Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RadarSEDiff-T: Radar Structured-Encoding Diffusion with Transformer Guidance for Radar Sensing Enhancement

Zijie Dai
Tianjin Normal University
Mingzhu Shi
shimz@tjnu.edu.cn

Tianjin Normal University
Xiaoxian Zhao
Tianjin Normal University
Meng Zhang
Tianjin Normal University
Jingcheng Li
Tianjin Normal University
Ziheng He
Tianjin Normal University
Shengquan Yin
Tianjin Normal University
Junteng Wang
Tianjin Normal University

Research Article

Keywords: Radar Sensing Enhancement, Conditional diffusion model, Latent diffusion model, Deep learning.

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11306819/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

RadarSEDiff-T: Radar Structured-Encoding Diffusion with Transformer Guidance for Radar Sensing Enhancement

Zijie Dai[1,2], Mingzhu Shi[1,2*†], Xiaoxian Zhao[1,2†], Meng Zhang[1,2†], Jingcheng Li[1,2†], Ziheng He[1,2†], Shengquan Yin[1,2†], Junteng Wang[1,2†]

[1*]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, 300387, China.
[2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, 300387, China.

*Corresponding author(s). E-mail(s): shimz@tjnu.edu.cn;
Contributing authors: 15651215723@163.com;
xiaoxian_zhao@outlook.com; 17760792638@163.com;
1960962061@qq.com; 2928123802@qq.com; 15264943848@163.com;
wjt18222632906@163.com;
†These authors contributed equally to this work.

Abstract

Leveraging the unique advantages of all-weather anti-interference capability and deep hardware reuse with 6G communication systems, millimeter-wave radar has emerged as the most promising native perception carrier for integrated sensing and communication (ISAC) scenarios. However, its inherent limitations of sparse point cloud distribution and low angular resolution prevent it from achieving LiDAR-level perception accuracy, which severely hinders its deployment in unmanned aerial vehicle (UAV) autonomous navigation applications. To address this issue, we propose RadarSEDiff-T, an optimized latent diffusion model

designed to reconstruct dense LiDAR-level point clouds from a single millimeter-wave radar frame. Specifically, we first design a point-cloud-like structure tailored to the unique properties of radar bird's-eye view (BEV) maps, and train a dedicated radar encoder RSE independently. We further replace the conventional U-Net architecture with stacked Transformer blocks, enabling multi-dimensional millimeter-wave radar features to guide the entire diffusion generation process through global modeling and attention modules. Moreover, we integrate a consistency model to accelerate the inference of the diffusion model, ensuring that the proposed method can be effectively deployed on resource-constrained UAV platforms. We evaluate our approach on the public RadarHD dataset. Experimental results demonstrate that our method outperforms state-of-the-art baselines on the millimeter-wave radar super-resolution task, and explicitly verifies the significant speed advantage of our enhanced diffusion framework.

Keywords: Radar Sensing Enhancement, Conditional diffusion model, Latent diffusion model, Deep learning.

1 Introduction

The evolution of 6G mobile communication is driving a profound transformation of network architecture from traditional communication-only frameworks to an AI-native integrated sensing and communication (ISAC) paradigm, where space-air-ground integrated ubiquitous intelligent perception has become one of the core capability objectives of 6G networks. In typical scenarios such as autonomous driving and UAV autonomous navigation, LiDAR has long served as the core environmental perception sensor by virtue of its high-density and high-precision point cloud output[1]. However, its exorbitant hardware cost and severe performance degradation under adverse weather conditions remain critical bottlenecks hindering large-scale deployment. Meanwhile, as the native hardware carrier for 6G millimeter-wave communication, millimeter-wave radar inherently features low cost, strong anti-interference

capability, and native hardware reusability with communication systems[2]. Nevertheless, its inherent sparse point cloud distribution and limited angular resolution have long prevented its perception accuracy from meeting LiDAR-level application requirements.

In existing cross-modal supervised pseudo-LiDAR point cloud generation paradigms[3–5], most methods rely on paired radar-LiDAR datasets for explicit constraint, which not only suffers from high sensitivity to sensor spatiotemporal calibration errors and poor cross-scene generalization, but also produces point clouds with widespread defects including broken topological structures and missing local details, making it difficult to reconstruct complete geometric contours of targets. Although recent studies have attempted to introduce architectures such as generative adversarial networks[6] and U-Net[7] into radar point cloud enhancement tasks, these models are prone to mode collapse, and the generated point clouds often exhibit scattered trajectories and blurred edges, failing to stably output LiDAR-level high-precision perception results.

Recently, benefiting from the progressive denoising generation property, diffusion models[8] can implicitly learn the fine geometric distribution of point clouds under sparse and high-noise radar inputs, opening a brand-new technical path to break through the accuracy upper bound of pure single-frame millimeter-wave radar point clouds. However, existing diffusion-based methods lack targeted optimization for radar map characteristics. As illustrated in Fig. 1(a) and (c), radar maps can be regarded as large black backgrounds with sparse white trajectories, which are essentially single-channel grayscale images rather than RGB images. Directly processing them as RGB

inputs will not only increase data dimensionality and raise training difficulty, but also introduce redundant artifacts from useless channels into final outputs, degrading generation quality.

Furthermore, most current diffusion models[9–11] adopt U-Net architecture for inference speed, which limits the influence of radar maps on the entire diffusion process-radar features can only guide generation via simple concatenation with model inputs at the very first layer.

To address all the above limitations, we propose RadarSEDiff-T, a conditional latent diffusion model that optimizes raw millimeter-wave radar measurements to LiDAR-level fine granularity. Specifically, as shown in Fig. 1(b), we design a 2D point-cloud-like structure for radar BEV maps, which consists of the coordinates and grayscale values of each radar pixel. This customized design enables our dedicated latent encoder, named RSE, to discard useless information such as color and natural image priors, and focus exclusively on the intrinsic characteristics of radar data.

Moreover, to fully exploit millimeter-wave radar features to modulate the entire diffusion process, we replace the conventional U-Net backbone with stacked Transformer blocks. This modification allows radar features to guide diffusion not only as conditional inputs at the initial stage, but also to be deeply embedded into the attention modules of Transformer blocks to provide more fine-grained generation guidance. In addition, we introduce a consistency model[12] to boost the inference efficiency of the framework. Through distillation on the full pretrained diffusion model, our method achieves high-fidelity generation within only a few steps, even single-step inference.

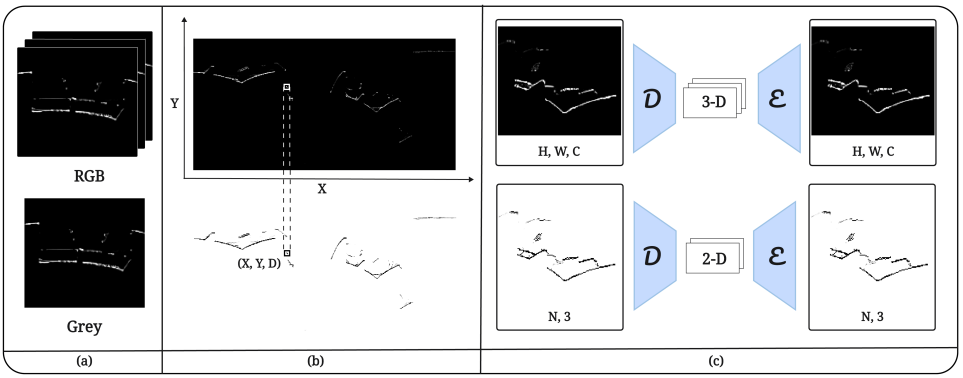


Fig. 1 (a) presents a comparative study between the RGB three-channel and grayscale single-channel representations of radar BEV maps; (b) illustrates the corresponding mapping relationship when converting the radar grayscale map to the 2D point-cloud-like structure; (c) provides a dimensionality comparison between the encoded features from the conventional image structure and the proposed 2D point-cloud-like structure.

This significant speedup ensures that the proposed RadarSEDiff-T can be effectively deployed on resource-constrained UAV platforms.

The main contributions of this paper are summarized as follows:

- We propose a 2D point-cloud-like structure tailored for radar map characteristics, and train a dedicated radar encoder RSE. Compared with general image encoders, it can fully exploit the intrinsic features of radar trajectories and eliminate interference from irrelevant redundant information.
- We present the RadarSEDiff-T conditional latent diffusion model, which adopts stacked Transformer blocks as the core backbone to enable comprehensive guidance from millimeter-wave radar features for the entire diffusion generation process. A consistency acceleration module is further integrated to boost the overall inference efficiency, supporting deployment on resource-limited edge platforms.

- Extensive experiments on the public RadarHD dataset demonstrate the strong competitiveness of our proposed model, especially its prominent inference speed advantage. Sufficient ablation studies quantitatively verify the effectiveness and performance gain of each individual module in the RadarSEDiff-T framework.

2 Related Work

Based on the research content and algorithm proposed in this paper, we conduct a comprehensive and in-depth investigation into deep learning-based radar perception enhancement algorithms and diffusion generative models.

2.1 Deep Learning Algorithms for Radar Perception Enhancement

Millimeter-wave radar, with its all-weather working capability that penetrates raindrops, fog, dust and other particulates, has become one of the core sensors for intelligent perception in complex environments. However, its reflected signals are highly susceptible to multipath effects and clutter interference. After being processed by traditional target detectors such as CFAR[13] and MUSIC[14], only sparse and noisy point clouds can be obtained, which are insufficient to support high-precision perception tasks. Early radar perception enhancement studies mostly focused on post-processing of sparse point clouds: removing outliers via RANSAC[15], completing SLAM mapping combined with velocity estimation[16], or introducing convolutional neural networks to fuse monocular images for cross-modal object detection[4]. These

schemes are inherently constrained by the sparsity of raw point clouds, making it impossible to achieve dense and fine-grained environmental characterization.

To break this limitation, the research community has gradually shifted to deep learning-based point cloud generation paradigms. Some studies take LiDAR point clouds as cross-modal supervision[3] to directly upsample the sparse point clouds output by traditional pipelines, but their enhancement performance has a clear upper bound since massive raw radar signal information has already been lost in pre-processing stages. Subsequent attempts learn the mapping from intermediate representations[17] such as range-Doppler heatmaps and azimuth-elevation maps to LiDAR representations via UNet and similar networks, yet limited by the generation capacity of the architecture, the final outputs are prone to losing sharp edges and fine geometric textures. Although the introduction of non-coherent integration and synthetic aperture integration[5] can improve point cloud density and angular resolution to a certain extent, these methods heavily rely on accurate ego-motion estimation, which makes them infeasible in extreme scenarios where other sensors fail.

The rapid evolution of generative models in recent years has opened a new path for this field. VAE-based methods[18] can convert sparse radar BEV representations into LiDAR-like dense outputs, but their generation results generally suffer from blurred edges. The cGAN architecture[19] can obtain high-definition enhanced outputs through adversarial training, but it is easily constrained by specific object categories and has obvious shortcomings in generalization, making it difficult to meet the radar perception enhancement requirements in complex open scenarios. These methods do not adopt specialized encoding for sparse radar BEV maps, but process them as

ordinary natural images. Due to the inherent differences between radar BEV maps and natural images, the enhancement performance of these methods is inevitably compromised to a certain extent.

2.2 Diffusion Generative Models

Proposed by Ho et al.[8], diffusion model is a generative model built on Markov chains, whose core idea is to learn data distribution by constructing two symmetric stochastic processes: the forward diffusion process and the reverse reconstruction process. It represents a new paradigm of generative models, and has recently surpassed previous state-of-the-art models on image generation tasks to achieve new SOTA performance.

In recent years, diffusion models, with their outstanding denoising and generation capabilities in general visual tasks, have provided a brand-new technical path for radar perception enhancement, and show remarkable application potential in the task of converting radar measurements to LiDAR-like representations. Among representative existing studies, Luan et al.[9] constructed a conditional denoising diffusion probabilistic model, which takes radar BEV images as condition to complete denoising and enhancement on temporally stacked corrupted LiDAR height maps, and optimizes the generation quality of sparse regions via weighted L1 loss. Zhang et al.[10] introduced the elucidated diffusion model and consistency sampling strategy to replace the traditional iterative denoising pipeline, and adopted LPIPS perceptual loss to improve generation quality while retaining the classic range-azimuth FFT preprocessing. Zhao et al.[11] introduced a pre-training framework from the monocular depth estimation

domain, and combined it with conditional latent diffusion models to realize single-frame radar perception enhancement, where both the encoder and the backbone U-Net are pre-trained to leverage the strong geometric priors of monocular depth estimation on walls, corners and object boundaries.

It is worth noting that all these methods tend to select U-Net as the backbone network of diffusion models. While this choice improves model efficiency and guarantees certain robustness, it severely limits the embedding paths for the conditional radar BEV information. This flaw degrades the final performance to a certain extent, especially under the single-frame setting, leading to significant performance loss in range and angular resolution. Drawing inspiration from these prior studies, we propose RadarSEDiff-T, which is designed to implement specialized encoding for radar BEV maps and integrate Transformer modules to better exploit the geometric spatial features of the conditional radar input. Furthermore, it incorporates a consistency model to significantly improve the overall inference efficiency of the framework.

3 Methodology

3.1 Overall Workflow

First, each LiDAR BEV image $X \in \mathbb{R}^{H\times W\times 3}$ is converted into a 2D point-cloud-like structure $X = \{x^i\}_{i=1}^n$ consisting of n points, where each point $x^i \in \mathbb{R}^3$ is constructed from the coordinates and grayscale value of the original pixel. The processed structure is then encoded by the proposed RSE module into the latent variable $Z \in \mathbb{R}^{N\times C}$. Unlike image diffusion, the generation probability of each point in point clouds is

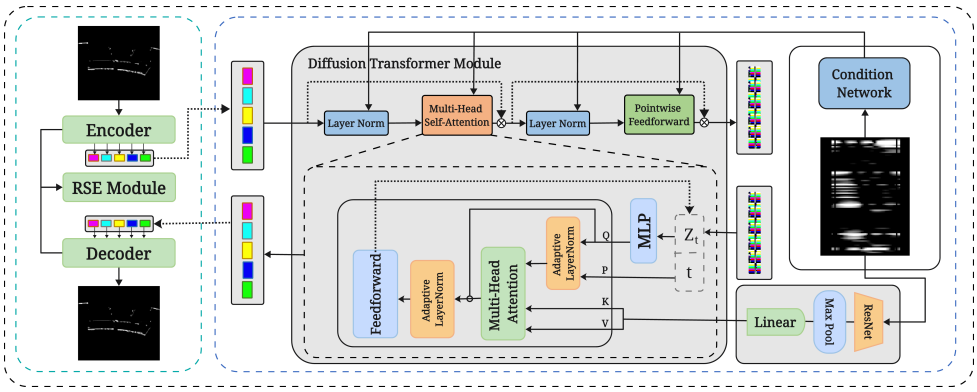


Fig. 2 The RadarSEDiff-T model architecture diagram. The left part corresponds to the RSE Module, which encodes the radar map converted into the point-cloud-like structure into latent variables via the pre-trained RSE encoder, and also implements the decoding operation for the latent variables generated by the model. The right part is the main diffusion backbone module, which performs training by progressively adding noise to the latent variables, then executes denoising to generate target latent variables, with the global features of the millimeter-wave radar map deeply embedded throughout the entire process.

mutually independent, and an additional shape latent variable c is required to control the overall point cloud distribution. The corresponding formulations for forward and reverse diffusion processes are defined as follows:

$$q(Z_{(1:T)}|Z_0) = \prod_{i=1}^{N} \prod_{t=1}^{T} q(z^i_{(t)}|z^i_{(t-1)}) \tag{1}$$

$$p_\theta(Z_{(0:T)}|c) = \prod_{i=1}^{N} p(z^i_{(T)}) \prod_{t=1}^{T} p_\theta(z^i_{(t-1)}|z^i_{(t)}, c) \tag{2}$$

Eq. (1) describes the forward diffusion process of point clouds, which starts from the initial distribution $q(Z_0)$ and gradually adds noise until it converges to the normal distribution $q(Z_T)$; Eq. (2) corresponds to the reverse diffusion process, which is

a progressive denoising generation process starting from the standard normal distribution $\mathcal{N}(0,1)$. Here, $q(z^i_{(t)}|z^i_{(t-1)})$ and $p_\theta(z^i_{(t-1)}|z^i_{(t)}, c)$ denote the Markov diffusion kernels, which add or remove noise for each point at the previous timestep and model the distribution of this point at the next timestep. The diffusion kernel is defined as:

$$q(z_{(t)}|z_{(t-1)}) = \mathcal{N}(z_{(t)}|\sqrt{1-\beta_t}z_{(t-1)}, \beta_t I) \tag{3}$$

$$p_\theta(z_{(t-1)}|z_{(t)}, c) = \mathcal{N}(z_{(t-1)}|\mu_\theta(z_{(t)}, t, c), \beta_t I) \tag{4}$$

where β_t is the variance scheduling hyperparameter that controls the diffusion rate of the process. In Eq. (4), μ_θ is the estimated mean implemented by a neural network parameterized by θ.

As illustrated in Fig. 2, the overall pipeline of RadarSEDiff-T consists of three core modules:

i) RSE Module: Encodes the 2D point-cloud-like structure converted from BEV images into the latent variable Z required for diffusion.

ii) Diffusion Transformer Module: Takes the latent variable encoded by the RSE module as input, extracts features from the radar map, and uses them as conditional context in the initial embedding stage and the attention modules of Transformer to control the entire diffusion generation process.

iii) Acceleration Module: Leverages the consistency model to reduce the diffusion process to only a few steps, even single-step generation, so as to significantly improve the overall inference efficiency of the model.

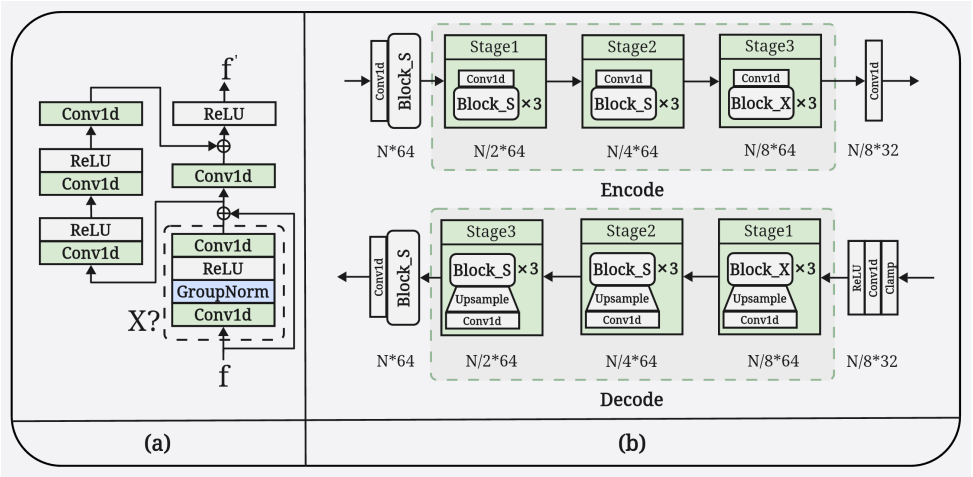


Fig. 3 The specific workflow of the RSE Encoder. (a) illustrates the detailed composition of the customized Block module; (b) demonstrates the specific modules traversed during the encoding and decoding processes, as well as the corresponding dimensional variation of variables.

3.2 RSE Module

We propose a novel radar feature-oriented encoding paradigm, the RSE Module, which converts radar BEV grayscale maps into 2D latent variables with point-cloud-like structures. This design not only reduces the complexity of latent representations, but also enables subsequent processing to focus more on the intrinsic radar trajectory characteristics rather than the natural image priors of BEV maps. The detailed architecture of this module is visualized in Fig. 3.

Specifically, as shown in Fig. 3(b), the backbone of the encoding process consists of three homologous feature aggregation stages, each containing one convolutional layer and three dedicated blocks. Given a point-cloud-like structure $X \in \mathbb{R}^{N\times 3}$ with N radar pixels, X is divided into overlapping patches at each feature aggregation stage, and then embedded into a new feature space. This operation halves the number of points

and adaptively adjusts the dimension of feature channels for hierarchical encoding. The detailed structure of each block is presented in Fig. 3(a), where the combination of convolution, normalization and activation function in the dashed box is selected according to the specific block type, to achieve the optimal aggregation effect at the final stage. The decoding process follows the symmetric reverse workflow.

3.3 Diffusion Transformer Module

Benefiting from the dedicated encoding of radar BEV maps by the RSE Module, our diffusion Transformer module directly trains the denoising process on the encoded latent variable z using the denoising diffusion probabilistic model (DDPM[8]), and guides the entire generation process with the extracted millimeter-wave radar features.

For the reverse process, the diffusion model is trained to learn a denoising network θ that inverts the forward diffusion process. The training objective is to maximize the variational lower bound of the negative log-likelihood involving all Z:

$$\begin{aligned} & L(\theta, \varphi) \\ =& \mathbb{E}_q[\sum_{t>1} D_{KL}(q(Z_{(t-1)}|Z_{(t)}, Z_{(0)}) \| p_\theta(Z_{(t-1)}|Z_{(t)}, c)) \\ & - \log p_\theta(Z_{(0)}|Z_{(1)}, c))] \end{aligned} \tag{5}$$

where $D_{KL}(.\|.)$ denotes the KL divergence that measures the distance between two distributions. Since both $q(Z_{(t-1)}|Z_{(t)}, Z_{(0)})$ and $p_\theta(Z_{(t-1)}|Z_{(t)}, c)$ are Gaussian distributions, we can reparameterize $\mu_\theta(z_{(t)}, t, c)$ to predict the noise term $\epsilon_\theta(Z_t, t, c)$. Then the training objective can be simplified to a simple mean squared loss between the model output $\epsilon_\theta(Z_t, t, c)$ and the ground-truth Gaussian noise ϵ: $\|\epsilon - \epsilon_\theta(Z_t, t, c)\|^2$.

After training $p_\theta(Z_{(t-1)}|Z_{(t)}, c)$, new latent variables are generated by progressive sampling $Z_{t-1} \sim p_\theta(Z_{(t-1)}|Z_{(t)}, c)$ with the reparameterization trick, where $Z_T \sim \mathcal{N}(0, 1)$ is initialized from standard normal distribution.

To construct a denoising network that can effectively capture long-range dependencies, we modify and adopt DiT as the neural backbone. The embedded Transformer architecture is capable of modeling long-term dependencies and capturing long-distance feature structures via self-attention mechanism.

Furthermore, to enhance the capability of removing useless noise from high-noise inputs, we introduce a dynamic position encoding to integrate the global features of the input map and the temporal information of noise timesteps. The dynamic Transformer is formulated as:

$$Q_t^{(l)} = Z_t^{(l)} W_Q^{(l)}, K_t^{(l)} = Z_t^{(l)} W_K^{(l)}, V_t^{(l)} = Z_t^{(l)} W_V^{(l)} \tag{6}$$

$$Z_t^{(l+1)} = AdaLN(MLP(softmax(\frac{Q_t^{(l)}(K_t^{(l)})^T}{\sqrt{D^{(l)}}})V_t^{(l)}), P_t) \tag{7}$$

where $Z_t^{(l)}$ denotes the embedding of diffusion voxel features at the l-th layer, $D^{(l)}$ is the dimension of features, P is the position encoding determined by the timestep t, and $AdaLN(., P)$ is the adaptive layer normalization. The AdaLN layer dynamically adjusts the mean and variance inserted at each layer via an MLP according to the position encoding P. Under the conditional generation setting, the input radar map first passes through a residual network and max-pooling layers to extract global features, which are taken as the shape latent variable c. Then $K_t^{(l)}$ and $V_t^{(l)}$ are computed

according to the latent variable c, to guide the generation process inside the attention modules.

3.4 Acceleration Module

After the diffusion model generates the latent variable and decodes it into the LiDAR BEV map, we adopt a consistency model[12] to distill the original full diffusion model. This optimization achieves few-step or even single-step generation, significantly boosting the overall inference efficiency of the proposed framework.

We integrate a consistency model distillation paradigm into RadarSEDiff-T, enabling high-fidelity single-step generation. Unlike standard DDPM requiring hundreds of iterative steps, the consistency model leverages the self-consistency property to map noisy latents at any timestep t directly to the clean data distribution, thereby bypassing sequential denoising.

We distill the pre-trained RadarSEDiff-T teacher network ϵ_θ into a student consistency network f_θ. The distillation objective enforces consistency between adjacent timesteps:

$$L_{consist} = \mathbb{E}_{Z_0,\epsilon,t}\left[\|f_\theta(Z_{t_n}, t_n, c) - f_\theta(Z_{t_{n-1}}, t_{n-1}, c)\|^2\right] \tag{8}$$

where t_n and t_{n-1} are adjacent noise levels. Crucially, the student network retains the full conditional embedding path, injecting the radar shape latent c into both initial embeddings and Transformer attention layers to prevent performance degradation often seen in accelerated methods.

Post-distillation, inference is reduced to a single forward pass: sampling $Z \sim \mathcal{N}(0, 1)$ and feeding it with c into f_θ yields the final clean latent. This process reduces inference latency by over 80% compared to the original DDPM, ensuring real-time deployability on onboard UAV computing units.

4 Experiments

In this section, extensive experiments are conducted to systematically validate the effectiveness of our proposed method. We first introduce the experimental setup, including the adopted dataset, evaluation metrics, and detailed implementation configurations. Then, quantitative comparison results are presented to demonstrate that our method outperforms other state-of-the-art approaches. Finally, comprehensive ablation studies of different categories are provided to verify the contribution of each key module in our framework.

4.1 Dataset and metrics

4.1.1 Dataset

We evaluate our method on RadarHD[18], a public indoor radar-LiDAR paired dataset captured by a single-chip millimeter-wave radar system. Specifically, RadarHD contains approximately 40k raw millimeter-wave radar signal data pairs, with high-resolution LiDAR ground truth collected across 67 distinct indoor and outdoor trajectories. In our experiments, we process the raw radar data following the exact same pipeline as RadarHD to generate radar BEV maps. The resolution of the

millimeter-wave radar BEV map is set to 64×256, while the corresponding LiDAR BEV map has a resolution of 512×256.

4.1.2 Evaluation Metrics

We extract 2D point clouds from both the generated BEV outputs and the corresponding LiDAR ground truth. For similarity measurement, we select Chamfer Distance (CD[20]) and Modified Hausdorff Distance (MHD[21]). Given two point clouds $S_1, S_2 \subseteq R^{N\times 3}$ with $|S_1| = |S_2|$, CD calculates the average distance between each point in S_1 and its nearest neighbor in S_2. which is widely adopted for its high computational efficiency, as formulated in Eq.(9).

$$CD = \frac{1}{|S_1|}\sum_{x\in S_1}\min_{y\in S_2}\|x-y\|_2 + \frac{1}{|S_2|}\sum_{y\in S_2}\min_{x\in S_1}\|y-x\|_2 \tag{9}$$

MHD reflects the average degree of mismatch between S_1 and S_2. A critical property of MHD is that it preserves the constraint on overall contour consistency, which means it exhibits higher sensitivity to severe local deviations and delivers more robust safety guarantees. The formal definition of MHD is given in Eq. (10). Lower values of both CD and MHD indicate better radar enhancement quality.

$$H_{MHD}(S_1, S_2) = \max(h_{MHD}(S_1, S_2), h_{MHD}(S_2, S_1)) \tag{10}$$

$$h_{MHD}(S_1, S_2) = \frac{1}{|S_1|}\sum_{s_1\in S_1}\min_{s_2\in S_2}\|s_1 - s_2\| \tag{11}$$

Here, $|S_1|$ denotes the total number of points contained in point set S_1, and min $\|s_1 - s_2\|$ represents finding the nearest distance from each point in S_1 to all points

in S_2. Following the prior work [4], both the prediction and the corresponding ground truth are scaled to a bounding box of $[-1, 1]^3$. This normalization allows the metrics to focus on the geometric shape of the point cloud rather than its absolute scale, leading to more accurate and unbiased measurements.

Table 1 Comparison of cutting-edge radar perception enhancement performance with prior work on RadarHD Dataset. The metrics are computed on 48 batch sizes. The best results are shown in bold. ↑ (resp. ↓) means the higher (resp. lower), the better

Method	Frames	CD-mean↓	MHD-mean↓	Time-cons↓
CFAR[13]	1	0.84	0.91	-
RadarHD[18]	41	0.44	0.34	4:54
RadarHD single-frame[18]	1	0.56	0.45	**4:23**
Luan et al.[9]	5	0.59	0.50	-
Zhang et al.[10]	1	0.38	0.29	42:33
RadarSFD[11]	1	**0.35**	0.28	2:08:56
Ours-EMD	1	0.36	**0.28**	13:15
Ours-CD	1	0.38	0.30	5:14

4.2 Implementation details

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 5070Ti GPU with 48GB memory. We train our model for 10,000 iterations using the Adam optimizer[22] with an initial learning rate of 1e-3, and the batch size is set to 32. During training, ResNet18[23] is adopted as the backbone for radar image feature

extraction in the Diffusion Transformer Module, and the dimension of the output feature vector is 384. Both the predicted mean and variance are represented as 384-dimensional vectors. The denoising model architecture we use is the standard DiT Block constructed in DiT[24].

4.3 Comparison with Existing Methods

To validate the performance of the proposed method, we conduct comparative experiments with five cutting-edge radar perception enhancement approaches from existing literature (References CFAR[13], RadarHD[18], Luan et al.[9], Zhang et al.[10], RadarSFD[11]). Among them, the methods proposed by Luan et al.[9], Zhang et al.[10] and RadarSFD[11] belong to the diffusion model framework. To ensure a fair comparison, the test batch size for all methods is uniformly set to 48. It is worth noting that all quantitative results of the compared methods are directly cited from their original publications, or obtained by running their officially released open-source codes and pre-trained models, with strictly identical parameter configurations consistent with our experiments.

To evaluate the generalization capability of our method across diverse scenarios, we train our model on different scene subsets of the RadarHD dataset. Table 1 reports the CD, MHD and inference latency of each method, while Fig. 5 presents the qualitative visualization results generated by different approaches. As can be observed from the comparison results, limited by the conversion loss introduced by the 2D point-cloud-like encoding, our method does not outperform all prior arts on every single metric. Nevertheless, our approach achieves a significant improvement in inference efficiency

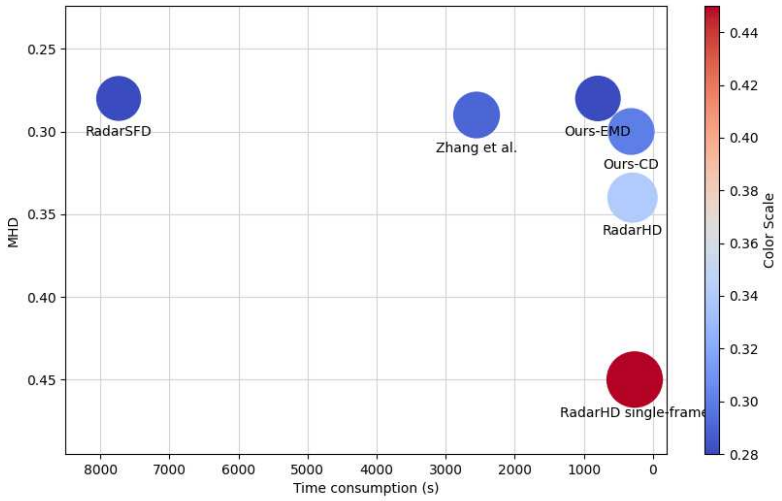


Fig. 4 the bubble chart of Table 1.

during the test phase, which is a critical performance indicator for deployment on resource-constrained platforms.

This conclusion is further visualized in the bubble chart in Fig. 4, where the x-axis denotes the time consumption, the y-axis represents the MHD value, and the size or color depth of each bubble corresponds to the CD value. A point located closer to the upper-right region indicates that the method achieves superior performance while maintaining high inference efficiency. It can be observed that our method achieves a dominant advantage in both MHD and CD metrics among all compared approaches. This benefit stems from our specially designed novel structure and dedicated encoder for radar BEV maps, which allows the model to avoid redundant computational overhead on useless dimensions and irrelevant features.

4.4 Ablation Studies

In this section, we conduct in-depth analysis to verify the effectiveness of our core modules, including the proposed novel structure encoding module and the two different image feature embedding strategies.

Effectiveness of the Novel Structure Encoding Module: To evaluate the performance of the proposed encoding module, we carry out ablation experiments on the RadarHD dataset. As shown in Table 2, we compare our RSE encoder with the pre-trained Marigold[25] image encoder adopted in RadarSFD, as well as the baseline that directly feeds raw inputs without any dedicated encoding. The baseline that directly applies the 2D point-cloud-like structure to the diffusion model fails to converge normally and cannot produce valid generation results, due to the absence of the shape latent variable c that governs the overall point cloud distribution. The experimental results demonstrate that, benefiting from the lower dimensionality of the 2D point-cloud-like structure compared to conventional image representations and its stronger focus on intrinsic radar features, our proposed scheme achieves superior performance in both quantitative evaluation metrics and inference speed.

Effectiveness of the Two Image Feature Embedding Schemes: As illustrated in Table 3, we not only test different image feature embedding strategies, but also evaluate different backbone architectures for the diffusion model. The results show that, compared with the concatenation-based feature embedding applied before the diffusion backbone, injecting conditional features into the attention modules inside Transformer blocks can more effectively guide the entire diffusion generation process, which makes our proposed scheme achieve better performance on all evaluation metrics.

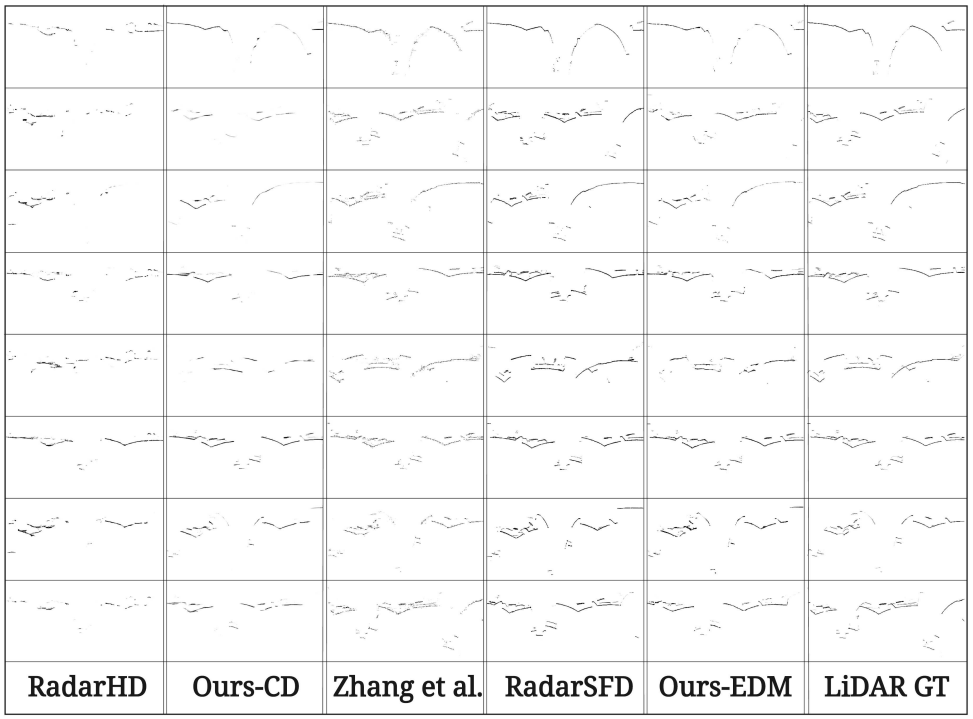


Fig. 5 Qualitative visualization results on RadarHD dataset.

5 Conclusion

In this paper, we present RadarSEDiff-T, an effective and feature-fusion enhanced framework that can accurately enhance the input millimeter-wave radar map to LiDAR-level precision. By combining the proposed novel structured encoding, dedicated feature embedding strategy and consistency model distillation, our method achieves the state-of-the-art trade-off between inference latency and generation quality, making it fully deployable on resource-constrained platforms.

We propose a brand-new structure and encoding paradigm specifically tailored for radar data. After this dedicated structured encoding, the dimensionality and computational complexity are significantly reduced, and the model can better focus on

Table 2 Ablation Study: Effectiveness of the Novel Structure Encoding Module. On RadarHD Dataset. The best results are shown in bold

VAE	Struct	CD-mean↓	MHD-mean↓	Time-cons↓
None	Image	0.461863	0.366447	43:34
	2D-pc	1.239864	1.739515	39:19
Marigold vae	Image	0.371886	0.295836	30:08
RSE Encoder	2D-pc	**0.365181**	**0.285765**	**13:15**

Table 3 Ablation Study:Effectiveness of the Two Image Feature Embedding Schemes. On RadarHD Dataset. The best results are shown in bold

Condition	Model Struction	CD-mean↓	MHD-mean↓
Concatenation	DiT	0.382657	0.303385
	U-Net	0.382241	0.321237
Cross-Attn	DiT	0.375276	0.308968
Concat + Attn	DiT	**0.365181**	**0.285765**

the intrinsic radar trajectory characteristics, which opens up new possibilities for its application in the field of single-frame radar perception enhancement. Researchers and practitioners in this domain can further extend our method to explore new directions of this specialized structure and encoding paradigm.

In future work, considering the unavoidable conversion loss introduced during the encoding transformation process, we plan to explore advanced techniques to further mitigate this loss.

Declarations

- This work was supported by Tianjin Research Innovation Project for Postgraduate Students (No.2026KYCX016F), the National Natural Science Foundation of China (No.61501328, No.62471335), Enterprise Joint Horizontal Science and Technology Project (No.53H21034).

- The authors declare that they have no competing interests.
- The dataset supporting the conclusions of this article is available in the RadarHD repository, https://github.com/akarsh-prabhakara/RadarHD.[18]
- Dai. was responsible for the primary innovations and the experimental work, Shi. provided the research ideas and guidance throughout all aspects of the study, Zhao. and Zhang. were responsible for drafting the manuscript text, Li., He., Yin. and Wang. assisted in the implementation of the experiments and the processing of data, All authors reviewed the manuscript.

References

[1] Bijelic, M., Gruber, T., Ritter, W.: A benchmark for lidar sensors in fog: Is detection breaking down? In: 2018 IEEE Intelligent Vehicles Symposium (IV) (2018). https://doi.org/10.1109/ivs.2018.8500543 . http://dx.doi.org/10.1109/ivs.2018.8500543

[2] Guan, J., Madani, S., Jog, S., Gupta, S., Hassanieh, H.: Through fog high-resolution imaging using millimeter wave radar. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/

10.1109/cvpr42600.2020.01148 . http://dx.doi.org/10.1109/cvpr42600.2020.01148

[3] Lu, C.X., Rosa, S., Zhao, P., Wang, B., Chen, C., Stankovic, J.A., Trigoni, N., Markham, A.: See through smoke: robust indoor mapping with low-cost mmwave radar. In: Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services. MobiSys '20, pp. 14–27. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3386901.3388945 . https://doi.org/10.1145/3386901.3388945

[4] Cheng, Y., Xu, H., Liu, Y.: Robust small object detection on the water surface through fusion of camera and millimeter wave radar. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15243–15252 (2021). https://doi.org/10.1109/ICCV48922.2021.01498

[5] Geng, R., Li, Y., Zhang, D., Wu, J., Gao, Y., Hu, Y., Chen, Y.: Dream-pcd: Deep reconstruction and enhancement of mmwave radar pointcloud. IEEE Transactions on Image Processing **33**, 6774–6789 (2024) https://doi.org/10.1109/TIP.2024.3512356

[6] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14, pp. 2672–2680. MIT Press, Cambridge, MA, USA (2014)

[7] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F.

(eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer, Cham (2015)

[8] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)

[9] Luan, K., Shi, C., Wang, N., Cheng, Y., Lu, H., Chen, X.: Diffusion-based point cloud super-resolution for mmwave radar data. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 11171–11177 (2024). https://doi.org/10.1109/ICRA57147.2024.10611026

[10] Zhang, R., Xue, D., Wang, Y., Geng, R., Gao, F.: Towards dense and accurate radar perception via efficient cross-modal diffusion model. IEEE Robotics and Automation Letters **9**(9), 7429–7436 (2024) https://doi.org/10.1109/LRA.2024.3426389

[11] Zhao, B., Garg, N.: Radarsfd: Single-frame diffusion with pretrained priors for radar point clouds. In: IEEE International Conference on Robotics and Automation (ICRA) (2026). https://doi.org/10.48550/arXiv.2509.18068

[12] Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)

[13] Miller, R.: Fundamentals of radar signal processing (richards, m.a.; 2005) [book review]. IEEE Signal Processing Magazine **26**(3), 100–101 (2009) https://doi.org/

10.1109/MSP.2009.932123

[14] Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation **34**(3), 276–280 (1986) https://doi.org/10.1109/TAP.1986.1143830

[15] Park, Y.S., Shin, Y.-S., Kim, J., Kim, A.: 3d ego-motion estimation using low-cost mmwave radars via radar velocity factor for pose-graph slam. IEEE Robotics and Automation Letters **6**(4), 7691–7698 (2021) https://doi.org/10.1109/LRA.2021.3099365

[16] Zhang, J., Zhuge, H., Wu, Z., Peng, G., Wen, M., Liu, Y., Wang, D.: 4dradarslam: A 4d imaging radar slam system for large-scale environments based on pose graph optimization. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 8333–8340 (2023). https://doi.org/10.1109/ICRA48891.2023.10160670

[17] Cheng, Y., Su, J., Jiang, M., Liu, Y.: A novel radar point cloud generation method for robot environment perception. IEEE Transactions on Robotics **38**(6), 3754–3773 (2022) https://doi.org/10.1109/TRO.2022.3185831

[18] Prabhakara, A., Jin, T., Das, A., Bhatt, G., Kumari, L., Soltanaghai, E., Bilmes, J., Kumar, S., Rowe, A.: High resolution point clouds from mmwave radar. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 4135–4142 (2023). https://doi.org/10.1109/ICRA48891.2023.10161429

[19] Guan, J., Madani, S., Jog, S., Gupta, S., Hassanieh, H.: Through fog high-resolution imaging using millimeter wave radar. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/cvpr42600.2020.01148 . http://dx.doi.org/10.1109/cvpr42600.2020.01148

[20] Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'77, pp. 659–663. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1977)

[21] Dubuisson, M.-P., Jain, A.K.: A modified hausdorff distance for object matching. In: Proceedings of 12th International Conference on Pattern Recognition, vol. 2, pp. 566567568–. IEEE Computer Society, Los Alamitos, CA, USA (1994). https://doi.org/10.1109/ICPR.1994.576361 . https://doi.ieeecomputersociety.org/10.1109/ICPR.1994.576361

[22] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)

[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016). https://doi.org/10.1109/cvpr.2016.90

[24] Peebles, W.S., Xie, S.: Scalable diffusion models with transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 4172–4182 (2022)

[25] Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–18 (2025) https://doi.org/10.1109/TPAMI.2025.3591076