Rascene: High-Fidelity 3D Scene Imaging with mmWave Communication Signals

Kunzhe Song Geo Jie Zhou Xiaoming Liu Huacheng Zeng
Department of Computer Science and Engineering, Michigan State University
{songkunz, geozhou, liuxm, hzeng}@msu.edu

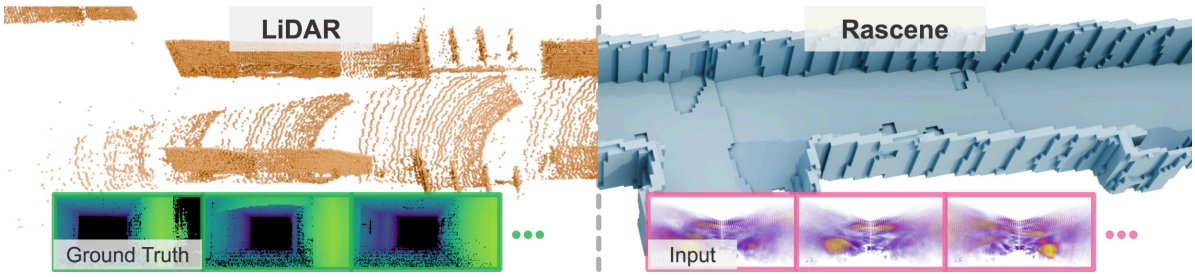

Figure 1. High-fidelity 3D imaging generated by Rascene from mmWave communication signals. We show that OFDM communication signals can support high-fidelity 3D imaging on a single device. Our multi-frame RF fusion suppresses multipath artifacts and integrates sparse observations into a complete 3D scene estimate (blue) that is closer to LiDAR ground truth (orange).

Abstract

Robust 3D environmental perception is critical for applications such as autonomous driving and robot navigation. However, optical sensors such as cameras and LiDAR often fail under adverse conditions, including smoke, fog, and non-ideal lighting. Although specialized radar systems can operate in these environments, their reliance on bespoke hardware and licensed spectrum limits scalability and cost-effectiveness. This paper introduces Rascene, an integrated sensing and communication (ISAC) framework that leverages ubiquitous mmWave OFDM communication signals for 3D scene imaging. To overcome the sparse and multipath-ambiguous nature of individual radio frames, Rascene performs multi-frame, spatially adaptive fusion with confidence-weighted forward projection, enabling the recovery of geometric consensus across arbitrary poses. Experimental results demonstrate that our method reconstructs 3D scenes with high precision, offering a new pathway toward low-cost, scalable, and robust 3D perception.

1. Introduction

Robust 3D environmental perception is critical for autonomous navigation and robotics. Existing 3D perception systems predominantly rely on cameras and LiDAR. However, camera-based methods [17, 24, 30, 33] are fundamentally constrained by non-ideal lighting and fail in the presence of visual obscurants like smoke, fog, and snow. LiDAR systems, while offering accurate geometric measurements [6, 7, 21], remain expensive, bulky, and are similarly susceptible to adverse weather and occluding materials.

Radar-based 3D imaging has emerged as a compelling alternative [10, 19, 23, 27], as it is robust to lighting variations and can sense through occlusions such as smoke and fog. However, practical deployment of *specialized* radar systems remains challenging. These systems often require ultra-wideband (multi-GHz) spectrum allocations [37, 47], which demand dedicated licenses or risk interference to incumbent systems. Furthermore, integrating bespoke sensing hardware increases cost, size, and power consumption [1, 3], limiting suitability for compact, energy-constrained platforms such as AR/VR headsets and home robots.

This paper introduces Rascene, an integrated sensing and communication (ISAC) framework that bridges this gap by utilizing millimeter-wave (mmWave) communication signals (*e.g.*, 5G and Wi-Fi) for high-fidelity 3D imaging (Fig. 1). Tab. 1 provides a comparison between Rascene and conventional sensing modalities. Unlike *specialized* radar, Rascene integrates sensing into communication systems. It leverages OFDM communication waveforms to extract fine-grained range and angle information without requiring dedicated sensing hardware or licensed spectrum. By leveraging communication network infrastructure, Rascene provides a scalable and low-cost pathway for robust 3D perception on commodity Wi-Fi and cellular devices.

The design of Rascene overcomes two fundamental challenges. The first is reliable RF data acquisition. Most communication-based sensing systems are *bistatic*, relying

Table 1. Comparison of Rascene and conventional modalities. [Keys: O.P.=*Obstacle Penetration*, O.R.=*Occlusion Resilience*, S.L.E.=*Spectrum License Exempt*, P.C.=*Power Consumption*.]

Tech.	Medium	Waveform	O.P.	O.R.	S.L.E.	Hardware	P.C.	Cost	Scalable
Camera	Light	—	No	Poor	—	Dedicated	Low	Low	High
LiDAR	Laser	Pulsed	No	Poor	—	Dedicated	High	High	Low
Radar	Radio	FMCW, etc.	Yes	Good	No	Dedicated	Med	Med	Low
Rascene	Radio	OFDM	Yes	Good	Yes	Reused	Low	∼Zero	High

on separate transmitter (Tx) and receiver (Rx). The dynamic relative poses of Tx and Rx devices introduce fundamental uncertainty. To address this, Rascene is built on our key finding (Sec. 3) that commodity mmWave devices can operate in full-duplex mode for *monostatic* sensing. The high directionality of phased-array antennas and short carrier wavelength provide sufficient Tx/Rx isolation. This capability allows Rascene to perform precise Channel Impulse Response (CIR) measurements using co-located antennas, eliminating pose uncertainties and enabling the generation of 3D point clouds analogous to those from FMCW radar.

The second challenge stems from the inherent nature of RF signals. Single-frame RF observations are low-resolution, sparse, and corrupted by multipath. To address this, we introduce a multi-frame 3D imaging network that fuses arbitrarily posed observations, where each frame contributes a latent feature volume with confidence-aware forward projection. This source-driven fusion enforces geometric consensus, suppresses multipath-induced ghost or hallucinated structures, and improves scene completeness.

We collect a large-scale dataset across diverse indoor environments to evaluate Rascene. Experiments demonstrate Rascene's strong cross-scene generalization. Under within-dataset evaluation, Rascene achieves the best overall performance among all baselines. Furthermore, Rascene benefits substantially from multi-frame inputs, validating the effectiveness of the proposed multi-frame fusion mechanism.

The main contributions of this work are as follows:

- We introduce Rascene, an ISAC framework that enables 3D imaging using mmWave OFDM communication signals, without dedicated sensing hardware or spectrum.
- We propose a multi-frame, confidence-aware fusion framework for RF signals to overcome the inherent challenges of noise, sparsity, and multipath corruption.
- Experiments on our collected dataset show that Rascene reconstructs scenes with high fidelity and outperforms baselines in within-dataset evaluation.

2. Related Work

2.1. RF Sensing

Prior RF sensing systems mainly relied on specialized radar hardware, with FMCW radar being a prevalent technology [1–4, 48]. However, the deployment of these solutions is constrained by their reliance on bespoke, ultra-wideband hardware. For instance, some radar imaging systems utilize bandwidths up to 4 GHz [10, 27, 37, 47], which require dedicated spectrum licenses from FCC.

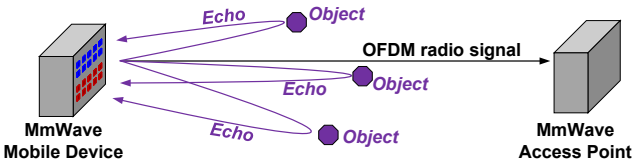

Figure 2. Illustration of monostatic sensing in a mmWave communication system. The mmWave device (left) simultaneously transmits and receives OFDM communication signal for sensing.

Recently, ISAC [9, 13, 14, 20, 25, 31, 41, 44] has emerged as a promising paradigm, which reuses existing communication signals and infrastructures for sensing. By exploiting Channel State Information (CSI) characteristics, sub-6 GHz Wi-Fi networks have been utilized for a variety of sensing tasks [26, 39, 40, 46]. While attractive for deployment, such systems are constrained by their bistatic setup, narrow bandwidth, and small antenna arrays, which makes high-fidelity 3D imaging challenging. In contrast, Rascene performs monostatic sensing in the mmWave band and introduces adaptive multi-frame fusion for 3D imaging.

2.2. Multi-View 3D Imaging

Imaging from Dense Visual Data. Multi-view 3D imaging is a central topic in computer vision. Classical methods [5, 16, 17, 34] exploit photometric consistency across calibrated images to estimate depth maps. More recently, neural implicit representations [24, 29, 33, 43] have achieved strong performance in novel view synthesis and scene reconstruction. A key assumption of these methods is the availability of dense and informative visual observations, where geometry can be inferred through reliable feature correspondence or photometric optimization across pixels.

Imaging from Sparse LiDAR Data. More closely related to our setting are methods that reconstruct scenes from sparse point clouds, typically captured by LiDAR sensors. While sparse, LiDAR directly measures scene geometry with high fidelity. Hence, aggregating multiple LiDAR frames, often registered using algorithms like ICP or SLAM [6, 15, 21, 35], is relatively straightforward. The fused point clouds are often dense enough to be directly processed [22, 38, 45, 49] or converted into other representations like TSDF [8, 11, 12, 18, 32, 36] for surface reconstruction.

3. RF Data Acquisition and Representation

MmWave communications have been widely adopted in both 5G and Wi-Fi standards to meet the ever-increasing demand for high data rates and low-latency connectivity. Consider a mobile mmWave communication device as shown in Fig. 2. When it transmits data packets, the emitted OFDM RF signals inherently illuminate the surrounding objects. If

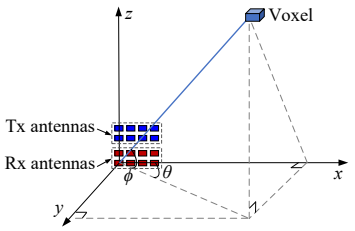

Figure 3. Illustration of angular estimation on a mmWave device.

the device is capable of transmitting and receiving simultaneously, it can capture the backscattered signals and estimate the Channel Impulse Response (CIR), thereby enabling precise object ranging without requiring dedicated sensing hardware, sensing waveform, or spectrum license. In practice, mmWave communication devices can indeed operate in full-duplex mode for *monostatic* sensing thanks to their highly directional phased-array antennas and short carrier wavelength, which together provide sufficient RF isolation between transmission and reception paths [42].

3.1. CIR-based Ranging

Consider an OFDM symbol embedded in a 5G or Wi-Fi data packet. Denote $X(k) \in \mathbb{C}$ as the data on OFDM subcarrier k, where $k = 0, 1, \ldots, K-1$ and K is the total number of subcarriers. Since the device is transmitting and receiving simultaneously, it receives a reflected copy of its transmitted OFDM signal from surrounding objects. Let $Y(k) \in \mathbb{C}$ denote the received data on subcarrier k. The channel response on subcarrier k can be estimated as $\hat{H}(k) = \frac{Y(k)}{X(k)}$. Then, the corresponding CIR can be computed by: $\hat{\mathbf{h}} \triangleq [\hat{h}(0), \hat{h}(1), \ldots, \hat{h}(K-1)] = \text{IFFT}\left([\hat{H}(0), \hat{H}(1), \ldots, \hat{H}(K-1)]\right)$, where $\hat{h}(n) \in \mathbb{C}$ denotes the coefficient of the n-th tap of the time-domain channel, $n = 0, 1, \ldots, K-1$.

Since the transmitter and receiver are co-located on the same device, their clock frequencies and data flow timing are synchronized. This synchronization allows the estimated CIR to be directly used for object ranging. Specifically, a strong magnitude $|\hat{h}(n)|$ indicates the presence of an object at a distance of $\frac{nc}{2B}$, where c is the light speed and B is the OFDM signal bandwidth. This capability fundamentally distinguishes the *monostatic* configuration adopted in Rascene from conventional *bistatic* systems, where the transmitter and receiver are physically separated and the estimated CIR cannot be used for object ranging.

3.2. Radio Imaging via Spatial Projection

Angular estimation is another critical component of radio imaging, as resolving the angular positions of reflections enables objects to be separated spatially. Commercial mmWave communication devices are widely equipped with phased-array antennas, which can be leveraged for angular estimation without requiring additional sensing hardware.

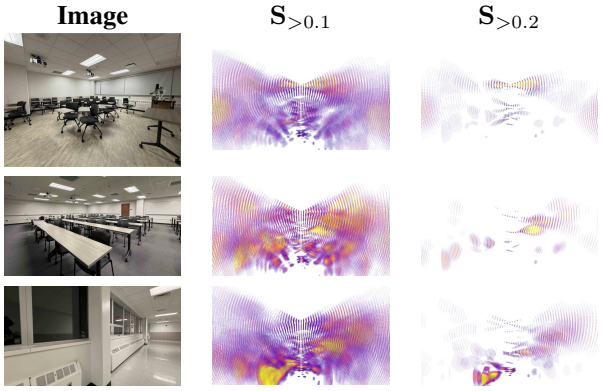

Figure 4. Examples of generated radio point clouds, where $\mathbf{S}_{>a} = \{s \in \mathbf{S} \mid s > a\}$. Threshold a is set to 0.1 and 0.2, and the 3D point clouds are projected onto 2D images for for ease of visualization.

Referring to Fig. 3, we define the patch antenna panel as the x-z plane, *i.e.*, the plane $y = 0$. Consider a voxel in the 3D space, and let θ and ϕ be its azimuth and elevation angles, respectively. The unit vector pointing from the antenna array to this voxel can then be written as:

$$\vec{u}(\theta, \phi) = [\cos(\theta)\cos(\phi),\ \cos(\theta)\sin(\phi),\ \sin(\phi)]. \quad (1)$$

Let $\mathcal{T}$ and $\mathcal{R}$ denote the sets of Tx and Rx antennas, respectively. Denote $\vec{p}_i = (x_i, 0, z_i)$ as the coordinate of Tx antenna $i \in \mathcal{T}$ and $\vec{q}_j = (x'_j, 0, z'_j)$ as the coordinate of Rx antenna $j \in \mathcal{R}$. When antenna i transmits radio signal and antenna j receives the backscattered signal from a voxel in direction (θ, ϕ), the additional signal propagation distance relative to the antenna array center is $\langle \vec{u}(\theta, \phi),\ \vec{p}_i + \vec{q}_j \rangle$, where $\langle \cdot, \cdot \rangle$ is inner product. Thus, the complex weight associated with the antenna pair (i, j) can be written as:

$$w_{i,j}(\theta, \phi) = \exp\big(-j\frac{2\pi}{\lambda}\langle \vec{u}(\theta, \phi),\ \vec{p}_i + \vec{q}_j \rangle\big), \quad (2)$$

where λ is the wavelength of carrier radio signal.

Let $\hat{\mathbf{h}}_{ij} = [\hat{h}_{ij}(0), \hat{h}_{ij}(1), \ldots, \hat{h}_{ij}(K-1)]$ be the CIR measurement when Rascene uses antenna pair (i, j) for signal transmission and reception. Then, the reflected signal strength of voxel (n, θ, ϕ) can be computed by:

$$s(n, \theta, \phi) = \left| \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{R}} w_{i,j}(\theta, \phi) \hat{h}_{ij}(n) \right|, \quad (3)$$

where n is the voxel distance index, with absolute distance being $r = \frac{nc}{2B}$. A large value of $s(n, \theta, \phi)$ indicates the presence of an object, and a small value indicates its absence.

Based on Eq. (3), a radio frame, represented by the point clouds in the spherical coordinate system, can be written as:

$$\mathbf{S} = \{s(n, \theta, \phi) \mid\ n \in \mathbf{N}, \theta \in \mathbf{\Theta}, \phi \in \mathbf{\Phi}\}. \quad (4)$$

Fig. 4 shows three radio frames of point clouds generated by a mmWave communication device.

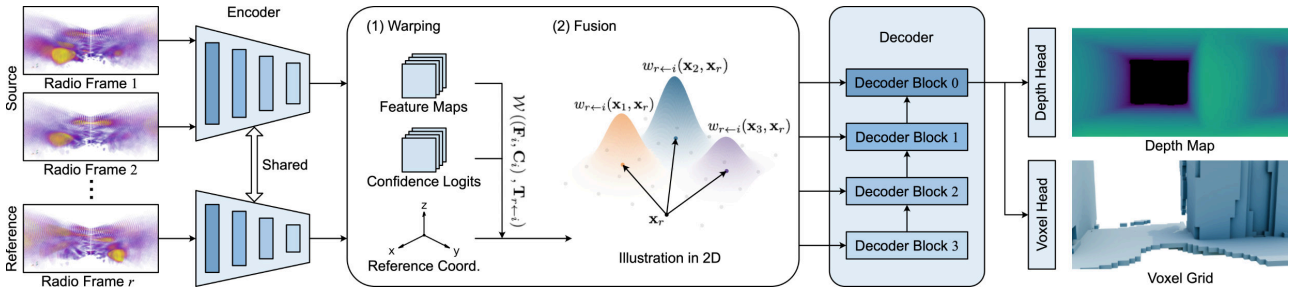

Figure 5. Overview of the multi-frame 3D RF imaging network. Given multiple radio frames and poses, a shared encoder predicts per-frame feature maps and confidence logits. We then warp all features to a reference frame and fuse them into a unified representation. A coarse-to-fine 3D decoder with voxel and depth heads outputs the reconstructed voxel grid and depth map.

4. Multi-Frame 3D RF Imaging

Given radio frames $\mathcal{S} = \{\mathbf{S}_i \in \mathbb{R}^{|\mathbf{N}|\times|\mathbf{\Theta}|\times|\mathbf{\Phi}|}\}_{i=1}^{N}$ captured at known poses $\mathcal{G} = \{\mathbf{G}_i \in SE(3)\}_{i=1}^{N}$, our objective is to learn a mapping function $\mathcal{F}$ that estimates a sequence of dense 3D voxel grids $\mathcal{V} = \{\hat{\mathbf{V}}_i \in \mathbb{R}^{X\times Y\times Z}\}_{i=1}^{N}$ and corresponding depth maps $\mathcal{D} = \{\hat{\mathbf{D}}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N}$ from the radio frames captured along arbitrary motion trajectories:

$$(\mathcal{V}, \mathcal{D}) = \mathcal{F}(\mathcal{S}, \mathcal{G}). \tag{5}$$

The main challenge lies in cross-frame fusion under sparse and multipath-corrupted RF observations. We address this with an RF imaging framework (Sec. 4.1) and a spatially adaptive warping and fusion module (Sec. 4.2) that enforces geometric consistency across arbitrary poses. An overview of the proposed framework is shown in Fig. 5.

4.1. RF Imaging Framework

Unlike conventional multi-view reconstruction, RF observations are neither texture-rich like RGB images nor geometrically explicit like LiDAR points. Instead, radio frames are strongly affected by attenuation and multipath (Fig. 4), which makes single-frame 3D imaging ill-posed.

Following Sec. 3, each $\mathbf{S}_i$ is transformed from spherical measurement space (n, θ, ϕ) into a Cartesian volume in the local sensor frame. For simplicity, we reuse $\mathbf{S}_i$ to denote this Cartesian representation. We then choose a reference frame $r \in \{1, \dots, N\}$ and transform all frames into its coordinate system. A shared encoder $\mathcal{E}$, parameterized by $\boldsymbol{\theta}_\mathcal{E}$, maps each frame to a latent feature volume $\mathbf{F}_i \in \mathbb{R}^{X\times Y\times Z\times C_F}$ and confidence logits $\mathbf{C}_i \in \mathbb{R}^{X\times Y\times Z}$:

$$(\mathbf{F}_i, \mathbf{C}_i) = \mathcal{E}(\mathbf{S}_i\,; \boldsymbol{\theta}_\mathcal{E}). \tag{6}$$

Since $(\mathbf{F}_i, \mathbf{C}_i)$ are defined in local coordinates, we warp them into the reference frame r using the relative rigid transformation $\mathbf{T}_{r\leftarrow i}$ and a warping operator $\mathcal{W}(\cdot, \mathbf{T})$:

$$\mathbf{T}_{r\leftarrow i} = \mathbf{G}_r^{-1}\mathbf{G}_i = \begin{bmatrix} \mathbf{R}_{r\leftarrow i} & \mathbf{t}_{r\leftarrow i} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{7}$$

$$(\tilde{\mathbf{F}}_i, \tilde{\mathbf{C}}_i) = \mathcal{W}\left((\mathbf{F}_i, \mathbf{C}_i), \mathbf{T}_{r\leftarrow i}\right). \tag{8}$$

After warping, the fusion operator $\mathcal{A}$ aggregates all warped features into a unified latent representation $\mathbf{Z}_r$, which captures geometric consensus across frames while suppressing multipath artifacts. Because this fused representation is still sparse, we employ a coarse-to-fine 3D decoder $\mathcal{O}$ to progressively densify $\mathbf{Z}_r$ into a dense feature volume $\mathbf{H}_r$. Finally, two task-specific heads, $\mathcal{H}_v$ and $\mathcal{H}_d$, predict the voxel grid $\hat{\mathbf{V}}_r$ and depth map $\hat{\mathbf{D}}_r$ from $\mathbf{H}_r$:

$$\mathbf{Z}_r = \mathcal{A}\left(\{(\tilde{\mathbf{F}}_i, \tilde{\mathbf{C}}_i)\}_{i=1}^{N}\right), \tag{9}$$

$$\mathbf{H}_r = \mathcal{O}(\mathbf{Z}_r\,; \boldsymbol{\theta}_\mathcal{O}), \tag{10}$$

$$\hat{\mathbf{V}}_r = \mathcal{H}_v(\mathbf{H}_r\,; \boldsymbol{\theta}_v), \quad \hat{\mathbf{D}}_r = \mathcal{H}_d(\mathbf{H}_r\,; \boldsymbol{\theta}_d). \tag{11}$$

The multi-frame 3D RF imaging framework, parameterized by $\mathbf{\Theta} = \{\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\mathcal{O}, \boldsymbol{\theta}_v, \boldsymbol{\theta}_d\}$, is optimized end-to-end. For a window of N frames, each frame is used once as the reference, and losses are summed over all references:

$$\mathcal{L} = \sum_{r=1}^{N}\left(\lambda_v \mathcal{L}_{\text{voxel}}^{(r)} + \lambda_d \mathcal{L}_{\text{depth}}^{(r)}\right), \tag{12}$$

where λ_v and λ_d are scalar hyperparameters that balance the two tasks. The voxel loss $\mathcal{L}_{\text{voxel}}^{(r)}$ is the binary cross-entropy loss between the predicted grid $\hat{\mathbf{V}}_r$ and its corresponding ground truth $\mathbf{V}_r^*$ over the volume Ω:

$$\mathcal{L}_{\text{voxel}}^{(r)} = -\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}[\mathbf{V}_r^*(\mathbf{x})\log(\hat{\mathbf{V}}_r(\mathbf{x})) + (1-\mathbf{V}_r^*(\mathbf{x}))\log(1-\hat{\mathbf{V}}_r(\mathbf{x}))], \tag{13}$$

and the depth loss $\mathcal{L}_{\text{depth}}^{(r)}$ is the L1 loss between the predicted depth map $\hat{\mathbf{D}}_r$ and its ground truth $\mathbf{D}_r^*$:

$$\mathcal{L}_{\text{depth}}^{(r)} = \frac{1}{HW}\sum_{(u,v)}\left|\hat{\mathbf{D}}_r(u,v) - \mathbf{D}_r^*(u,v)\right|. \tag{14}$$

4.2. Spatially Adaptive Warping and Fusion

Instantiating the warping $\mathcal{W}$ in Eq. (8) and fusion $\mathcal{A}$ in Eq. (9) operators is the core challenge of our framework.

Unlike LiDAR points, which are geometrically explicit, RF observations provide only indirect and highly ambiguous measurements of scene geometry. As a result, direct cross-frame warping and aggregation can easily accumulate inconsistent evidence and spurious reflections.

To address this issue, we formulate both $\mathcal{W}$ and $\mathcal{A}$ as sparse projection and aggregation operators. Rather than querying values at target voxels, each source voxel $\mathbf{x}_i \in \Omega_i$ is transformed into the reference frame and contributes only within a local support region in the target grid. This strategy avoids repeated sampling of empty target locations and better preserves sparse yet informative RF responses that are critical for reconstruction. Given frame i, the continuous reference-frame coordinate of source voxel $\mathbf{x}_i$ is:

$$\tilde{\mathbf{x}}_{r\leftarrow i} = \mathbf{T}_{r\leftarrow i}\mathbf{x}_i. \tag{15}$$

The source feature $\mathbf{F}_i(\mathbf{x}_i)$ is then distributed onto the target grid Ω_r by contributing to neighboring voxels $\mathbf{x} \in \mathcal{N}(\tilde{\mathbf{x}}_{r\leftarrow i})$ within a given radius R. Since our transformations are purely rigid, without anisotropic scaling or shearing, we employ a simple but effective isotropic Gaussian kernel K_σ to model geometric contribution, controlled by a parameter σ:

$$K_\sigma(\tilde{\mathbf{x}}_{r\leftarrow i}, \mathbf{x}) = \exp\left(-\frac{\|\tilde{\mathbf{x}}_{r\leftarrow i} - \mathbf{x}\|^2}{2\sigma^2}\right). \tag{16}$$

The fusion operator $\mathcal{A}$ leverages this projection by performing a spatially adaptive, confidence-weighted average. We define an adaptive weight $w_{r\leftarrow i}(\mathbf{x}_i, \mathbf{x}_r)$ for the contribution of source voxel $\mathbf{x}_i$ in frame i to target voxel $\mathbf{x}_r$ in frame r. This weight combines the geometric proximity kernel K_σ with predicted signal reliability at the source, derived from the confidence logits $\mathbf{C}_i(\mathbf{x}_i)$:

$$w_{r\leftarrow i}(\mathbf{x}_i, \mathbf{x}_r) = K_\sigma(\tilde{\mathbf{x}}_{r\leftarrow i}, \mathbf{x}_r) \cdot [\alpha(\mathbf{C}_i(\mathbf{x}_i))]^\eta, \tag{17}$$

where $\alpha(c) = \log(1 + e^c)$ maps confidence logits c to a non-negative reliability score. We raise this reliability score to the power η, a hyperparameter that controls the sharpness of confidence weighting.

The final unified representation $\mathbf{Z}_r$ is then computed as a normalized weighted average. By weighting each contribution using the learned non-linear reliability, this formulation ensures that the aggregate consensus is dominated by high-confidence geometric signals, yielding stable and robust fusion:

$$\mathbf{Z}_r(\mathbf{x}_r) = \frac{\sum_{i=1}^{N}\sum_{\mathbf{x}_i\in\Omega_i} w_{r\leftarrow i}(\mathbf{x}_i, \mathbf{x}_r)\mathbf{F}_i(\mathbf{x}_i)}{\sum_{i=1}^{N}\sum_{\mathbf{x}_i\in\Omega_i} w_{r\leftarrow i}(\mathbf{x}_i, \mathbf{x}_r) + \varepsilon}, \tag{18}$$

where $w_{r\leftarrow i}(\mathbf{x}_i, \mathbf{x}_r) = 0$ when $\mathbf{x}_r \notin \mathcal{N}(\tilde{\mathbf{x}}_{r\leftarrow i})$, and $\varepsilon > 0$ ensures numerical stability in low-support regions. Ω_i denotes the grid of source voxels in frame i.

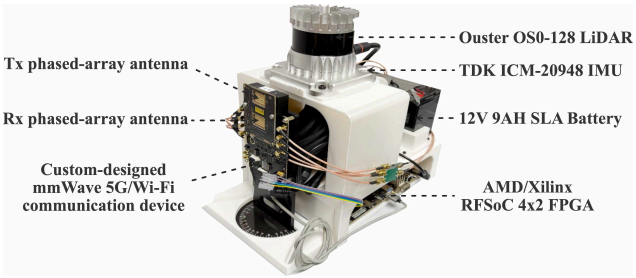

Figure 6. Rascene data collection platform.

5. Implementation

Hardware Setup. We built a prototype of Rascene as shown in Fig. 6 to evaluate its performance in realistic scenarios. The prototyped Rascene comprises three main sensors: (a) a custom-designed mmWave 5G/Wi-Fi communication device that supports full-duplex operation for monostatic sensing, (b) a commercial LiDAR sensor, and (c) a commercial IMU. Our custom-designed communication device operates at 60 GHz with a bandwidth of 1.2288 GHz, featuring 16 transmit (Tx) and 16 receive (Rx) antenna elements. It provides an effective sensing range of 7 meters, covering a field of view (FoV) of 120° horizontally and 60° vertically. For ground truth, we use an Ouster 128-beam LiDAR to capture 3D point clouds. For 6-DoF pose acquisition, we integrate a TDK ICM-20948 9-axis IMU.

Model Architecture. Each RF frame is projected to a Cartesian voxel grid of size $64 \times 64 \times 32$ (12 cm voxel size). The encoder $\mathcal{E}$ and decoder $\mathcal{O}$ both use four convolutional stages, with channel multipliers $(1, 2, 4, 8)$ relative to the stem channels. Following Sec. 4.2, we perform stage-wise warping $\mathcal{W}$ and fusion $\mathcal{A}$ after each encoder stage to align and aggregate multi-view evidence in the reference frame. For the sparse projection operator, we use neighborhood radius $R = 2$ and confidence sharpness $\eta = 3$.

Data Collection. We collected synchronized RF-LiDAR frame pairs across 20 indoor environments. The LiDAR and ISAC streams are recorded at 10 Hz while the platform moves at 0.5 m/s. We sample one frame every 2 s and group five sampled frames into a window, which preserves sufficient overlap for geometric consistency while providing meaningful viewpoint diversity.

6. Experiments

6.1. Main Results

Quantitative Results. We first evaluate cross-scene generalization by training on 12 environments and testing on 8 unseen environments (A–H), with results in Tab. 2. Across all test scenes, Rascene achieves strong average performance on both depth estimation and voxel reconstruction. Although error varies with scene difficulty (*e.g.*, "E"/"F"

Table 2. Quantitative evaluation results of Rascene's cross-scene generalization performance. The model is trained on 12 scenarios and evaluated on 8 distinct, unseen test scenarios (A-H). We report metrics for both depth estimation (AbsRel, MAE, RMSE) and 3D voxel reconstruction (CD, CD_{Diag}). Lower values are better for all metrics.

Metric		A	B	C	D	E	F	G	H	Avg.
Depth	AbsRel (%)	10.1	4.9	4.5	3.3	17.3	19.4	7.4	8.7	9.4
	MAE (cm)	22.9	11.5	10.2	7.3	36.2	37.4	16.8	19.9	20.2
	RMSE (cm)	43.2	21.2	22.2	17.7	63.3	66.9	34.4	38.1	38.0
Voxel	CD (cm)	20.4	15.4	13.0	9.1	37.0	26.5	16.8	16.9	19.7
	CD_{Diag}(%)	2.5	1.7	1.5	1.1	3.7	3.3	2.0	2.1	2.3

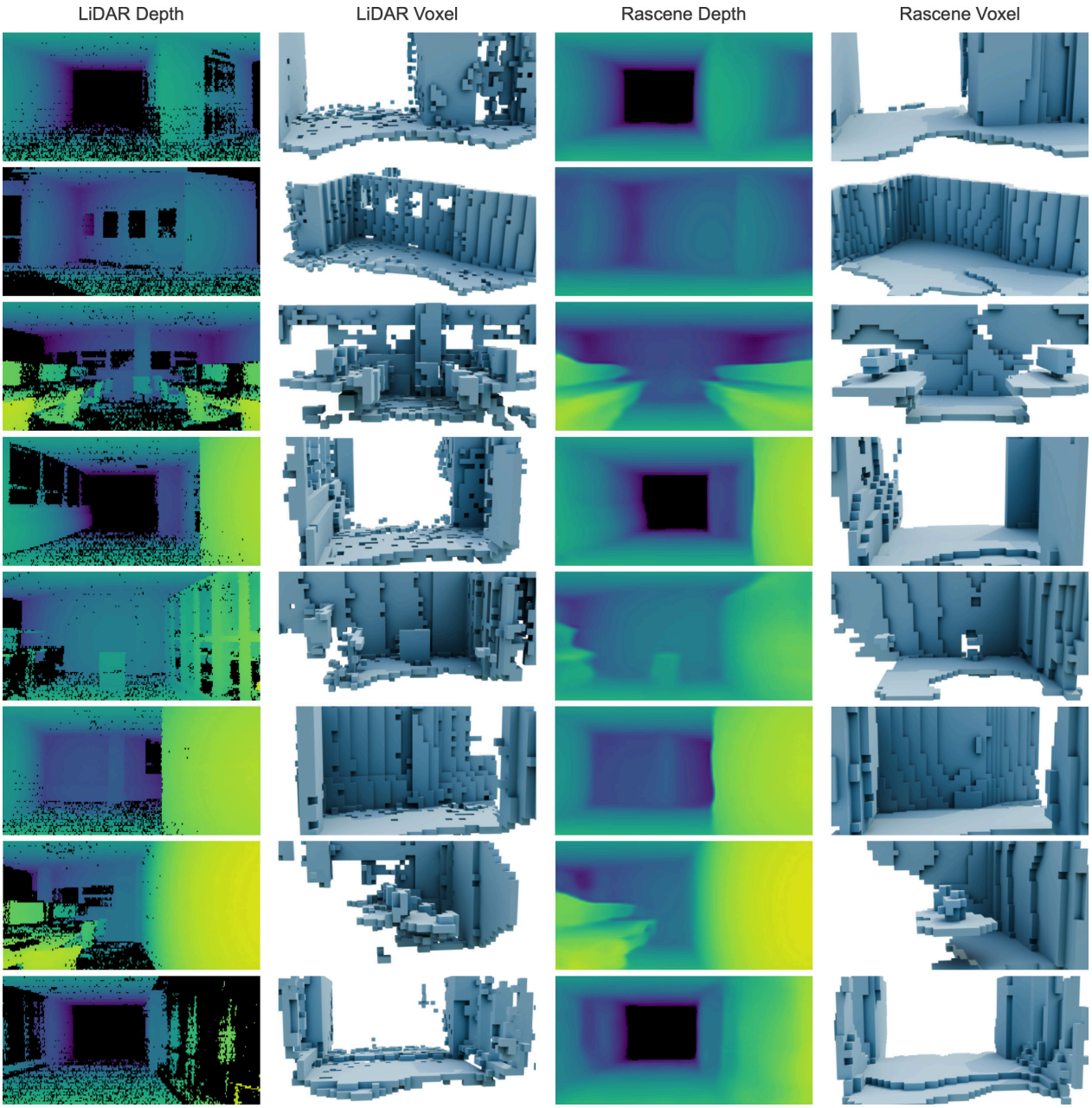

Figure 7. Qualitative results of our Rascene system. For each row, we show the ground truth depth map and voxel grid derived from the LiDAR sensor (Cols 1-2), and our corresponding predictions generated using mmWave communication signal (Cols 3-4).

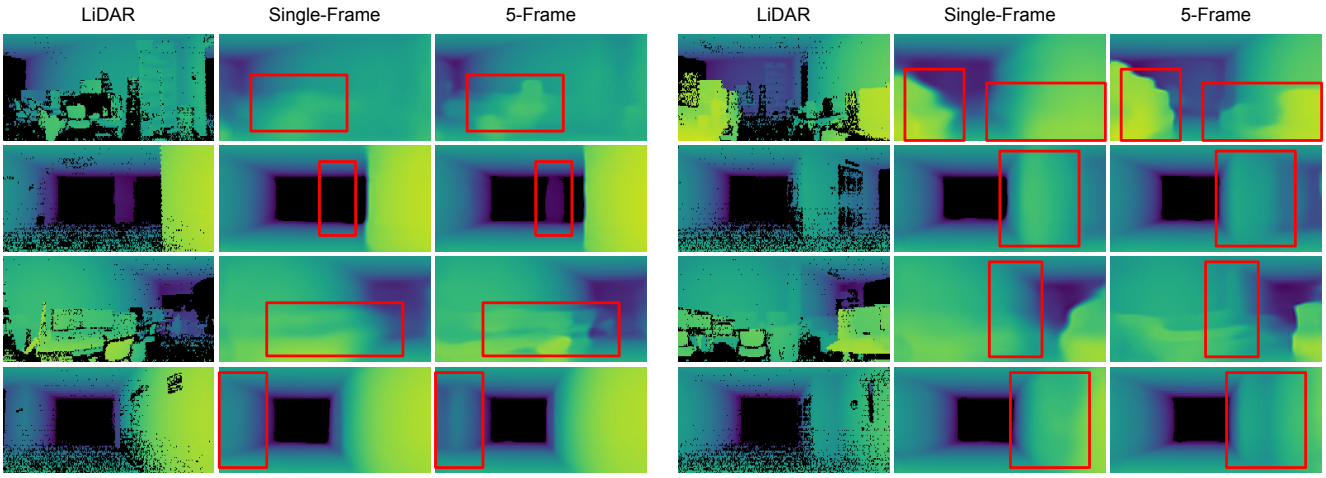

Figure 8. Qualitative comparison of single-frame and 5-frame predictions. Multi-frame fusion reduces missed detections and suppresses hallucinated structures, yielding more complete and geometrically consistent reconstructions closer to the LiDAR ground truth.

Table 3. Within-dataset comparison with baselines.

Methods	# of frame	AbsRel	MAE	CD	CD_{Diag}
PanoRadar [27]	1	14.7%	34.1	32.2	3.8%
CartoRadar [28]	5	—	—	26.8	3.1%
Rascene (Ours)	1	14.1%	32.9	31.6	3.6%
Rascene (Ours)	**5**	**9.4%**	**20.2**	**19.7**	**2.3%**

are harder than "C"/"D"), the normalized Chamfer Distance CD_{Diag} remains consistently low (1.1%–3.7%), indicating stable geometric quality across environments.

Qualitative Results. Fig. 7 provides qualitative comparisons between LiDAR targets (Cols 1–2) and Rascene predictions from mmWave communication signals (Cols 3–4). In many scenes, LiDAR targets contain no-return regions (black pixels), commonly caused by absorption on low-albedo surfaces (*e.g.*, dark carpets) and specular reflection on smooth materials (*e.g.*, glass). Despite these challenging regions, Rascene recovers coherent scene geometry, highlighting the complementary robustness of RF sensing to optical material failure modes.

Comparison with Baselines. For a fair comparison, we implement the models of PanoRadar [27] and CartoRadar [28] and evaluate all methods on the same within-dataset split. As shown in Tab. 3, Rascene with 5-frame fusion achieves the best overall performance. More importantly, compared with CartoRadar, our model benefits more from multi-frame, validating the effectiveness of our multi-frame RF fusion module.

6.2. Ablation Study

Impact of Multi-Frame Fusion. We study the effect of the number of fused frames N in Eq. (9). Tab. 4 shows that increasing N from 1 to 5 steadily improves both depth and voxel metrics. The most significant gain occurs when moving from 1 to 2 frames, indicating that even one additional viewpoint introduces strong geometric constraints for disambiguation. Fig. 8 further shows that multi-frame fusion suppresses hallucinated structures and alleviates missed detections, resulting in more complete geometry.

Table 4. Impact of the number of fused radio frames (N in Eq. (9)).

Frame #	Depth			Voxel	
	AbsRel	MAE	RMSE	CD	CD_{Diag}
1	14.1%	32.9	56.2	31.6	3.6%
2	11.1%	24.6	44.2	26.0	3.0%
3	9.8%	21.8	40.4	21.9	2.5%
5	9.4%	20.2	38.0	19.7	2.3%

Table 5. Impact of pose inaccuracies (rotation and translation perturbations added at test time).

Error Range		Depth			Voxel	
		AbsRel	MAE	RMSE	CD	CD_{Diag}
Rot.	$0^\circ \sim 5^\circ$	11.7%	25.8	46.8	23.1	2.6%
	$5^\circ \sim 10^\circ$	18.4%	40.0	66.8	31.2	3.6%
	$10^\circ \sim 15^\circ$	18.8%	42.0	67.7	34.3	3.9%
Trans.	$0 \sim 5$ cm	9.4%	20.2	38.1	19.7	2.3%
	$5 \sim 10$ cm	9.4%	20.4	38.2	19.8	2.3%
	$10 \sim 15$ cm	9.5%	20.6	38.5	20.0	2.3%
No Perturbation		9.4%	20.2	38.0	19.7	2.3%

Robustness to Pose Inaccuracies. We evaluate robustness to pose noise by perturbing ground-truth poses at test time. As shown in Tab. 5, our model remains highly stable under translation perturbations while being more sensitive to rotation errors. This trend is geometrically expected, since angular errors are amplified with increasing range. For example, a rotation error of 5° induces an offset of 61.1 cm at a distance of 7 m.

Cumulative Error Analysis. Fig. 9 reports the cumulative distributions of per-pixel absolute and relative depth er-

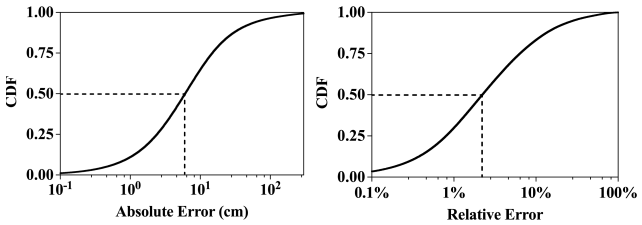

Figure 9. Cumulative distribution functions illustrating absolute and relative depth estimation errors.

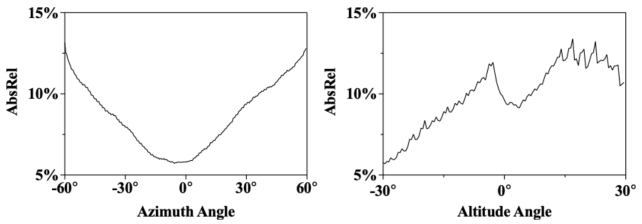

Figure 10. Evaluation of depth estimation accuracy across the azimuth and altitude dimensions of the sensor field of view.

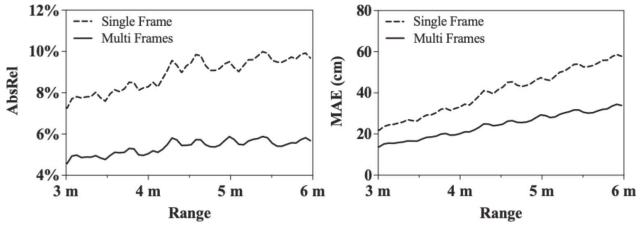

Figure 11. Depth error as a function of range for single-frame and multi-frame inference.

rors. The median absolute error is 6.1 cm, and 90% of pixels are below 37.6 cm; the median relative error is 2.2%, with a 90th-percentile value of 16.2%. This long-tail distribution indicates that most pixels are reconstructed accurately, while the overall mean error is affected by a small fraction of large-error outliers, which typically located near depth discontinuities or in regions with severe signal attenuation.

Field of View Error Analysis. We further analyze how depth accuracy varies across the sensor field of view. Fig. 10 shows the AbsRel across azimuth and altitude bins.

Along the azimuth axis, the error follows a U-shape: it is lowest near boresight ($\sim$5.9% at 0°) and increases toward the boundaries ($\sim$13% at $\pm 60^\circ$), consistent with reduced beamforming gain at the edges. Along the altitude axis, the error is lower near the floor ($\sim$5.7% near -30°), peaks around mid-altitude ($\sim$12%), and then slightly decreases toward the ceiling, suggesting that scene complexity is the dominant factor along the vertical direction.

The mid-altitude region typically contains cluttered objects, which are inherently more difficult to reconstruct. In contrast, the floor and ceiling are dominated by large planar structures, which are easier to reconstruct due to their geometric simplicity, even under lower antenna gain.

Table 6. Evaluation of Rascene's robustness to common occluders.

Occlusion	Depth			Voxel	
	AbsRel	MAE	RMSE	CD	CD_{Diag}
No Occ.	5.6%	13.7	24.8	18.4	2.0%
Paper Sheet	6.1%	14.4	25.2	19.1	2.1%
Styrofoam	6.4%	15.7	27.4	21.2	2.3%

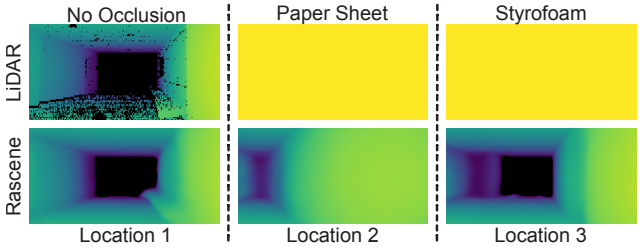

Figure 12. Representative examples of different sensors' occlusion resilience in a corridor scene at different locations.

Range-Dependent Error Analysis. We analyze depth error as a function of range for single-frame and multi-frame inference in Fig. 11. While the error increases with distance due to signal attenuation, our multi-frame RF fusion method substantially alleviates this trend. From 3 to 6 m, the absolute relative error increase is 2.4% for single-frame inference but only 1.1% for multi-frame inference. Over the same range, multi-frame fusion reduces mean absolute error from 41 cm to 24 cm, demonstrating long-range robustness.

Robustness to Occlusion. Finally, we evaluate Rascene's robustness to occlusion, a key advantage of RF sensing over optical sensors (Tab. 1). We place two common occluders, a paper sheet and a styrofoam box, in front of the device. As shown in Tab. 6, the performance degradation is minimal, indicating that Rascene can effectively sense through these materials. Fig. 12 further provides a qualitative comparison. While LiDAR is completely blocked by occluders, Rascene still reconstructs the underlying scene geometry.

7. Conclusion

In this paper, we presented Rascene, a monostatic ISAC framework that enables high-fidelity 3D scene imaging on individual mmWave communication devices. By leveraging CIR measurements from full-duplex OFDM communication devices and a confidence-aware multi-frame fusion strategy across arbitrary poses, Rascene mitigates the sparsity, noise, and multipath ambiguity of single-frame RF observations. Experiments across diverse indoor environments demonstrate strong cross-scene generalization, consistent gains from multi-frame fusion, and robustness to common occlusions. These results highlight the potential of using commodity communication infrastructure for low-cost, scalable, and robust 3D perception, especially in scenarios where optical sensing is unreliable.

References

[1] Omid Abari, Dinesh Bharadia, Austin Duffield, and Dina Katabi. Enabling {high-quality} untethered virtual reality. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 531–544, 2017. 1, 2

[2] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 3d tracking via body radio reflections. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, pages 317–329, 2014.

[3] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)*, 34(6):1–13, 2015. 1

[4] Fadel Adib, Zachary Kabelac, and Dina Katabi. {Multi-Person} localization via {RF} body reflections. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, pages 279–292, 2015. 2

[5] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics*, 37(6):1874–1890, 2021. 2

[6] Xieyuanli Chen, Andres Milioto, Emanuele Palazzolo, Philippe Giguere, Jens Behley, and Cyrill Stachniss. Suma++: Efficient lidar-based semantic slam. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4530–4537. IEEE, 2019. 1, 2

[7] Xieyuanli Chen, Thomas Läbe, Andres Milioto, Timo Röhling, Olga Vysotska, Alexandre Haag, Jens Behley, and C. Stachniss. Overlapnet: Loop closing for lidar-based slam. *ArXiv*, abs/2105.11344, 2020. 1

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5939–5948, 2019. 2

[9] Xiang Cheng, Dongliang Duan, Shijian Gao, and Liuqing Yang. Integrated sensing and communications (isac) for vehicular communication networks (vcn). *IEEE Internet of Things Journal*, 9(23):23441–23451, 2022. 2

[10] Yuwei Cheng, Jingran Su, Mengxin Jiang, and Yimin Liu. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics*, 38(6):3754–3773, 2022. 1, 2

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 2

[12] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2

[13] J Dilling, R Krücken, and G Ball. Isac overview. *Hyperfine Interactions*, 225(1):1–8, 2014. 2

[14] Fuwang Dong, Fan Liu, Yuanhao Cui, Wei Wang, Kaifeng Han, and Zhiqin Wang. Sensing as a service in 6g perceptive networks: A unified framework for isac resource allocation. *IEEE Transactions on Wireless Communications*, 22 (5):3522–3536, 2022. 2

[15] David Droeschel and Sven Behnke. Efficient continuous-time slam for 3d lidar-based online mapping. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5000–5007. IEEE, 2018. 2

[16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2

[17] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1, 2

[18] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7154–7164, 2019. 2

[19] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through fog high-resolution imaging using millimeter wave radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11464–11473, 2020. 1

[20] Zhenyao He, Wei Xu, Hong Shen, Derrick Wing Kwan Ng, Yonina C. Eldar, and Xiaohu You. Full-duplex communication for isac: Joint beamforming and power optimization. *IEEE Journal on Selected Areas in Communications*, 41(9): 2920–2936, 2023. 2

[21] Wolfgang Hess, Damon Kohler, Holger Rapp, and Daniel Andor. Real-time loop closure in 2d lidar slam. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 1271–1278. IEEE, 2016. 1, 2

[22] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8479–8488, 2022. 2

[23] Tianshu Huang, Akarsh Prabhakara, Chuhan Chen, Jay Karhade, Deva Ramanan, Matthew O'toole, and Anthony Rowe. Towards foundational models for single-chip radar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24655–24665, 2025. 1

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2

[25] Musa Furkan Keskin, Mohammad Mahdi Mojahedian, Jesus O Lacruz, Carina Marcus, Olof Eriksson, Andrea Giorgetti, Joerg Widmer, and Henk Wymeersch. Fundamental trade-offs in monostatic isac: A holistic investigation towards 6g. *IEEE Transactions on Wireless Communications*, 2025. 2

[26] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM conference on special interest group on data communication*, pages 269–282, 2015. 2

[27] Haowen Lai, Gaoxiang Luo, Yifei Liu, and Mingmin Zhao. Enabling visual recognition at radio frequency. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 388–403, 2024. 1, 2, 7

[28] Haowen Lai, Zhiwei Zheng, and Mingmin Zhao. Rf-based 3d slam rivaling vision approaches. In *Proceedings of the 31th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 170–185, 2025. 7

[29] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8456–8465, 2023. 2

[30] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10486–10496, 2025. 1

[31] Francesca Meneghello, Cheng Chen, Carlos Cordeiro, and Francesco Restuccia. Toward integrated sensing and communications in ieee 802.11 bf wi-fi networks. *IEEE Communications Magazine*, 61(7):128–133, 2023. 2

[32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[34] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 2

[35] Yue Pan, Pengchuan Xiao, Yujie He, Zhenlei Shao, and Zesong Li. Mulls: Versatile lidar slam via multi-metric linear least square. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11633–11640. IEEE, 2021. 2

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[37] 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. 1, 2

[38] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7205–7218, 2021. 2

[39] Elahe Soltanaghaei, Avinash Kalyanaraman, and Kamin Whitehouse. Multipath triangulation: Decimeter-level wifi localization and orientation with a single unaided receiver. In *Proceedings of the 16th annual international conference on mobile systems, applications, and services*, pages 376–388, 2018. 2

[40] Kunzhe Song, Qijun Wang, Shichen Zhang, and Huacheng Zeng. Siwis: Fine-grained human detection using single wifi device. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1439–1454, 2024. 2

[41] Kunzhe Song, Maxime Zingraff, and Huacheng Zeng. Spectrum shortage for radio sensing? leveraging ambient 5g signals for human activity detection. *arXiv preprint arXiv:2603.03579*, 2026. 2

[42] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. mmtrack: Passive multi-person localization using commodity millimeter wave radio. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 2400–2409. IEEE, 2020. 3

[43] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2

[44] Kai Wu, Jacopo Pegoraro, Francesca Meneghello, J Andrew Zhang, Jesus O Lacruz, Joerg Widmer, Francesco Restuccia, Michele Rossi, Xiaojing Huang, Daqing Zhang, et al. Sensing in bistatic isac systems with clock asynchronism: A signal processing perspective. *IEEE Signal Processing Magazine*, 41(5):31–43, 2024. 2

[45] Zhaoyang Xia, You-Chen Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Y. Qiao. Scpnet: Semantic scene completion on point cloud. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17642–17651, 2023. 2

[46] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 969–978, 2024. 2

[47] Ruibin Zhang, Donglai Xue, Yuhan Wang, Ruixu Geng, and Fei Gao. Towards dense and accurate radar perception via efficient cross-modal diffusion model. *IEEE Robotics and Automation Letters*, 2024. 1, 2

[48] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd annual international conference on mobile computing and networking*, pages 95–108, 2016. 2

[49] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9934–9943, 2020. 2

Rascene: High-Fidelity 3D Scene Imaging with mmWave Communication Signals

Supplementary Material

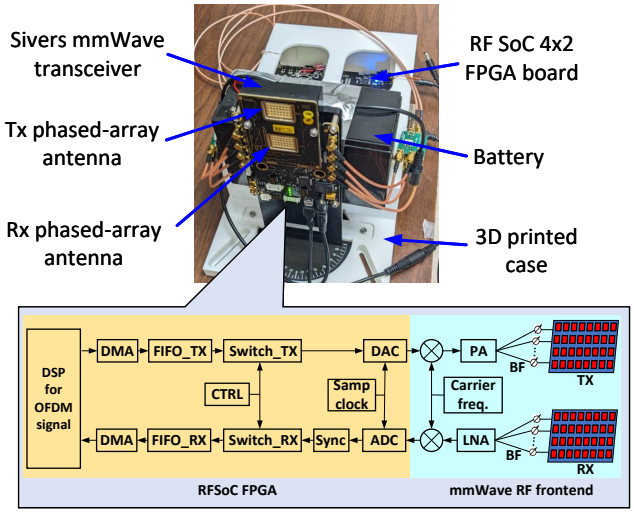

Figure 13. Our prototyped monostatic ISAC device.

8. Monostatic ISAC Hardware

We built a monostatic ISAC prototype using commercial off-the-shelf (COTS) components, enabling joint communication and sensing within a compact device.

Fig. 13 shows our monostatic ISAC prototype, with its parameters summarized in Tab. 7. The system consists of two primary COTS modules: (i) an AMD/Xilinx RFSoC 4x2 FPGA board, and (ii) a Sivers mmWave transceiver with Tx/Rx phased-array antennas. The RFSoC FPGA implements an OFDM signal processing pipeline compatible with 5G and Wi-Fi protocols, while the mmWave transceiver handles 60 GHz radio transmission and reception. Within the FPGA, the transmission and reception pipelines are jointly optimized and precisely calibrated to ensure timing and phase alignment required for monostatic sensing. Phased-array antenna control is seamlessly integrated into the processing pipeline, enabling beam steering to be synchronized with signal transmission.

Fig. 14 shows an example of simultaneous sensing and communication using the prototyped Rascene. During sensing data collection, Rascene simultaneously sends data packets to another device, supporting continuous video streaming. Both sensing and communication share the same hardware, spectrum band, and radiated energy. Our implementation demonstrates that a monostatic ISAC device can be realized without specialized sensing hardware. Furthermore, because the design leverages COTS communication components, it can be readily integrated into existing 5G and Wi-Fi mmWave communication devices through firmware and software upgrades.

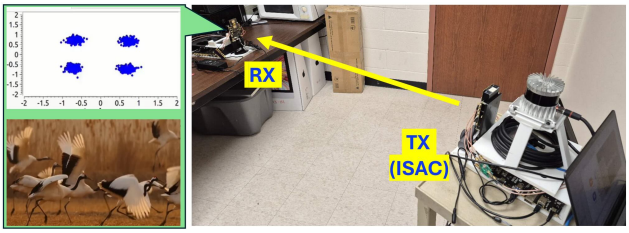

Figure 14. Illustration of video streaming data communication during sensing data collection.

Table 7. The parameters of our monostatic ISAC device.

Hardware parameters	
Sampling rate	1.2288 GSPS
Number of Tx antennas	16
Number of Rx antennas	16
Center frequency	60 GHz
Transmission power	20 dBm
Communication parameters	
Waveform	OFDM
FFT points	1024
Number of valid subcarriers	900
Cyclic prefix length	276
OFDM symbol duration	1.057 μs
Number of OFDM symbols per frame	$16 \times 16 = 256$
Supporting protocols	5G and Wi-Fi
Sensing parameters	
Detection time of a frame	67 μs
Number of frames per second	10
Theoretical detection range	30 m
Practical detection range	10 m
Number of effective antennas (horizontal)	8
Number of effective antennas (elevation)	4
Horizontal antenna spacing	0.5 wavelength
Elevation antenna spacing	0.5 wavelength
Horizontal field of view	[-60°, 60°]
Elevation field of view	[-30°, 30°]

9. Data Collection

Platform. To collect paired RF-LiDAR data, we mounted our custom-designed ISAC device, an Ouster OS0-128 LiDAR, and a TDK ICM-20948 IMU on a movable cart. The final dataset contains synchronized RF-LiDAR frame pairs collected from 20 indoor environments spanning diverse layouts, clutter levels, and construction materials such as drywall, glass, and metal. For each environment, we recorded approximately 10-20 minutes of data while manually moving the cart along unconstrained trajectories. Ex-

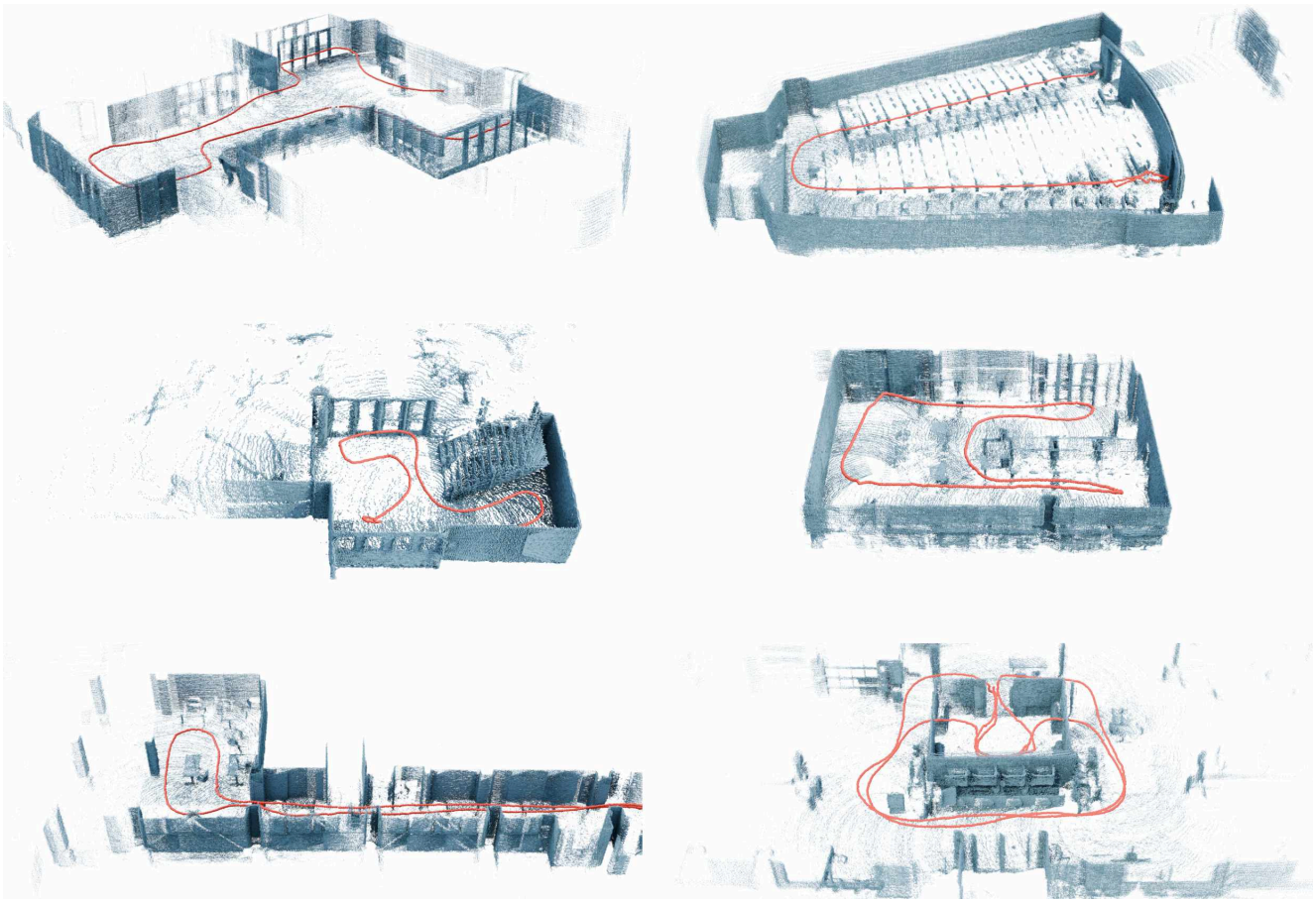

Figure 15. Sample trajectory segments (shown in red) from different scenes, visualized on the ground truth LiDAR point clouds.

ample trajectory segments and scene snapshots are shown in Fig. 15 and Fig. 16, respectively. All sensing modalities, including radio frames, LiDAR scans, and IMU measurements, were synchronized using timestamps from a shared clock source.

Ground Truth. The Ouster OS0-128 LiDAR provides a 360° horizontal and 90° vertical field of view, whereas our ISAC sensor covers 120° horizontally and 60° vertically. To align the two modalities, we calibrated the fixed extrinsic transformation between the rigidly mounted sensors and used it to crop the panoramic LiDAR observations to the field of view of the ISAC sensor. The cropped high-resolution LiDAR point clouds are then used to derive the ground-truth 3D geometry $\mathbf{V}^*$ and depth maps $\mathbf{D}^*$ for training and evaluation.

Temporal Sampling Strategy. Both the LiDAR and ISAC streams are recorded at 10 Hz, while the platform moves at an average speed of 0.5 m/s. Directly using consecutive frames would yield only a small spatial baseline and limited parallax, making it difficult to disambiguate true scene structure from multipath artifacts. We therefore adopt a sparse temporal sampling strategy: one frame is selected every 2 s from the continuous streams, and five sampled frames are grouped into one input window. This design preserves sufficient spatial overlap for cross-frame geometric consensus while introducing enough viewpoint variation to provide useful parallax and more diverse multipath observations. After warping the frames into a shared reference coordinate system, true scene structures remain more consistent across views than multipath artifacts, which makes the fusion process more reliable.

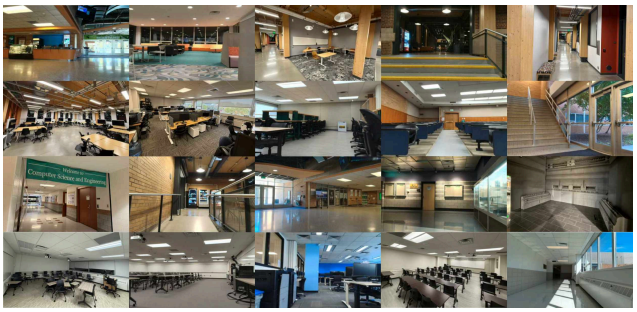

Figure 16. Example snapshots from the 20 distinct indoor environments included in our dataset.