SDDiff: Boosting Radar Perception via Spatial-Doppler Diffusion

Shengpeng Wang[1] , **Xin Luo**[1] , **Yulong Xie**[1] and **Wei Wang**[2] *

[1]Huazhong University of Science and Technology

[2]Wuhan University

{wsp666, l_xin, yulong_xie}@hust.edu.cn, wangw@whu.edu.cn

Abstract

Point cloud extraction (PCE) and ego velocity estimation (EVE) are key capabilities gaining attention in 3D radar perception. However, existing work typically treats these two tasks independently, which may neglect the interplay between radar's spatial and Doppler domain features, potentially introducing additional bias. In this paper, we observe an underlying correlation between 3D points and ego velocity, which offers reciprocal benefits for PCE and EVE. To fully unlock such inspiring potential, we take the first step to design a **S**patial-**D**oppler **Diff**usion (SDDiff) model for simultaneously dense PCE and accurate EVE. To seamlessly tailor it to radar perception, SDDiff improves the conventional latent diffusion process in three major aspects. First, we introduce a representation that embodies both spatial occupancy and Doppler features. Second, we design a directional diffusion with radar priors to streamline the sampling. Third, we propose Iterative Doppler Refinement to enhance the model's adaptability to density variations and ghosting effects. Extensive evaluations show that SDDiff significantly outperforms state-of-the-art baselines by achieving 59% higher in EVE accuracy, 4× greater in valid generation density while boosting PCE effectiveness and reliability. The code and dataset will be available on https://github.com/StellarEsti/SDDiff.

1 Introduction

Millimeter-wave radar for all-weather perception is increasingly attracting widespread attention in robotics, computer vision, augmented reality, and autonomous driving [Harlow *et al.*, 2024; Zhang *et al.*, 2022; Xu *et al.*, 2021; Adolfsson *et al.*, 2021]. Point cloud extraction (PCE) and ego velocity estimation (EVE) are crucial pillars of radar perception. PCE acts as a *low-level sensory process*, extracting fundamental object information from reflected radar signals, including position, reflectivity, and Doppler velocity. Conversely, EVE serves as a *high-level cognitive process*, leveraging elemental point clouds to infer the radar's ego velocity. As Fig. 1 shows, comprehensive PCE and accurate EVE form a solid perceptual foundation for downstream tasks, including object detection, simultaneous localization and mapping (SLAM), path planning, and autonomous navigation.

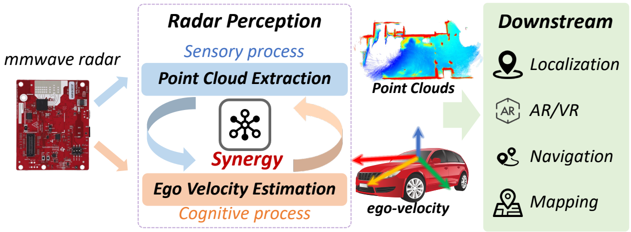

Figure 1: Illustration of our work's objective: simultaneously extracting point clouds and estimating ego velocity to enhance radar perception.

Existing studies have started to focus on PCE [Prabhakara *et al.*, 2023; Zhang *et al.*, 2024] and EVE [Pang *et al.*, 2024] separately, yielding promising results. These tasks are typically treated independently, with PCE relying solely on signal intensity and EVE anchored in sparse, chaotic points processed by onboard systems. However, addressing these tasks in isolation may overlook the interplay between radar's spatial and Doppler domain features, potentially introducing additional bias. Specifically, on the one hand, non-object regions may display high reflection intensity due to the notorious "multi-path effect" [Yataka *et al.*, 2024]. Consequently, relying solely on signal intensity may mislead PCE, impeding the balance between detection density and clutter suppression. On the other hand, sparse, noisy points with poor vertical angle resolution severely degrade EVE performance.

One of our key observations is that there is a synergy between PCE and EVE, which can reciprocally evoke potential gains from each other. We conduct a feasibility study using the Coloradar Dataset [Kramer *et al.*, 2022], as shown in Fig. 2(a). The orientation and Doppler velocity of 3D point clouds jointly define a surface, with its precise parameters determined by the radar's ego velocity. Given various points, we estimate these parameters by Random Sample Consensus (RANSAC) [Fischler and Bolles, 1981]. As shown in Fig. 2(b), sparse, noisy points after onboard CFAR

*Corresponding author

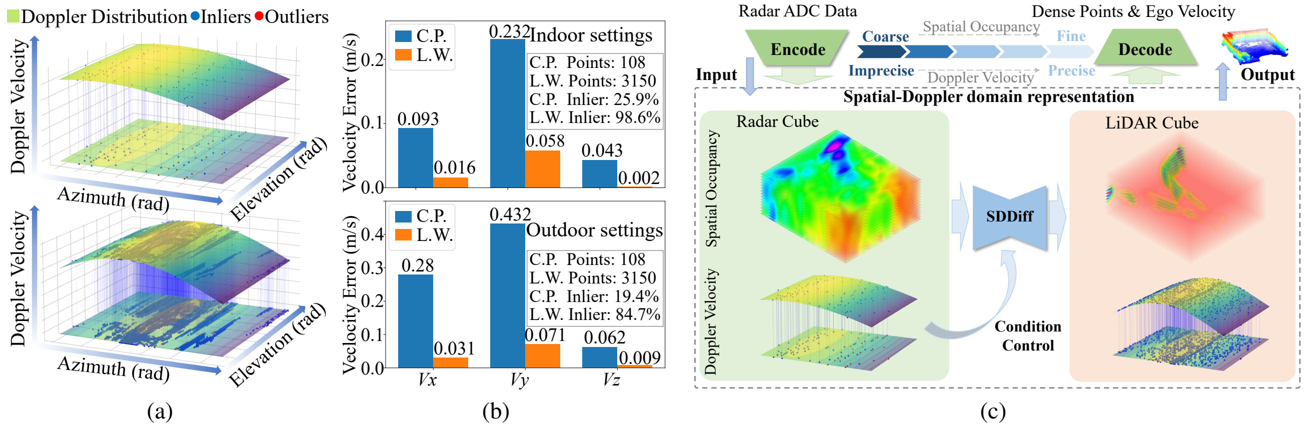

Figure 2: (a) The upper figure presents Doppler velocity distribution derived from sparse points after onboard CFAR post-processing, while the lower figure displays points warped by LiDAR. (b) Given CFAR-processed (C.P.) points or LiDAR-warped (L.W.) points, velocity estimation errors are observed in both indoor and outdoor settings. (c) The brief illustration of SDDiff, where PCE and EVE are jointly refined through the SDDR Purification Process, sculpting coarse representations into fine ones.

post-processing lead to an awkward inlier rate (about 25.9% at a threshold of 0.08). Points warped by dense LiDAR point clouds show over 5× improvement in EVE, with the inlier rate reaching 98.6% indoors. This demonstrates that the synergy offers reciprocal benefits: accurate ego velocity helps filter or refine points aligned with Doppler velocity. In turn, robust and dense PCE can further enhance EVE accuracy.

To fully unlock such inspiring reciprocal potential, we introduce a Spatial-Doppler Domain Representation (SDDR) for 3D radar perception. As shown in Fig. 2(c), the radar analog-to-digital converter (ADC) signal is encoded into a 3D tensor that embodies both spatial occupancy and Doppler velocity. PCE and EVE are jointly refined through SDDR Purification Process, sculpting coarse, ghost-prone representations into fine, uncontaminated ones.

Inspired by remarkable generative capabilities of diffusion models [Ho *et al.*, 2020], we design a **S**patial-**D**oppler **Diff**usion (SDDiff) model to generate an enhanced SDDR by effectively aligning spatial and Doppler features. Specifically, we propose two key designs that allow SDDiff to seamlessly tailor to 3D radar perception. On the one hand, naive diffusion model aims to transform multi-modality in latent space, requiring a plethora of sampling steps. This high demand is attributed to sampling from a standard Gaussian distribution, which is well-suited for image synthesis. However, it introduces sampling wastage and ambiguous outcomes for the SDDR purification task. To address this, we design a directional diffusion with radar priors, transforming consistent modality in SDDR. This significantly reduces the sampling steps and facilitates finer spatial occupation. On the other hand, we incorporate iteratively refined Doppler velocity profiles into the model as conditions and apply a physical constraint to bridge PCE and EVE based on the synergy. This enables the model to learn adaptive feature representations that are resilient to density variations and ghost effects.

We summarize the contributions below:

- We propose a novel Spatial-Doppler Diffusion model to sculpt coarse representations into fine ones for 3D radar perception. To the best of our knowledge, this is the first attempt to simultaneously attain dense point clouds and accurate ego velocity from Spatial-Doppler Domain Representation.
- We design a directional Spatial-Doppler diffusion with radar priors to streamline the sampling. This deeply reduces sampling steps and mitigates ambiguous outcomes.
- We propose Iterative Doppler Refinement, leveraging the Doppler-consistency of refined spatial occupancy to enhance the model's adaptability to density variations and ghosting effects.
- Experimental results demonstrate that our method outperforms the previous state of the arts and achieves reciprocal benefits across PCE and EVE. Additionally, we will make our self-collected dataset publicly available to the research community. This dataset facilitates further advancements in 3D radar perception.

2 Related Work

2.1 Traditional Model for Radar Detection

In early radar detection, low-pass filtered intermediate frequency (IF) signals from millimeter-wave radar were processed via FFT, angle estimation [Schmidt, 1986; Li *et al.*, 2003; Roy and Kailath, 1989], and Constant-False-Alarm-Rate (CFAR) [Nitzberg, 1972], resulting in sparse, artifact-laden point clouds. Such sparse points are insufficient for downstream tasks such as state estimation, mapping, and navigation. To generate denser radar points, some research [Lai *et al.*, 2024; Qian *et al.*, 2020] leverages radar motion to construct virtual antenna arrays, similar to synthetic aperture radar (SAR) imaging [Ausherman *et al.*, 1984], to enhance angular resolution. However, these methods depend on precise motion estimation or predetermined trajectory, hindering their translation into real-world implementation. Other studies [Cen and Newman, 2019; Cen and Newman, 2018] concentrate on eliminating clutter points caused by multipath effects to enhance the quality of radar point clouds. However, they are limited to high-resolution mechanical scanning radars and unsuitable for commercial off-the-shelf radar.

2.2 Generative Model for Radar Detection

Recently, numerous studies have focused on generating denser radar points through cross-modal supervision from high-resolution sensors, such as LiDAR or depth camera. Guan et al. [Guan *et al.*, 2020] leverages a conditional Generative Adversarial Network (cGAN) [Goodfellow *et al.*, 2014; Sun *et al.*, 2021] to achieve object-level imaging and recover the high-frequency shape of the specific object. To materialize scene-level imaging, milliMap [Lu *et al.*, 2020] and RadarHD [Prabhakara *et al.*, 2023] employ a frame-stacking strategy, combining multiple CFAR-filtered Cartesian map patches or planar Range-Azimuth (RA) maps as U-Net [Ronneberger *et al.*, 2015] inputs, with maps projected by LiDAR points for supervision. Similarly, Zhang et al. [Zhang *et al.*, 2024] apply a diffusion model [Ho *et al.*, 2020] to restore radar RA maps under LiDAR supervision, subsequently generating radar points in bird's-eye view (BEV). However, these intensity-only methods overlook the significant contributions of Doppler features to the spatial point cloud's contextual understanding. To enhance 3D point clouds, Luan et al. [Luan *et al.*, 2024] encode CFAR-processed 3D sparse point clouds into BEV images for diffusion model. Unlike the direct approaches, other studies [Cheng *et al.*, 2022; Fan *et al.*, 2024] predict Range-Doppler (RD) maps, and then use conventional angle estimation to derive 3D points. Nevertheless, they still resort to traditional angle estimation.

2.3 Ego Velocity Estimation for Radar

Some methods estimate ego velocity by solving a transformation between two consecutive exteroceptive measurements. General registration techniques, such as ICP [Besl and McKay, 1992], NDT [Biber and Straßer, 2003], and their variants [Censi, 2008; Segal *et al.*, 2009], are often applied for this. However, these sensor-agnostic approaches struggle with noisy radar point clouds, which provide poor point-to-point correspondence. While [Cen and Newman, 2019] introduces a keypoint extraction and data association scheme, it is unsuitable for low-resolution commercial radars. More methods leverage Doppler velocities for 2D ego velocity estimation, typically relying on RANSAC [Kellner *et al.*, 2013; Kellner *et al.*, 2014] or end-to-end networks like RadarEVE [Pang *et al.*, 2024]. However, these approaches are built on sparse and chaotic points processed by the radar's onboard system. In open environments with few points, such methods suffer significant performance degradation or even complete failure. To address these challenges, we propose SDDiff, the first method, to the best of our knowledge, that enables both dense point cloud extraction (PCE) and accurate 3D ego velocity estimation (EVE) directly from the raw ADC data of a single-chip radar.

3 Spatial-Doppler Diffusion Model

In this section, we present the Directional Spatial-Doppler Diffusion Model with radar priors in Sec. 3.1 and Iterative Doppler Refinement in Sec. 3.2, followed by an overview of SDDiff model as illustrated in Sec. 3.3.

3.1 Directional diffusion with radar priors

Spatial-Doppler Domain Representation We focus on radar's information-rich ADC data rather than the sparse points from onboard systems. To simultaneously achieve dense PCE and accurate EVE, we introduce a Spatial-Doppler Domain Representation (SDDR) designed around two key principles: 1) Capturing both spatial occupancy and Doppler velocity with minimal information loss. 2) Eliminating redundant data to reduce computational cost and memory usage when feasible. Specifically, we transform raw radar data into a 4D tensor $\boldsymbol{C'} \in \mathbb{R}^{R\times A\times E\times D}$ via fast Fourier transform (FFT), where the dimensions correspond to range, azimuth, elevation, and Doppler. Higher values in the tensor typically indicate a higher likelihood of a point's existence. For a spatial position $\boldsymbol{s}_{k,i,j} = (r_k, a_i, e_j)$, the Doppler velocity $v^r_{k,i,j}$ is uniquely determined. Empirically, the maximum value along the Doppler axis is typically at least 6× greater than the second when a point exists at $\boldsymbol{s}_{k,i,j}$. Using this property, we extract the indices of peak values along the Doppler axis to determine Doppler velocities as shown in Fig. 3. Subsequently, intensity $\boldsymbol{u}$ and Doppler velocity $\boldsymbol{v}$ are concatenated into a polar SDDR $\boldsymbol{C} = [\boldsymbol{u}; \boldsymbol{v}] \in \mathbb{R}^{R\times A\times E\times 2}$, where intensity represents the spatial occupancy and the index corresponds to the Doppler velocity. This can significantly reduce the data volume and computational overhead.

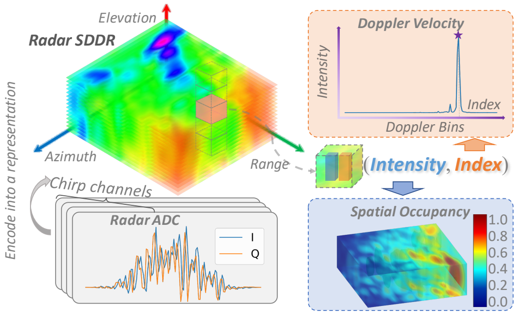

Figure 3: Diagram of Spatial-Doppler Domain Representation.

Similarly, LiDAR's spatial occupancy can be represented by wrapping the points into a polar format consistent with the radar configuration. Refining PCE and EVE is defined as the SDDR Purification Process, transforming coarse, ghost-prone representations into fine, uncontaminated ones. The conventional diffusion process typically transforms the target distribution into a standard Gaussian distribution over long steps, and subsequently iteratively samples a new target starting from the Gaussian noise given conditional embeddings. However, it requires a plethora of sampling steps and produces ambiguous outcomes for the SDDR purification task. In response, we define a directional diffusion process, starting from an initial state with a homogeneous radar representation. The forward and reverse processes are comprehensively presented and analyzed as follows.

Forward process Similar to DDPM [Ho *et al.*, 2020], we construct a parameterized Markov chain $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ to model the transition between radar and LiDAR distributions as shown in Fig. 4. Given radar's spatial occupancy $\boldsymbol{u}_\mathbf{0}$ aligned with LiDAR's $\boldsymbol{x}_\mathbf{0}$, the forward diffusion process is defined as gradually adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ to the target representation along the direction toward the radar prior $\boldsymbol{u}_\mathbf{0}$ as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{u}_0) := \mathcal{N}(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_{t-1} + (1-\alpha_t)\boldsymbol{u}_0, \lambda_t^2 \mathbf{I}) \quad (1)$$

where α_t, λ_t denote pre-defined weight, variance schedule.

Theorem 1. *Given the directional diffusion process with control signal $\boldsymbol{u}_\mathbf{0}$, sampling the state $\boldsymbol{x}_t$ at an arbitrary timestep t are expressed in closed form:*

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{u}_0) := \mathcal{N}(\boldsymbol{x}_t; \prod_{k=1}^{t} \alpha_k \boldsymbol{x}_0 + (1 - \prod_{k=1}^{t} \alpha_k)\boldsymbol{u}_0, \beta_t^2 \mathbf{I}) \quad (2)$$

where $\beta_t^2 = \sum_{k=1}^{t} \bar{\alpha}_t^2 \lambda_k^2 / \bar{\alpha}_k^2$ and $\bar{\alpha}_n = \prod_{i=1}^{n} \alpha_i$.

The proof of Theorem 1 is included in the appendices of the supplementary material. Eqn. 1 shows that directional diffusion process starts sampling from the radar prior distribution and diffuses toward the LiDAR SSDR. This facilitates fewer sampling steps and mitigates ambiguous outcomes.

Reverse process The key of the reverse process is to estimate the posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{u}_0)$ with the radar prior $\boldsymbol{u}_0$. Following most of the literature on generative models [Rombach *et al.*, 2022], we aim to choose a tractable distribution $\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \boldsymbol{u}_0), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{u}_0))$ by optimizing the variational bound on negative log likelihood of $p(\boldsymbol{x}_0|\boldsymbol{u}_0)$.

$$\min_\theta \sum_{t=1}^{T} \mathbb{E}\left[D_{KL}\left[q\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{u}_0\right) || p_\theta\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{u}_0\right)\right]\right] \quad (3)$$

where $D_{KL}[\cdot||\cdot]$ denotes the Kullback-Leibler (KL) divergence.

Theorem 2. *For directional diffusion, with the forward process defined in Eqn. 2, the posterior distribution of the latent variable is tractable when conditioned on $\boldsymbol{x}_0$ and radar's prior $\boldsymbol{u}_0$, and is given in explicit form as follows:*

$$p\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{u}_0\right) \propto \exp\left(-\frac{\left(\boldsymbol{x}_{t-1} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{t}}(\boldsymbol{x}_{0,t}, \boldsymbol{u}_0)\right)^2}{2\sigma_t^2}\right) \quad (4)$$

$$\textit{where } \tilde{\boldsymbol{\mu}}_{\boldsymbol{t}} = \frac{\alpha_t \beta_{t-1}^2}{\beta_t^2}\boldsymbol{x}_t + \frac{\beta_t^2 - \alpha_t \beta_{t-1}^2 - \lambda_t^2 \bar{\alpha}_{t-1}}{\beta_t^2}\boldsymbol{u}_0 \quad (5)$$
$$+ \lambda_t^2 \bar{\alpha}_{t-1} \boldsymbol{x}_0 / \beta_t^2 \quad \textit{and} \quad \sigma_t^2 = \lambda_t^2 \beta_{t-1}^2 / \beta_t^2$$

Derivation details on Theorem 2 are provided in the supplementary material appendix due to space limitations.

According to Eqn. (2)(4)(5), the mean $\tilde{\boldsymbol{\mu}}_{\boldsymbol{t}}(\boldsymbol{x}_t, \boldsymbol{u}_0)$ of the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{u}_0)$ conditioned by the previous latent state $\boldsymbol{x}_t$ and the radar prior $\boldsymbol{u}_0$ is further parameterized as follows:

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{u}_0) = \frac{1}{\alpha_t}\boldsymbol{x}_t + \frac{\alpha_t - 1}{\alpha_t}\boldsymbol{u}_0 + \frac{\lambda_t^2}{\alpha_t \beta_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}} \quad (6)$$

Therefore, given homogeneous radar prior $\boldsymbol{u}_0$, the objective function in Problem 3 is simplified as:

$$\mathcal{L}_{\text{Spatial}} = \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{u}_0}\left[\frac{\lambda_t^2}{2\alpha_t^2 \beta_{t-1}^2}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{u}_0)||_2^2\right] \quad (7)$$

3.2 Iterative Doppler Refinement

Point cloud density variations across indoor and outdoor environments and multi-path-induced ghost points hinder model generalization. Drawing from Fig. 2, we design Iterative Doppler Refinement to mitigate ghost artifacts while progressively enhancing spatial representations for more accurate ego-motion estimation. Specifically, Doppler velocity, *i.e.* relative radial velocity $v_{i,j}^r$, is determined by the radar's ego-velocity $\boldsymbol{v}^{ego}$ and the target's azimuth a_i and elevation angle e_j when the target is stationary relative to the ground.

$$v_{i,j}^r = [\cos a_i \cos e_j, \sin a_i \cos e_j, \sin e_j]\left[v_x^{ego}, v_y^{ego}, v_z^{ego}\right]^T \quad (8)$$

According to the Cauchy-Schwarz inequality, the peak height of the distribution surface of v^r *w.r.t.* a_i and e_j corresponds to the magnitude of the radar's velocity, while the peak position indicates the azimuth and elevation angles of the radar's motion as shown in Fig. 4(d). Considering that static targets dominate the radar field of view in scene-level PCE, we utilize Doppler-consistency as a critic to refine spatial occupancy. Based on Eqn. 8 we train a differentiable model $f_\psi(\boldsymbol{v})$ for velocity regulation. Given a progressively refined spatial representation $\boldsymbol{x}_t$, a reduction operation along range followed by softmax is applied to obtain a soft mask $\boldsymbol{M}_t$. Further, we define a Doppler-consistency loss to jointly refine spatial occupancy and ego velocity.

$$\mathcal{L}_{\text{Doppler}} = \mathop{\mathbb{E}}_{\boldsymbol{M}_t, \boldsymbol{v}_t}\left[\frac{\lambda_t^2 \bar{\alpha}_{t-1}^2}{2\beta_t^2 \beta_{t-1}^2}||\boldsymbol{v}^{ego} - f_\psi(\boldsymbol{M}_t \odot \boldsymbol{v})||_2^2\right] \quad (9)$$

where $\odot$ donates Hadamard product, the weights are determined by Eqn. 5.

3.3 Overview of Our SDDiff

Fig. 4 presents the overview of our SDDiff. During training, the LiDAR representations are softened with a Gaussian kernel to transition to a probabilistic spatial occupancy $\boldsymbol{x}_0$. Subsequently, they and radar priors $\boldsymbol{u}_0$ are fed into the directional spatial-Doppler diffusion module. We adopt SDDNet as the denoising model, designed on a standard 3D U-Net architecture with paired up-down blocks incorporating ResNet and attention mechanism. To optimize computational efficiency, radar Doppler profiles $\boldsymbol{v}$ are embedded *w.r.t.* chirp channels before injected into each 3D U-Net layer for cross-attention. Subsequently, Iterative Doppler Refinement with the latent variable $\boldsymbol{x}_t$ and Doppler profile $\boldsymbol{v}$ as inputs, is initialized to mitigate density variations and ghosting effects. The overall training loss for SDDNet is formulated as:

$$\mathcal{L}_{\text{SDDNet}} = \mathcal{L}_{\text{Spatial}} + \omega \mathcal{L}_{\text{Doppler}} \quad (10)$$

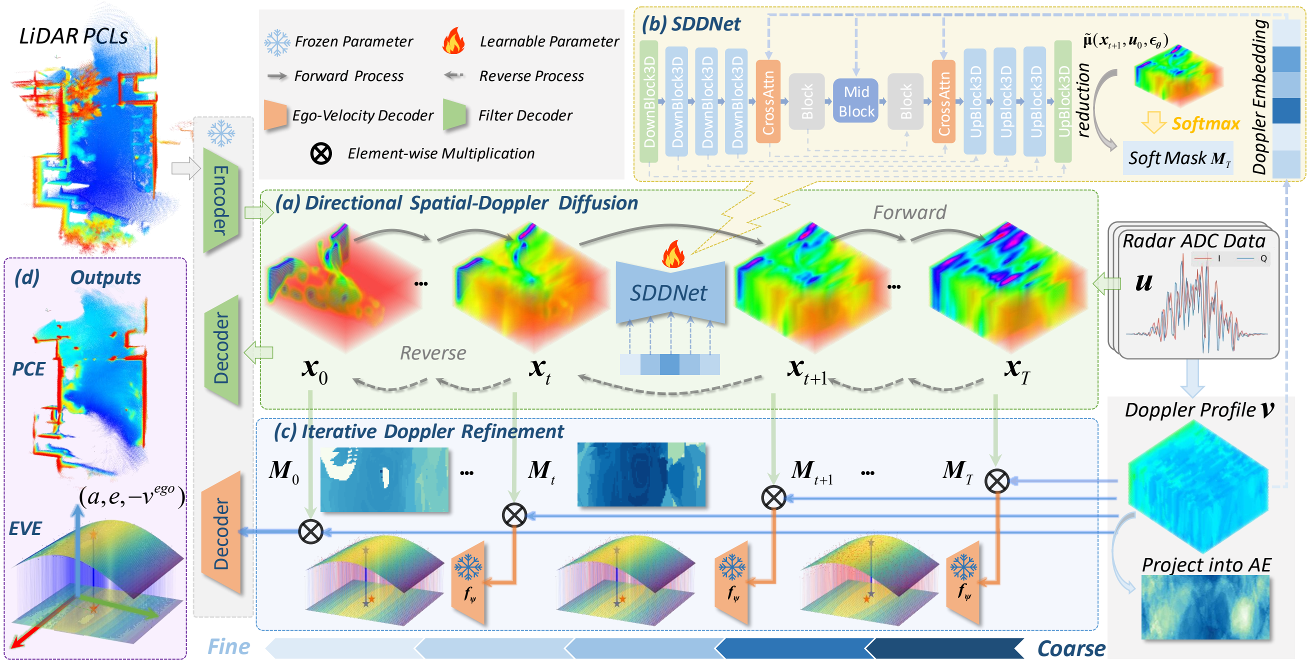

Figure 4: The overview of our SDDiff.

where ω denotes the weight balancing spatial and Doppler losses. Notably, only the parameters of SDDNet are optimized during training. The inference process resembles Stable Diffusion, iteratively sampling the SDDR through Markov chains. The refined SDDR is subsequently employed for ego velocity estimation.

4 Experiments Setup

4.1 Dataset

We evaluate the proposed method using both the publicly available Coloradar dataset and a self-collected dataset across different indoor and outdoor scenarios.

ColoRadar Dataset. We conduct our method on ColoRadar Dataset [Kramer *et al.*, 2022], which includes raw ADC sample data for single-chip radar, dense LiDAR point clouds, and odometry pose information. We use LiDAR point clouds as the ground truth for PCE and velocity solved through odometry for EVE. The ColoRadar contains both indoor and outdoor scenes with a total of 52 sequences. For a fair comparison with other learning-based baselines, we select the same 36 sequences as the training set and others for testing.

Self-Collected Dataset. We collected real-world data from diverse environments using the platform shown in Fig. 5 to assess the model's generalization. The dataset comprises 10,371 frames, with 10% used for fine-tuning and 90% for testing. Reliable odometry for EVE ground truth was obtained using Fast-Livo [Zheng *et al.*, 2022].

4.2 Implementation Details

We implement our SDDiff using Pytorch 1.11.0 with CUDA 12.4. The parameters ω of the weighted spatial and Doppler loss are set to 0.01. We apply Gaussian filtering with sigma values of [0.2, 0.5, 1] as the filter encoder and perform offline training for the associated decoder. The Ego-Velocity Decoder is trained according to Eqn. 8, with the Azimuth-Elevation (AE) Doppler profile as input. The forward process variances are set to constants increasing linearly from $\bar{\alpha}_1 = 0.01$ to $\bar{\alpha}_T = 0.99$ and the noise scale λ_k is set to 0.1. To illustrate the effectiveness of the directional diffusion strategy, we set the sampling step size $T = 20$. We train SDDNet for 100 epochs on ColoRadar Dataset with AdamW optimizer and a learning rate 10^{-4}. It takes about 5 days to train our model with a machine using three NVIDIA GeForce RTX 4090 GPUs and Intel Xeon Gold 6226R CPU.

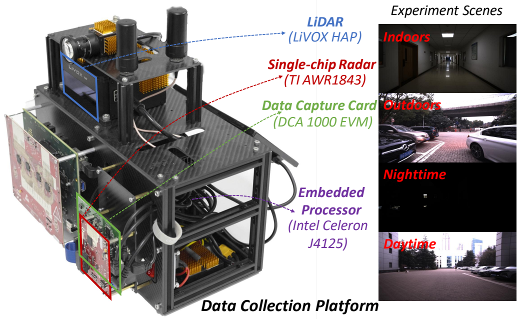

Figure 5: Our customized hand-held data collection platform.

4.3 Evaluation Metric

Baseline. We evaluate the PCE performance of SDDiff compared to traditional method OS-CFAR [1988], as well as generative methods RPDNet [2022], RadarHD [2023] and

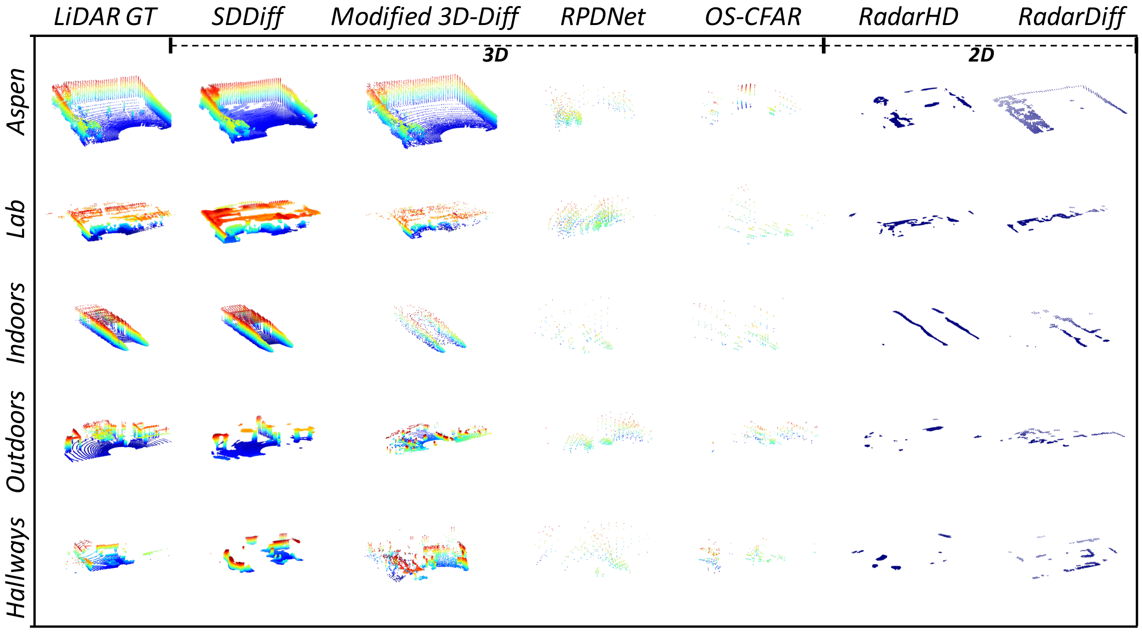

Figure 6: Qualitative results of different methods and ours on the ColoRadar dataset.

RadarDiff [2024]. Since no generative models exist for scene-level 3D PCE, we modify the naive diffusion model with radar spatial occupancy as input for comparison with our approach, denoted as modified 3D-Diff. For EVE performance evaluation, we compare our method with ICP [1992], RANSAC [2013], and RadarEVE [2024].

PCE Metrics. We use Earth Mover's Distance (EMD) and Chamfer Distance (CD) to quantify the similarity between the extracted point clouds $\mathcal{P}$ and the ground truth $\mathcal{Q}$. Additionally, we adopt the definition of clutter points as provided in RPDNet [Cheng *et al.*, 2022], as follows:

$$\mathcal{P}_{\text{clutter}} = \{\boldsymbol{p} \in \mathcal{P},\ s.t,\ d\left(\boldsymbol{q},\boldsymbol{p}\right) > \tau_1,\ \forall \boldsymbol{q} \in \mathcal{Q}\} \tag{11}$$

where $d\left(\boldsymbol{q},\boldsymbol{p}\right)$ denotes the Euclidean Distance between $\boldsymbol{p}$ and $\boldsymbol{q}$. Similarly, we define the 'shot' LiDAR points as those generated points that can effectively represent the scene.

$$\mathcal{Q}_{\text{shot}} = \{\boldsymbol{q} \in \mathcal{Q},\ \exists \boldsymbol{p} \in \mathcal{P},\ s.t,\ d\left(\boldsymbol{p},\boldsymbol{q}\right) < \tau_2\} \tag{12}$$

We define scene representation level and valid point ratio for point cloud generation quality assessment, and effective generation density for density evaluation.

- *Valid Point Ratio (VPR)* is defined as $1 - |\mathcal{P}_{\text{clutter}}|/|\mathcal{P}|$, reflecting the reliability of the generated point clouds.
- *Scene Representation Level (SRL)* is defined as $|\mathcal{Q}_{\text{shot}}|/|\mathcal{Q}|$, quantifying the effectiveness of the point cloud generation.
- *Effective Generation Density (EGD)* is characterized as $|(\mathcal{P} - \mathcal{P}_{\text{clutter}}|)/|\mathcal{Q}_{\text{shot}}|$, referring to the density of the point clouds.

EVE Metrics. We evaluate EVE using Mean Absolute Error (MAE) and cumulative velocity error density in both indoor and outdoor scenarios.

5 Results and Analysis

5.1 Point Cloud Extraction Results

Qualitatively, we showcase the PCE results of various approaches, as shown in Fig. 6. Our method demonstrates superior density and accuracy in PCE compared to alternatives. Quantitatively, we compare the point clouds generated by different schemes on the Coloradar Dataset against the LiDAR ground truth, using metrics EMD and CD. As presented in Tab. 1, points generated by our method exhibit the closest alignment with the LiDAR ground truth. We also evaluate our method on the self-collected dataset to demonstrate its generalization. As shown in Tab. 2, our approach consistently outperforms others when applied to new scenarios.

Reliability and Effectiveness Analysis: Different models show variations in the number of points and clutter at different thresholds. To comprehensively assess the reliability and effectiveness of the generated point clouds, we evaluate the clutter rate, representation level, and density of various PCE methods. As illustrated in Fig. 7, 2D methods exhibit high VPR and low SRL due to the lack of elevation information. our method achieves a balanced trade-off between effectiveness and reliability, outperforming the state-of-the-art baseline with a 30% improvement in VPR and a 33% increase in SRL. Moreover, our method ensures the highest valid point cloud density compared to other approaches. This is attributed to the reciprocal benefit mentioned in Fig. 2, which enhances the model's robustness to density variations and ghosting effects.

5.2 Ego Velocity Estimation Results

As shown in Tab. 3, our method improves EVE by 30% and 59% over the state-of-the-art in indoor and outdoor scenarios, respectively. This is attributed to our PCE process, which

Object Dense Detection Methods	Classroom		Armyroom		Hallways		Labroom		Aspen_room		Outdoors		Longboard	
	EMD ↓	CD ↓	EMD ↓	CD ↓	EMD ↓	CD ↓	EMD	CD ↓	EMD ↓	CD ↓	EMD ↓	CD ↓	EMD ↓	CD↓
OS-CFAR [1988]	1.03	1.09	1.06	1.05	1.15	1.18	1.07	1.29	1.19	1.28	1.54	1.37	4.26	3.57
RPDNet [2022]	0.97	0.85	1.31	1.07	1.40	1.10	1.09	1.01	1.30	1.08	1.69	1.24	2.29	1.67
RadarHD† [2023]	0.57	0.53	0.77	0.76	1.63	1.29	1.40	1.11	0.85	0.77	3.20	2.28	4.55	3.95
RadarDiff† [2024]	0.72	0.54	0.87	0.68	1.46	1.03	1.13	0.83	1.32	0.89	2.14	1.45	4.16	2.73
Modified 3D-Diff	0.58	0.56	0.77	0.77	0.85	0.84	0.90	0.89	0.88	0.88	1.32	1.29	1.20	1.12
SDDiff (Ours)	**0.25**	**0.24**	**0.31**	**0.31**	**0.56**	**0.64**	**0.47**	**0.53**	**0.38**	**0.38**	**0.75**	**0.80**	**0.81**	**0.88**

Table 1: Point Cloud Extraction results on the Coloradar Dataset test split. † indicates 2D methods for distinction.

Object Dense Detection Methods	Indoors						Outdoors					
	EMD↓ (50%)	CD↓ (50%)	MHD↓ (50%)	EMD↓ (90%)	CD↓ (90%)	MHD↓ (90%)	EMD↓ (50%)	CD↓ (50%)	MHD↓ (50%)	EMD↓ (90%)	CD↓ (90%)	MHD↓ (90%)
OS-CFAR [1988]	0.78	1.01	0.88	1.23	1.62	1.54	2.32	2.21	2.36	4.72	4.57	5.51
RPDNet [2022]	0.96	0.91	0.83	1.46	1.53	1.34	1.51	1.39	1.30	2.63	2.01	2.49
RadarHD† [2023]	0.55	0.51	0.36	1.03	0.89	0.64	1.91	1.53	1.57	4.43	3.24	5.32
RadarDiff† [2024]	0.54	0.77	0.31	1.22	2.01	1.56	1.63	1.68	1.68	3.35	3.25	3.33
Modified 3D-Diff	0.42	0.41	0.33	1.14	0.98	0.81	0.79	0.76	0.64	2.05	1.16	1.86
SDDiff (Ours)	**0.33**	**0.33**	**0.21**	**0.79**	**0.79**	**0.44**	**0.62**	**0.61**	**0.50**	**0.85**	**0.84**	**0.75**

Table 2: Point Cloud Similarity Comparison at different percentiles on datasets collected from real-world scenarios.

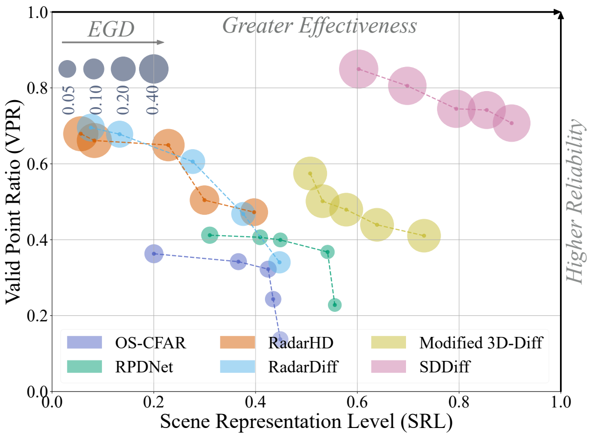

Figure 7: The evaluation results for effectiveness, reliability, and density of PCE. A larger scatter radius corresponds to a higher EGD.

provides more accurate and reliable point clouds for EVE. In contrast, other methods degrade outdoors due to relying on the sparse points processed by the data capture card.

EVE Methods	MAE↓ (· m/s)	Cumulative Density of Error/% (Indoors/Outdoors) ↑				
		≤0.05m/s	≤0.1m/s	≤0.15m/s	≤0.2m/s	≤0.25m/s
ICP	0.60/0.77	5/6	12/10	17/15	23/20	29/25
RANSAC	0.31/0.57	**35**/13	55/27	63/37	66/44	67/47
RadarEVE	0.13/0.27	29/8	51/14	66/22	79/36	87/50
SDDiff	**0.09/0.11**	31/**29**	**59/55**	**78/74**	**89/86**	**95/93**

Table 3: The EVE performance results

5.3 Ablation Study

We conduct an ablation study to validate the gains introduced by PCE module *i.e.* Directional Spatial-Doppler Diffusion and EVE module *i.e.* Iterative Doppler Refinement. As shown in Tab. 4, the PCE without Doppler information incurs a 15% reduction in PCE performance compared to the whole SDDiff. Nevertheless, it still outperforms Modified 3D-Diff, which suffers from ambiguous outcomes due to its diffusion scheme. Moreover, it speeds up inference by 3.13× compared to Modified 3D-Diff, owing to the more efficient sampling prior distribution. The EVE without reliable spatial occupancy leads to a 1.2× and 3.4× increase in indoor and outdoor errors, respectively. This demonstrates that SDDiff effectively harnesses the potential gains from the synergy between PCE and EVE.

PCE Module	EVE Module	VPR (%)↑	SRL (%)↑	EGD (×)↑	V.E. (m/s)↓	Speed (×)↑
✔	✗	65.5	69.2	0.98	-	3.13
✗	✔	-	-	-	0.11/0.37	-
✔	✔	77.1	79.6	1.17	0.09/0.11	2.57

Table 4: Ablation quantitative results on EVE and PCE.

6 Conclusion

This paper introduces SDDiff, a directional Spatial-Doppler diffusion model to simultaneously enable dense point cloud extraction and accurate ego velocity estimation. To reduce sampling wastage and mitigate ambiguous outcomes, we design a directional diffusion with radar priors to streamline the sampling. Additionally, we design Iterative Doppler Refinement to defend against density variations and ghosting effects. The experiments demonstrate that SDDiff enhances EVE accuracy while improving both the effectiveness and reliability of PCE.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China with Grant 62471194, and the Fundamental Research Funds for the Central Universities.

References

[Adolfsson *et al.*, 2021] Daniel Adolfsson, Martin Magnusson, Anas Alhashimi, Achim J Lilienthal, and Henrik Andreasson. Cfear radarodometry-conservative filtering for efficient and accurate radar odometry. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5462–5469. IEEE, 2021.

[Ausherman *et al.*, 1984] Dale A Ausherman, Adam Kozma, Jack L Walker, Harrison M Jones, and Enrico C Poggio. Developments in radar imaging. *IEEE Transactions on Aerospace and Electronic Systems*, (4):363–400, 1984.

[Besl and McKay, 1992] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[Biber and Straßer, 2003] Peter Biber and Wolfgang Straßer. The normal distributions transform: A new approach to laser scan matching. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 3, pages 2743–2748. IEEE, 2003.

[Blake, 1988] Stephen Blake. Os-cfar theory for multiple targets and nonuniform clutter. *IEEE transactions on aerospace and electronic systems*, 24(6):785–790, 1988.

[Cen and Newman, 2018] Sarah H Cen and Paul Newman. Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6045–6052. IEEE, 2018.

[Cen and Newman, 2019] Sarah H Cen and Paul Newman. Radar-only ego-motion estimation in difficult settings via graph matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 298–304. IEEE, 2019.

[Censi, 2008] Andrea Censi. An icp variant using a point-to-line metric. In *2008 IEEE international conference on robotics and automation*, pages 19–25. Ieee, 2008.

[Cheng *et al.*, 2022] Yuwei Cheng, Jingran Su, Mengxin Jiang, and Yimin Liu. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics*, 38(6):3754–3773, 2022.

[Fan *et al.*, 2024] Cong Fan, Shengkai Zhang, Kezhong Liu, Shuai Wang, Zheng Yang, and Wei Wang. Enhancing mmwave radar point cloud via visual-inertial supervision. *arXiv preprint arXiv:2404.17229*, 2024.

[Fischler and Bolles, 1981] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Guan *et al.*, 2020] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through fog high-resolution imaging using millimeter wave radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11464–11473, 2020.

[Harlow *et al.*, 2024] Kyle Harlow, Hyesu Jang, Timothy D Barfoot, Ayoung Kim, and Christoffer Heckman. A new wave in robotics: Survey on recent mmwave radar applications in robotics. *IEEE Transactions on Robotics*, 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Kellner *et al.*, 2013] Dominik Kellner, Michael Barjenbruch, Jens Klappstein, Jürgen Dickmann, and Klaus Dietmayer. Instantaneous ego-motion estimation using doppler radar. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 869–874. IEEE, 2013.

[Kellner *et al.*, 2014] Dominik Kellner, Michael Barjenbruch, Jens Klappstein, Jürgen Dickmann, and Klaus Dietmayer. Instantaneous ego-motion estimation using multiple doppler radars. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1592–1597. IEEE, 2014.

[Kramer *et al.*, 2022] Andrew Kramer, Kyle Harlow, Christopher Williams, and Christoffer Heckman. Coloradar: The direct 3d millimeter wave radar dataset. *The International Journal of Robotics Research*, 41(4):351–360, 2022.

[Lai *et al.*, 2024] Haowen Lai, Gaoxiang Luo, Yifei Liu, and Mingmin Zhao. Enabling visual recognition at radio frequency. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 388–403, 2024.

[Li *et al.*, 2003] Jian Li, Petre Stoica, and Zhisong Wang. On robust capon beamforming and diagonal loading. *IEEE transactions on signal processing*, 51(7):1702–1715, 2003.

[Lu *et al.*, 2020] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pages 14–27, 2020.

[Luan *et al.*, 2024] Kai Luan, Chenghao Shi, Neng Wang, Yuwei Cheng, Huimin Lu, and Xieyuanli Chen. Diffusion-based point cloud super-resolution for mmwave radar data.

In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11171–11177, 2024.

[Nitzberg, 1972] Ramon Nitzberg. Constant-false-alarm-rate signal processors for several types of interference. *IEEE Transactions on Aerospace and Electronic Systems*, (1):27–34, 1972.

[Pang *et al.*, 2024] Changsong Pang, Xieyuanli Chen, Yimin Liu, Huimin Lu, and Yuwei Cheng. Radarmoseve: A spatial-temporal transformer network for radar-only moving object segmentation and ego-velocity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4424–4432, 2024.

[Prabhakara *et al.*, 2023] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4135–4142. IEEE, 2023.

[Qian *et al.*, 2020] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 3d point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–23, 2020.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Roy and Kailath, 1989] Richard Roy and Thomas Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 37(7):984–995, 1989.

[Schmidt, 1986] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[Segal *et al.*, 2009] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009.

[Sun *et al.*, 2021] Yue Sun, Zhuoming Huang, Honggang Zhang, Zhi Cao, and Deqiang Xu. 3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 1–8. IEEE, 2021.

[Xu *et al.*, 2021] Jingao Xu, Guoxuan Chi, Zheng Yang, Danyang Li, Qian Zhang, Qiang Ma, and Xin Miao. Followupar: Enabling follow-up effects in mobile ar applications. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 1–13, 2021.

[Yataka *et al.*, 2024] Ryoma Yataka, Pu Wang, Petros Boufounos, and Ryuhei Takahashi. Sira: Scalable interframe relation and association for radar perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15024–15034, 2024.

[Zhang *et al.*, 2022] Mingjin Zhang, Chengyu He, Jing Zhang, Yuxiang Yang, Xiaoqi Peng, and Jie Guo. Sar-to-optical image translation via neural partial differential equations. In *IJCAI*, pages 1644–1650, 2022.

[Zhang *et al.*, 2024] Ruibin Zhang, Donglai Xue, Yuhan Wang, Ruixu Geng, and Fei Gao. Towards dense and accurate radar perception via efficient cross-modal diffusion model. *IEEE Robotics and Automation Letters*, 9(9):7429–7436, 2024.

[Zheng *et al.*, 2022] Chunran Zheng, Qingyan Zhu, Wei Xu, Xiyuan Liu, Qizhi Guo, and Fu Zhang. Fast-livo: Fast and tightly-coupled sparse-direct lidar-inertial-visual odometry. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4003–4009. IEEE, 2022.