

# LiDAR-to-4D Radar Synthesis for Building Large-Scale Tensor Datasets

## Supplementary Material

Section A provides a detailed explanation, along with experiments, of the reason for choosing the cGAN-based [14] L2RDaS Generator. Section B compares and analyzes the datasets used in this study, confirming that the L2RDaS Generator can be applied regardless of the data collection environment or sensor hardware. Section C presents additional qualitative results to visually evaluate the synthesis quality. Section D addresses the absence of the Doppler dimension and explores its synthesis feasibility. Section E further validates the framework’s generalization, 3D tensor effectiveness, and synthesis fidelity. Section F discusses the applicability and potential scalability of L2RDaS based on the preceding analyses.

### A. Diffusion-Based L2RDaS Experiments

The reason for choosing a cGAN-based L2RDaS Generator instead of a diffusion-based model in this study is that, under limited training data (about 6000 frames available for training), the cGAN method may be more suitable for learning high-dimensional data such as the 3D tensors  $\mathbf{T}_{C-RAE}$  or higher-dimensional forms. Diffusion models estimate the probability density gradients (score function) defined over the entire high-dimensional space [10]. However, real data are known to concentrate on a low-dimensional manifold that occupies only a small portion of the high-dimensional space [21]. Therefore, learning the score over the wide regions outside this manifold requires massive amounts of data and computation due to the curse of dimensionality [24].

In contrast, cGANs operate as implicit density models that do not explicitly estimate probability densities, but instead learn a direct mapping function that projects from the low-dimensional latent space to the data manifold [9]. This method reduces the burden of exploring unnecessary regions in the ambient space, allowing the model to focus only on the support region where real data exist, thereby enabling more efficient learning [1]. Therefore, for the  $\mathbf{T}_{C-RAE}$  synthesis task in this study, the cGAN method appears to be a more practical choice, particularly given the limited amount of training data.

We conducted experiments by synthesizing the 3D tensors  $\mathbf{T}_{C-RAE}$  and comparing the results between the two methods. We first designed the diffusion-based [10] L2RDaS model as described below. In this section, we denote  $\mathbf{T}_{C-RAE}$  as  $\mathbf{Y}$ .

Since radar tensor power exhibits a wide dynamic range, we convert  $\mathbf{Y}$  into a log-domain latent variable  $\mathbf{X}_0$  for nu-

merical stability and range control:

$$\mathbf{X}_0 = \log(\max(\mathbf{Y}, \varepsilon) + \varepsilon), \quad (1)$$

where  $\varepsilon = 10^{-5}$  is a small constant.

The forward process of the conditional diffusion model is defined as:

$$\mathbf{X}_t = \sqrt{\bar{\alpha}_t} \mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

where  $\bar{\alpha}_t$  is the cumulative product of  $(1 - \beta_t)$ , and  $\beta_t$  is determined by the cosine-based noise schedule [15].  $\mathbf{X}_t$  denotes the latent tensor at timestep  $t$  with mixed noise, and  $\boldsymbol{\epsilon}$  represents pure Gaussian noise. The forward process is designed to define a data distribution during training.

At each timestep  $t$ , we construct an input feature  $\mathbf{C}_t$  using the current state  $\mathbf{X}_t$ , the conditional tensor  $\mathbf{V}_{\text{LiDAR}}$ , and the time embedding  $\gamma(t)$ . The constructed  $\mathbf{C}_t$  is fed into the L2RDaS Generator to predict the v-parameterization [19] output  $\mathbf{v}_\theta(\mathbf{C}_t)$ . The target tensor is defined as:

$$\mathbf{v}^* = \sqrt{\bar{\alpha}_t} \boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t} \mathbf{X}_0. \quad (3)$$

The model minimizes a mean squared error (MSE) loss to enforce  $\mathbf{v}_\theta(\mathbf{C}_t) \approx \mathbf{v}^*$  and additionally applies an L1 loss on  $\hat{\mathbf{Y}} = \exp(\hat{\mathbf{X}}_0)$ , which is obtained by converting the reconstructed  $\hat{\mathbf{X}}_0$  back to the data domain.

During inference, synthesis is performed using only  $\mathbf{V}_{\text{LiDAR}}$ , without real radar  $\mathbf{Y}$ . Starting from a Gaussian distribution,

$$\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (4)$$

the reverse process iterates from  $t = T \rightarrow 0$ , yielding  $\hat{\mathbf{X}}_0$ , and the final synthesized radar tensor  $\mathbf{T}_{C-RAE} = \hat{\mathbf{Y}}$  is obtained via exponential transformation.

Table 3. Performance comparison between cGAN-based and diffusion-based L2RDaS Generators

Method	Metrics	
	PSNR (dB) $\uparrow$	SSIM $\uparrow$
cGAN-based [14]	<b>31.006</b>	<b>0.897</b>
Diffusion-based (DDPM) [10, 15, 19]	18.264	0.782
Diffusion-based (XCube) [18]	30.826	0.868

As shown in Tab. 3, diffusion-based L2RDaS exhibited lower performance compared to the cGAN-based method, with decreases of 12.74 dB in PSNR and 0.115 in SSIM. These correspond to reductions of approximately 41% and

Table 4. Comparison of autonomous driving datasets used for training L2RDaS and for dataset expansion experiments. The L2RDaS Generator was trained on K-Radar, while KITTI, nuScenes, Dual Radar, and VoD were used for dataset expansion.

Dataset	Location	LiDAR Sensor	
		Model Name	Channels
K-Radar [16]	Daejeon / Gangwon, South Korea	Ouster os2-64	64
KITTI [8]	Karlsruhe, Germany	Velodyne HDL-64E	64
nuScenes [2]	Boston, USA / Singapore	Velodyne HDL-32E	32
Dual Radar [25]	China	RoboSense Ruby Lite	80
VoD [17]	Delft, Netherlands	Velodyne HDL-64E S3	64

12%, respectively, indicating that the diffusion model did not achieve sufficient quality in synthesizing the 3D tensors  $\mathbf{T}_{C-RAE}$ . Efficiency differences were also substantial: the diffusion-based model required approximately 9.06 seconds per synthesis, whereas the cGAN-based generator required only 0.09 seconds—more than 100× faster. These quantitative results demonstrate that, at present, the cGAN-based L2RDaS is more suitable for 3D tensor synthesis in terms of both synthesis quality and computational efficiency.

Furthermore, to mitigate the curse of dimensionality inherent in high-dimensional tensor synthesis, we conducted additional validation using XCube [18]. XCube is a diffusion-based model that integrates a Variational Autoencoder (VAE) [11] with a hierarchical structure. In our experiment, we configured the XCube model to synthesize 4D radar tensors by utilizing LiDAR sparse tensors as conditioning inputs.

As shown in Tab. 3, the quantitative evaluation reveals that the XCube-trained model yielded a PSNR 0.18 dB lower and an SSIM 0.029 lower than those of the cGAN-based model. Consequently, although the dimensionality reduction method using a VAE demonstrated some effectiveness for tensor synthesis, the overall performance still fell short of that of the cGAN. These results clearly substantiate the validity of our architectural choice under limited dataset conditions. According to XCube [18], the model was trained on large-scale datasets (e.g., Objaverse [6] 800K, ShapeNet [4] 57K). In contrast, the available K-Radar data in this study is limited to approximately 6k frames. Due to this data scarcity and the inherently noisy characteristics of radar tensors, diffusion models struggle to converge properly, confirming that the cGAN is a more effective generative model for this task.

Moreover, a disparity exists in computational efficiency. Synthesizing tensors using XCube requires approximately 13.57 seconds per frame. GT-Aug, one of the methods utilizing L2RDaS, augments data by synthesizing it online at every epoch; therefore, a shorter synthesis time is crucial for its effectiveness and practical applicability. A diffusion model requiring over 13 seconds per frame excessively in-

flates the total training time, rendering such online augmentation unfeasible. In contrast, the cGAN, with its rapid inference speed of 0.09 seconds, is perfectly suited for real-time online augmentation. Finally, the cGAN demonstrated superior GPU memory efficiency. It required only 9,492 MB during training and 1,543 MB during inference. Conversely, the baseline diffusion model (DDPM) [10, 15, 19] required 12,484 MB for training and 1,844 MB for inference, while XCube [18] demanded significantly more resources, recording 22,646 MB for training and 2,411 MB for inference.

## B. Comparison of Datasets Used in the Experiments

The dataset used to train L2RDaS Generator is K-Radar [16]. External datasets that were not used during training—KITTI [8], nuScenes [2], Dual Radar [25], and VoD [17]—were utilized for dataset expansion experiments. The data collection locations and LiDAR hardware configurations of each dataset are summarized in Tab. 4.

Despite differences in driving environments (*i.e.*, location) and various LiDAR hardware configurations—including multiple sensor models such as Ouster OS2-64, Velodyne HDL-64E, and HDL-32E—the synthesized 4D radar tensors  $\mathbf{T}_{C-RAE}$  reproduce key radar characteristics, such as sidelobes and strong reflections around objects, as shown in Fig. 4. This indicates that the L2RDaS Generator does not rely on a specific LiDAR hardware model; rather, it captures the structural patterns of LiDAR inputs and learns the generative rules underlying these radar characteristics in  $\mathbf{T}_{C-RAE}$ . For example, although L2RDaS was trained with 64-channel Ouster LiDAR data, it successfully synthesized  $\mathbf{T}_{C-RAE}$  for nuScenes, which uses a 32-channel Velodyne LiDAR, while preserving the aforementioned radar characteristics.

## C. Additional Qualitative Results

This section presents additional qualitative results of the 4D radar tensors  $\mathbf{T}_{C-RAE}$  synthesized by L2RDaS. Fig. 5 shows that L2RDaS reproduces key radar characteristics



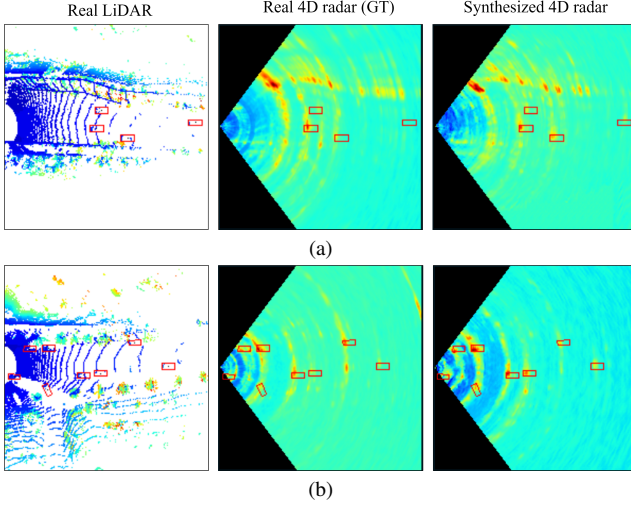


Figure 5. Qualitative synthesis results on the K-Radar test split.

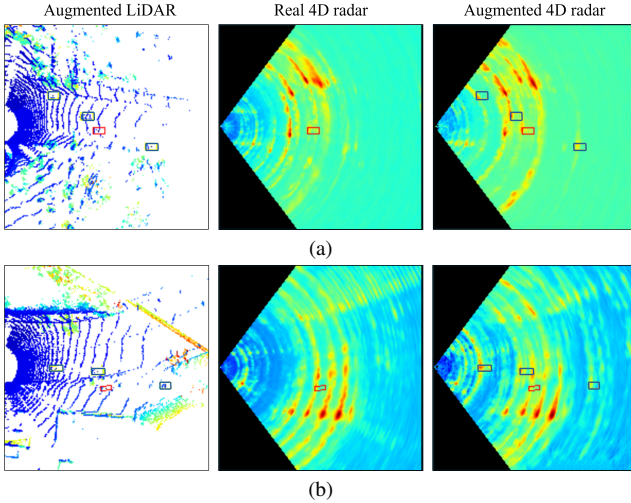


Figure 6. Qualitative samples of 4D radar tensor GT-Aug results synthesized by L2RDAS.

such as sidelobes and strong reflections around objects. L2RDAS also supports 4D radar tensor GT-Aug: when LiDAR point clouds are augmented with additional objects, the model synthesizes consistent  $\mathbf{T}_{C-RAE}$  for the augmented frames, as shown in Fig. 6.

Fig. 7 illustrates failure samples. In frames with densely clustered vehicles or extremely narrow road geometries, the synthesized tensors may become blurred and fail to capture detailed reflection patterns, suggesting limitations in learning complex multi-reflection interactions.

Despite these difficulties, using the synthesized  $\mathbf{T}_{C-RAE}$  for dataset expansion (Fig. 8) or additionally applying L2RDAS to 4D radar GT-Aug leads to measurable performance improvements compared to training solely on real

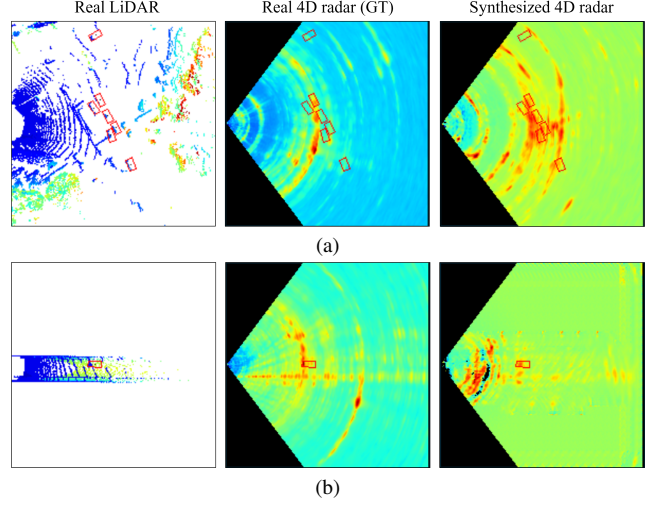


Figure 7. Failure samples on K-Radar test split. Frames with densely clustered vehicles and narrow-road conditions lead to blurred structures or incomplete reflection patterns.

datasets, as shown in Tab. 2. This demonstrates that, even when synthesis quality is imperfect in certain frames, the  $\mathbf{T}_{C-RAE}$  synthesized by L2RDAS effectively broadens the diversity and distributional range of training data, thereby enhancing overall generalization performance.

## D. Synthesizing Doppler Dimension

The L2RDAS framework focuses on synthesizing 3D spatial tensors (Range, Azimuth, Elevation), excluding the Doppler dimension. This represents a clear limitation, as it cannot synthesize complete 4D radar tensors.

This exclusion stems from the inherent limitations of the K-Radar [16] dataset and practical objectives. As noted in RTNH+ [12], K-Radar exhibits a very narrow Doppler span (-1.93 to 1.87 m/s) compared to VoD [17] (-26.5 to 26.5 m/s). This restricted range makes it prone to overflow issues and severely hinders the generative model from learning meaningful dynamic characteristics. Furthermore, omitting the Doppler dimension was an inevitable design choice to secure the rapid training and inference speeds essential for large-scale dataset construction.

Unlike cameras and LiDARs, 4D radar uniquely provides radial velocity information through the Doppler dimension. Leveraging this distinct advantage, numerous recent object detection models actively incorporate Doppler dimension to enhance their perception capabilities. Consequently, the absence of the Doppler dimension in our synthesized tensors inevitably leads to a degradation in downstream object detection performance. For instance, in an ablation study using the VoD [17], the removal of the Doppler dimension caused the  $mAP_{3D}$  within the driving area (ROI)

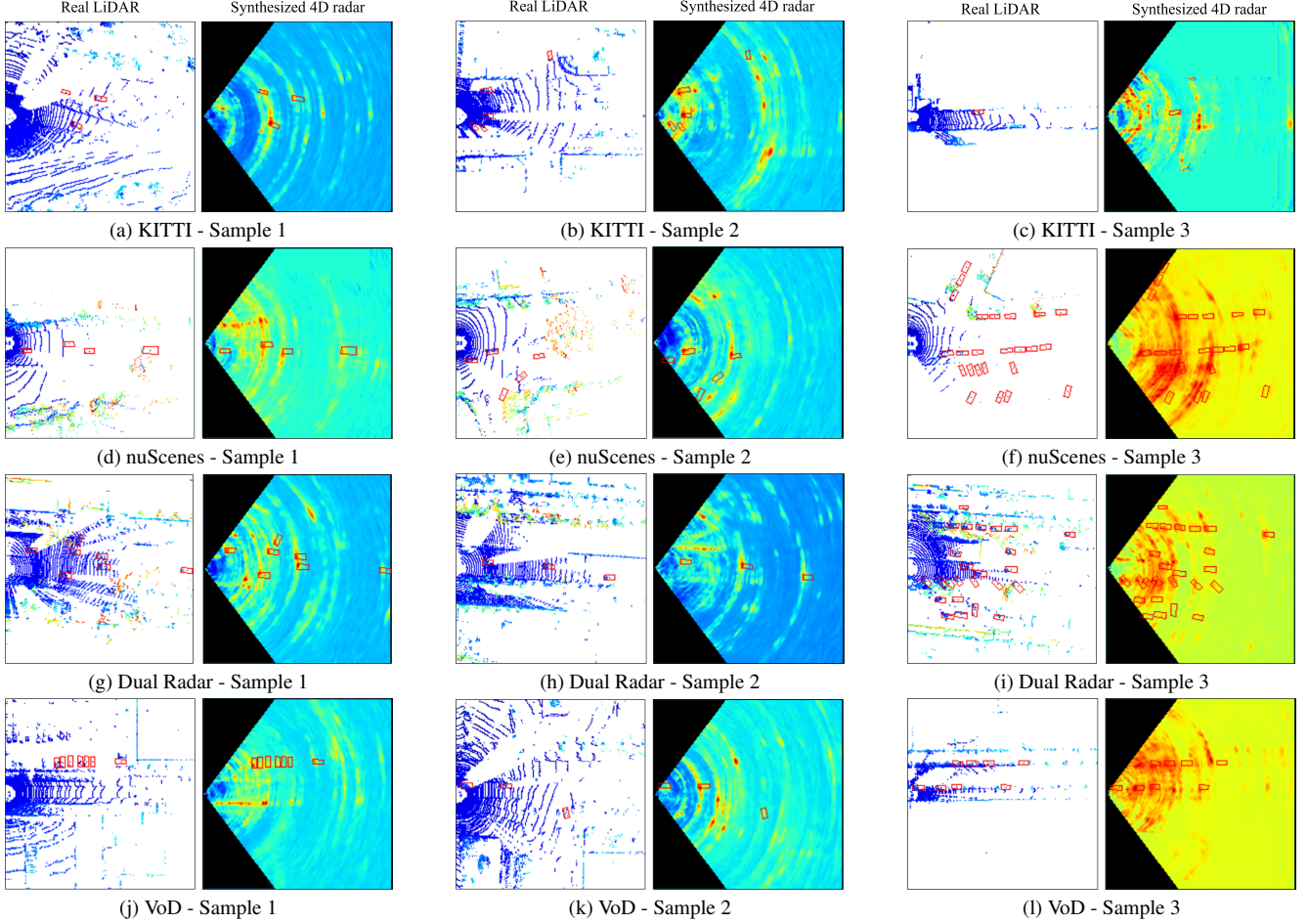


Figure 8. Synthesis results on external datasets not used during training (KITTI, nuScenes, Dual Radar, and VoD). The first two columns illustrate representative successful samples, while the third column presents failure samples.

of RadarPillarNet [26] to drop from 62.34% to 54.51%, and that of DaDan [23] to drop from 70.22% to 66.2%. Additionally, MFNet [22], which directly utilizes the Doppler dimension for multi-frame aggregation, cannot be applied to our synthesized data.

To verify the technical feasibility of Doppler synthesis, we conducted a preliminary experiment on a single sequence by integrating ConvLSTM [20] into the existing encoder to learn temporal correlations. The results showed an SSIM improvement from 0.821 to 0.850 compared to single-frame synthesis. However, this method critically degrades computational efficiency: GPU memory usage surged approximately fourfold (from 4,624 MB to 19,098 MB), and the training time increased approximately 30-fold, requiring 22 seconds per iteration. Therefore, we confirmed that incorporating the Doppler dimension is currently unsuitable for practical.

Despite this limitation, our research provides significant contributions by successfully synthesizing 3D spatial ten-

sors from 2D spaces, thereby maximizing spatial advantages, and by demonstrating its utility as a simulator, as will be discussed in Section F. If datasets with comprehensive Doppler spans become available in the future, the proposed ConvLSTM-based method can serve as a reference for complete 4D tensor synthesis; however, further research to improve computational efficiency is essential.

## E. Additional Experiments and Discussions

We present additional experimental results to comprehensively evaluate the performance of the proposed L2RDAs framework and provide an in-depth analysis of the validity of the data representation and synthesis strategies adopted in this study. Furthermore, we review various prior studies that utilize 4D radar tensors.

### E.1. Generalization on Diverse Object Classes

Although this study primarily demonstrated its performance focusing on Sedan class, we conducted multi-class object

detection experiments, including the Bus/Truck class, to verify the generalization capability of the framework. As shown in Tab. 5, when the model was trained with the addition of synthesized data ( $R_{All}^{Syn}$ ), the overall  $mAP$  and the detection performance for Sedans improved.

Conversely, the performance improvement for the Bus/Truck class was marginal or even showed a slight decrease. This is attributed to the severe class imbalance within the K-Radar dataset (19,275 Sedans vs. 3,681 Buses/Trucks, an approximate ratio of 5.2:1), which resulted in insufficient data for the generative model to adequately learn the structural characteristics of the minority class. It is expected that if a dataset with a balanced class distribution is acquired in the future, the proposed framework will be capable of successfully synthesizing diverse objects as well.

## E.2. Effectiveness of 3D Tensor Representation

We experimentally validate the rationale for utilizing 3D spatial information over 2D representations, and tensor formats over point clouds.

First, we compare the differences between 2D and 3D tensors. According to prior research (K-Radar [16]), utilizing 3D tensors (RAE) instead of 2D radar (RA)—which lacks elevation information—improves object detection performance by approximately 18%. L2RDaS maximizes this spatial advantage by completely and accurately synthesizing this 3D spatial information.

Second, we analyze the performance disparity based on the data representation method (Point Cloud vs. Tensor). As shown in Tab. 6 (a), despite utilizing the identical dataset (Seq 1 58), the 3D C-RAE tensor-based model achieved a relative improvement of 7.27% in  $AP_{3D}$  and 6.38% in  $AP_{BEV}$  compared to the Point cloud-based model extracted via CA-CFAR. This is because the tensor method fundamentally prevents the inevitable information loss that occurs during the point cloud conversion process. In other words, it demonstrates the inherent superiority of tensor representation, which perfectly preserves the 3D structure without any information loss.

## E.3. Impact of Synthesis Fidelity vs. Data Quantity

We conducted a controlled experiment to verify whether merely increasing the volume of training data guarantees an improvement in object detection performance, or if synthesis fidelity is an essential prerequisite. To isolate ‘synthesis quality’ as the sole independent variable, we augmented the identical amount of data ( $R_{All}^{Syn}$ ) using the low-fidelity generator (Exp 1) and the highest-fidelity L2RDaS generator (Exp 3) from Table 1 of the main text, and compared their performances.

As a result, as shown in Tab. 6 (b), even though the dataset size was increased equally, training with low-quality

data resulted in a 6.33% drop in  $AP_{BEV}$  and a 5.77% decrease in  $AP_{3D}$  compared to utilizing the highest quality data (L2RDaS). This clearly demonstrates that the high object detection performance achieved in this study is not merely due to a naive increase in data quantity, but rather stems directly from the high synthesis fidelity guaranteed by L2RDaS.

## E.4. Extended Related Works

To efficiently utilize massive 4D radar tensor data, prior studies have attempted various downstream methods. K-Radar [16], RTNH+ [12], and 3D-LRF [3] extract and utilize points with high power based on percentiles, whereas DPFT [7] and EchoFusion [13] convert the data into 2D projections (e.g., RA, AE) to reduce computational complexity. Meanwhile, methods such as CenterRadarNet [5] propose directly feeding the C-RAE tensor into neural networks. Consequently, the 3D tensors synthesized in our study provide a versatile data format that is seamlessly compatible with all of these diverse downstream utilization methods.

## F. Discussion on L2RDaS Application

As shown in Section 3, L2RDaS is a framework that synthesizes 4D radar tensors  $T_{C-RAE}$  from LiDAR point clouds  $X_{LiDAR}$  contained in existing autonomous driving datasets. Through the proposed synthesis procedure, we confirmed both qualitatively and quantitatively that the synthesized tensors successfully reproduce real radar reflection characteristics and distributions (Figs. 4 and 5, Tab. 1).

In particular, L2RDaS demonstrates high generalizability, as it can synthesize radar tensors  $T_{C-RAE}$  even for datasets that were not used during training, regardless of the data collection environment or LiDAR sensor model (Tab. 4). This property is also verified through the diverse synthesis samples presented (Figs. 4 and 8). Furthermore, by adding object LiDAR points to the LiDAR point clouds and synthesizing them using L2RDaS, 4D radar tensor-based GT-Aug can be performed (Figs. 2, 4 and 6). These synthesized results not only follow the real radar distribution closely while preserving radar characteristics, but also lead to performance improvements when applied to object detection model training, confirming that L2RDaS provides practically meaningful synthesis quality (Tab. 2).

L2RDaS can also be utilized as a simulation tool. For example, various scenarios can be synthesized by adding or removing object point clouds to or from LiDAR points in existing datasets collected across different environments. This enables the creation of infinitely expandable synthetic datasets across diverse environments and conditions, demonstrating the practical value of L2RDaS as a simulation tool for synthesizing 4D radar tensors  $T_{C-RAE}$ .



Table 5. Object detection performance across multiple object classes.  $R_{All}^{Syn}$  denotes the combined dataset of  $R_{KITTI}^{Syn}$ ,  $R_{nuScenes}^{Syn}$ ,  $R_{Dual-Radar}^{Syn}$ , and  $R_{VoD}^{Syn}$ . An IoU threshold of 0.3 is applied.

Method		BEV (%)			3D (%)		
Detection model	Train data	$mAP$	Sedan $AP$	Bus/Truck $AP$	$mAP$	Sedan $AP$	Bus/Truck $AP$
RTNH[16]	$R_{Kr}^{Real}$	38.94	47.19	<b>30.68</b>	34.98	44.60	<b>25.36</b>
	$R_{Kr}^{Real} + R_{All}^{Syn}$	<b>42.07</b>	<b>53.31</b>	30.83	<b>35.31</b>	<b>45.57</b>	25.05

Table 6. Ablation studies evaluating the impact of data representation and synthesis fidelity on Sedan object detection.

Method	$AP_{BEV}$ (%)	$AP_{3D}$ (%)
<b>(a) Effectiveness of Data Representation (Seq 1~58)</b>		
Point Cloud (CA-CFAR)	52.78	44.86
3D Tensor (C-RAE)	<b>56.15</b>	<b>48.12</b>
<b>(b) Impact of Synthesis Fidelity (<math>R_{Kr}^{Real} + R_{All}^{Syn}</math>)</b>		
Low Fidelity Generator (Exp 1)	50.18	43.27
High Fidelity Generator (Exp 3)	<b>53.57</b>	<b>45.92</b>

## References

- [1] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017. 4
- [2] 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 multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 5
- [3] Yujeong Chae, Hyeonseong Kim, and Kuk-Jin Yoon. Towards robust 3d object detection with lidar and 4d radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15162–15172, 2024. 8
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [5] Jen-Hao Cheng, Sheng-Yao Kuan, Hou-I Liu, Hugo Latapie, Gaowen Liu, and Jenq-Neng Hwang. Centerradarnet: Joint 3d object detection and tracking framework using 4d fmcw radar. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 998–1004. IEEE, 2024. 8
- [6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 5
- [7] Felix Fent, Andras Palffy, and Holger Caesar. Dpft: Dual perspective fusion transformer for camera-radar-based object detection. *arXiv preprint arXiv:2404.03015*, 2024. 8
- [8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 5
- [9] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 4
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4, 5
- [11] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 5
- [12] Seung-Hyun Kong, Dong-Hee Paek, and Sangyeon Lee. Rtnh+: Enhanced 4d radar object detection network using two-level preprocessing and vertical encoding. *IEEE Transactions on Intelligent Vehicles*, 2024. 6, 8
- [13] Yang Liu, Feng Wang, Naiyan Wang, and ZHAO-XIANG ZHANG. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. *Advances in Neural Information Processing Systems*, 36:53964–53982, 2023. 8
- [14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 4
- [15] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 4, 5
- [16] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022. 5, 6, 8, 9
- [17] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian FP Kooij, and Dariu M Gavrilă. Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. 5, 6
- [18] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4209–4219, 2024. 4, 5
- [19] Tim Salimans and Jonathan Ho. Progressive distillation



for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. [4](#), [5](#)

- [20] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. [7](#)
- [21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. [4](#)
- [22] Bin Tan, Zhixiong Ma, Xichan Zhu, Sen Li, Lianqing Zheng, Sihan Chen, Libo Huang, and Jie Bai. 3-d object detection for multiframe 4-d automotive millimeter-wave radar point cloud. *IEEE Sensors Journal*, 23(11):11125–11138, 2022. [7](#)
- [23] Xingzheng Wang, Jiahui Li, Jianbin Wu, Shaoyong Wu, and Lihua Li. Dadan: Dynamic-augmented and density-aware network for accurate 3d object detection with 4d radar. *IEEE Sensors Journal*, 2025. [7](#)
- [24] Ruofeng Yang, Bo Jiang, Cheng Chen, Baoxiang Wang, Shuai Li, et al. Few-shot diffusion models escape the curse of dimensionality. *Advances in Neural Information Processing Systems*, 37:68528–68558, 2024. [4](#)
- [25] Xinyu Zhang, Li Wang, Jian Chen, Cheng Fang, Guangqi Yang, Yichen Wang, Lei Yang, Ziyang Song, Lin Liu, Xiaofei Zhang, et al. Dual radar: A multi-modal dataset with dual 4d radar for autonomous driving. *Scientific Data*, 12(1):439, 2025. [5](#)
- [26] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rc-fusion: Fusing 4-d radar and camera with bird’s-eye view features for 3-d object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023. [7](#)