiple conditioning inputs.

In this paper, we propose a novel research problem in the context of CAV: generating realistic point clouds for reference agents. This direction offers significant potential for the research community, addressing the critical challenge of data collection in CAV, which is inherently difficult and costly. Unlike existing works focused on ego-centric scene generation, our approach shifts the perspective to collaborative scenarios. We view this as a complementary research direction and are open to enhancing our proposed solution by adopting advancements in efficiency and controllability from ongoing work in 3D scene generation.

Diffusion-based Generative Models. Diffusion models (DMs)[35] have made significant advancements in various domains, particularly in generating high-quality images. Initially, DMs were applied directly to raw pixel data, achieving remarkable results[8, 17, 21]. To improve efficiency, Latent Diffusion Models (LDMs) [32] operate in a compressed latent space, preserving visual quality while significantly reducing computational requirements. These approaches have found widespread application across diverse tasks, including 3D scene generation, as discussed in the previous section.

Controllable Diffusion Models. Many existing works focus on controlling generative processes through text prompts, particularly in text-to-image (T2I) synthesis [9, 12, 28, 29, 32, 33]. The predominant strategy involves performing denoising in feature space while integrating text conditions into the denoising process via a cross-attention mechanism. While these approaches achieve impressive synthesis quality, text prompts often lack reliable structural guidance for precise generation.

To address this limitation, several works improve structural control during generation. For instance, [3, 11, 16, 42] explore methods to enhance structure guidance in text-driven synthesis. Meanwhile, works like [22, 26, 49] introduce additional trainable modules built upon pre-trained T2I models to provide more targeted and controllable outputs.

In this paper, we leverage the approach proposed by [26] during the second stage of our framework. This stage grounds the generation process, ensuring that the outputs align with given semantic cues.

Domain Adaptation. Unsupervised domain adaptation (UDA) has been extensively studied. A common approach for domain adaptation is to learn domain-invariant embeddings by minimizing the distributional differences between source and target domains [24, 36, 38, 39]. More recently, adversarial training methods have gained popularity for bridg-

Datasets	Venue	Real?	Agent			# Cls	# Frames	Mod.
			dynamic	static	#			
OPV2V	ICRA'22	x	o	x	2-7	1	11.5k	C, L
V2X-Sim	RA-L'22	x	o	o	2-5	2	10k	L
V2XSet	ECCV'22	x	o	o	2-5	1	11.5k	C, L
DAIR-V2X	CVPR'22	o	x	o	2	10	39k	C, L
V2V4Real	CVPR'23	o	o	x	2	1	20k	L
MARS	CVPR'24	o	o	x	2	x	15k	C, L
TYP's motivation		semi-real	o	o	∞	1	-	L

Table S1. **Existing datasets for CAV.** Real-world datasets are limited by the challenges of data collection. Our proposed research problem aims to address this issue.

ing domain gaps effectively [2, 13, 14, 20, 23, 40]. These methods leverage a discriminator to distinguish between domains, encouraging the generator to produce features or outputs that are indistinguishable across domains.

In this paper, we adopt a discriminator inspired by adversarial training-based approaches to reduce the domain gap in embedded features between multi-agent and single-agent datasets. This step ensures that the domain-adapted embeddings provide robust guidance for generation training on single-agent datasets in the second stage of our proposed method.

Neural Radiance Fields. Neural Radiance Fields (NeRF) have significantly advanced 2D novel view synthesis (NVS) by encoding scenes as implicit volumetric functions optimized through ray-marching [25]. While effective in generating high-quality novel views, NeRF requires dense multi-view images and suffers from high computational costs [4, 5]. Extensions such as Mip-NeRF [4] improve aliasing, and depth-supervised variants reduce multi-view dependency [7, 31]. In 3D LiDAR-based NVS, NeRF-inspired methods like LiDAR-NeRF [37], Neural LiDAR Fields [19], and NeRF-LiDAR [48] adapt implicit representations to synthesize novel LiDAR views. These approaches enhance reconstruction but struggle with sparse data, large-scale outdoor scenes, and dynamic objects, as ray-marching is inefficient for LiDAR's discrete nature [37].

Our work shares similarities with NeRF-based LiDAR generation [19, 37, 48, 50] as we also synthesize LiDAR point clouds. However, unlike these methods focused on scene reconstruction from multiple samples (*e.g.*, views, time frames), TYP generates collaborative driving data from a single frame. Instead of modeling implicit densities, TYP directly generates LiDAR point clouds with spatial consistency even at a long distance, enabling single-agent datasets to be converted into multi-agent data for autonomous driving.

B. Existing Datasets for CAV

We summarize existing CAV datasets in Tab. S1, highlighting the current state of CAV research. At the time of this paper, no real-world dataset includes both dynamic and static agents and supports more than two agents, primarily due to the challenges of real-world data collection. Additionally, some datasets are limited to vehicle-only labels or a single sensor modality. These limitations drive our work, pushing boundaries and introducing a new research direction.

Figure S1. **Visualization with validation data of OPV2V.** The generated point clouds are well-aligned with the ground-truth bounding boxes and follow the physics (*e.g.*, occluded areas).

As shown in Fig. 1 and Fig. S2, TYP demonstrates strong potential to scale up the number of agents—both static and dynamic—through the proposed generation framework. Empirical results in Tab. 3 further validate that the generated point clouds can enhance CAV development.

C. Results on OPV2V

In Tab. 1 of the main paper, we validate the quality of the generated point clouds by replacing the ground-truth point clouds of the reference agents with the generated ones on the OPV2V dataset [46]. In this supplemental material, we investigate the impact of having access to a limited amount of labeled data.

Setting. As outlined in Sec. 4.3 of the main paper, the original training set of 44 scenes was split into two halves: the first 22 scenes were used to train the generation model, while the remaining 22 scenes were used for inference to generate point clouds of reference agents. Here, we further utilize the first split as a source of limited labeled data.

Results. Firstly, the results in Tab. S2 exhibit a consistent trend with Tab. 1 in the main paper, demonstrating that using generated point clouds achieves results comparable to those obtained with ground-truth point clouds (oracle). Secondly, the results in Tab. S2 highlight that incorporating additional limited labeled data further reduces the gap between using ground-truth and generated point clouds. For example, in Early Fusion with 22 additional labeled scenes, the performance with generated point clouds matches that of ground-truth point clouds (*i.e.*, both achieve 0.78).

D. Statistical Results of Experiments

In Tab. 1 of the main paper, we validate the quality of the generated point clouds by replacing the ground-truth point clouds of the reference agents with the generated ones on the OPV2V dataset [46]. In this supplemental material, we extend this evaluation by conducting two additional runs (*i.e.*, three in total) and present the statistical results in Tab. S3. The results consistently demonstrate that using the generated point clouds achieves performance comparable to that

Method	Train Data	0 Add. Scene				5 Add. Scene				10 Add. Scene				22 Add. Scene			
		s	m	l	all	s	m	l	all	s	m	l	all	s	m	l	all
No Fusion	ego's gt only	0.67	0.41	0.13	0.40	0.85	0.66	0.22	0.57	0.82	0.60	0.20	0.53	0.87	0.71	0.25	0.60
Early Fushion [6]	baseline					0.01	0.00	0.00	0.01	0.38	0.06	0.04	0.16	0.87	0.59	0.41	0.61
	+gt (**oracle**)	0.76	0.42	0.31	0.49	0.89	0.65	0.48	0.66	0.88	0.63	0.48	0.65	0.96	0.82	0.62	0.78
	+TYP (**ours**)	0.75	0.38	0.29	0.46	0.87	0.63	0.46	0.65	0.89	0.66	0.50	0.67	0.96	0.80	0.63	0.78
Late Fushion [46]	baseline					0.15	0.10	0.03	0.09	0.54	0.33	0.19	0.35	0.91	0.77	0.50	0.71
	+gt (**oracle**)	0.74	0.57	0.36	0.55	0.93	0.82	0.52	0.74	0.89	0.75	0.49	0.70	0.95	0.85	0.56	0.77
	+TYP (**ours**)	0.71	0.49	0.32	0.50	0.90	0.75	0.47	0.69	0.84	0.68	0.43	0.63	0.95	0.84	0.53	0.75
AttFuse [46]	baseline					0.66	0.31	0.28	0.41	0.86	0.60	0.48	0.64	0.94	0.76	0.56	0.74
	+gt (**oracle**)	0.94	0.80	0.62	0.77	0.96	0.84	0.69	0.82	0.96	0.82	0.67	0.79	0.98	0.87	0.69	0.82
	+TYP (**ours**)	0.90	0.73	0.56	0.72	0.95	0.81	0.65	0.79	0.94	0.79	0.62	0.77	0.98	0.88	0.75	0.86
V2X-ViT [45]	baseline					0.66	0.44	0.28	0.48	0.83	0.53	0.37	0.60	0.91	0.68	0.46	0.70
	+gt (**oracle**)	0.87	0.71	0.50	0.71	0.88	0.73	0.57	0.74	0.91	0.77	0.63	0.78	0.94	0.80	0.66	0.81
	+TYP (**ours**)	0.84	0.65	0.40	0.65	0.88	0.72	0.50	0.71	0.90	0.74	0.55	0.74	0.94	0.79	0.59	0.78

Table S2. **Results on OPV2V with limited labeled data.** Using generated point clouds consistently achieves results comparable to oracles, demonstrating the quality of the generation. With additional labeled scenes, the gap is further minimized.

Method	Train Data	Average ($\pm$ *std.*)			
		s	m	l	all
Early Fusion [6]	gt (**oracle**)	0.78 ($\pm$ *0.02*)	0.44 ($\pm$ *0.04*)	0.35 ($\pm$ *0.04*)	0.52 ($\pm$ *0.02*)
	TYP (**single**)	0.63 ($\pm$ *0.13*)	0.34 ($\pm$ *0.15*)	0.24 ($\pm$ *0.09*)	0.40 ($\pm$ *0.11*)
	TYP (**ours**)	0.75 ($\pm$ *0.04*)	0.40 ($\pm$ *0.09*)	0.30 ($\pm$ *0.08*)	0.47 ($\pm$ *0.06*)
Late Fusion [46]	gt (**oracle**)	0.77 ($\pm$ *0.07*)	0.61 ($\pm$ *0.10*)	0.40 ($\pm$ *0.06*)	0.58 ($\pm$ *0.08*)
	TYP (**single**)	0.75 ($\pm$ *0.03*)	0.55 ($\pm$ *0.05*)	0.35 ($\pm$ *0.04*)	0.55 ($\pm$ *0.04*)
	TYP (**ours**)	0.79 ($\pm$ *0.07*)	0.60 ($\pm$ *0.10*)	0.37 ($\pm$ *0.05*)	0.58 ($\pm$ *0.07*)
AttFuse [46]	gt (**oracle**)	0.93 ($\pm$ *0.01*)	0.78 ($\pm$ *0.02*)	0.62 ($\pm$ *0.02*)	0.77 ($\pm$ *0.02*)
	TYP (**single**)	0.90 ($\pm$ *0.02*)	0.70 ($\pm$ *0.03*)	0.54 ($\pm$ *0.02*)	0.70 ($\pm$ *0.02*)
	TYP (**ours**)	0.91 ($\pm$ *0.00*)	0.72 ($\pm$ *0.02*)	0.56 ($\pm$ *0.01*)	0.72 ($\pm$ *0.00*)

Table S3. **Statistical Results on OPV2V.** We report the mean and standard deviation from multiple runs of the same experiment, demonstrating the consistency of the results. Additionally, we include the performance of point clouds generated using single-stage training, which is consistently worse than the two-stage approach, highlighting the generation quality of two-stage training.

of the ground-truth (oracle) point clouds, highlighting the robustness, consistency, and reproducibility of our approach.

E. Single-Stage *vs*. Multi-Stage Training

In Tab. 5 of the main paper, we compare the quality of generated point clouds between single-stage and the proposed multi-stage training by evaluating the distance between generated and ground-truth samples. In this supplemental material, we extend this analysis by conducting CAV training using point clouds generated by the single-stage training model.

The results in Tab. S3 demonstrate that the performance of single-stage training consistently lags behind the proposed multi-stage approach, particularly in scenarios that directly rely on point clouds (*i.e*., early fusion). Furthermore, the multi-stage method remains essential for translating single-agent datasets into collaborative versions, underscoring its critical role in the proposed framework (*cf*. Secs. 3.4 and 4.5 in the main paper).

F. More Qualitative Results

We present additional examples of TYP in Fig. S2. These examples demonstrate the ability to designate *any location* as a reference, effectively simulating both static and dynamic agents communicating with the ego vehicle. This flexibility overcomes the limitations of existing real-world CAV datasets, which are often constrained by specific communication types (*i.e*., vehicle-to-vehicle or vehicle-to-infrastructure) and a limited number of agents. Furthermore, the generated point clouds are both realistic and semantically consistent with the ego agent's perception.

In Fig. S3, we provide more examples of the collaborative version of the Waymo dataset (*i.e*., ColWaymo), which was utilized to pre-train the detector for CAV tasks, as discussed in Sec. 4.5 and Tab. 3 of the main paper. These examples further highlight the high quality of the point clouds generated

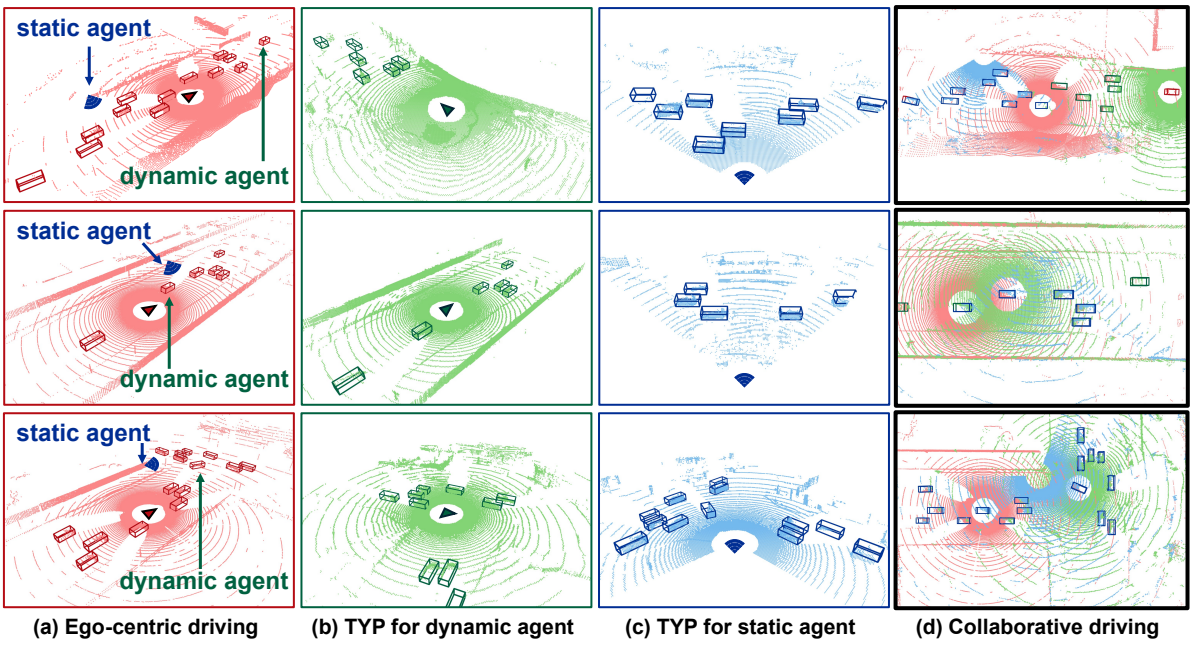

Figure S2. **Illustration of the proposed problem and solution, Transfer Your Perspective (TYP).** (a) A given sensory data captured by the ego-car (red triangle). (b) A generated sensory data by TYP, seeing from the viewpoint of another vehicle (green triangle) in the same scene. (c) A generated sensory data, seeing from an imaginary static agent like roadside units (blue icon). (d) Putting all the sensory data together, given or generated, TYP enables the development of collaborative perception with little or no real collaborative driving data.

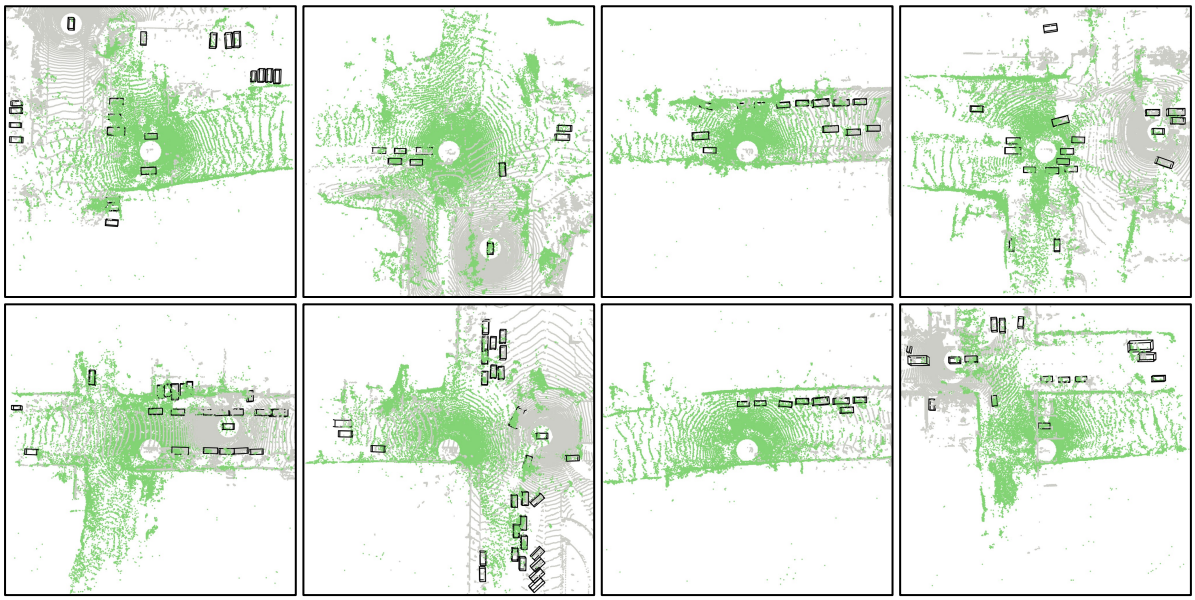

Figure S3. **Qualitative results on Collaborative Waymo.** The gray point clouds are from the original single-agent dataset and the green are generated by TYP conditioning on them.

by TYP and underscore its potential to significantly scale up datasets for CAV research.

G. Future Work

This paper follows the existing benchmark [46, 47] to focus on vehicle-like objects. However, TYP is scalable and can extend to broader object categories when semantic informa-

tion is available (*cf*. Sec. 3.1 and Sec. 3.2). We are also open to exploring cross-modality generation in future research.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*, 2018. 1

[2] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 2

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *CVPR*, 2021. 2

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[6] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019. 3

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 1

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 1

[11] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 1

[12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 1

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[18] Qianjiang Hu, Zhimin Zhang, and Wei Hu. Rangeldm: Fast realistic lidar point cloud generation. In *ECCV*, 2025. 1

[19] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. In *ICCV*, 2023. 2

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 1

[22] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2025. 1

[23] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 2

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 1

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*. ACM New York, NY, USA, 2021. 2

[26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 1

[27] Kazuto Nakashima and Ryo Kurazume. Lidar data synthesis with denoising diffusion probabilistic models. In *ICRA*, 2024. 1

[28] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 1

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1

[30] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *CVPR*, 2024. 1

[31] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 2022. 2

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

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael

Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[34] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *ICCV*, 2019. 1

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016. 1

[37] Tang Tao, Longfei Gao, Guangrun Wang, Yixing Lao, Peng Chen, Hengshuang Zhao, Dayang Hao, Xiaodan Liang, Mathieu Salzmann, and Kaicheng Yu. Lidar-nerf: Novel lidar view synthesis via neural radiance fields. In *ACM MM*, 2024. 2

[38] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1

[39] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 1

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[41] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning localized generative models for 3d point clouds via graph convolution. In *ICLR*, 2018. 1

[42] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1

[43] Yang Wu, Kaihua Zhang, Jianjun Qian, Jin Xie, and Jian Yang. Text2lidar: Text-guided lidar point cloud generation via equirectangular transformer. In *ECCV*. Springer, 2024. 1

[44] Yuwen Xiong, Wei-Chiu Ma, Jingkang Wang, and Raquel Urtasun. Ultralidar: Learning compact representations for lidar completion and generation. In *CVPR*, 2023. 1

[45] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *ECCV*, 2022. 3

[46] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *ICRA*, 2022. 2, 3, 4

[47] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *CVPR*, 2023. 4

[48] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. In *AAAI*, 2024. 2

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 1

[50] Zehan Zheng, Fan Lu, Weiyi Xue, Guang Chen, and Changjun Jiang. Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis. In *CVPR*, 2024. 2

[51] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *ECCV*, 2022. 1

[52] Vlas Zyrianov, Henry Che, Zhijian Liu, and Shenlong Wang. Lidardm: Generative lidar simulation in a generated world. *arXiv preprint arXiv:2404.02903*, 2024. 1