Dense Indoor 3D Scene Recovery from Radar via Geometric Prior Distillation

Samah Hussein
samah.husseinyoussef@epfl.ch
Haitham Hassanieh
haitham.alhassanieh@epfl.ch

École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

Abstract

Dense 3D scene understanding has long relied on LiDAR and RGB cameras, sensors that raise privacy concerns and fail under poor visibility conditions such as smoke, fog, or darkness. Millimeter wave radar offers a robust, privacy-preserving alternative, but translating its sparse and noisy point clouds into dense scene geometry is severely underconstrained: radar observations are corrupted by multipath reflections and specularity effects, and the same returns are consistent with many possible scene geometries. We propose PRISM (Prior-guided Radar Indoor Scene Mapping), a framework that extends the application of mmWave radar to dense indoor 3D scene recovery. The core challenge is that this mapping requires simultaneously acquiring the geometric structure of indoor environments and interpreting sparse and noisy radar returns. We address this by decoupling the two: a generative decoder is first trained on LiDAR data to internalize the geometric structure of indoor scenes and then frozen while a radar encoder is trained to activate it. Reconstruction and distillation objectives guide the radar encoder to produce latent representations compatible with the frozen decoder's geometric prior, bridging the modality gap without requiring LiDAR at inference. On real test scenes from unseen buildings, our approach achieves a 31% improvement in Chamfer Distance and a 26% improvement in F-score at 0.6 m over the best adapted prior radar method, while requiring approximately 3× less training data than end-to-end alternatives. Check out our project page for a demo video and more.

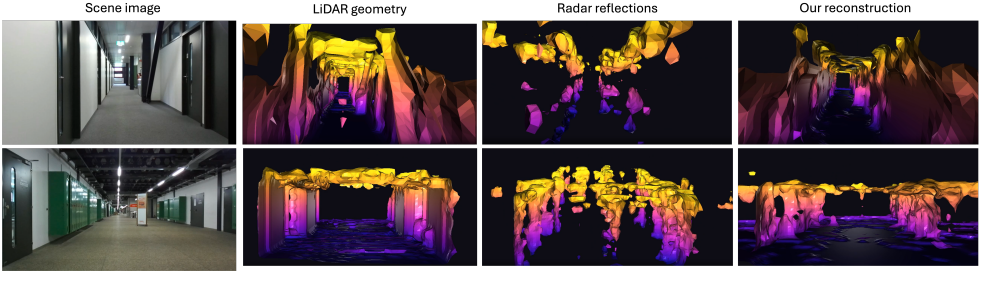

Figure 1: Dense 3D scene recovered from mmWave radar point clouds (colour encodes height) alongside photographs of the corresponding scenes. All scenes are from entirely unseen buildings. No camera or LiDAR is used at inference.

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

Indoor 3D scene understanding is fundamental to autonomous navigation, augmented reality, and assistive robotics. It is a safety-critical requirement in scenarios such as firefighting, where robots must navigate smoke-filled environments with no visual reference. Current approaches rely on LiDAR and RGB cameras, which offer high spatial resolution but raise privacy concerns and fail precisely under degraded visibility conditions. Millimeter wave (mmWave) radar is an attractive alternative: it is robust to lighting and visibility, privacy-preserving, and penetrates occlusions such as smoke and fog. Yet, recovering dense 3D scene geometry from radar is particularly challenging indoors, where dense furniture, narrow corridors, and reflective surfaces amplify multipath propagation and clutter. Many surfaces produce specular reflections that never reach the receiver, leaving large portions of the geometry unobserved. Past work has explored coarser 2D scene representations [20, 24, 25]. However, moving from 2D to full volumetric 3D recovery further compounds the above difficulties: the underdetermination of the mapping grows with the output dimensionality, and no single radar sweep can constrain a dense 3D volume.

Prior radar-based perception mainly focused on object-level reconstruction [8, 13, 21] and human pose estimation [16, 34]. Indoor scene recovery remains limited to 2D representations [20, 24, 25]. DREAM-PCD [10] addresses radar point cloud denoising rather than scene reconstruction, while PanoRadar [15] achieves near-LiDAR depth quality but relies on a motorized rotating radar and produces 2D range projections rather than volumetric 3D. Dense 3D scene recovery of full indoor environments from forward-facing commodity radar point clouds remains unaddressed. The fundamental difficulty is that this mapping requires simultaneously learning the geometric structure of indoor scenes and learning to interpret sparse, specular, and multipath-corrupted radar observations; two distinct problems that end-to-end approaches must solve together from limited radar data.

In this paper, we propose PRISM (Prior-guided Radar Indoor Scene Mapping), a framework that extends the application of mmWave radar to dense indoor 3D scene recovery. PRISM decouples the above two challenges. LiDAR provides dense, reliable geometric measurements of indoor scenes, making it well-suited for learning geometric structure independently of radar. We first train a generative autoencoder on dense LiDAR-derived volumes; the decoder internalizes the spatial statistics of walls, furniture, and room structure from this stable supervision. We then freeze this decoder and train a radar encoder to activate it via reconstruction and knowledge distillation objectives. These objectives guide the radar encoder toward latent representations compatible with the frozen LiDAR prior, transferring geometric knowledge across the modality gap without requiring LiDAR at inference. The frozen decoder acts as a strong geometric inductive bias over the radar encoder's output space, constraining it toward plausible scene geometries that are consistent with the training distribution. We find this design is not a convenience but a necessity: jointly fine-tuning from a LiDAR checkpoint causes catastrophic forgetting, with cross-modal radar gradients eroding the geometric prior and degrading performance below random initialisation.

We evaluate PRISM on a real indoor dataset collected across multiple buildings, with test sequences in entirely unseen environments and co-located LiDAR ground truth. We show that dense indoor 3D scene recovery from mmWave radar point clouds is learnable and generalises across environments. Against the best prior radar method evaluated on our setting, our approach achieves a Chamfer Distance of 1.97 versus 2.84 and an F-score at 0.6 m of 0.776 versus 0.618. Against end-to-end training of the same architecture without a geometric prior, our distillation framework achieves CD 1.97 versus 2.15 and F-score 0.776

versus 0.764, and reaches comparable performance with approximately 3× fewer training sequences, demonstrating that the frozen LiDAR prior provides geometric knowledge that cannot be recovered from radar observations alone. The dataset and code will be released to support further research.

2 Related Work

Radar-based 3D Perception. Early work on radar-based 3D perception focused on recovering object geometry from sparse, noisy point clouds in controlled settings. RF-Capture [1] and RFPose3D [32] introduced through-wall human figure recovery from RF reflections. HawkEye [11] and 3DRIMR [27] subsequently demonstrated object-level depth estimation from mmWave radar using generative models. RFConstruct[13] uses two orthogonal Cheap-Off-The-Shelf (COTS) mmWave radars with odometry-aware temporal fusion to reconstruct 3D shapes of outdoor objects. GeRaF [21] takes a neural approach, modeling raw RF signal physics through implicit representations to recover object-level geometry.

More recently, Radar Fields [3] learns a frequency-space neural scene representation for FMCW radar, while RadarSplat [14] adapts 3D Gaussian Splatting to scanning radar for high-fidelity reconstruction. Both are per-scene inverse-rendering methods on outdoor driving data, whereas we learn a single feed-forward model that predicts geometry for new scenes in one pass. At the scene level, RadarOcc [7] introduces 3D occupancy prediction from 4D imaging radar for autonomous driving, operating on raw radar tensors.

Most relevant to our work are [4, 10, 15, 24, 32]. RadarHD [24] demonstrated high-resolution indoor point cloud generation using a U-Net based super-resolution approach. However, it only operates on 2D projections without elevation information. DREAM-PCD [10] uses a hardware setup similar to ours, but only addresses radar point cloud denoising. RaLD [32] generates high-resolution 3D radar point clouds from raw radar spectra via latent diffusion, and MilliPCD [4] generates scene-scale indoor point clouds from a handheld mmWave device; both produce point clouds rather than dense volumetric occupancy, and RaLD's iterative diffusion sampling is substantially slower than our single forward pass. PanoRadar [15] focuses on depth estimation to achieve near-LiDAR quality but requires a specialized motorized rotating radar setup that can be impractical for deployment. In contrast, our work targets *dense volumetric geometric scene recovery* at the indoor scene level from accumulated commodity mmWave radar point clouds, predicted feed-forward across unseen scenes rather than optimized per scene — a task and problem formulation unaddressed by prior work.

Cross-Modal Knowledge Distillation for 3D Perception. Cross-modal knowledge distillation has been studied in autonomous driving to bridge the performance gap between cameras and LiDAR. DistillBEV [31], BEVDistill [5], and UniDistill [35] train camera-based student detectors to imitate features of a pretrained LiDAR teacher in BEV space, with no LiDAR required at inference. More recently, radar has become a target modality for such distillation: RadarDistill [2] transfers LiDAR feature knowledge to a radar-based *object detector* via cross-modal alignment modules, while CRKD [33] trains a camera-radar student to imitate a LiDAR-camera teacher in shared BEV feature space. LiCROcc [22] extends this to 3D semantic occupancy, distilling from a fused LiDAR-camera teacher to radar-only and radar-camera student models for outdoor scenes. RadOcc [7] demonstrates that naive feature or logit alignment is insufficient for cross-modal occupancy distillation and proposes rendering-assisted consistency losses as a more effective alternative. Our work differs from

all of these prior works in two fundamental ways. First, we operate in indoor environments with significantly more multipath reflections and clutter than in outdoor automotive settings. Second, and more importantly, rather than distilling features from an active teacher network, we freeze the decoder itself as a learned generative prior — a structural design choice we show is empirically necessary, as joint fine-tuning corrupts the geometric knowledge encoded in the pretrained decoder.

Generative Priors for 3D Reconstruction. Several works have explored using pretrained autoencoder decoders as shape priors for point cloud completion. [30] trains an autoencoder on complete point clouds and uses feature alignment losses to guide partial-cloud encoders toward the learned latent space. MV-DeepSDF [19] applies a frozen pretrained DeepSDF [23] decoder to multi-sweep LiDAR data for 3D vehicle reconstruction. More closely aligned with our approach, RaPD [9] proposes a two-stage reconstruction-aware prior distillation framework for semi-supervised point cloud completion: a shape prior is learned in stage one via reconstruction pretraining on complete clouds, and then distilled into a completion network in stage two. However, RaPD operates within a single modality on the completion task, whereas our approach performs cross-modal transfer from a LiDAR-trained generative decoder to a radar encoder, using reconstruction loss alone as the alignment signal and requiring no explicit latent matching. More recently, GenPC[17] leverages pretrained 3D generative models as zero-shot priors for point cloud completion, and diffusion-based approaches have used pretrained text-to-image models as shape priors for completing partial point clouds. Our work instantiates the frozen generative prior idea specifically for the cross-modal radar-to-LiDAR setting, where the decoder encodes the geometric structure of full indoor scenes learned from dense LiDAR supervision, and the radar encoder learns to activate it from impoverished radar observations.

3 Problem Formulation

We consider the task of dense indoor 3D occupancy estimation from mmWave radar. Let $\mathcal{P}_R = \{p_i \in \mathbb{R}^3\}_{i=1}^{N}$ denote an accumulated radar point cloud obtained by aggregating returns over a trajectory of $T \approx 20$ frames from two orthogonally-mounted COTS mmWave radars. The desired output is a binary occupancy volume $\mathcal{V} \in \{0,1\}^{H \times W \times D}$ defined over a voxel grid at resolution $\delta = 0.15$ m/voxel, covering a spatial extent of $[-20,+20]$ m laterally (X), $[0,+20]$ m forward (Y), and $[-20,+20]$ m vertically (Z), yielding a grid of $267 \times 134 \times 267$ voxels. The forward-only Y range reflects the physical field of view of our sensor configuration. Each voxel $v_{ijk} = 1$ if any surface is present within that cell, and $v_{ijk} = 0$ otherwise.

The key constraint is that LiDAR is available only during training. At inference, the system must produce $\mathcal{V}$ from $\mathcal{P}_R$ alone. This is challenging for three reasons. First, radar point clouds are extremely sparse relative to the scene volume. Second, radar returns are noisy and subject to multipath reflections, producing spurious observations with no geometric correspondence to real surfaces. Third, the mapping from sparse, noisy radar observations to dense volumetric geometry is severely underconstrained — it is extremely challenging to learn scene geometry from radar returns alone.

We evaluate predicted occupancy volumes against LiDAR-derived ground truth using Chamfer Distance (CD), F-score, and Intersection over Union (IoU).

4 PRISM: Cross-Modal Geometric Distillation

Dense indoor 3D occupancy estimation from radar point clouds poses two fundamental questions. First, is the task learnable at all — can a network trained on sparse, noisy radar observations acquire sufficient supervisory signal to produce geometrically coherent dense predictions that generalise to entirely unseen environments? Second, given the inherently underconstrained nature of the radar-to-geometry mapping, how can we best leverage the rich geometric knowledge available from LiDAR to go beyond what radar point clouds alone can deliver? We address both questions through PRISM a two-stage framework, illustrated in Figure 2, that decouples the two challenges at the heart of this problem: learning the geometric structure of indoor scenes, and learning to interpret sparse, noisy radar observations. In Stage 1 (Section4.1), we train a LiDAR autoencoder whose decoder internalises a strong geometric prior of indoor scenes. In Stage 2 (Section4.3), we freeze this decoder and train a radar encoder to interface with it via reconstruction and feature distillation objectives. Section 4.2 establishes what can be achieved without a geometric prior via end-to-end cross modal training, and directly motivates our frozen-prior framework.

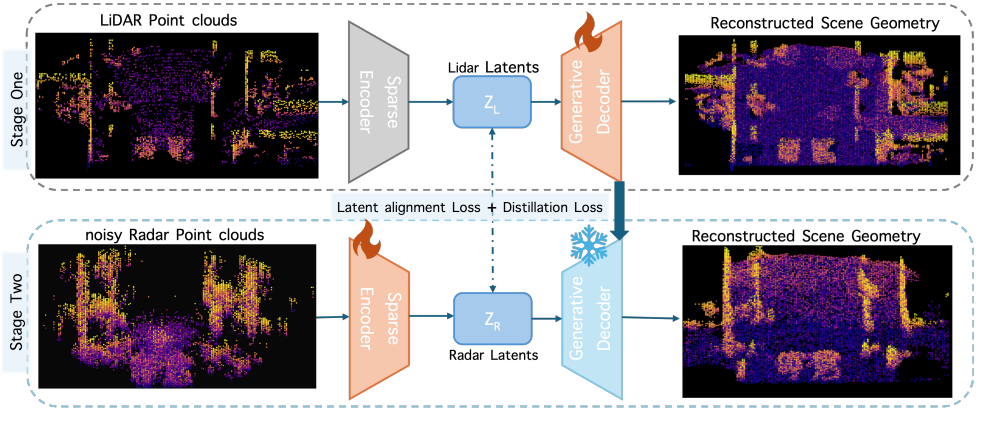

Figure 2: Overview of PRISM. *Stage 1:* A LiDAR autoencoder is trained on dense occupancy volumes; the decoder internalises the geometric structure of indoor scenes. *Stage 2:* The decoder is frozen and a radar encoder is trained to produce latent codes that activate it via reconstruction and distillation losses. Latent distillation losses additionally encourage the radar bottleneck z_R to remain compatible with the LiDAR bottleneck z_L. At inference, only the radar encoder and frozen decoder are used.

4.1 Stage 1: LiDAR Geometric Prior

LiDAR-derived occupancy volumes are obtained by voxelising accumulated LiDAR point clouds at resolution $\delta = 0.15$ m. The encoder compresses the input into a compact bottleneck representation z_L, and the decoder, equipped with generative transposed convolutions, reconstructs a dense binary occupancy volume from z_L alone. Through this process, the decoder internalizes the geometric structure of indoor scenes: it learns not merely to project latent codes to output voxels, but to complete and generate plausible geometry consistent with the training distribution.

The network is supervised by a Z-weighted binary cross-entropy loss. Since the floor plane dominates the voxel distribution and is trivially predictable, we reduce its contribution: voxels within a range above the estimated floor baseline receive a loss weight of 0.1,

while all other voxels are weighted at 1.0. This encourages the model to focus on semantically meaningful structures such as furniture and obstacles. To further guide convergence of the generative decoder layers and prevent unconstrained voxel generation in early training, we additionally supervise each intermediate decoder stage with an auxiliary Binary Cross-Entropy (BCE) loss weighted at $\alpha = 0.2$.

Critically, we apply a skip connection dropout curriculum during training. Skip features from the encoder are dropped with increasing probability p_{skip} across training epochs, progressively forcing the decoder to reconstruct geometry from the bottleneck z_L alone. This serves three purposes: in early training, skip features provide rich multi-scale context that guides the generative decoder layers towards plausible reconstructions before the bottleneck representation has matured; progressively dropping them forces the decoder to internalise a self-sufficient geometric prior independent of encoder features; and it directly prepares the decoder for Stage 2, where skip connections carry radar rather than LiDAR features and their reliability cannot be assumed.

4.2 Cross-Modal Learning Baselines

We train two end-to-end variants using the same sparse 3D convolutional architecture described in Section 4.4 and an asymmetric focal loss applied to the final occupancy output. In both cases, the encoder and decoder are jointly optimised on radar point cloud inputs with LiDAR-derived occupancy volumes as supervision — no stage-wise training, no frozen components.

E2E Random Initialisation. The first variant initialises all weights randomly and trains the full network on radar inputs from scratch. This directly probes whether the radar-to-occupancy mapping is learnable under pure reconstruction supervision, without any geometric prior. We find that it is: the model converges to produce geometrically coherent occupancy predictions and generalises to test sequences recorded in entirely unseen buildings, answering our first question affirmatively. This establishes a strong baseline on what radar supervision alone can achieve and serves as the primary reference point for evaluating the benefit of geometric prior distillation.

E2E Fine-Tuning from LiDAR Checkpoint. The second variant attempts to leverage LiDAR pretraining by initialising both encoder and decoder from the Stage1 LiDAR checkpoint and jointly fine-tuning on radar input. Rather than improving over random initialisation, this actually *degrades* performance, achieving a Chamfer Distance of 2.488 compared to 2.178 for random initialisation. We attribute this to catastrophic forgetting, where the decoder is subjected to cross-modal gradients from the noisy and sparse radar distribution during joint fine-tuning. These gradients progressively overwrite the learned geometric structure, eroding the very prior that makes LiDAR pretraining valuable. The decoder's geometric knowledge is not robust to simultaneous optimisation under a mismatched input modality.

In order to successfully leverage LiDAR pretraining, we trust that the decoder has internalised a useful geometric prior, and we must first, shield it from destabilising radar gradients to preserve this knowledge, and second, train a radar encoder that is able to activate this geometric prior successfully.

4.3 Stage 2: Radar Encoder Training and Dual Distillation

With the decoder frozen, a radar encoder is trained from random initialisation. Only the encoder parameters are updated; the decoder weights remain fixed throughout. The encoder receives accumulated radar point clouds $\mathcal{P}_R$ as input and must learn to produce bottleneck representations z_R that, when passed through the frozen decoder, reconstruct geometrically coherent occupancy volumes. The primary training signal is an asymmetric focal loss [18] applied to the final occupancy output:

$$\mathcal{L}_{\text{focal}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \,(1-p_i)^{\gamma}\,\text{BCE}(s_i, y_i) \tag{1}$$

where $p_i = \exp(-\text{BCE}(s_i, y_i))$, $\alpha_i \in \{\alpha_{\text{pos}}, \alpha_{\text{neg}}\}$ assigns per-voxel class weights, and $\gamma = 2.0$ down-weights easy examples. We set $\alpha_{\text{pos}} = 0.75$ and $\alpha_{\text{neg}} = 0.25$ to address the severe class imbalance between occupied and empty voxels. Intermediate decoder outputs are additionally supervised with the same focal loss, denoted $\mathcal{L}_{\text{inter}}$ and weighted at $\lambda_{\text{inter}} = 0.2$, providing multi-scale gradient signal to the radar encoder through the frozen decoder's generative stages.

Reconstruction supervision alone, however, leaves the radar encoder's latent space unconstrained, leaving the encoder free to arrive at any representation that the frozen decoder can decode, without any guarantee that it will be compatible at the feature level with the geometric prior encoded during Stage 1. To actively guide the encoder towards representations compatible with the frozen decoder's geometric prior, we augment the objective with two complementary distillation losses, both measuring the discrepancy between the radar bottleneck z_R and the LiDAR bottleneck z_L, one operating over a dense bird's-eye-view projection of the scene, the other enforcing direct feature correspondence at matched 3D locations. The full training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{inter}}\,\mathcal{L}_{\text{inter}} + \lambda_{\text{BEV}}\,\mathcal{L}_{\text{BEV}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}} \tag{2}$$

with $\lambda_{\text{inter}} = 0.2$, $\lambda_{\text{BEV}} = 0.5$ and $\lambda_{\text{align}} = 1.0$.

Macro-spatial distillation ($\mathcal{L}_{\text{BEV}}$). We project both z_R and z_L onto a 64×64 bird's-eye-view grid by collapsing the vertical axis via feature pooling, and minimise their MSE:

$$\mathcal{L}_{\text{BEV}} = \|\text{BEV}(z_R) - \text{BEV}(z_L)\|_F^2 \tag{3}$$

where $\text{BEV}(\cdot)$ denotes the vertical projection and $\|\cdot\|_F$ is the Frobenius norm. This provides a dense spatial supervision signal over the full scene footprint, forcing the radar encoder to place large structures — walls, furniture, room boundaries — in the correct physical locations in the BEV plane, even when the sparse radar point cloud provides only partial coverage.

Micro-feature alignment ($\mathcal{L}_{\text{align}}$). At the bottleneck level, we minimise the cosine distance between spatially matched radar and LiDAR feature vectors:

$$\mathcal{L}_{\text{align}} = \frac{2}{|\mathcal{M}|}\sum_{(i,j)\in\mathcal{M}} \Big(1 - \cos(z_R^i, z_L^j)\Big) \tag{4}$$

where $\mathcal{M}$ denotes spatially matched voxel pairs and z_L^j is detached from the computation graph. This cosine form equals the squared Euclidean distance between ℓ_2-normalised codes

$\hat{z} = z/\|z\|$ (hence the factor 2). Matching is performed via a collision-free integer hash over $(batch, x, y, z)$ coordinates. While the sparse bottleneck limits the size of $\mathcal{M}$, this loss directly pulls radar feature vectors towards their LiDAR counterparts at precise 3D locations, complementing the coarser BEV signal.

The two losses are designed to be complementary: $\mathcal{L}_{\text{BEV}}$ provides broad, dense spatial supervision over the scene layout, while $\mathcal{L}_{\text{align}}$ enforces precise feature-level correspondence at matched bottleneck locations. The mean cosine similarity between z_R and z_L at convergence is 0.143, whereas for the end-to-end baseline it is effectively zero (0.002); the two similarity distributions are in fact disjoint. This suggests that the combined objective meaningfully draws the two representations together: the radar encoder does not replicate the LiDAR latent space — impossible given the modality gap — but aligns sufficiently to activate the frozen decoder's geometric prior. The ablation in Section 5.3 isolates the contribution of each loss.

4.4 Architecture Details

Both the LiDAR encoder (Stage 1) and radar encoder (Stage 2) share an identical sparse 3D convolutional architecture built on MinkowskiEngine [6]. The encoder consists of seven blocks: an initial $3\times3\times3$ convolution projecting the scalar occupancy input to 16 channels, followed by six strided blocks each comprising a $2\times2\times2$ strided convolution for downsampling and a $3\times3\times3$ convolution for refinement, with batch normalisation and ELU activations throughout. Channel widths double at each scale: $[16, 32, 64, 128, 256, 512, 1024]$, producing a bottleneck $z \in \mathbb{R}^{1024}$ per active voxel at stride 64.

The decoder mirrors the encoder using six generative transposed convolution blocks as implemented in MinkowskiEngine [6]. Each block upsamples the sparse tensor via a $2\times2\times2$ generative transposed convolution, optionally fuses skip features from the corresponding encoder level via element-wise addition, refines with a $3\times3\times3$ convolution, and prunes geometrically implausible voxels via a learned $1\times1\times1$ occupancy head before passing to the next stage, preventing unbounded growth of the active voxel set during upsampling. The final decoder block omits pruning, and a separate $1\times1\times1$ convolution produces the per-voxel occupancy logits. Channel widths in the decoder follow the symmetric schedule $[16, 32, 64, 128, 256, 512]$ from fine to coarse.

Training details. Floor voxels are downweighted during training (loss weight 0.1) and stochastically removed from the input with probability 0.5 in both stages, preventing the trivially predictable floor plane from dominating the loss.

In Stage 1, skip connections are dropped with a scheduled probability that increases throughout training, progressively forcing the decoder to rely on the bottleneck representation alone. In Stage 2, skip dropout is held fixed at $p_{\text{skip}} = 0.5$ throughout. Stage 1 is trained for 50 epochs using AdamW with learning rate 10^{-2}, weight decay 10^{-4}, and exponential learning rate decay with $\gamma = 0.95$, taking 5 h on a single GH200 GPU. Stage 2 is trained for 50 epochs using AdamW with learning rate 10^{-4}, weight decay 0.01, and a cosine annealing schedule, taking 10 h. For reference, the end-to-end baseline trains in 8.5 h on the same hardware. At inference, PRISM runs a single forward pass at 24 ms per scan ($\sim$42 FPS) for its 58.4M parameters.

5 Experiments & Discussion

5.1 Dataset and Setup



Figure 3: Sensor rig: two orthogonal TI cascaded mmWave radars (horizontal + vertical), a 128-channel HESAI OT128 LiDAR, and a ZED 2i stereo camera.

Data collection. We introduce a new indoor mmWave radar dataset collected across multiple buildings over 7 recording days. The dataset comprises 66 experiments across 49 training, 2 validation, and 15 test sequences, totalling 2.41 km of trajectory. The sensor rig is illustrated in Figure 3 and consists of two TI cascaded 4-chip mmWave radars [28] mounted inside a shared enclosure, one oriented horizontally and one vertically, both facing forward. This orthogonal configuration provides complementary azimuth-elevation coverage: the horizontal radar resolves targets in the forward-lateral plane, while the vertical radar resolves targets in the forward-elevation plane, together producing denser and more complete angular coverage than a single unit. A 128-channel HESAI OT128 LiDAR [12] mounted on top of the enclosure provides dense geometric ground truth, and a ZED 2i stereo camera [26] is co-located on the front face for future multimodal work. Ego-motion for both point cloud accumulation and ground truth generation is estimated from LiDAR scans using KISS-ICP [29].

Train/val/test split. Training sequences were recorded across 6 days in September 2025, covering 49 experiments with a mean trajectory of 40.3 m per experiment. The validation set consists of 2 sequences from the same recording period. For testing, 15 test sequences were recorded on a separate day in May 2026 in entirely unseen buildings, providing a clean evaluation of generalisation to new environments.

Input representation. At both training and inference, the radar input $\mathcal{P}_R$ is formed by accumulating returns over 20 frames randomly sampled from a sliding window of 60 frames centred on the current position (± 30 frames), with ego-motion compensation applied via KISS-ICP [29] odometry. Ground truth occupancy volumes are derived by voxelizing the full LiDAR trajectory accumulated over each experiment, cropped and transformed into the radar frame of reference. This asymmetry is intentional: rather than supervising the model to match single-frame LiDAR distributions (which inherit the sensor's own gaps and occlusions), we train against full-trajectory accumulations that best approximate the complete scene geometry the model is asked to predict.

Evaluation metrics. We evaluate using Chamfer Distance (CD), Intersection over Union (IoU), and F-score at 30 cm, and 60 cm thresholds (F@30, F@60). All predicted and ground truth volumes are evaluated at the same voxel resolution of $\delta = 0.15$ m.

Baselines We compare against two prior methods that represent the closest existing work on radar-based 3D reconstruction. DREAM-PCD [10] is a radar point cloud denoising method that jointly addresses specular loss, low angular resolution, and noise, using a hardware configuration similar to ours (two orthogonal radars co-located with LiDAR). PanoRadar [15] achieves near-LiDAR depth estimation quality and supports recognition tasks such as segmentation and object detection, but is designed for a specialised motorised rotating radar and panoramic scanning. Neither method targets dense volumetric occupancy estimation in indoor environments; both are adapted to our forward-facing setup (†) and serve as a reference for what radar-based methods designed for adjacent tasks achieve when applied to ours.

Table 1: Quantitative comparison on our indoor mmWave radar dataset. CD = Chamfer Distance (lower is better). F@30 and F@60 denote F-score at 30 cm and 60 cm thresholds respectively. † denotes methods adapted for our indoor forward-facing setup (see Section 5.1). For E2E Random Init and PRISM, CD, IoU and F@30 are reported as mean ± std over 5 training seeds; F@60 is from a single run. **Bold** = best, underline = second best.

Method	CD ↓	IoU ↑	F@30 ↑	F@60 ↑
Prior methods (adapted)				
DREAM-PCD [10]†	5.0545	0.2015	0.3337	0.5751
PanoRadar [15]†	2.8440	0.0368	0.4068	0.6175
cross-modal baseline				
E2E Random Init	2.15 ± 0.12	0.201 ± 0.001	0.625 ± 0.007	0.764
Ours				
PRISM	**1.97 ± 0.10**	**0.210 ± 0.002**	**0.646 ± 0.009**	**0.776**
Oracle				
LiDAR Autoencoder	0.917	0.286	0.815	0.899

5.2 Main Results

Table 1 reports quantitative results on the test set. Prior methods adapted to our setup perform substantially worse overall. DREAM-PCD produces poorly localised and sparse volumes reflected in a high Chamfer Distance (5.05), due to the nature of their denoising task, while PanoRadar — designed for depth estimation from panoramic scanning — struggles to generalise to the forward-facing nature of the data, in addition to predicting depth voxels only, achieving near-zero IoU (0.037) despite moderate F-scores. Neither method produces well-localised dense reconstructions from indoor radar point clouds.

Our method, PRISM, achieves the best result across all reported metrics: CD 1.97, IoU 0.210, F@30 0.646, and F@60 0.776. The End-to-End randomly initialised (E2E Random Init.) baseline is the strongest cross-modal baseline (CD 2.15, F@30 0.625), confirming that the task is learnable from radar point clouds alone and generalises to entirely unseen environments. Our method improves upon it consistently across all metrics, with the most notable gain in Chamfer Distance, reflecting more accurate geometric localisation enabled by the frozen LiDAR prior. A third baseline, an end-to-end model fine-tuned from LiDAR pretraining, is discussed in Table 2 (Section 5.3), where jointly fine-tuning encoder and decoder causes catastrophic forgetting and degrades performance below random initialisation.

The gap relative to the LiDAR Autoencoder oracle (CD 0.917) indicates substantial room for improvement, particularly in recovering fine-grained geometry that the radar encoder cannot reliably activate from sparse point cloud observations.

5.3 Ablation Study

Table 2 ablates the key design choices: the handling of the LiDAR-pretrained checkpoint, the decoder freezing strategy, and the distillation objective.

LiDAR fine-tuning causes catastrophic forgetting. Fine-tuning initialises both encoder and decoder from the Stage 1 LiDAR checkpoint and jointly optimises on radar input, the initial approach for leveraging LiDAR pretraining. Rather than improving over random initialisation, this degrades performance to CD 2.488 compared to 2.178 for E2E Random Init. The LiDAR-pretrained decoder is destabilised by cross-modal gradients from the noisy radar

Table 2: Ablation study on freezing strategy and distillation objective. All models use the same radar encoder architecture trained for 50 epochs with AdamW ($lr = 10^{-4}$). Each configuration is a single training run, so values differ slightly from the seed-averaged results in Table 1.

Model	**Frozen**	**Distillation**	**CD** ↓	**F@30** ↑
End-to-end baselines				
E2E Random Init	✗	—	2.178	0.632
E2E Fine-tuning (LiDAR init)	✗	—	2.488	0.626
Frozen decoder variants				
No distillation	✓	—	1.961	0.625
Bottleneck align only	✓	$\mathcal{L}_{\text{align}}$	2.048	0.617
BEV Distil. only	✓	$\mathcal{L}_{\text{BEV}}$	1.998	0.636
PRISM (Both combined)	✓	$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{BEV}}$	1.895	0.637

distribution, progressively overwriting the geometric prior it encodes. This confirms that joint fine-tuning is not a viable path to leveraging LiDAR pretraining, and directly motivates the frozen-decoder design.

Freezing the decoder is beneficial even without distillation. Comparing No distillation (frozen decoder, no auxiliary losses) against E2E Random Init isolates the effect of freezing alone. Contrary to what one might expect, freezing without any distillation guidance already outperforms end-to-end training (CD 1.961 vs 2.178), demonstrating that preserving the geometric prior is itself valuable, and shows that the frozen decoder provides a strong inductive bias that constrains the radar encoder's output space towards geometrically plausible reconstructions. However, the encoder must still be guided towards representations that activate the frozen decoder effectively; without this, F@30 (0.625) remains below E2E Random Init (0.632), indicating incomplete utilisation of the prior. The distillation losses address this.

Both distillation losses are complementary. Among the frozen decoder variants, the combined objective in PRISMachieves the best performance (CD 1.895, F@30 0.637), outperforming both BEV distillation alone (CD 1.998, F@30 0.636) and bottleneck alignment alone (CD 2.048, F@30 0.617). Neither loss alone is sufficient: bottleneck alignment alone is limited by the sparse nature of the bottleneck, where few matched voxel pairs exist between radar and LiDAR at stride 64, restricting the gradient signal to a small set of precisely localised but infrequent correspondences. BEV distillation alone provides a denser and more stable supervision signal over the full scene footprint, explaining its stronger individual performance, however, it operates on projected features and lacks the precise 3D feature-level correspondence that alignment provides. Together, the two losses operate at complementary granularities: BEV enforces the correct spatial layout of large structures across the full scene, while bottleneck alignment pulls the radar encoder's feature representations towards those of the LiDAR encoder at matched 3D locations. The cosine similarity between z_R and z_L at convergence (0.143) confirms that the combined objective meaningfully draws the two representations together without collapsing them.

Latent space alignment. Figure 4 measures cosine similarity between each model's radar bottleneck z_R and the LiDAR oracle bottleneck z_L across data regimes. The alignment hierarchy directly mirrors the performance hierarchy of Table 2: the combined distillation objective in PRISM achieves the highest similarity to the LiDAR oracle (0.143 at 49 experiments), followed by alignment-only (0.097), BEV-only (0.050), and frozen without distilla-

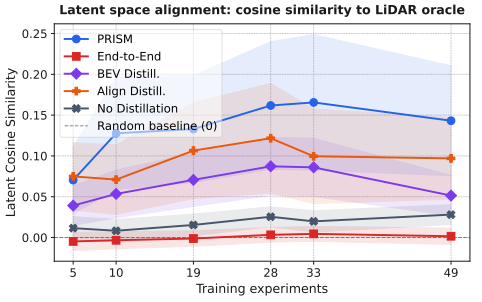

Figure 4: Cosine similarity between z_R and z_L across data regimes. Shaded regions indicate standard deviation across test scans.

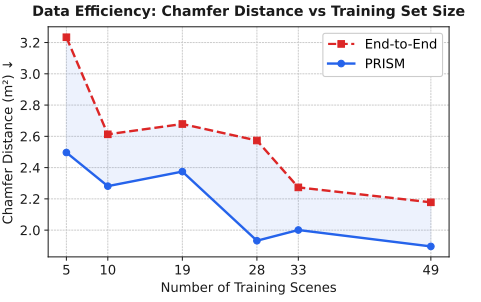

Figure 5: Chamfer Distance as a function of training set size. Distillation trained on 10 scenes matches End-to-End trained on 33 scenes (3× reduction in required data).

tion (0.028). End-to-end training remains near zero throughout, confirming that its encoder develops representations essentially orthogonal to the LiDAR prior regardless of how much radar data is available. Notably, even the frozen decoder without distillation losses induces a small but consistent alignment above random, confirming that the frozen prior's structure provides implicit guidance to the encoder. The combined distillation objective achieves substantially higher alignment than either loss alone, corroborating the complementarity observed in the ablation: explicit distillation is necessary for meaningful alignment, and both losses together draw the radar encoder's representations closest to the geometric prior.

5.4 Data Efficiency

Figure 5 compares PRISM Geometric Prior Distillation and End-to-End training as a function of training set size. PRISM consistently achieves lower Chamfer Distance across all data regimes. The advantage is largest at low data — at 5 training scenes the gap reaches 0.74 CD (2.50 vs 3.24), where the frozen decoder's geometric prior compensates for the near-total absence of radar-specific supervision. End-to-End training, which must learn the geometric structure of indoor scenes entirely from radar observations, degrades substantially at low data, reflecting the difficulty of this underconstrained mapping without a geometric prior.

The practical implication is a 3× reduction in required training data: distillation trained on 10 scenes achieves comparable Chamfer Distance (2.28) to End-to-End trained on 33 scenes (2.28). This gain stems directly from the frozen decoder supplying the geometric knowledge of indoor scenes that End-to-End must acquire from scratch. This requires substantially more radar data to discharge. The advantage persists even at the full dataset (49 scenes, CD 1.90 vs 2.18), confirming that the geometric prior provides a lasting benefit beyond the low-data regime.

5.5 Qualitative Results

Figure 6 shows occupancy predictions (visualized via Marching Cubes) on test scenes from entirely unseen buildings. Our method produces geometrically coherent completions that recover room boundaries, walls, and large structures from radar reflections that cover only a small fraction of the scene, demonstrating that the frozen decoder's geometric prior successfully fills in plausible geometry beyond what the sparse radar input directly observes.

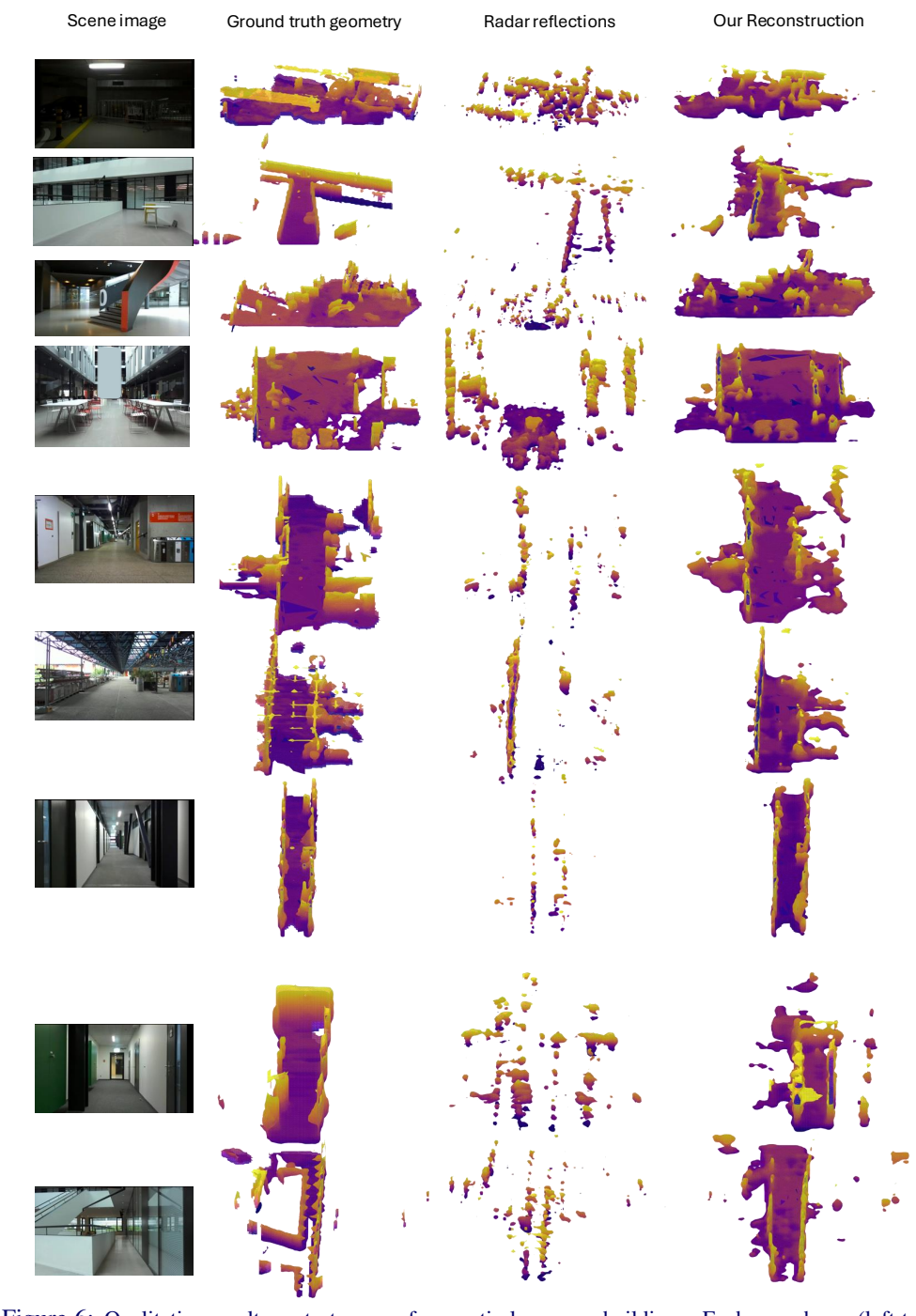

Figure 6: Qualitative results on test scenes from entirely unseen buildings. Each row shows (left to right): the real scene photograph, LiDAR-derived ground truth occupancy, accumulated radar reflections, and our dense 3D reconstruction. Colour encodes height (purple = floor, yellow = ceiling). Our method recovers geometrically coherent room structure from observations that cover only a small fraction of the scene.

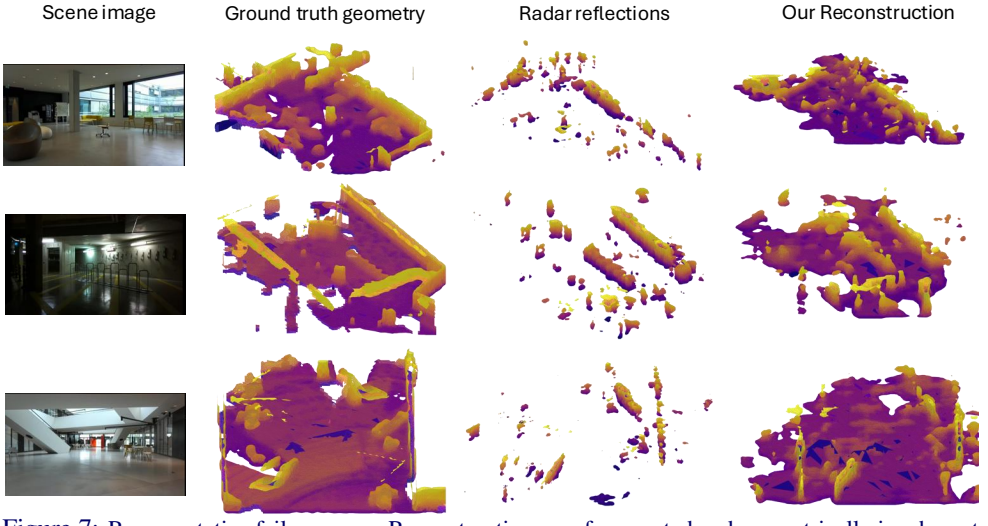

Figure 7: Representative failure cases. Reconstructions are fragmented and geometrically incoherent in scenes with large open spaces or high ceilings, where radar returns are too sparse to reliably activate the frozen decoder's geometric prior.

Figure 7 shows representative failure cases. The primary failure mode is incomplete recovery of geometry in large open spaces or high-ceiling environments, where radar returns are too sparse to activate the decoder's geometric prior effectively, resulting in fragmented reconstructions. A secondary failure mode occurs at the boundary of the forward-facing field of view, where accumulated radar coverage is thinner and the prior is asked to extrapolate beyond observed geometry.

6 Conclusion

We have presented PRISM, a framework for dense indoor 3D occupancy estimation from mmWave radar point clouds. A LiDAR-trained decoder is frozen to preserve its geometric prior, and a radar encoder is trained to interface with it via reconstruction and distillation objectives. The key practical findings are that end-to-end training from scratch is a viable baseline that generalises to unseen environments, that joint fine-tuning from a LiDAR checkpoint causes catastrophic forgetting and should be avoided, and that the frozen prior reduces the radar training data requirement by roughly threefold.

The gap to the LiDAR oracle remains substantial, and scenes with large open spaces or sparse radar coverage remain difficult. We hope the baselines, dataset, and analysis provided here are useful to future work on radar-based scene understanding.

Acknowledgements

We thank Björn Rathgeber for his help with data collection, and the anonymous reviewers for their constructive feedback, which improved this paper. This work was supported under project ID a0216 as part of the Swiss AI Initiative, through a small grant from the ETH Domain and computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure.

References

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Trans. Graph.*, 34(6), nov 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818072. URL https://doi.org/10.1145/2816795.2818072.

[2] Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features. *arXiv preprint arXiv:2403.05061*, 2024.

[3] David Borts, Erich Liang, Tim Broedermann, Andrea Ramazzina, Stefanie Walz, Edoardo Palladin, Jipeng Sun, David Brueggemann, Christos Sakaridis, Luc Van Gool, Mario Bijelic, and Felix Heide. Radar fields: Frequency-space neural scene representations for fmcw radar. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657510. URL https://doi.org/10.1145/3641519.3657510.

[4] Pingping Cai and Sanjib Sur. Millipcd: Beyond traditional vision indoor point cloud generation via handheld millimeter-wave devices. 6(4), January 2023. doi: 10.1145/3569497. URL https://doi.org/10.1145/3569497.

[5] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-modal BEV distillation for multi-view 3d object detection. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-2zfgNS917.

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[7] Fangqiang Ding, Xiangyu Wen, Yunzhou Zhu, Yiming Li, and Chris Xiaoxuan Lu. Radarocc: Robust 3d occupancy prediction with 4d imaging radar. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 101589–101617. Curran Associates, Inc., 2024. doi: 10.52202/079017-3222. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b81d83165e3145a2e7d33bb5e33ea913-Paper-Conference.pdf.

[8] Laura Dodds, Maisy Lam, Waleed Akbar, Yibo Cheng, and Fadel Adib. Wave-former: Through-occlusion 3d reconstruction via wireless shape completion. *arXiv preprint arXiv:2511.14152*, 2025.

[9] Zhaoxin Fan, Yulin He, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Reconstruction-aware prior distillation for semi-supervised point cloud completion. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 726–735. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/81. URL https://doi.org/10.24963/ijcai.2023/81. Main Track.

[10] Ruixu Geng, Yadong Li, Dongheng Zhang, Jincheng Wu, Yating Gao, Yang Hu, and Yan Chen. Dream-pcd: Deep reconstruction and enhancement of mmwave radar point-cloud, 2023.

[11] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through fog high-resolution imaging using millimeter wave radar. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11470, 2020. doi: 10.1109/CVPR42600.2020.01148.

[12] Hesai Technology. OT128 Automotive-Grade 360° Long-Range Lidar, 2026. URL https://www.hesaitech.com/product/ot128/.

[13] Samah Hussein, Junfeng Guan, Swathi Shree Narashiman, Saurabh Gupta, and Haitham Al Hassanieh. 3d shape reconstruction from autonomous driving radars. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_399/paper.pdf.

[14] Pou-Chun Kung, Skanda Harisha, Ram Vasudevan, Aline Eid, and Katherine A Skinner. Radarsplat: Radar gaussian splatting for high-fidelity data synthesis and 3d reconstruction of autonomous driving scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27596–27606, 2025.

[15] 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*, ACM MobiCom '24, page 388–403, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704895. doi: 10.1145/3636534.3649369. URL https://doi.org/10.1145/3636534.3649369.

[16] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5715–5724, January 2023.

[17] An Li, Zhe Zhu, and Mingqiang Wei. Genpc: Zero-shot point cloud completion via 3d generative priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1308–1318, 2025.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] Yibo Liu, Kelly Zhu, Guile Wu, Yuan Ren, Bingbing Liu, Yang Liu, and Jinjun Shan. Mv-deepsdf: Implicit modeling with multi-sweep point clouds for 3d vehicle reconstruction in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8306–8316, October 2023.

[20] 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*, MobiSys '20, page 14–27, New

York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379540. doi: 10.1145/3386901.3388945. URL https://doi.org/10.1145/3386901.3388945.

[21] Jiachen Lu, Hailan Shanbhag, and Haitham Al Hassanieh. Geraf: Neural geometry reconstruction from radio frequency signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=z3PMVmzoya.

[22] Yukai Ma, Jianbiao Mei, Xuemeng Yang, Licheng Wen, Weihua Xu, Jiangning Zhang, Xingxing Zuo, Botian Shi, and Yong Liu. Licrocc: Teach radar for accurate semantic occupancy prediction using lidar and camera. *IEEE Robotics and Automation Letters*, 10(1):852–859, 2025. doi: 10.1109/LRA.2024.3511427.

[23] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] 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, 2023. doi: 10.1109/ICRA48891.2023.10161429.

[25] Emerson Sie, Xinyu Wu, Heyu Guo, and Deepak Vasisht. Radarize: Enhancing radar slam with generalizable doppler-based odometry. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, MOBISYS '24, page 331–344, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705816. doi: 10.1145/3643832.3661871. URL https://doi.org/10.1145/3643832.3661871.

[26] Stereolabs Inc. ZED 2. https://www.stereolabs.com/products/zed-2, 2022. [Online; accessed jun-20-2024].

[27] 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, 2021. doi: 10.1109/IPCCC51483.2021.9679394.

[28] Texas Instruments Inc. MMWCAS-RF-EVM mmwave cascade imaging radar rf evaluation module. https://www.ti.com/tool/MMWCAS-RF-EVM, 2024. [Online; accessed jun-24-2024].

[29] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023. doi: 10.1109/LRA.2023.3236571.

[30] Xiaogang Wang, Marcelo H Ang, and Gim Hee Lee. Point cloud completion by learning shape priors. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10719–10726. IEEE, 2020.

[31] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[32] Ruijie Zhang, Bixin Zeng, Shengpeng Wang, Fuhui Zhou, and Wei Wang. Rald: Generating high-resolution 3d radar point clouds with latent diffusion. *arXiv preprint arXiv:2511.07067*, 2025.

[33] Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. 2024.

[34] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018. doi: 10.1109/CVPR.2018.00768.

[35] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.