2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

RadarSFD: Single-Frame Diffusion with Pretrained Priors for Radar Point Clouds

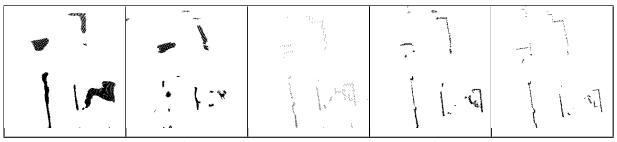
Bin Zhao, Nakul Garg

***Abstract*— Millimeter-wave radar provides perception robust to fog, smoke, dust, and low light, making it attractive for size, weight, and power constrained robotic platforms. Current radar imaging methods, however, rely on synthetic aperture or multi-frame aggregation to improve resolution, which is impractical for small aerial, inspection, or wearable systems. We present *RadarSFD*, a conditional latent diffusion framework that reconstructs dense LiDAR-like point clouds from a single radar frame without motion or SAR. Our approach transfers geometric priors from a pretrained monocular depth estimator into the diffusion backbone, anchors them to radar inputs via channel-wise latent concatenation, and regularizes outputs with a dual-space objective combining latent and pixel-space losses. On the RadarHD benchmark, *RadarSFD* achieves state-of-the-art performance against baseline models. Qualitative results show recovery of fine walls and narrow gaps, and experiments across new environments confirm strong generalization. Ablation studies highlight the importance of pretrained initialization, radar BEV conditioning, and the dual-space loss. Together, these results establish the first practical single-frame, no-SAR mmWave radar pipeline for dense point cloud perception in compact robotic systems. The project page is available at https://phi-lab-rice.github.io/RadarSFD/**

I. Introduction

Millimeter-wave (mmWave) radar is emerging as a practical alternative to LiDAR for robotic perception. Unlike LiDAR, mmWave signals penetrate fog, smoke, dust, and low light, making them robust in conditions where optical sensors struggle [1], [2]. This robustness is especially attractive for platforms with tight size, weight, and power (SWaP) budgets, such as autonomous inspection drones in confined spaces, compact robots for search-and-rescue, or wearable devices [3], [4], [5], [6]. These platforms cannot afford bulky rotating rigs, linear actuators, or multi-view scanning mechanisms. Instead, they need compact modules that can deliver single-frame spatial perception without any ego-motion or synthetic aperture. In this work, we ask: *can a single radar capture yield a dense, LiDAR-like point cloud suitable for robotic perception?*

Traditional radar imaging faces a fundamental limitation: small apertures result in low angular resolution [9]. Classical processing pipelines such as beamforming and CFAR (Constant False Alarm Rate) extract reliable points [10], but the resulting point clouds remain sparse and miss fine structures. To improve the resolution, existing approaches use synthetic aperture radar (SAR), which requires carefully controlled ego-motion or temporal stacking of multiple frames. Recently, learning-based methods have advanced the field significantly. For instance, RadarHD [7] introduced an asymmetric VAE that collects 40 consecutive past frames to produce denser point clouds. Luan et al. [11] reduces the temporal stacks to 5 frames, while Zhang et al. [8] adapts diffusion to 2D (Bird's eye view) BEV images. These are important works that demonstrate the promise of AI-based radar perception. However, as we show in Figure 1, when temporal stacks or SAR are removed, RadarHD degrades significantly under the single-frame setting, producing blurred and fragmented outputs. The result from Zhang et al., while single-frame, shows visibly coarser resolution compared to LiDAR. In contrast, our approach recovers fine walls and gaps from a single frame without sacrificing radar resolution, showing that single-frame, no-SAR radar remains an open and impactful challenge.



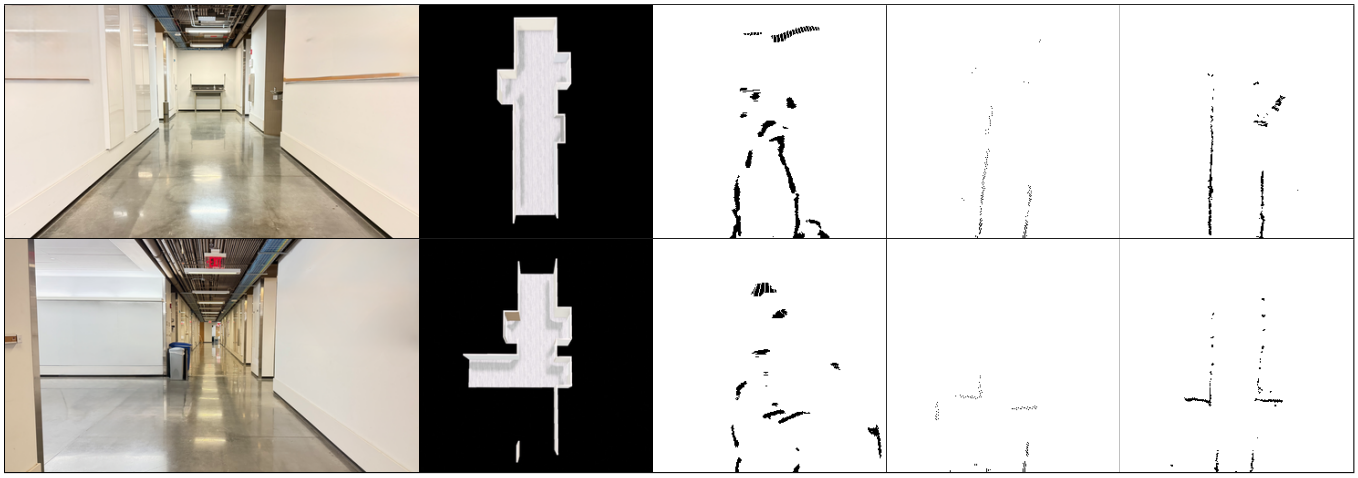
(a) (b) (c) (d) (e)

Fig. 1: Estimated point cloud reconstructions: (a) RadarHD (41 input frames) [7], (b) RadarHD single-frame, (c) Zhang et al. [8], (d) *RadarSFD* (Ours), and (e) LiDAR ground truth.

Our approach builds on the intuition that radar alone does not need to learn the entire structure of the world from scratch. Instead, we can transfer the structural priors encoded in large pretrained generative models. Recent works in monocular depth estimation show that diffusion models like Marigold [12] learn strong geometric priors about walls, corners, and object boundaries. Inspired by this, we propose to condition a latent diffusion model on the raw radar BEV. The pretrained priors act as a "world model", while the radar input anchors these priors to the actual scene. Similar cross-modality transfer has shown success in domains such as underwater restoration [13], supporting the idea that pretrained diffusion backbones can be adapted across sensing modalities. We also build on prior work's insight that lightly thresholded radar BEVs preserve weak reflections and sidelobes that carry useful information [7]. Together, these insights form the basis of our method.

We introduce *RadarSFD* (Radar Single-Frame Diffusion), a conditional latent diffusion pipeline for single-frame radar-to-LiDAR translation. Our model operates in the latent space of a frozen Stable Diffusion VAE for efficiency. We

Bin Zhao and Nakul Garg are with Rice University, Houston TX 77005 USA. Email:{bz35@rice.edu, nakul@rice.edu}.



(a) RGB image of unseen building (b) 3D layout BEV (c) RadarHD [7] (d) Zhang et al. [8] (e) *RadarSFD* (ours)

Fig. 2: Real world test for generalization with completely unseen data in our campus building. All models are trained on the same radar dataset from RadarHD [7]: (a) RGB image of unseen environment, (b) 3D floor plan layouts, (c) RadarHD single-frame baseline, (d) Zhang et al. diffusion baseline [8], and (e) *RadarSFD* (ours) single-frame latent diffusion.

initialize the U-Net from Marigold to inject pretrained depth and shape priors. Conditioning is applied by channel-wise concatenation of radar BEV latents and noisy LiDAR BEV latents, aligning the input geometry directly with the generative process. To ensure scene faithfulness, we combine the standard latent space MSE loss with pixel-space L1, SSIM [14], and LPIPS [15]. This design mitigates the common failure mode of diffusion models – hallucination/producing plausible but incorrect scenes – by tethering generations to the specific radar input.

To evaluate, we use Chamfer distance (CD) [16] and Modified Hausdorff distance (MHD) [17], which measure point-to-point and structural similarity between generated point clouds and the LiDAR ground truth. Compared to the single-frame RadarHD baseline, our method reduces CD from 56 cm to 35 cm and MHD from 45 cm to 28 cm, showing much sharper reconstructions from a single radar frame. Our performance is also close to the recent Zhang et al. diffusion method [8], which reports 38 cm CD and 29 cm MHD, but importantly that approach achieves its results by lowering range resolution, whereas our method maintains the native 4 cm resolution. Further, even against the multi-frame RadarHD with 41 stacked frames, our method remains competitive and slightly better, with about a 20% reduction in CD and an 18% reduction in MHD. As Figure 1 shows, this enables us to recover fine walls and narrow gaps even in static, no-SAR settings. Figure 2 further demonstrates strong generalization across new environments. Our ablation studies confirm that pretrained initialization, BEV concatenation, and the dual-space loss are key to achieving this performance.

In summary, this paper makes three contributions:

- A no-SAR radar perception pipeline that generates dense, LiDAR-like point clouds from single-frame radar, suitable for SWaP-constrained platforms.
- A conditional latent diffusion model that transfers monocular-depth priors into radar-to-LiDAR translation and achieves state-of-the-art accuracy.
- A dual-space objective function and ablation study to make design choices in pretraining, conditioning, and input representation, which provides a practical reproducible recipe for radar-to-LiDAR point clouds.

II. Related Work

Conventional mmWave radar signal processing pipelines start with transforming the raw I/Q signals into a Range-Doppler-Angle data cube. Standard angle estimation is done via beamforming with an FFT across the antenna array, yet its angular resolution is fundamentally limited by the physical aperture, a constraint known as the Rayleigh limit [18]. To surpass this limitation, super-resolution algorithms like MVDR, ESPRIT and MUSIC are often employed, providing much finer angular estimates at the cost of higher computational complexity [19], [20], [21], [22]. Separately, the range resolution is constrained by the chirp bandwidth, which can cause objects close in distance to merge into a single detection [23], [24]. After the data cube is formed, CFAR algorithms are typically applied to detect targets by dynamically setting thresholds against the local clutter [10], [25], [26], [27], [28].

To address these resolution limitations, synthetic aperture radar (SAR) techniques are commonly employed [29], [30], [31]. While SAR significantly improves angular resolution, it requires precise motion control and temporal accumulation. This makes SAR impractical for SWaP-platforms or real-time applications where motion is constrained.

Recent advances in machine learning have opened new avenues in radar perception enhancement. RadarHD [7] pioneered the use of asymmetric variational autoencoders (VAE) to transform sparse radar range-azimuth heatmaps into dense, LiDAR-like representations. RadarHD utilizes lightly thresholded 2D range-azimuth BEV images from range and angle FFT outputs, using temporal stacks of 40 frames as synthetic aperture and dense LiDAR BEV

images as supervision. While effective, the VAE architecture produces relatively blurry outputs compared to more advanced generative models. Similar VAE-based approaches have demonstrated deployment feasibility on UAV platforms [32], though they still rely on multi-frame inputs.

Conditional Generative Adversarial Networks (cGAN) have also been explored for radar perception enhancement, particularly for domain-specific applications. HawkEye [2] designs a cGAN architecture to recover high-resolution depth maps from conditioning on SAR-aided mmWave heatmaps, specifically for vehicle imaging in fog conditions. Similarly, Mi-Shape [33] applies cGAN for human pose estimation from mmWave signals, generating silhouettes and predicting joint locations. While GAN-based approaches can produce sharp and high-quality outputs with adversarial training, these radar cGANs are typically optimized for specific object categories. This domain specificity limits their generalizability compared to more recent generative approaches.

Recent diffusion approaches have advanced radar perception significantly. Diffusion models demonstrate superior denoising and generative capabilities across general-purpose computer vision tasks [34], [35], [36], making them attractive for radar-to-LiDAR translation. The Luan et al. work [11] introduced a conditional denoising diffusion probabilistic model (DDPM) that processes 5 temporally-stacked LiDAR BEVs with pixel values representing heights conditioned on the radar BEVs to denoise the corrupted LiDAR BEVs and produce enhanced LiDAR-like radar outputs, employing a weighted-L1 loss to handle sparse regions effectively. The Zhang et al. method [8] adopted advanced diffusion formulations including Elucidated Diffusion Models (EDM) [37] and consistency sampling [38] to replace iterative DDPM denoising. Their preprocessing pipeline retains traditional range-azimuth FFT processing while incorporating perceptual losses such as LPIPS [15], which is widely employed in computer vision for measuring perceptual similarity. Despite these advances, existing diffusion-based methods depend on temporal stacks or motion cues, and those that achieve single-frame processing sacrifice range and angle resolution.

Our work addresses this gap by employing a latent diffusion architecture that combines a pretrained VAE encoder-decoder with a denoising U-Net. The VAE encoder maps 2D LiDAR and radar BEV images to latent space, where our proposed U-Net performs denoising operations before the decoder reconstructs enhanced range-azimuth BEV outputs. This approach leverages the superior denoising capabilities of diffusion models while operating on single-frame radar inputs, eliminating the need for temporal aggregation while maintaining full range and image resolution (Table I).

TABLE I: Summary of related work

Methods	Arch.	Input	Single-frame (No SAR)
RadarHD [7]	VAE	2D BEV	✗
RadCloud [32]	VAE	2D BEV	✗
HawkEye [2]	cGAN	3D Voxel	✗
MiShape [33]	cGAN	2D FEV	✓
Luan et al. [11]	Diffusion	3D Point Cloud	✗
Zhang et al. [8]	Diffusion	2D BEV	✓
RadarSFD (Ours)	Latent Diffusion	2D BEV	✓

III. Our Approach

We leverage the denoising and generative capabilities of diffusion models to address the cross-modality challenge of transforming sparse, noisy radar measurements into dense LiDAR-like representations. Radar data naturally contains noise and artifacts, making it well matched to diffusion's iterative denoising process. By conditioning the generation on radar inputs, we can guide the model to recover fine geometric detail while remaining faithful to the observed scene. This section describes our *RadarSFD* pipeline, starting from the fundamentals of diffusion models, moving to the radar input representation, and finally the key architectural choices that enable single-frame radar-to-LiDAR translation.

A. Primer on Diffusion Models

1) Denoising Diffusion Models: Diffusion models are probabilistic generative models that learn to reverse a noise corruption process [34]. In our setting, the model learns to transform sparse radar inputs into high-resolution, LiDAR-like bird's-eye view (BEV) images.

The framework has two stages: a fixed forward process q and a learned reverse process p_θ. The forward process gradually corrupts a clean LiDAR BEV $\mathbf{x}_0$ with Gaussian noise over T timesteps:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t controls the variance at each step. As $t \to T$, the distribution converges to pure noise, erasing all structure from the original LiDAR image.

The reverse process p_θ, parameterized by a U-Net, learns to invert this corruption. Starting from pure noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the model iteratively predicts and removes noise to reconstruct $\hat{\mathbf{x}}_0$. Conditioning on radar representation $\mathbf{c}$ guides the denoising so that the output reflects the true scene. Rather than directly predicting clean data, the network predicts the noise $\boldsymbol{\epsilon}_\theta(\cdot)$ added in the forward process.

Training minimizes the mean squared error between the predicted and true noise:

$$L(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2\right], \tag{2}$$

where $\mathbf{x}_t$ is sampled from the forward process. This objective anchors the denoising trajectory so that the final output is both geometrically sharp and consistent with the radar input.

2) Latent Diffusion Models (LDM): Standard diffusion models operate directly in pixel space, requiring significant computational resources that scale poorly with image resolution. Latent diffusion models [35] address this by performing diffusion in a learned latent space. A pretrained VAE with encoder $\mathcal{E}$ and decoder $\mathcal{D}$ compresses images

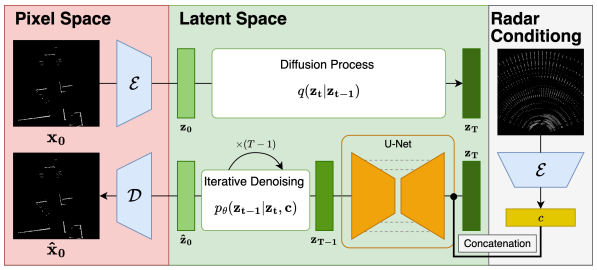


Fig. 3: Overview of the *RadarSFD*'s diffusion architecture. The radar and LiDAR images are first encoded into latent space. The radar latent $\mathbf{c}$ is concatenated with the noisy LiDAR latent $\mathbf{z}_t$ and fed into a pretrained U-Net denoiser. The U-Net iteratively removes noise while preserving structure as guided by the radar condition. After denoising, the decoder reconstructs a LiDAR-like point cloud with sharp geometry from a single radar frame.

to latent representations, reducing computational cost from $O(H \times W)$ to $O(h \times w)$ where $h, w \ll H, W$.

$$z = \mathcal{E}(x), \quad \hat{x} = \mathcal{D}(z) \tag{3}$$

The diffusion framework now operates on latent vectors $\mathbf{z}$ instead of raw pixels. The forward process gradually adds noise to the clean latent $\mathbf{z_0}$ over T timesteps, and the reverse process learns to predict the additive noise and denoise it from $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The training objective is adapted accordingly

$$L_{LDM}(\theta) = \mathbb{E}_{t,\mathcal{E}(\mathbf{x}_0),\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|^2\right] \tag{4}$$

where $\mathbf{z_t}$ is the noisy LiDAR latent representation at timestep t, $\mathbf{c}$ is the radar latent condition. Once the iterative denoising is finished, the decoder $\mathcal{D}$ maps the final, clean latent vector $\hat{\mathbf{z_0}}$ back into high-dimensional pixel space to produce the high-resolution LiDAR-like BEV $\mathbf{x_0}' = \mathcal{D}(\hat{\mathbf{z_0}})$.

B. Radar Input Representation

Our radar signal processing transforms raw I/Q signals into information-rich and structured representations compatible with diffusion models. We first apply range FFT and azimuth FFT to raw I/Q samples, generating raw 2D range-azimuth BEV images. Following RadarHD's preprocessing justification, we apply only light static thresholding (5% of the magnitude) to preserve information from radar sidelobes and multipath reflections that typically camouflage as noise but contain valuable structural information. This preprocessing strategy integrates effectively with diffusion's denoising capabilities. Traditional algorithms discard weak signals and artifacts as noise, but these components often carry environmental cues about scene geometry. By providing rich, lightly-filtered heatmaps as input, we enable the diffusion model to distinguish salient features from true noise through learned priors.

C. LDM Design Choices

Our conditional latent diffusion model is built to recover fine-grained LiDAR-like structure from a single radar frame. Figure 3 illustrates the overall architecture: radar and LiDAR BEVs are first encoded by the frozen VAE into latent space, concatenated, and then denoised by a pretrained U-Net. Here we explain four key design aspects that enable this pipeline: the VAE architecture, the choice of pretrained U-Net backbone, the conditioning strategy, and the training objective.

1) VAE Architecture: We use the distilled Tiny AutoEncoder for Stable Diffusion (TAESD), which compresses inputs by a factor of $8\times$ while preserving fine detail. This allows the diffusion process to operate on compact latent representations ($4 \times 32 \times 64$ for our BEVs) rather than high-dimensional images, reducing computation by orders of magnitude. The VAE remains frozen during both training and inference. This provides three benefits: (1) it leverages robust pretrained image priors learned at web scale, (2) it ensures consistent latent properties for both radar and LiDAR BEVs, and (3) it keeps inference efficient with negligible added cost compared to a full VAE. Encoding both modalities with the same frozen VAE ensures their latents are directly compatible, enabling simple channel-wise concatenation for conditioning.

2) Choice of Pre-trained U-Nets: Training a diffusion backbone from scratch is infeasible for radar–LiDAR translation due to limited dataset size and high compute demands. Instead, we initialize from *Marigold* [12], a state-of-the-art monocular depth estimator. Marigold's U-Net has been optimized to infer scene depth and geometry from RGB images, which transfers well to our task of mapping sparse radar returns to dense LiDAR structure. We hypothesize that these "shape-preserving" priors provide a stronger starting point than training on radar–LiDAR pairs alone. For comparison, we also explore Stable Diffusion v2 (SDv2), which encodes broad semantic priors but less geometric bias, in our ablation studies. Our results confirm that depth-focused priors from Marigold provide superior reconstructions, while SDv2 offers a useful baseline for analyzing generalization.

3) Conditioning Strategy: A central question is how to inject radar information into the denoising U-Net. We explore and study two strategies inspired by Marigold and SDv2:

- **Channel-wise concatenation (Marigold-style).** The radar BEV is encoded by the frozen VAE into latent $\mathbf{c}$, which is concatenated with the noisy LiDAR latent $\mathbf{z}_t$ along the channel dimension. This provides direct spatial alignment, anchoring diffusion to the observed radar geometry and preserving structural fidelity.
- **Cross-attention (SDv2-style).** The radar BEV is first projected into a sequence of embeddings, which are then injected into the U-Net at each layer through cross-attention. This provides more flexibility, allowing the model to learn higher-level correlations across modalities.

We find that each strategy presents a trade-off. Concatenation is simple and provides strong geometric guidance but may limit generalization. Cross-attention enables more abstract relationships but loses explicit spatial alignment and requires training an additional encoder. In this work, we

adopt concatenation as the primary conditioning scheme, and evaluate cross-attention in ablation studies to quantify the trade-off.

4) Training Objective: A standard LDM is trained with a latent noise prediction loss Eq. 4, which encourages the predicted noise to match the true Gaussian noise added during the forward process. While this ensures distributional consistency, it does not guarantee structural accuracy after decoding. As a result, the model may generate "LiDAR-like" outputs that look plausible but correspond to the wrong scene.

To address this, we add a pixel-space reconstruction loss L_p to tether generations to the true LiDAR structure:

$$L_p = \lambda_{L1} L_{L1} + \lambda_{SSIM} L_{SSIM} + \lambda_{LPIPS} L_{LPIPS}. \quad (5)$$

Here, L_{L1} is the Mean absolute error that enforces per-pixel accuracy, L_{SSIM} is the loss from SSIM that preserves structural integrity, and L_{LPIPS} is the LPIPS loss that aligns perceptual similarity with human judgments.

The final training objective is a weighted combination of latent and pixel-space losses:

$$L_{total} = L_{LDM} + \lambda_p L_p. \quad (6)$$

This dual-space objective balances distribution matching with per-scene fidelity. It reduces hallucinations, enforces sharper geometry, and ensures reconstructions remain consistent with the radar input.

IV. Evaluation

We evaluate *RadarSFD* on the RadarHD dataset, comparing against traditional and learning-based baselines. Our analysis covers dataset setup, evaluation metrics, baseline descriptions, and point cloud comparisons.

A. Dataset

Most publicly available radar datasets target automotive applications with outdoor environments [39], [40], [41]. Indoor-focused datasets like RaDiCal [42] and RadarRGBD [43] provide raw I/Q signals but lack corresponding LiDAR ground truth. View-of-Delft (VoD) [44] offers radar-LiDAR pairs but only provides processed radar point clouds rather than raw signals, limiting preprocessing flexibility. Additionally, VoD focuses on outdoor driving scenarios.

ColoRadar [45] comes closest to our requirements, providing both raw radar signals and LiDAR ground truth. However, indoor samples comprise only 15k of the dataset. RadarHD dataset addresses this gap with a larger collection of indoor radar-LiDAR pairs captured using single-chip mmWave systems. The dataset contains approximately 40k data pairs of raw mmWave radar signals with high-resolution LiDAR ground truth collected across 67 diverse trajectories indoor and outdoor. Given its primary focus on indoor scenarios and availability of raw radar representations, we adopt RadarHD as our training and evaluation dataset.

Evaluation is performed on the official test split, which contains trajectories not used during training. The split is designed to measure generalization under three conditions: (i) New trajectories in the same environment, (ii) trajectories from similar indoor environments, and (iii) trajectories from unseen and visually distinct environments.

TABLE II: 2D Radar Super-Resolution Performance

Work	# of Frames	Mean CD↓	Mean MHD↓
CFAR	1	0.84	0.91
RadarHD [7]	41	0.44	0.34
RadarHD [7] single-frame	1	0.56	0.45
Luan et al. [11]	5	0.59	0.50
Zhang et al. [8]	1	0.38	0.29
***RadarSFD* (Ours)**	1	**0.35**	**0.28**

B. Metrics

We extract 2D point clouds from both the generated BEVs and the corresponding LiDAR ground truth, and compute two widely used similarity metrics: (i) **Chamfer Distance (CD)**: the average bidirectional distance between each point and its nearest neighbor in the other point cloud. (ii) **Modified Hausdorff Distance (MHD)**: the mean nearest-neighbor distance, more robust to outliers than the classical Hausdorff distance. Lower values indicate closer agreement with the LiDAR reference.

C. Baselines

We compare *RadarSFD* against both traditional and recent learning-based methods:

- **CFAR** [26]: A conventional CA-CFAR detector with 5 dB threshold, representing non-learning approaches.
- **RadarHD** [7]: A VAE-based model trained with 41 stacked radar frames (temporal SAR).
- **RadarHD (single-frame)**: The same model evaluated with only one frame repeated 41 times, to measure performance without SAR cues.
- **Luan et al.** [11]: A diffusion-based method using 5 temporal frames, primarily targeting 3D point cloud generation.
- **Zhang et al.** [8]: A recent diffusion framework designed for single-frame inference.

D. Point Cloud Comparison

Table II summarizes the mean CD and MHD across methods. As expected, CFAR has the highest error due to

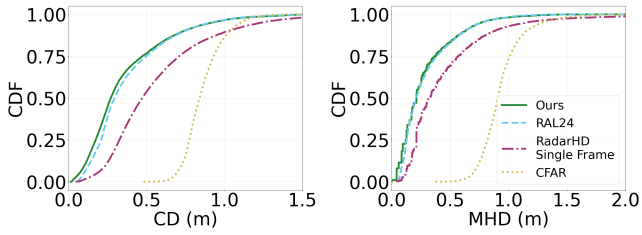
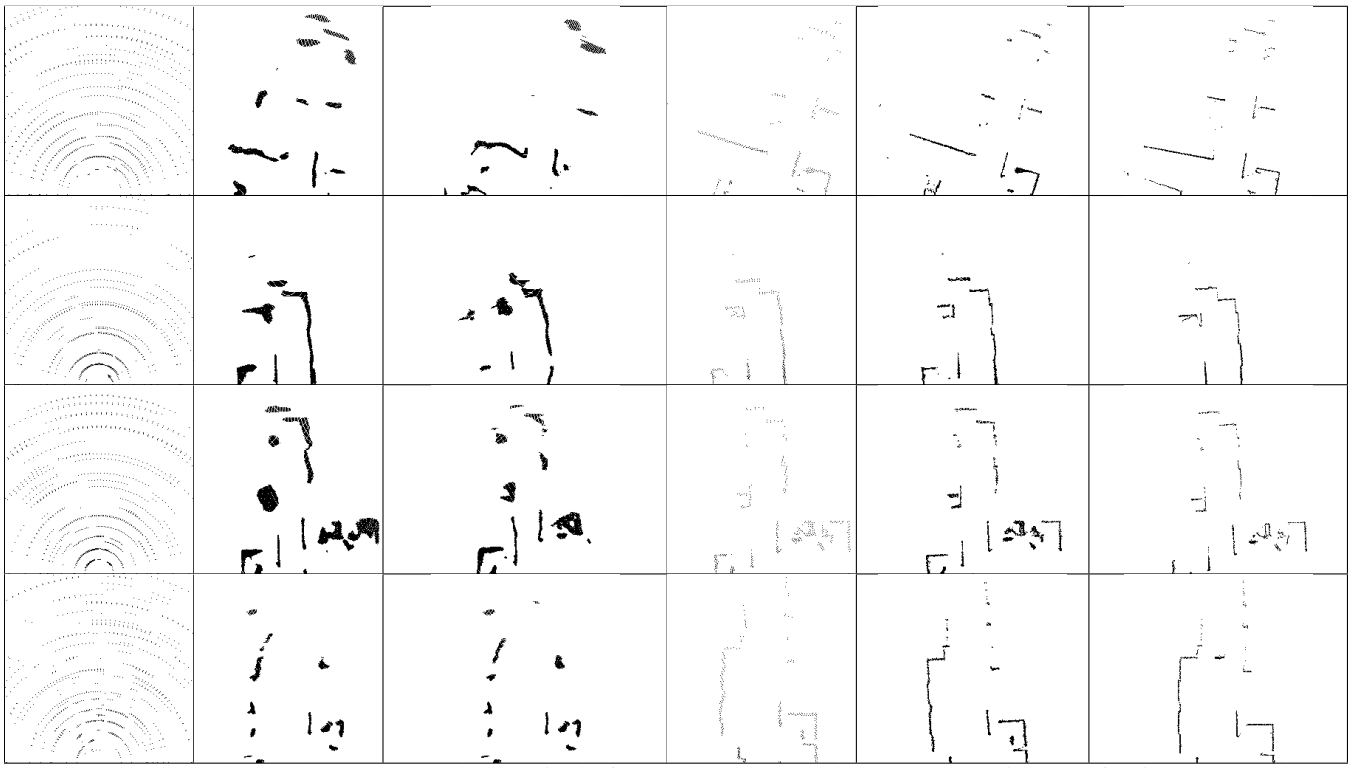


Fig. 4: CDFs of reconstruction error (CD, MHD). *RadarSFD* achieves the lowest errors, outperforming Zhang et al., RadarHD single-frame, and CFAR, with most samples under 0.5 m CD and 0.4 m MHD.



(a) CA-CFAR (b) RadarHD (c) RadarHD single-frame (d) Zhang et al. (e) Ours (*RadarSFD*)(f) LiDAR ground truth

Fig. 5: Qualitative comparison of point cloud reconstructions on four representative scenes with varying complexity. CA-CFAR is sparse and noisy; RadarHD [7] (41 frames) captures geometry but with blurry edges and clutter; RadarHD single-frame misses structures; Zhang et al. [8] yields cleaner but low-resolution (128×128) outputs. Our method achieves sharp, complete reconstructions at 256×512, closely matching ground-truth LiDAR and showing clearer wall boundaries with fewer hallucinations. All results are shown in Cartesian coordinates for direct comparison.

sparse detections. RadarHD achieves strong results with 41 stacked frames but degrades sharply in the single-frame case (0.56 m CD, 0.45 m MHD), underscoring its dependence on SAR. Both diffusion-based methods improve in the single-frame setting: Zhang et al. achieves 0.38 m CD and 0.29 m MHD, while *RadarSFD* achieves 0.35 m CD and 0.28 m MHD, a reduction of 8% and 3% respectively. At first glance, the numbers suggest parity, but the two methods differ fundamentally in design. Zhang et al. operates directly in pixel space, which makes inference slower (2.4 s per frame) and training less data-efficient. In contrast, *RadarSFD* performs diffusion in latent space, compressing inputs through a frozen VAE. This reduces inference to 1.3 s per frame, lowers compute requirements, and enables transfer of pretrained priors for generalization to unseen environments (shown in Fig. 2)

CDF plots for CD and MHD in Figure 4 illustrate the error distribution across test samples. Our method consistently shifts the curves leftward compared to baselines, indicating lower errors across the dataset. Notably, the *RadarSFD* curve overtakes the Zhang et al. baseline [8], showing that the majority of scenes benefit from our latent-diffusion approach. This effect is especially clear in the MHD plot, where *RadarSFD* achieves lower tail errors, highlighting improved robustness to outliers. These results confirm that transferring pretrained depth priors into the radar-to-LiDAR pipeline leads to more faithful reconstructions.

Figure 5 provides qualitative comparison of reconstruction across baselines. We exclude Luan et al. [11] since its implementation is not public and its reported visualizations focus on the VoD dataset. RadarHD with 41 temporal frames captures coarse scene contours but often produces blurred boundaries and misses fine structural detail. For example, in the first row, RadarHD introduces clutter near the scene center, while its single-frame version drops structures near the lower right region. These artifacts reflect the limitations of its VAE backbone, which lacks the denoising strength of diffusion models. By contrast, diffusion-based methods generate sharper and more complete geometry.

Both Zhang et al. [8] and *RadarSFD* recover the overall shape of the scene. However, Zhang et al. operates on downsampled 128×128 BEVs, producing coarser details, while *RadarSFD* reconstructs directly at 256×512 resolution, yielding finer structural fidelity. Qualitatively, this is visible in sharper wall boundaries and cleaner gaps in our reconstructions.

Diffusion models can also introduce artifacts. In the second row, both Zhang et al. and *RadarSFD* produce hallucinated points near the lower-left region. Such failure cases - minor hallucinations or occasional missing geometry - likely

TABLE III: Ablation study of *RadarSFD* on RadarHD dataset.

Experiment	Radar Input	Backbone	Conditioning	Mean CD [m]↓	Mean MHD [m]↓
Zero-threshold BEV	BEV (zero-thresh)	Marigold (pretrained)	Concatenation	0.36	0.29
Raw I/Q input	I/Q signals	Marigold (pretrained)	Cross-attn	*0.81*	*0.74*
Random init	BEV (light-thresh)	Marigold (random)	Concatenation	*0.85*	*1.08*
Alt. pretraining (SDv2)	BEV (light-thresh)	SDv2 (pretrained)	Cross-attn	0.39	0.32
Latent-only loss	BEV (light-thresh)	Marigold (pretrained)	Concatenation	*1.12*	*1.12*
L1 only	BEV (light-thresh)	Marigold (pretrained)	Concatenation	0.35	0.28
L1 + SSIM	BEV (light-thresh)	Marigold (pretrained)	Concatenation	0.42	0.34
L1 + LPIPS	BEV (light-thresh)	Marigold (pretrained)	Concatenation	0.35	0.28
RadarSFD	BEV (light-thresh)	Marigold (pretrained)	Concatenation	**0.35**	**0.28**

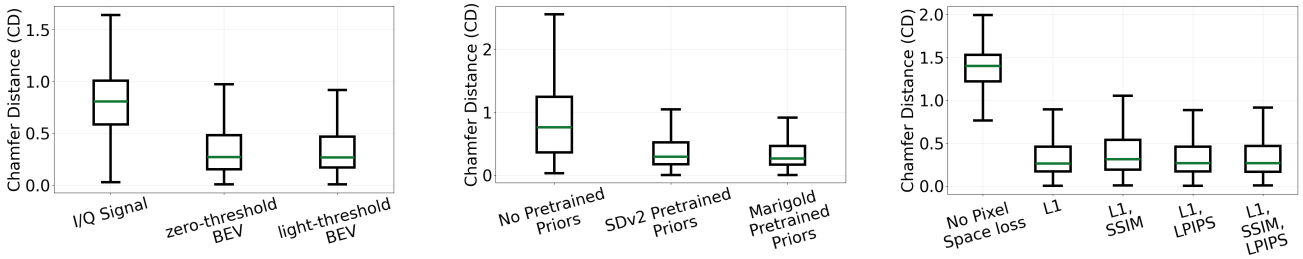


Fig. 6: Ablation box plots (Chamfer Distance). Left to right: input representation, pretrained priors, training losses. Thresholded BEV inputs outperform raw I/Q; depth-pretrained priors (Marigold, SDv2) beat random initialization; adding pixel-space L1 drives most of the gain, with SSIM and LPIPS providing only marginal benefit.

stem from the limited scale of the RadarHD dataset. These observations suggest that larger and more diverse training corpora would further improve the reliability of single-frame radar-to-LiDAR translation.

E. Ablation Study

We evaluate the impact of key design choices in *RadarSFD*: radar input representation, pretrained priors with conditioning, and training objectives. Results are summarized in Table III and Figure 6.

1) Radar Input Representation: We compare light-threshold BEV (our default), zero-threshold BEV, and raw I/Q inputs. Zero-threshold BEV performs nearly on par with light-threshold, showing that mild noise filtering is sufficient for diffusion models. In contrast, directly using raw I/Q signals with a custom CNN encoder increases Chamfer Distance by 2.3×, reflecting the difficulty of learning radar physics from scratch without priors optimized for image-like inputs. This confirms that lightly preprocessed BEV heatmaps are the most effective representation for conditioning.

2) Pretrained Priors and Conditioning: Next we evaluate the role of pretrained weights and conditioning strategies. Without pretrained priors, performance degrades by nearly 3× and variance increases sharply, with errors exceeding 2.5 m. This highlights that geometric priors learned by Marigold's depth estimation training are essential for cross-modal translation. Initializing from Stable Diffusion v2 yields reasonable results but underperforms Marigold, validating that task-specific geometric priors outperform generic semantic ones.

We also tested radar conditioning via cross-attention, as in SDv2. While flexible, cross-attention underperforms concatenation across both I/Q inputs and SDv2 priors. Channel-wise concatenation provides explicit spatial alignment between radar and LiDAR latents, leading to sharper reconstructions and stronger geometric fidelity.

3) Training Objectives: Finally, we analyze training objectives. Latent-only supervision performs worst, with errors exceeding 1 m, confirming that latent alignment alone does not guarantee structural fidelity after decoding. Adding pixel-space losses dramatically improves performance. L1-only training already matches our full system, while combinations with SSIM or LPIPS provide complementary benefits. Our complete formulation (L1+SSIM+LPIPS) offers the best consistency across scenes, balancing numerical accuracy, structural similarity, and perceptual fidelity.

V. Conclusion

We introduce *RadarSFD*, a latent diffusion model that reconstructs dense LiDAR-like point clouds from single mmWave radar frames without requiring SAR. By transferring pretrained depth priors into latent diffusion, *RadarSFD* achieves state-of-the-art accuracy while remaining efficient and practical for SWaP-constrained robotic platforms. Our results and ablations highlight the importance of pretrained priors, BEV conditioning, and dual-space objectives, offering a general recipe for cross-modal sensor translation.

References

[1] M. Bijelic, T. Gruber, and W. Ritter, "A benchmark for lidar sensors in fog: Is detection breaking down?" in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018, pp. 760–767.

[2] J. Guan, S. Madani, S. Jog, S. Gupta, and H. Hassanieh, "Through fog high-resolution imaging using millimeter wave radar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 464–11 473.

[3] N. J. Sanket, C. M. Parameshwara, C. D. Singh, A. V. Kuruttukulam, C. Fermüller, D. Scaramuzza, and Y. Aloimonos, "Evdodgenet: Deep dynamic obstacle dodging with event cameras," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 651–10 657.

[4] N. Garg and N. Roy, "Sirius: A self-localization system for resource-constrained iot sensors," in *Proceedings of the 21st annual international conference on mobile systems, applications and services*, 2023, pp. 289–302.

[5] Y. Bai, N. Garg, and N. Roy, "Spidr: Ultra-low-power acoustic spatial sensing for micro-robot navigation," in *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, 2022, pp. 99–113.

[6] Y. P. Talwekar, A. Adie, V. Iyer, and S. B. Fuller, "Towards sensor autonomy in sub-gram flying insect robots: A lightweight and power-efficient avionics system," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9675–9681.

[7] A. Prabhakara, T. Jin, A. Das, G. Bhatt, L. Kumari, E. Soltanaghaei, J. Bilmes, S. Kumar, and A. Rowe, "High resolution point clouds from mmwave radar," *arXiv preprint arXiv:2206.09273*, 2022.

[8] R. Zhang, D. Xue, Y. Wang, R. Geng, and F. Gao, "Towards dense and accurate radar perception via efficient cross-modal diffusion model," *IEEE Robotics and Automation Letters*, 2024.

[9] Y. Bai, N. Garg, and N. Roy, "Spidr: Microstructure-assisted vision for ubiquitous tiny robots," *Commun. ACM*, Feb. 2026, online First. [Online]. Available: https://doi.org/10.1145/3772712

[10] M. A. Richards *et al.*, *Fundamentals of radar signal processing*. Mcgraw-hill New York, 2005, vol. 1.

[11] K. Luan, C. Shi, N. Wang, Y. Cheng, H. Lu, and X. Chen, "Diffusion-based point cloud super-resolution for mmwave radar data," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 171–11 177.

[12] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 9492–9502.

[13] J. Wu, T. Wang, M. A. B. Siddique, M. J. Islam, C. Fermuller, Y. Aloimonos, and C. A. Metzler, "Single-step latent diffusion for underwater image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[15] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[16] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," Tech. Rep., 1977.

[17] M.-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Proceedings of 12th international conference on pattern recognition*, vol. 1. IEEE, 1994, pp. 566–568.

[18] R. C. Hansen, "Fundamental limitations in antennas," *Proceedings of the IEEE*, vol. 69, no. 2, pp. 170–182, 2005.

[19] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 2005.

[20] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 2002.

[21] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[22] N. Garg, Y. Bai, and N. Roy, "Owlet: Enabling spatial information in ubiquitous acoustic devices," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 255–268.

[23] N. Garg, I. Shahid, R. K. Sheshadri, K. Sundaresan, and N. Roy, "Large network {UWB} localization: Algorithms and implementation," in *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, 2025, pp. 1187–1203.

[24] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu *et al.*, "Radar perception in autonomous driving: Exploring different data representations," *IEEE Transactions on Intelligent Vehicles*, 2023.

[25] M. Barkat, S. Himonas, and P. Varshney, "Cfar detection for multiple target situations," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 136, no. 5. IET, 1989, pp. 193–209.

[26] G. Minkler and J. Minkler, "Cfar: the principles of automatic radar detection in clutter," *Nasa sti/recon technical report a*, vol. 90, p. 23371, 1990.

[27] P. P. Gandhi and S. A. Kassam, "Analysis of cfar processors in nonhomogeneous background," *IEEE Transactions on Aerospace and Electronic systems*, vol. 24, no. 4, pp. 427–445, 2002.

[28] H. Rohling, "Radar cfar thresholding in clutter and multiple target situations," *IEEE transactions on aerospace and electronic systems*, no. 4, pp. 608–621, 2007.

[29] M. E. Yanik and M. Torlak, "Near-field mimo-sar millimeter-wave imaging with sparsely sampled aperture data," *Ieee Access*, vol. 7, pp. 31 801–31 819, 2019.

[30] A. V. Muppala, A. Y. Nashashibi, E. Afshari, and K. Sarabandi, "Fast-fourier time-domain sar reconstruction for millimeter-wave fmcw 3-d imaging," *IEEE Transactions on Microwave Theory and Techniques*, vol. 72, no. 12, pp. 7028–7038, 2024.

[31] K. Qian, Z. He, and X. Zhang, "3d point cloud generation with millimeter-wave radar," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–23, 2020.

[32] D. Hunt, S. Luo, A. Khazraei, X. Zhang, S. Hallyburton, T. Chen, and M. Pajic, "Radcloud: Real-time high-resolution point cloud generation using low-cost radars for aerial and ground vehicles," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 269–12 275.

[33] A. Adhikari, H. Regmi, S. Sur, and S. Nelakuditi, "Mishape: Accurate human silhouettes and body joints from commodity millimeter-wave devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–31, 2022.

[34] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[36] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[37] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.

[38] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," 2023.

[39] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 6433–6438.

[40] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[41] I. Roldan, A. Palffy, J. F. Kooij, D. M. Gavrila, F. Fioranelli, and A. Yarovoy, "A deep automotive radar detector using the radelft dataset," *IEEE Transactions on Radar Systems*, 2024.

[42] T.-Y. Lim, S. A. Markowitz, and M. N. Do, "Radical: A synchronized fmcw radar, depth, imu and rgb camera data dataset with low-level fmcw radar signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 941–953, 2021.

[43] T. Song, J. Ye, A. Guo, G. He, and B. Yang, "Radarrgbd a multi-sensor fusion dataset for perception with rgb-d and mmwave radar," *arXiv preprint arXiv:2505.15860*, 2025.

[44] A. Palffy, E. Pool, S. Baratam, J. F. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[45] A. Kramer, K. Harlow, C. Williams, and C. Heckman, "Coloradar: The direct 3d millimeter wave radar dataset," *The International Journal of Robotics Research*, vol. 41, no. 4, pp. 351–360, 2022.