SPIRIT: SPARSE PHYSICS-INFORMED INVERSE RENDERING FOR SAR IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Radio-frequency (RF) and wave-based imaging systems play a critical role in sensing under adverse conditions, but are fundamentally constrained by a trade-off between collection speed and spatial resolution. Reducing measurements to enable rapid collection of fragment geometric information across sensing dimensions, leading to highly non-injective inverse problems that destabilize conventional image-domain reconstruction and super-resolution methods. Synthetic aperture radar (SAR) exemplifies this challenge in sparse-aperture sampling. We introduce SPIRIT, a physics-informed inverse rendering framework for sparse-aperture SAR. Rather than enhancing formed images, SPIRIT reconstructs an explicit three-dimensional surface. By adopting an optimization strategy and explicitly modeling visibility, phase coherence, and wave propagation, SPIRIT aggregates distributed geometric features across sparse aperture views into a coherent reconstruction. Extensive experiments on synthetic and real-world data demonstrate that SPIRIT recovers accurate geometry and signal under extreme sparsity, reducing data collection by over 90% while achieving low geometric and signal error and consistently outperforming baselines.

1 INTRODUCTION

High-resolution radio-frequency (RF) imaging fundamentally depends on large effective apertures. In practice, physically large arrays are often impractical, and modern systems instead rely on sensor motion to synthesize virtual apertures, as in synthetic aperture radar (SAR) Vachon (2005); Soumekh (1999). While effective, high-fidelity SAR reconstruction typically requires dense spatial sampling, which can lead to long data-collection times even for relatively small objects Yanik et al. (2020). Reducing measurements is therefore attractive for fast sensing, but under severe aperture undersampling the degradation is not merely a loss of image sharpness. Sparse-aperture SAR becomes a coherent inverse problem in which large portions of the spatial-frequency domain are unobserved, producing interference, sidelobes, and geometric ambiguity Patel et al. (2010); Potter et al. (2010). In this regime, the task is better formulated as sparse-aperture inverse imaging: infer a scene whose coherent forward response explains the retained measurements while regularizing components that are not uniquely determined by the sparse aperture.

Existing image super-resolution paradigms are poorly matched to this setting. Image-domain methods recover plausible high-frequency content from learned priors Dong et al. (2015); Liang et al. (2021); Duan et al. (2025), but they do not explicitly enforce the geometry-dependent phase relationships of coherent RF sensing. Recent neural radar methods instead optimize implicit or explicit scene representations through differentiable radar models Borts et al. (2024); Lu et al. (2025); Takawale & Roy (2025); Kung et al. (2025). These systems establish the value of physics-aware radar reconstruction, but their acquisition geometries and reconstruction targets differ from the extreme scan-line sparsity studied here. When most aperture support is removed, the forward operator admits a large null space: several geometrically distinct scenes can be compatible with the retained measurements.

Sparse coherent RF reconstruction is difficult for two coupled reasons. First, severe aperture sparsity creates non-injectivity because missing angular and spatial-frequency support cannot be recovered from the measurements alone Potter et al. (2010); Li & Stoica (2008). Second, phase-sensitive optimization is highly non-convex: at millimeter-wave frequencies, sub-millimeter geometric pertur-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bations can induce large phase changes, so direct complex-valued fitting can converge to measurement-consistent but geometrically incorrect solutions Antun et al. (2020); Arridge et al. (2019). Analytic renderers introduce a third practical issue: real measurements contain hardware response, material-dependent scattering, calibration error, and multipath that are only approximately represented by tractable forward models.

We introduce **SPIRIT**, a sparse physics-informed inverse-rendering framework for SAR imaging. Rather than enhancing a formed image, SPIRIT reconstructs an explicit three-dimensional surface whose forward-rendered measurements are constrained by the observed sparse signals. The framework combines a structured surface representation, a differentiable single-scattering RF renderer with visibility and phase modeling, and a two-stage optimization schedule. Stage I establishes a geometry-stable hypothesis using objectives that are less sensitive to phase wrapping. Stage II then introduces complex-valued consistency together with a constrained residual pathway for bounded model mismatch.

Our formulation targets severe scan-line sparsity, where only a small subset of aperture positions is retained. The missing support makes the inverse problem fundamentally non-unique; retained views provide complementary evidence but do not preserve all information available from the full aperture. SPIRIT therefore should be interpreted as selecting a structured, measurement-consistent solution under explicit geometric and physical inductive biases, not as recovering absent information without assumptions.

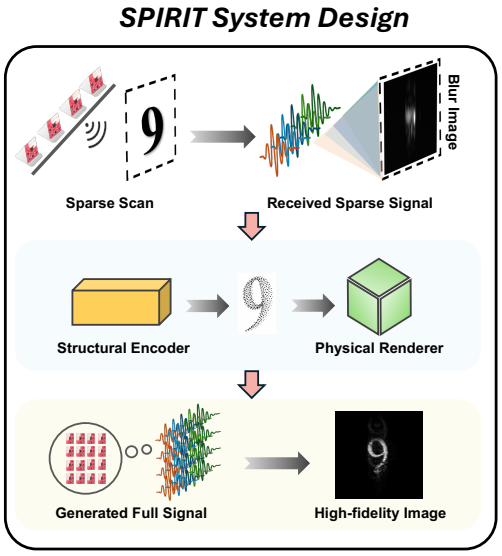


Figure 1: SPIRIT formulates sparse-aperture SAR reconstruction as physics-informed inverse rendering.

Contributions. Our contributions are threefold: (i) we formulate extreme sparse-aperture SAR as a structure-centric inverse-rendering problem that makes non-injectivity and prior dependence explicit; (ii) we introduce a geometry-first reconstruction architecture that combines a structured surface prior, differentiable RF rendering, and staged phase-sensitive refinement; and (iii) through object-disjoint controlled experiments, we show that measurement fit and geometric correctness can decouple under severe sparsity which motivates separating structural hypothesis formation from phase-sensitive refinement. We further evaluate the final system on an object-disjoint MITO test split in Appendix F.3. Stage I geometry learning uses synthetic measurements, while a disjoint real training split is used only for Stage II residual adaptation; the resulting experiment tests physical measurements with hardware, calibration, material, and multipath effects not reproduced exactly by the synthetic pipeline. Because the full pipeline uses real adaptation, we do not describe this experiment as zero-shot synthetic-to-real transfer.

2 Related Work

Sparse-aperture reconstruction lies at the intersection of image super-resolution, neural RF scene reconstruction, and classical sparse inverse problems. The important distinction is not simply model family, but what is observed, what latent scene representation is optimized, and which ambiguities are resolved by data versus priors.

2.1 Image Super-Resolution

Image super-resolution methods learn mappings from degraded or low-resolution images to visually sharper outputs. CNN/Transformer regressors such as SRCNN and SwinIR Dong et al. (2015); Liang et al. (2021) and recent diffusion-based systems Duan et al. (2025); Wan et al. (2024) can synthesize

Table 1: Scope comparison of representative reconstruction paradigms. Entries summarize the published input domain, scene representation, forward model, and primary reconstruction regime; the table is descriptive rather than a performance ranking.

Method	Input	Representation	Forward / Prior Model	Primary Target / Regime
SwinIR Liang et al. (2021)	formed image	pixel/Transformer features	learned image prior	image restoration / super-resolution
DiT4SR Duan et al. (2025)	formed image	diffusion Transformer	learned generative prior	real-world image super-resolution
Radar Fields Borts et al. (2024)	raw FMCW frequency data	implicit neural field	physics-informed FMCW sensor model	radar scene reconstruction from measured trajectories
GeRaF Lu et al. (2025)	RF measurements	neural SDF / reflectance field	physics-based RF volumetric rendering	near-range 3D geometry reconstruction
SpINRv2 Takawale & Roy (2025)	complex FMCW spectrum	implicit neural representation	differentiable frequency-domain synthesis	volumetric FMCW reconstruction
RadarSplat Kung et al. (2025)	radar images + poses	Gaussian primitives	radar rendering + noise/multipath model	driving-scene reconstruction and synthesis
Back-Projection Diffusion Zhang et al. (2025)	wideband scattered measurements	conditional diffusion in BP latent	wave-based back-projection structure	wideband inverse scattering, including sub-Nyquist settings
SPIRIT (Ours)	sparse FMCW phase history + poses	structured explicit surface	differentiable single-scattering RF renderer	isolated-object geometry under extreme scan-line sparsity

plausible high-frequency detail. Sparse SAR differs because aperture removal changes the sensing operator itself: the task is to infer a scene consistent with complex-valued coherent measurements, not merely to restore an already formed image. We therefore use image-domain super-resolution as a conceptual contrast rather than a direct solution to the problem studied here.

2.2 Neural RF Inverse Rendering

Physics-aware radar reconstruction increasingly combines neural scene representations with differentiable forward models. Radar Fields learns implicit geometry and reflectance directly from raw FMCW measurements Borts et al. (2024); GeRaF reconstructs geometry using neural implicit representations and RF volumetric rendering Lu et al. (2025); SpINRv2 supervises an implicit representation in the complex frequency domain Takawale & Roy (2025); and RadarSplat uses Gaussian primitives with explicit radar-noise modeling for driving scenes Kung et al. (2025). Differentiable SAR rendering similarly enables gradient-based reconstruction from simulated SAR imagery Wilmanski & Tamir (2023). These works establish that RF reconstruction benefits from sensor-aware forward models, but they use different representations, acquisition support, and scene regimes. Our focus is narrower: isolated-object reconstruction when scan-line support is deliberately reduced to an extreme degree.

2.3 Sparse Aperture and Inverse Scattering

Sparse SAR/ISAR has long been studied through compressed sensing and structured regularization Patel et al. (2010); Potter et al. (2010); Bi et al. (2023). Learning-based approaches include unfolded sparse-aperture imaging and autofocusing Wei et al. (2021), learned TomoSAR inversion Qian et al. (2022), regularization-by-denoising for sparse 3D SAR Wang et al. (2024), and self-supervised sparse ISAR reconstruction Wang et al. (2025). Wideband Back-Projection Diffusion further combines wave-based back-projection structure with a learned diffusion prior for inverse scattering Zhang et al. (2025). SPIRIT differs in where the prior is imposed: the latent variable is an explicit surface with deterministic neighborhood structure, and measurement fitting is mediated by a differentiable coherent renderer. Our experiments test whether this geometry-first inductive bias is effective in the controlled, predominantly connected-object regime; we do not claim that a fixed-topology surface is universally preferable to implicit fields, Gaussian primitives, surfels, or hybrid representations. Accordingly, the quantitative baselines in Section 5 are matched controls under a common sparse-aperture protocol rather than claimed faithful reproductions of every cited radar system.

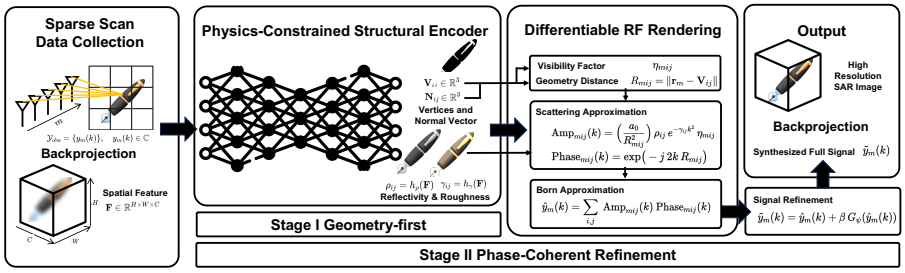


Figure 2: Overview of SPIRIT. Sparse coherent measurements are encoded into an explicit surface representation and evaluated through a differentiable RF forward model. The renderer produces predicted measurements at observed and unobserved aperture locations; consistency with the observed sparse measurements constrains the recovered geometry, while the learned structural prior regularizes unobserved components of the scene.

3 Problem Formulation

We formulate sparse-aperture RF reconstruction as a physics-constrained inverse-imaging problem. Let $\mathbf{Y}_{\text{obs}}$ denote coherent measurements retained at a sparse subset of aperture positions. Because this sampling removes portions of the spatial-frequency and angular support, the corresponding forward operator is non-injective: multiple scene configurations may explain the same sparse observations. Our objective is to seek a physically plausible scene representation whose forward response is consistent with the observed measurements while being regularized by explicit structural priors. Because the operator is non-injective, this objective does not identify a unique ground-truth scene from the sparse measurements alone; it selects one measurement-consistent hypothesis under the representation and regularization assumptions.

3.1 Structure-Centric Inverse Imaging

Conventional super-resolution methods cast reconstruction as a pixel-domain mapping $f : I_{\text{LR}} \to I_{\text{HR}}$. In RF imaging, however, apparent resolution loss arises from limited aperture coverage, leaving large portions of the spectrum unobserved under sparse scanning, as detailed in Appendix B.3.

As detailed in Appendices B.1.8 and B.2, RF signals encode scene geometry primarily through the phase accumulated along propagation paths. When aperture coverage is sparse, the forward operator admits a non-trivial null space, and multiple geometrically distinct scenes may explain the same or nearly indistinguishable sparse measurements. In this regime, the missing components are not uniquely determined by formed-image intensities alone; any reconstruction of them necessarily relies on additional structural, physical, or learned assumptions.

We therefore solve the sparse-aperture inverse problem in an explicit, structure-centric scene representation.

$$\mathbf{Y} \approx \mathcal{F}(\mathcal{S}), \tag{1}$$

where $\mathcal{F}$ denotes a physics-based forward operator modeling coherent wave propagation, whose derivation is provided in Appendix B.1. The scene representation encodes explicit *surface-centric physical structure* rather than image pixels:

$$\mathcal{S} = \{\mathbf{P}, \mathbf{N}, \boldsymbol{\rho}, \boldsymbol{\gamma}\}. \tag{2}$$

Here, $\mathbf{P} \in \mathbb{R}^{N\times 3}$ denotes a structured set of surface points in 3D space. Surface normals $\mathbf{N} \in \mathbb{R}^{N\times 3}$ are deterministically derived from $\mathbf{P}$ and encode local surface orientation. Material properties are modeled using two complementary parameters: $\boldsymbol{\rho} \in \mathbb{R}^N$ represents surface reflectivity, capturing radar cross-section magnitude. In contrast, $\boldsymbol{\gamma} \in \mathbb{R}^N$ encodes surface roughness and coherence loss due to sub-wavelength micro-geometry.

3.2 Physics-Constrained Forward Model

The mapping from scene structure $\mathcal{S}$ to received signal $\mathbf{Y}$ is governed by electromagnetic wave propagation. Starting from Maxwell's equations under time-harmonic excitation, the scattered field

satisfies the Helmholtz equation, detailed in Appendix B.1. Specifically, we adopt a Born-style single-scattering surface model, implemented in a Physical-Optics-like form as a discrete sum over surface elements.

We model the measurement process using a differentiable physics-based forward operator $\mathcal{F}_{\mathrm{PO}}$:

$$\begin{aligned}\hat{y}(\mathbf{r}_m, k) &= \mathcal{F}_{\mathrm{PO}}(\mathcal{S}) \\ &\approx \sum_{i=1}^{N} \mathcal{V}(\mathbf{x}_i, \mathbf{r}_m)\, A(\mathbf{x}_i)\, \boldsymbol{\sigma}_i\, e^{-j2k\|\mathbf{r}_m - \mathbf{x}_i\|},\end{aligned} \tag{3}$$

where $\mathbf{r}_m$ denotes the radar platform position at aperture index m, k is the wavenumber, and $\mathbf{x}_i \in \mathbf{P}$ is the i-th surface point. The term $\mathcal{V}(\mathbf{x}_i, \mathbf{r}_m)$ is a differentiable visibility function modeling occlusion, and $A(\mathbf{x}_i)$ accounts for antenna gain and distance-dependent path loss. The complex scattering coefficient $\boldsymbol{\sigma}_i$ is parameterized by surface reflectivity $\boldsymbol{\rho}_i \in \mathbb{R}$ and roughness $\boldsymbol{\gamma}_i \in \mathbb{R}$. The approximation arises from a Born-style single-scattering assumption with a Physical-Optics-like surface discretization, under which higher-order multipath and material-dependent effects are neglected for tractability, as detailed in Appendix B.1 and Appendix B.3. Throughout this paper, "physics-constrained" means consistency with this explicit approximate forward model; it does not imply exact electromagnetic simulation of all hardware, material, polarization, or multipath effects.

3.3 Inverse Objective

Recovering $\mathcal{S}$ from sparse signals is posed as a constrained energy minimization problem:

$$\min_{\mathcal{S}} \; \mathcal{L}_{\mathrm{data}}\big(\mathcal{F}_{\mathrm{PO}}(\mathcal{S}), \mathbf{Y}_{\mathrm{obs}}\big) + \lambda\, \mathcal{L}_{\mathrm{struct}}(\mathcal{S}), \tag{4}$$

where $\mathcal{L}_{\mathrm{data}}$ measures complex-valued discrepancy in both amplitude and phase, and $\mathcal{L}_{\mathrm{struct}}$ enforces geometric regularity such as surface smoothness and normal consistency Gropp et al. (2020). Equation equation 4 describes the scene-level inverse objective conceptually. In SPIRIT, this inverse map is amortized by the structural encoder: network parameters are learned across the training corpus, while the predicted surface for each input is evaluated through the same forward-consistency and structural terms.

4 SPIRIT Design

4.1 Overview

SPIRIT reconstructs an explicit scene surface from sparse coherent measurements and uses that surface to predict the corresponding SAR response through a structure-first, physics-constrained inverse-imaging pipeline. The method consists of three tightly coupled components. First, a *physics-aligned structural encoder* maps sparse measurements into an explicit surface-based scene representation by injecting known collection geometry and physical constraints. Second, a *differentiable RF forward renderer* enforces phase-coherent wave propagation, enabling signal-domain supervision to constrain geometry directly. Third, a *two-stage optimization strategy with signal refinement* stabilizes learning under extreme phase sensitivity and compensates for inevitable model–reality mismatch in analytic RF rendering. The design addresses two coupled sources of difficulty: sparse aperture sampling leaves a non-trivial null space, while coherent phase makes optimization highly sensitive to small geometric errors. The structured surface prior restricts the ambiguous solution space, and staged optimization delays phase-sensitive fitting until a stable geometric initialization has been established.

4.2 Physics-Constrained Structural Encoder

SPIRIT addresses sparse-aperture non-injectivity by reconstructing an explicit surface rather than a pixel grid, volumetric density field, or unconstrained point set. Under sparse apertures, high-capacity representations can fit measurement ambiguity without uniquely fixing geometry. We therefore use a structured surface prior with stable local neighborhoods and a compact hypothesis class. This choice is an inductive bias for the isolated, predominantly connected objects studied here rather than a claim of universal superiority over volumetric, implicit, or Gaussian representations.

Given sparse complex frequency-modulated continuous-wave (FMCW) SAR signals $\mathbf{Y}_{\text{obs}}$ and the corresponding platform poses $\{\mathbf{r}_m\}$ (background in Appendix B.1.4), SPIRIT first applies a fixed differentiable backprojection operator to expose coarse spatial support:

$$\mathbf{F} = \mathcal{B}(\mathbf{Y}_{\text{obs}}; \{\mathbf{r}_m\}). \tag{5}$$

Backprojection maps the signal domain onto a spatial grid, detailed in Appendix C.1. On a predefined two-dimensional reference plane, the operator evaluates how strongly a scatterer at each grid location (x, y) would contribute to the received echoes under the known propagation distances and platform geometry; Algorithm 1 gives the discrete implementation. The output of $\mathcal{B}$ is a feature tensor $\mathbf{F} \in \mathbb{R}^{B\times 3\times H\times W}$, whose channels store the accumulated real part, imaginary part, and magnitude of the backprojected signal, forming a flattened, defocused spatial energy map.

From the backprojected feature map $\mathbf{F}$, SPIRIT applies a convolutional encoder–decoder to extract structured geometric features from the defocused spatial representation. We adopt a convolutional encoder–decoder with skip connections, whose downsampling path aggregates global object extent under heavy sidelobe corruption, while the upsampling path restores spatial alignment. The network outputs a dense feature map $\mathbf{Z} \in \mathbb{R}^{B\times C\times H\times W}$, where each grid location encodes a local surface descriptor that will subsequently be lifted into an explicit three-dimensional vertex.

To lift the two-dimensional feature grid into an explicit three-dimensional surface, SPIRIT employs a structured grid deformation strategy. Rather than predicting an unordered point cloud, we interpret the $H \times W$ feature grid as a deformable surface mesh with fixed topology. To prevent degenerate solutions under sparse scans, we enforce bounded support via a global–local structure. A global head predicts a coarse axis-aligned bounding box $(\mathbf{v}_{\min}, \mathbf{v}_{\max}) \in \mathbb{R}^3 \times \mathbb{R}^3$ that captures the overall spatial extent of the object. Conditioned on the dense feature map $\mathbf{Z}$, a local head predicts normalized coordinates $\boldsymbol{\delta}_{i,j} \in [0, 1]^3$ for each grid cell. Final vertex positions are then obtained by

$$\mathbf{V}_{i,j} = \mathbf{v}_{\min} + \boldsymbol{\delta}_{i,j} \odot (\mathbf{v}_{\max} - \mathbf{v}_{\min}), \tag{6}$$

which constrains all vertices to the predicted object support and reduces unbounded geometric drift under the sparse-aperture null space. The resulting inductive bias, implementation details, and representation limitations are discussed in Appendix C.2 and Appendix G.

Surface normals are computed deterministically from the grid topology. Local tangents are obtained via finite differences along the grid axes, and per-vertex normals are computed by cross-products and normalization:

$$\mathbf{N}_{i,j} = \frac{(\mathbf{V}_{i+1,j} - \mathbf{V}_{i-1,j}) \times (\mathbf{V}_{i,j+1} - \mathbf{V}_{i,j-1})}{\|(\mathbf{V}_{i+1,j} - \mathbf{V}_{i-1,j}) \times (\mathbf{V}_{i,j+1} - \mathbf{V}_{i,j-1})\|_2 + \epsilon}. \tag{7}$$

This construction enforces local geometric consistency and stable orientation. In addition, SPIRIT predicts per-vertex reflectivity $\rho_{i,j}$ and roughness $\gamma_{i,j}$, allowing residual signal variation to be explained without forcing geometry to absorb all mismatch. Normal computation, gradient flow, and regularization are described in Appendix C.3.

4.3 Differentiable RF Rendering and Signal Refinement

SPIRIT enforces physical consistency by synthesizing complex-valued radar signals from the predicted surface and directly comparing them to the observed sparse signals. The renderer implements a differentiable approximation of the SAR forward operator under a single-bounce scattering model grounded in wave propagation and phase accumulation, shown in Appendix C.4. For platform position $\mathbf{r}_m$ and wavenumber k, the synthesized return is

$$\begin{aligned} \hat{y}_m(k) = &\sum_{i,j} \alpha_{m,i,j}(k)\, \sigma_{i,j}(k)\, \eta_{m,i,j} \\ &\cdot \exp\big(-j\, 2k\, \|\mathbf{r}_m - \mathbf{V}_{i,j}\|\big), \end{aligned} \tag{8}$$

where $\alpha_{m,i,j}(k)$ captures distance-dependent attenuation and known system scaling, and $\sigma_{i,j}(k)$ is the complex scattering coefficient parameterized by the predicted reflectivity $\rho_{i,j}$ and roughness $\gamma_{i,j}$. The visibility term $\eta_{m,i,j} \in [0, 1]$ is computed via soft backface culling:

$$\begin{aligned} \eta_{m,i,j} &= \sigma\big(\kappa\, \langle \mathbf{N}_{i,j}, \mathbf{d}_{m,i,j} \rangle\big), \\ \mathbf{d}_{m,i,j} &= \frac{\mathbf{r}_m - \mathbf{V}_{i,j}}{\|\mathbf{r}_m - \mathbf{V}_{i,j}\|_2}, \end{aligned} \tag{9}$$

where σ denotes the sigmoid. Soft visibility provides continuous gradients with respect to surface orientation, whereas a hard visibility decision would introduce zero-gradient discontinuities during optimization. Renderer gradients and implementation details are given in Appendix C.4.

Even with accurate geometry, analytic RF rendering inevitably deviates from real measurements due to unmodeled effects such as antenna patterns, hardware response, material-dependent scattering, and weak multipath interactions. This model–reality mismatch, analyzed in Appendix B.3, can bias geometric inference if strict signal consistency is enforced too early or too rigidly.

To compensate for these discrepancies without compromising physical interpretability, SPIRIT introduces a signal refinement layer formulated strictly as a residual correction, rather than a primary generative model:

$$\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} + \beta\, \mathcal{G}_\psi(\hat{\mathbf{Y}}). \tag{10}$$

Here, $\mathcal{G}_\psi$ is a lightweight residual network applied only to the analytically rendered signal, and β controls the magnitude of the residual correction. We initialize β near zero and regularize the residual toward the identity mapping so that the analytic renderer remains the dominant source of the predicted signal. This design restricts, but does not mathematically eliminate, non-physical corrections. The residual branch does not itself parameterize scene geometry, but the joint Stage II objective can still couple signal refinement and geometry updates. The refiner may therefore absorb signal components not represented by the analytic renderer, so signal improvement is interpreted jointly with geometric accuracy rather than as independent evidence of geometric correctness. The residual formulation and identity regularization are therefore intended to keep corrections small relative to the analytic prediction, but they do not constitute a hard physical guarantee. Architecture and regularization details are provided in Appendix C.5.

4.4 Two-Stage Optimization Strategy

Stage I: Geometry-first bootstrapping. We first optimize the encoder–decoder to recover globally consistent geometry and bounded support while avoiding phase-induced local minima. Stage I emphasizes objectives that constrain coarse structure but are less sensitive to phase wrapping, including image-domain losses on backprojected reconstructions, magnitude-based signal terms, and geometric regularizers on surface distribution and normals. The signal refinement layer is turned off at this stage to prevent the model from explaining missing diversity by learning arbitrary signal corrections.

Stage II: Phase-coherent alignment with controlled refinement. After Stage I establishes a geometry-stable initialization, Stage II introduces complex-valued signal consistency and activates the residual refinement module. In the final SPIRIT configuration, the geometry backbone remains trainable but is updated with a substantially reduced learning rate together with geometric regularization, while the material parameters and residual refiner are jointly optimized. This controlled joint optimization differs from unconstrained signal fitting. Among the tested variants, full signal-driven unfreezing without the refiner substantially degrades geometry, whereas the final joint configuration preserves the Stage I geometric solution while improving formed-image and signal metrics (Table 3). Loss composition and training schedules are detailed in Appendix C.6.

5 Experimental Setup

Dataset and Measurements. Our primary controlled evaluation uses synthetic sparse-aperture SAR measurements, for which both reference geometry and full-aperture coherent measurements are available. After filtering incomplete instances, the benchmark contains 1,957 object–measurement instances, partitioned into 1,567 training, 195 validation, and 195 test instances; test objects are disjoint from the training objects. We simulate a planar 77 GHz FMCW radar operating over 77.5–80.5 GHz, with center frequency 79.0 GHz, bandwidth 3.0 GHz, and 512 samples per chirp. The synthetic aperture covers $200\,\mathrm{mm} \times 100\,\mathrm{mm}$ and contains 13 scan lines as the full-aperture reference. Sparse inputs retain 1, 3, or 6 scan lines and are represented as $\mathbf{Y}_{\mathrm{in}} = \mathbf{Y}_{\mathrm{full}} \odot \mathbf{M}$, where $\mathbf{M}$ is a binary scan-line mask and missing entries are zero-padded. Training uses randomized scan-line masks, whereas validation and test results use fixed center-line masks for reproducible comparison. The synthetic data generator and SPIRIT renderer are distinct implementations: the generator operates on reference meshes with hard visibility and prescribed scattering components, whereas SPIRIT

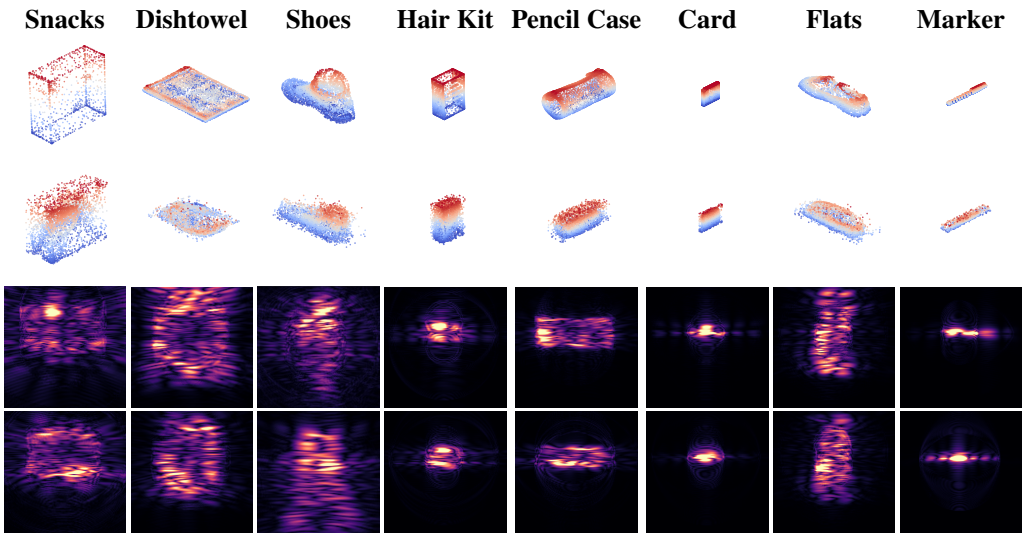


Figure 3: Held-out synthetic reconstructions under extreme 1-scan sparsity. Columns correspond to representative test objects excluded from the training split; rows show (top to bottom) ground-truth point clouds, SPIRIT-predicted geometry, sparse SAR observations, and predicted renderings. These examples visualize geometry recovery in the controlled synthetic setting; held-out physical-measurement results are reported separately in Appendix F.3.

renders from its predicted structured surface with soft visibility and learned reflectivity/roughness. Both retain first-order geometry-based approximations, so the synthetic benchmark is best interpreted as a controlled attribution setting rather than complete evidence of real-world electromagnetic generalization. Appendix F.3 evaluates a 60-object MITO subset with an object-disjoint 40/10/10 split; both captures of each object stay in one partition. Stage I uses synthetic measurements only, and the disjoint real training split adapts only the Stage II residual branch, so the full protocol is not zero-shot. Registered reference geometry is shared across methods without method-specific post-alignment, and the held-out physical test split uses geometry-, signal-, and image-domain metrics.

Matched Baselines. We compare against backprojection, FISTA compressed sensing, an image-domain U-Net, a height-map prior, and implicit/volumetric neural-field controls using identical measurements, splits, masks, reconstruction volume, and metrics. FISTA uses 20 iterations ($\lambda = 0.01$, step 0.5), while the neural-field controls use a fixed 100-iteration per-scene budget; settings are fixed across test objects with no test-set tuning. These are matched controls rather than exhaustive reproductions of recent radar systems with different inputs or forward operators (Appendix E.8).

Implementation Details. Differentiable backprojection converts sparse complex measurements into real, imaginary, and magnitude feature channels. An asymmetric ResNet–UNet predicts a 128×128 structured surface through the global–local parameterization in Section 4.2, and the renderer uses the single-bounce model in Section 4.3 with visibility sharpness $\kappa = 10.0$. Full architecture details are provided in Appendix E.

Training Protocol. Stage I trains for 300 epochs with AdamW at 10^{-4}; Stage II jointly optimizes geometry, material parameters, and the refiner at 3×10^{-6} with geometry regularization. Splits/masks use seed 42; Table 2 aggregates three independent SPIRIT runs, each on one RTX 8000 (Appendix E.9).

Evaluation Metrics. We report CD (mm) for geometry, phase-history ℓ_1 error for signal fidelity, and PSNR/SSIM on backprojected magnitude images. Geometry is the primary target; signal/image metrics are complementary. Definitions, forward projection for image-domain controls, and alignment procedures are given in Appendix E.11.

Table 2: Controlled synthetic comparison under 1-, 3-, and 6-scan sparsity. CD is reported in millimeters; PSNR/SSIM use backprojected magnitude images. SPIRIT is mean ± standard deviation over three seeds; controls use fixed schedules.

	1 Scan				3 Scans				6 Scans			
Method	CD (mm) ↓	Sig. Err. ↓	PSNR ↑	SSIM ↑	CD (mm) ↓	Sig. Err. ↓	PSNR ↑	SSIM ↑	CD (mm) ↓	Sig. Err. ↓	PSNR ↑	SSIM ↑
Backprojection	135.0	**0.175**	12.4	0.421	123.0	**0.187**	13.1	0.455	141.0	0.185	14.2	0.484
Compressed Sensing	82.0	0.208	13.5	0.421	64.0	0.208	12.5	0.427	55.0	**0.162**	13.6	0.422
Implicit Neural Field	52.0	0.351	19.5	0.743	61.0	0.358	19.2	0.736	52.0	0.318	24.2	0.766
Volumetric Neural Field	48.0	0.321	20.2	0.785	44.0	0.302	21.1	0.774	41.0	0.282	22.5	0.792
Image-Domain UNet	75.0	0.684	12.4	0.681	63.0	0.512	19.1	0.705	52.0	0.556	19.3	0.721
Height-Map Prior	81.0	0.324	20.5	0.521	57.0	0.256	21.7	0.621	32.0	0.232	23.3	0.675
SPIRIT (Ours)	**23.1** ± **1.2**	0.233 ± 0.015	**25.3** ± **0.3**	**0.829** ± **0.0063**	**24.3** ± **1.5**	0.242 ± 0.0032	**26.8** ± **0.4**	**0.842** ± **0.0039**	**15.2** ± **1.9**	0.235 ± 0.0042	**27.5** ± **0.3**	**0.854** ± **0.0041**

Table 3: Core 1-scan ablation. Full unfreezing without the refiner causes geometric drift; the final joint configuration gives the best tested joint result.

Configuration	**CD (mm) ↓**	**SSIM ↑**	**PSNR ↑**	**SigErr ↓**
Stage I (Best)	23.5	0.8273	17.1	0.2212
Stage I (Base)	29.1	0.8110	19.2	0.2328
Stage II: Frozen Backbone + Refiner	29.8	0.8150	21.3	0.2138
Stage II: Partial Geometry Freeze	38.3	0.8070	20.3	0.2146
Stage II: Full Unfreeze (No Refiner)	53.3	0.8160	22.1	0.2174
Stage II: Full Unfreeze + Refiner	**23.1**	**0.8290**	**25.3**	**0.2050**

5.1 Geometry Recovery under Sparse Aperture

Table 2 evaluates the object-disjoint synthetic test set under 1-, 3-, and 6-scan inputs. SPIRIT obtains the lowest CD and highest PSNR/SSIM among the matched controls at all three sparsity levels. Its phase-history error is not uniformly lowest, showing that lower residual need not imply better geometry. Figure 3 shows representative held-out 1-scan reconstructions; the matched representation controls contextualize the complete system without establishing universal optimality of fixed-topology surfaces.

5.2 What Does Staged Optimization Contribute?

Table 3 isolates geometry initialization from phase-sensitive refinement under 1-scan sparsity. Full unfreezing without the refiner increases CD to 53.3 mm despite higher PSNR, whereas the final *Full Unfreeze + Refiner* configuration obtains CD 23.1 mm, SSIM 0.8290, PSNR 25.3, and signal error 0.2050. Thus, within the tested variants, structural regularization and phase-sensitive fitting play distinct roles, and measurement fit alone is not a reliable surrogate for scene correctness. Appendix F provides additional coverage, real-data, noise, and regularization analyses.

Real-world measurements. On object-disjoint MITO test objects after residual-only adaptation on a disjoint real split, SPIRIT has the lowest CD (3.15 mm) and highest PSNR/SSIM (20.45 dB/0.805) among matched controls; a volumetric neural field has a slightly lower signal residual (0.353 vs. 0.362). This is a controlled near-field, non-zero-shot experiment (Appendix F.3).

6 Conclusion

SPIRIT formulates severe sparse-aperture SAR as a non-injective inverse-imaging problem, combining a structured surface prior, differentiable RF rendering, and staged geometry–signal optimization. Controlled experiments show that lower measurement residual need not imply better geometry, motivating structural stabilization before phase-sensitive refinement; held-out MITO measurements provide complementary evidence after residual-only adaptation on a disjoint real split. The validated scope remains isolated, object-scale near-field sensing under approximate single-scattering physics, and the results support separating structural hypothesis selection from measurement refinement without claiming universal optimality of fixed topology.

REFERENCES

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

Hui Bi, Yufan Song, Yanjie Yin, Jingjing Zhang, Wen Hong, and Yirong Wu. A novel sparse sar unambiguous imaging method based on mixed-norm optimization. *Science China Information Sciences*, 66(11):219302, 2023.

David Borts, Erich Liang, Tim Broedermann, Andrea Ramazzina, Stefanie Walz, Edoardo Palladin, Jipeng Sun, David Brueggemann, Christos Sakaridis, Luc Van Gool, et al. Radar fields: Frequency-space neural scene representations for fmcw radar. pp. 1–10, 2024.

Laura Dodds, Tara Boroushaki, Cusuh Ham, and Fadel Adib. Mito: A millimeter-wave dataset and simulator for non-line-of-sight perception. *arXiv preprint arXiv:2502.10259*, 2025.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. volume 38, pp. 295–307. IEEE, 2015.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *arXiv preprint arXiv:2002.10099*, 2020.

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*, pp. 27596–27606, 2025.

Jian Li and Petre Stoica. *MIMO radar signal processing*. John Wiley & Sons, 2008.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

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.

Vishal M Patel, Glenn R Easley, Dennis M Healy, and Rama Chellappa. Compressed synthetic aperture radar. *IEEE Journal of selected topics in signal processing*, 4(2):244–254, 2010.

Lee C. Potter, Emre Ertin, Jason T. Parker, and Müjdat Cetin. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98(6), 2010.

Kun Qian, Yuanyuan Wang, Yilei Shi, and Xiao Xiang Zhu. γ-net: Superresolving sar tomographic inversion via deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

Mehrdad Soumekh. *Synthetic aperture radar signal processing*, volume 3. New York: Wiley, 1999.

Harshvardhan Takawale and Nirupam Roy. Spectral domain neural reconstruction for passband fmcw radars. *arXiv preprint arXiv:2506.08163*, 2025.

Paris W Vachon. Digital processing of synthetic aperture radar data: algorithms and implementation. *Geomatica*, 59(3):353–355, 2005.

Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. Clearsr: Latent low-resolution image embeddings help diffusion-based real-world super resolution models see clearer. *arXiv preprint arXiv:2410.14279v1*, 2024.

Yangyang Wang, Xu Zhan, Jing Gao, Jinjie Yao, Shunjun Wei, and Jiansheng Bai. Array sar 3-d sparse imaging based on regularization by denoising under few observed data. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Ziwen Wang, Jianping Wang, Pucheng Li, Yifan Wu, and Zegang Ding. Self-supervised-isar-net enables fast sparse isar imaging. *IEEE Transactions on Aerospace and Electronic Systems*, 2025.

Shunjun Wei, Jiadian Liang, Mou Wang, Jun Shi, Xiaoling Zhang, and Jinhe Ran. Af-ampnet: A deep learning approach for sparse aperture isar imaging and autofocusing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Michael C Wilmanski and Jonathan I Tamir. Differentiable rendering for synthetic aperture radar imagery. *IEEE Transactions on Aerospace and Electronic Systems*, 60(1):82–93, 2023.

Muhammet Emin Yanik, Dan Wang, and Murat Torlak. Development and demonstration of mimo-sar mmwave imaging testbeds. *IEEE Access*, 8:126019–126038, 2020.

Borong Zhang, Martin Guerra, Qin Li, and Leonardo Zepeda-Núñez. Back-projection diffusion: Solving the wideband inverse scattering problem with diffusion models. *Computer Methods in Applied Mechanics and Engineering*, 443:118036, 2025.

A Appendix

This appendix provides additional technical background, implementation details, and extended analyses that complement the main paper. It is organized as follows:

- **Appendix B** reviews the radar and SAR fundamentals.
- **Appendix C** presents mathematical formulations, operator definitions, and optimization analysis of SPIRIT.
- **Appendix D** describes preliminary experiments and motivating observations that inform the proposed formulation.
- **Appendix E** provides comprehensive experimental details.
- **Appendix F** reports extensive ablation studies and qualitative analyses.
- **Appendix G** discusses limitations, future research directions, and broader implications.
- **Appendix I** AI Use Statement.

B Extensive Radar and SAR Imaging Background

B.1 Electromagnetic Wave Propagation and Scattering Model

This appendix establishes the electromagnetic (EM) foundations underlying our radar forward model. We clarify the physical assumptions implicitly adopted by analytic RF rendering and delineate their validity regimes, which directly motivate the signal refinement strategy introduced in SPIRIT.

B.1.1 From Maxwell's Equations to the Helmholtz Equation

Maxwell's equations govern the propagation of electromagnetic waves. In a source-free, linear, isotropic, and non-dispersive medium, the time-domain equations are given by

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{x}, t) &= -\mu \frac{\partial \mathbf{H}(\mathbf{x}, t)}{\partial t}, \\ \nabla \times \mathbf{H}(\mathbf{x}, t) &= \epsilon \frac{\partial \mathbf{E}(\mathbf{x}, t)}{\partial t}, \\ \nabla \cdot \mathbf{E}(\mathbf{x}, t) = 0, &\quad \nabla \cdot \mathbf{H}(\mathbf{x}, t) = 0, \end{aligned} \tag{11}$$

where $\mathbf{E}$ and $\mathbf{H}$ denote the electric and magnetic fields, and ϵ and μ are the permittivity and permeability of the medium.

Assuming time-harmonic fields of the form $\mathbf{E}(\mathbf{x}, t) = \Re\{\mathbf{E}(\mathbf{x})e^{j\omega t}\}$, Maxwell's equations reduce to the frequency-domain vector Helmholtz equation,

$$\nabla^2 \mathbf{E}(\mathbf{x}) + k^2 \mathbf{E}(\mathbf{x}) = 0, \tag{12}$$

where $k = \omega/c$ is the wavenumber and $c = 1/\sqrt{\mu\epsilon}$ is the speed of light. An analogous equation holds for the magnetic field. This formulation provides the starting point for most high-frequency radar imaging models.

B.1.2 Single-Frequency Wave Propagation and Green's Function

For a monochromatic point source, the solution to the scalar Helmholtz equation can be expressed using the free-space Green's function,

$$G(\mathbf{x}, \mathbf{x}') = \frac{e^{-jk\|\mathbf{x}-\mathbf{x}'\|}}{4\pi\|\mathbf{x} - \mathbf{x}'\|}, \tag{13}$$

which describes spherical wave propagation from source location $\mathbf{x}'$ to observation point $\mathbf{x}$. In radar systems, both the transmitted and received fields can be modeled as superpositions of spherical waves, with the phase determined by the propagation distance. This distance-dependent phase term underlies coherent radar imaging and directly links geometry to the received signal.

B.1.3 SCATTERING MODEL AND BORN APPROXIMATION

To model the interaction between incident waves and a scene, we adopt a first-order single-scattering linearization, in which the received field is modeled as a coherent superposition of independent surface-element contributions. This Born-style assumption neglects higher-order multiple interactions, but preserves phase coherence and the geometric dependence of propagation.

Under a monostatic configuration, the complex-valued received signal at radar location $\mathbf{r}_m$ and wavenumber k can be written in the same form as the main text as a discrete surface sum:

$$\hat{y}(\mathbf{r}_m, k) \approx \sum_{i,j} \mathcal{V}(\mathbf{V}_{i,j}, \mathbf{r}_m)\, A(\mathbf{V}_{i,j})\, \boldsymbol{\sigma}_{i,j}\, \exp\big(-j\, 2k\|\mathbf{r}_m - \mathbf{V}_{i,j}\|\big), \tag{14}$$

where $\mathbf{V}_{i,j}$ and $\mathbf{N}_{i,j}$ denote surface vertices and normals, $\mathcal{V}(\cdot)$ is a (differentiable) visibility term, and $A(\cdot)$ absorbs system scaling and distance-dependent attenuation. The complex scattering coefficient $\boldsymbol{\sigma}_{i,j}$ is parameterized by per-vertex reflectivity $\rho_{i,j}$ and roughness $\gamma_{i,j}$.

The Born approximation is valid when scattering is weak, and multiple reflections between scene elements can be neglected. This assumption typically holds for electrically small or moderately reflective objects in free space, and is commonly adopted in radar imaging literature for tractable modeling and inversion.

Limitations of the Single-Scattering Assumption While the single-scattering (single-bounce) surface-sum model enables efficient forward modeling and differentiable rendering, it does not fully capture the complexity of real-world RF signals. Its validity degrades in several common scenarios:

- **Strong reflectors and specular surfaces**, where higher-order reflections dominate the measured signal.
- **Multipath propagation**, such as corner reflectors or urban environments, where waves undergo multiple bounces before returning to the sensor.
- **Material-dependent scattering and polarization effects**, which introduce angle- and frequency-dependent responses not captured by a scalar reflectivity model.
- **Hardware and calibration imperfections**, including antenna coupling and phase noise.

These effects introduce systematic discrepancies between analytic RF renderers and real signals. As a result, enforcing strict signal consistency with an idealized forward model can bias the reconstruction or amplify modeling errors, particularly under sparse-aperture sampling. This inherent model–reality mismatch motivates the signal refinement mechanism introduced in SPIRIT, which compensates for unmodeled scattering effects while preserving physical interpretability.

B.1.4 FMCW RADAR SIGNAL MODEL

Frequency Modulated Continuous Wave (FMCW) radar is widely used in modern robotic and automotive sensing systems due to its low power consumption, compact hardware, and simultaneous range and velocity estimation capability. This section formalizes the FMCW signal model and clarifies the physical meaning of the measured intermediate-frequency (IF) signal used throughout this work.

B.1.5 TRANSMIT CHIRP SIGNAL

An FMCW radar transmits a continuous waveform whose instantaneous frequency increases linearly over time. The complex baseband representation of a single chirp can be written as

$$s_{\text{tx}}(t) = \exp\big(j2\pi\left(f_c t + \tfrac{1}{2}Kt^2\right)\big), \quad t \in [0, T], \tag{15}$$

where f_c denotes the carrier frequency, T is the chirp duration, and $K = B/T$ is the frequency sweep rate determined by the bandwidth B.

This linear frequency modulation encodes distance information as a frequency shift in the received signal, enabling efficient range estimation without pulsed transmission.

B.1.6 Received Signal with Propagation Delay

Consider a point scatterer located at range R from the radar. The transmitted signal experiences a round-trip propagation delay $\tau = 2R/c$, yielding the received signal

$$s_{\text{rx}}(t) = \sigma\, s_{\text{tx}}(t-\tau) = \sigma \exp\big(j2\pi \left[f_c(t-\tau) + \tfrac{1}{2}K(t-\tau)^2\right]\big)\,, \tag{16}$$

where σ denotes the complex reflectivity incorporating material response and propagation loss.

Direct digitization of $s_{\text{rx}}(t)$ would require sampling at the carrier frequency f_c, which is infeasible for mmWave radar systems due to analog-to-digital converter (ADC) bandwidth and power constraints.

B.1.7 Dechirping and Intermediate-Frequency Signal

To overcome this limitation, FMCW radar performs *dechirping* by mixing the received signal with the conjugate of the transmitted signal,

$$s_{\text{IF}}(t) = s_{\text{rx}}(t)\, s_{\text{tx}}^{*}(t). \tag{17}$$

Substituting the expressions above and neglecting higher-order terms in τ, the resulting IF signal simplifies to

$$s_{\text{IF}}(t) \approx \sigma \exp(-j2\pi K\tau t)\, \exp(-j2\pi f_c \tau)\,. \tag{18}$$

This operation shifts the signal from radio frequency to baseband while preserving phase coherence. The IF signal is a complex sinusoid whose frequency

$$f_b = K\tau = \frac{2KR}{c} \tag{19}$$

is directly proportional to the target range.

The use of the IF signal is essential: it reduces the required ADC sampling rate from the carrier frequency ($\sim$f$_c$) to the beat frequency ($\sim$MHz), while retaining both amplitude and phase information necessary for coherent imaging. In our implementation, FMCW measurements provide a band of frequencies (equivalently, wavenumbers k); The forward model and losses are applied per-k and aggregated across the band, matching the notation in the main text.

B.1.8 Phase, Distance, and Geometric Sensitivity

This section analyzes the relationship between phase, propagation distance, and geometric sensitivity in mmWave radar, providing the mathematical basis for the phase instability challenges encountered in sparse-aperture reconstruction.

Phase as a Function of Distance The round-trip propagation of a monochromatic wave induces a phase shift proportional to path length. For a scatterer at distance R, the accumulated phase is

$$\phi(R) = 2kR = \frac{4\pi R}{\lambda}, \tag{20}$$

where $k = 2\pi/\lambda$ is the wavenumber and λ denotes the carrier wavelength.

This expression highlights that the phase is not a nuisance term but a direct geometric encoding. Any change in distance ΔR induces a corresponding phase change

$$\Delta\phi = 2k\,\Delta R. \tag{21}$$

Phase Sensitivity at Millimeter Wavelengths At millimeter-wave frequencies, wavelengths are on the order of a few millimeters. As a result, even sub-millimeter geometric perturbations produce significant phase variation. In particular,

$$\Delta R = \frac{\lambda}{4} \quad \Rightarrow \quad \Delta\phi = \pi, \tag{22}$$

which corresponds to a complete sign inversion of the complex signal.

This extreme sensitivity implies that minor geometric errors can induce large oscillations in phase-based loss functions, yielding highly non-convex optimization landscapes with dense local minima. Such behavior is a fundamental property of coherent wave physics and cannot be eliminated solely through regularization or architectural design.

Phase is Structure, Not Noise Unlike incoherent imaging modalities, where phase is often discarded, phase in radar imaging encodes precise geometric structure. While noise may perturb phase signals, the underlying phase response remains deterministic and directly linked to propagation distance. Discarding or smoothing phase, therefore, removes essential structural information, undermining the ability to resolve fine geometry from coherent measurements.

Range Resolution and Bandwidth Range resolution in FMCW radar is determined by the signal bandwidth rather than carrier frequency. Specifically, the minimum resolvable distance difference is

$$\Delta R = \frac{c}{2B}, \tag{23}$$

where B denotes the chirp bandwidth.

While increasing bandwidth improves range resolution, it does not mitigate phase-induced instability. This decoupling between range resolution and phase sensitivity underlies the ill-posed nature of sparse-aperture RF inverse imaging.

B.2 Synthetic Aperture Radar Forward Model

Synthetic Aperture Radar (SAR) extends single-scan radar sensing by coherently combining multiple observations acquired from distinct spatial locations. This section formalizes the SAR forward model used throughout this work, explicitly defining the signal operator whose inversion under sparse sampling motivates our approach.

B.2.1 Signal Notation

We consider a monostatic SAR configuration in which a single radar platform serves as both transmitter and receiver. Let $\mathbf{r}_m \in \mathbb{R}^3$ denote the radar position at signal index m, corresponding to the m-th spatial location along the synthetic aperture trajectory. The scene is modeled as a continuous spatial reflectivity function $\sigma(\mathbf{x}) \in \mathbb{C}$, where $\mathbf{x} \in \mathbb{R}^3$ denotes a spatial point.

Each SAR signal captures the coherent superposition of echoes scattered from all points in the scene under a single-scattering assumption.

B.2.2 SAR Forward Operator

For a given wavenumber $k = 2\pi/\lambda$, the received baseband signal at signal index m can be expressed as

$$y_m(k) \approx \sum_{i,j} \mathcal{V}(\mathbf{V}_{i,j}, \mathbf{r}_m)\, A(\mathbf{V}_{i,j})\, \boldsymbol{\sigma}_{i,j}\, \exp\big(-j\, 2k\|\mathbf{r}_m - \mathbf{V}_{i,j}\|\big). \tag{24}$$

This expression accounts for the round-trip propagation phase accumulated between the radar position $\mathbf{r}_m$ and the scatterer at $\mathbf{x}$. Amplitude effects such as antenna gain and propagation loss are omitted here for clarity, as they do not alter the fundamental conditioning of the inverse problem.

Equation (24) defines the single-scattering SAR forward operator $\mathcal{F}$ used in this work, mapping the structured surface field $\mathcal{S}$ to complex measurements indexed by platform position and wavenumber.

$$\mathbf{y} = \mathcal{F}(\sigma). \tag{25}$$

Each signal $y_m(k)$ is a global projection of the scene, and meaningful spatial discrimination emerges only through coherent aggregation across the synthetic aperture.

B.3 Resolution and Sparse Aperture Limitation

This section analyzes the fundamental resolution limits of SAR imaging and explains why sparse aperture sampling leads to non-injective inverse scattering.

B.3.1 Rayleigh Resolution and Aperture Length

In SAR, cross-range resolution is governed by the effective aperture length rather than pixel sampling density. Under standard assumptions, the Rayleigh cross-range resolution Δx can be approximated as

$$\Delta x \approx \frac{\lambda R}{2L}, \tag{26}$$

where λ is the wavelength, R is the range to the target, and L denotes the synthetic aperture length.

Increasing aperture length directly improves spatial resolution by increasing angular diversity and phase discrimination. This principle underlies the effectiveness of dense SAR trajectories for high-resolution imaging.

B.3.2 Sparse Aperture Limitation

Sparse-aperture sampling induces a range of characteristic artifacts arising from incomplete spectral coverage rather than noise or algorithmic deficiencies.

k-Space Coverage Interpretation From a spectral perspective, SAR signals correspond to samples of the scene's spatial frequency content. Each radar position $\mathbf{r}_m$ contributes a distinct slice in spatial frequency (k-space), and dense aperture coverage fills this domain sufficiently to enable stable inversion.

Sparse-aperture sampling, however, results in incomplete and irregular k-space coverage. Large regions of the spatial frequency domain remain unobserved, resulting in information loss that cannot be recovered by post-processing.

Null Space and Non-Injectivity When aperture coverage is sparse, the SAR forward operator $\mathcal{F}$ admits a non-trivial null space. there exist distinct scene surfaces $\mathcal{S}_1 \neq \mathcal{S}_2$ such that

$$\mathcal{F}(\mathcal{S}_1) = \mathcal{F}(\mathcal{S}_2). \tag{27}$$

This many-to-one mapping implies that the inverse problem is non-injective: no unique solution exists even in the absence of noise. As a result, sparse-scan SAR reconstruction is fundamentally ill-posed, and resolution cannot be recovered without introducing additional structural or physical constraints.

Grating Lobes and Ghost Targets Irregular or undersampled apertures produce grating lobes in the point spread function, causing energy from a single scatterer to replicate at multiple spatial locations. These ghost targets are coherent artifacts that persist even under noiseless signals.

Energy Leakage and Smearing Incomplete aperture coverage causes spectral leakage, spreading energy across the image domain and smearing fine structural details. This leakage masks weak scatterers and degrades contrast, particularly in high-dynamic-range scenes.

Spectral Leakage and Windowing Finite-aperture sampling induces spectral leakage in the spatial-frequency domain, leading to sidelobes in the reconstructed image. Windowing functions, such as Hamming or Kaiser windows, trade resolution for sidelobe suppression, thereby modifying the practical point spread function (PSF). In sparse-aperture scenarios, however, windowing alone cannot resolve ambiguities arising from missing spectral support.

C Mathematical and Algorithmic Details of SPIRIT

This appendix provides the mathematical formulations, operator definitions, and optimization analysis corresponding directly to the components introduced in Section 4.

C.1 Differentiable Backprojection Encoder

SPIRIT employs a differentiable backprojection encoder that maps sparse, complex-valued SAR measurements into a spatially organized feature representation. Backprojection can be understood as a

per-location coherent accumulation process. For each spatial grid cell (or voxel) in the reconstruction volume, the encoder computes the expected propagation delay from every radar aperture position to that location, applies the corresponding phase compensation to the received complex signal, and coherently sums the contributions across apertures. When a grid cell corresponds to a true scattering center, phase-aligned signals from multiple views constructively interfere, producing a strong response; otherwise, phase mismatch leads to destructive cancellation. In this way, the image is formed by aggregating energy independently at each spatial location. Unlike classical backprojection used purely for image formation, our implementation preserves complex-valued phase information and is explicitly designed to be differentiable, allowing gradients to backpropagate through the coherent accumulation process and enabling end-to-end optimization under the SAR forward model.

Forward operator. Let $\mathbf{Y}_{\text{obs}} = \{y_m(k)\}_{m\in\mathcal{M}}$ denote sparse FMCW SAR signals acquired at platform locations $\{\mathbf{r}_m\}_{m\in\mathcal{M}}$, indexed by wavenumber k. For a candidate spatial location $\mathbf{x} \in \mathbb{R}^3$, the backprojection encoder is defined as

$$\mathcal{B}(\mathbf{x}) = \sum_{m\in\mathcal{M}} \int y_m(k)\, \exp(j\, 2k\, \|\mathbf{r}_m - \mathbf{x}\|)\, dk, \tag{28}$$

which coherently aligns the phase histories associated with $\mathbf{x}$. This operation can be interpreted as a matched filtering process in the spatial domain and corresponds to the adjoint of the SAR forward operator (Appendix B.2).

In practice, the integral over k is implemented via discrete summation, and signal values are obtained through differentiable interpolation when required. The operator naturally supports arbitrary sparse apertures, as missing azimuth positions correspond to omitted summation terms.

Algorithmic evaluation. Algorithm 1 summarizes the forward evaluation of the differentiable backprojection encoder over a discrete spatial grid.

Algorithm 1 Forward Differentiable Backprojection Encoder

Require: Sparse signals $\{y_m(k)\}_{m\in\mathcal{M}}$, platform poses $\{\mathbf{r}_m\}$
Require: Spatial grid $\{\mathbf{x}_j\}_{j=1}^{N}$
Ensure: Complex-valued feature map $\mathcal{B}(\mathbf{x})$

```
for each spatial location x_j do
    B(x_j) ← 0
    for each signal index m ∈ M do
        for each wavenumber sample k do
            φ_mj(k) ← exp(j 2k ‖r_m − x_j‖)
            B(x_j) += y_m(k) · φ_mj(k)
        end for
    end for
end for
return B(x)
```

Differentiability and gradient flow. All operations in Eq. (28) are differentiable with respect to both the input signals and spatial coordinates. Let $\mathcal{L}$ denote a scalar loss defined on downstream representations. Gradients propagate through both the signal values and the geometric phase terms, yielding

$$\frac{\partial \mathcal{L}}{\partial y_m(k)} = \sum_j \frac{\partial \mathcal{L}}{\partial \mathcal{B}(\mathbf{x}_j)} \exp(j\, 2k\, \|\mathbf{r}_m - \mathbf{x}_j\|)\,, \tag{29}$$

and

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_j} = \sum_{m,k} \frac{\partial \mathcal{L}}{\partial \mathcal{B}(\mathbf{x}_j)}\, y_m(k)\, j\, 2k\, \frac{\mathbf{x}_j - \mathbf{r}_m}{\|\mathbf{x}_j - \mathbf{r}_m\|} \exp(j\, 2k\, \|\mathbf{r}_m - \mathbf{x}_j\|)\,. \tag{30}$$

These expressions ensure that gradients explicitly respect wave-propagation geometry and phase coherence.

Complex feature construction. The backprojection output $\mathcal{B}(\mathbf{x})$ is complex-valued. To interface with standard neural architectures while preserving coherent information, we construct a multi-channel real feature representation.

$$\mathbf{Z}(\mathbf{x}) = \big[\Re\{\mathcal{B}(\mathbf{x})\}, \Im\{\mathcal{B}(\mathbf{x})\}, |\mathcal{B}(\mathbf{x})|\big].$$

This representation retains both phase-sensitive and magnitude-based cues, which are critical under sparse and non-injective collection regimes (Appendix B.3).

Discussion. The differentiable backprojection encoder serves as a physics-aligned signal transform that bridges sparse SAR observations and spatially structured representations. By exposing coherent phase alignment to gradient-based optimization, SPIRIT can reason directly in the space of physically realizable SAR images rather than relying on pixel-domain heuristics.

C.2 Structured Grid Scene Parameterization

SPIRIT represents geometry as a structured $H \times W$ surface grid deformed into 3D space.

Bounding-box parameterization. Vertices are generated via

$$\mathbf{V}_{i,j} = \mathbf{v}_{\min} + \boldsymbol{\delta}_{i,j} \odot (\mathbf{v}_{\max} - \mathbf{v}_{\min}), \qquad \boldsymbol{\delta}_{i,j} \in [0,1]^3. \tag{31}$$

This parameterization constrains all vertices to a compact spatial domain.

Stability under sparse inverse scattering. Under sparse aperture coverage, the SAR forward operator admits a large null space (Appendix B.3). Unbounded point-based representations can drift arbitrarily while still matching subsets of signals. Bounding-box constraints eliminate scale-free degeneracies and stabilize optimization by restricting the hypothesis space to physically plausible geometry.

Why a structured grid. Unlike unstructured point clouds, the grid induces explicit topology and consistent neighborhood relations. This provides consistent local neighborhoods for normal computation, surface orientation, and gradient propagation during wave-consistent rendering under sparse signals.

C.3 Normal Computation and Gradient Flow

Surface normals are derived deterministically from the grid geometry.

Finite-difference formulation. For interior vertices,

$$\mathbf{t}^x_{i,j} = \mathbf{V}_{i+1,j} - \mathbf{V}_{i-1,j}, \tag{32}$$

$$\mathbf{t}^y_{i,j} = \mathbf{V}_{i,j+1} - \mathbf{V}_{i,j-1}, \tag{33}$$

and the normal is

$$\mathbf{N}_{i,j} = \frac{\mathbf{t}^x_{i,j} \times \mathbf{t}^y_{i,j}}{\|\mathbf{t}^x_{i,j} \times \mathbf{t}^y_{i,j}\|_2 + \epsilon}. \tag{34}$$

Differentiability. This construction is fully differentiable. Gradients from the renderer propagate through normals to vertex positions, encouraging locally smooth and physically consistent surface orientation. Learning normals independently was empirically unstable and produced non-integrable surfaces.

C.4 Differentiable RF Renderer

The renderer implements a discrete, differentiable approximation of the SAR forward operator (Appendix B.2).

Signal synthesis. For platform position $\mathbf{r}_m$ and wavenumber k, the synthesized signal is

$$\hat{y}_m(k) = \sum_{i,j} A(\mathbf{V}_{i,j}, \mathbf{r}_m, k)\, \sigma_{i,j}(k)\, \eta_{m,i,j}\, \exp\big(-j\, 2k\, \|\mathbf{r}_m - \mathbf{V}_{i,j}\|\big), \tag{35}$$

where $A(\cdot)$ models distance-dependent attenuation and known system scaling, while $\sigma_{i,j}(k)$ denotes the complex scattering coefficient parameterized by the predicted material descriptors. In our implementation, the attenuation term includes a $1/R$ decay factor, consistent with the experimental setup.

Soft visibility. Visibility is approximated as

$$\eta_{m,i,j} = \sigma\big(\kappa\, \langle \mathbf{N}_{i,j}, \mathbf{d}_{m,i,j} \rangle\big), \tag{36}$$

where $\mathbf{d}_{m,i,j}$ is the view direction. This soft formulation avoids zero-gradient discontinuities and provides continuous orientation cues for optimization.

Geometric gradients. The full renderer gradient contains contributions from distance attenuation, soft visibility, surface normals, and phase. The phase-dependent component for a single surface contribution $\hat{y}_{m,i,j}(k)$ is proportional to

$$\left.\frac{\partial \hat{y}_{m,i,j}(k)}{\partial \mathbf{V}_{i,j}}\right|_{\text{phase}} = -j2k\, \hat{y}_{m,i,j}(k) \frac{\mathbf{V}_{i,j} - \mathbf{r}_m}{\|\mathbf{V}_{i,j} - \mathbf{r}_m\|}, \tag{37}$$

with additional terms from the geometry dependence of attenuation and visibility. Automatic differentiation uses the complete renderer; the displayed phase term isolates the rapidly oscillatory component that drives the non-convexity discussed in Appendix B.3.

C.5 Signal Refinement as Constrained Residual Learning

Analytic RF rendering cannot capture all real-world effects (Appendix B.3). SPIRIT introduces a signal refinement module to compensate for this model–reality mismatch.

Residual formulation. Given analytically rendered signals $\hat{\mathbf{Y}}$, the refined prediction is

$$\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} + \beta\, \mathcal{G}_\psi(\hat{\mathbf{Y}}), \tag{38}$$

where $\mathcal{G}_\psi$ is a lightweight residual network and β is initialized small.

Identity regularization and schedule. The refiner is regularized toward a near-identity correction and is activated only after coarse geometry convergence. The residual branch receives the analytically rendered signal rather than scene parameters and does not provide a direct residual-gradient path to the geometry encoder. This separation limits the extent to which residual learning can mask geometric error, but it is not a hard physical guarantee: the refiner may still absorb signal structure omitted by the approximate renderer.

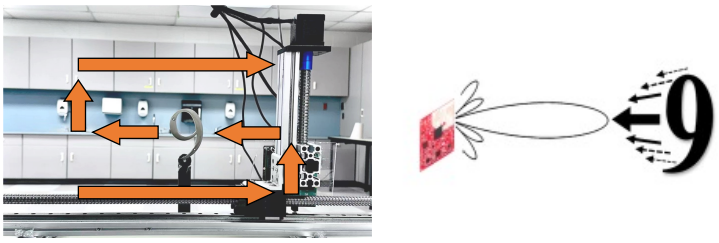
C.6 Optimization Landscape and Two-Stage Training

Phase-induced non-convexity. The RF phase $\phi = 2kR$ implies that a displacement $\Delta R = \lambda/4$ induces a π phase shift (Appendix B.1.8). This produces a highly oscillatory loss surface with many local minima and cycle-skipping behavior.

Continuation interpretation. Let

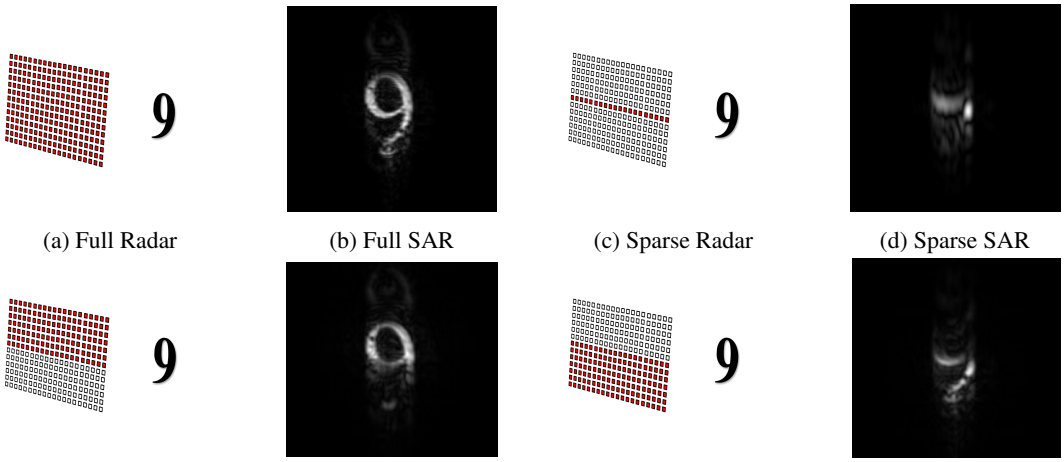
$$\mathcal{L}(\mathbf{V}) = \mathcal{L}_{\text{mag}}(\mathbf{V}) + \lambda\, \mathcal{L}_{\text{phase}}(\mathbf{V}).$$

Stage I sets the weight of phase-sensitive terms to zero, with a smoothed objective dominated by magnitude- and geometry-based losses. Stage II increases λ after the geometry estimate has entered a stable local basin, reintroducing phase-sensitive consistency only after coarse structure has been established. This continuation-style strategy reduces the tendency of early phase gradients to drive large geometric errors under sparse apertures; it does not imply convergence to a unique or global optimum.



(a) Data Collection Setup (b) Propagation Illustration

Figure 4: Real-world SAR data collection and signal propagation illustration. (a) Physical scanning setup and Z-shaped raster trajectory used for planar SAR collection, consistent with the simulator configuration. (b) Conceptual illustration of wave propagation and angle-dependent scattering, highlighting that while the beam covers the entire region of interest, different aperture locations emphasize different geometric components of the target.



(a) Full Radar (b) Full SAR (c) Sparse Radar (d) Sparse SAR

(e) Upper Radar (f) Upper SAR (g) Lower Radar (h) Lower SAR

Figure 5: Real-world SAR observations of a metallic digit 9 under different aperture configurations. Black boxes mark radar aperture locations, and the red marker denotes the active position. From left to right: full-aperture, single-pass sparse aperture, upper-aperture subset, and lower-aperture subset. Full-aperture integration produces the most complete formed image, whereas restricted aperture subsets emphasize different portions of the target. The comparison illustrates viewpoint-dependent scattering and the loss of angular/spatial-frequency support under sparse acquisition.

D PRELIMINARY

We begin with a real-world SAR collection to motivate the sparse-aperture setting and illustrate how aperture position changes the formed image under a fixed field of view. This preliminary study is qualitative: it demonstrates viewpoint-dependent scattering and incomplete angular support, rather than claiming that sparse sampling preserves all information available from the dense aperture. Figures 4–5 use conventional formed SAR observations for problem motivation; they are not final SPIRIT reconstructions and are not part of the quantitative MITO evaluation in Appendix F.3.

Data are collected using a Commercial Off-the-Shelf (COTS) mmWave radar, the TI IWR 1642, with a mechanical setup comprising the SAR setup shown in Figure 4. The radar performs a planar scan over a rectangular region of size $200\,\mathrm{mm} \times 100\,\mathrm{mm}$ following a Z-shaped raster trajectory, repeated 13 times. A metallic digit 9 is placed within the scanning region as the target. The target remains fully illuminated throughout the collection, as the radar beamwidth covers the entire region of interest at all aperture locations. This setup keeps the target within the nominal field of view across the scan, allowing us to separate field-of-view coverage from the loss of angular and spatial-frequency support caused by using fewer aperture positions.

The aggregated reconstruction obtained by coherently combining all 13 scans is shown in Fig. 5b. With dense aperture coverage, the characteristic geometry of the digit 9 becomes clearly visible. However, a reconstruction from a single scan, shown in Fig. 5d, produces incomplete and spatially ambiguous structures. The single-scan reconstruction is incomplete because the restricted aperture provides substantially less angular and spatial-frequency support than the full collection. The backscattered energy is also highly viewpoint dependent. Only surfaces that are specularly aligned with the viewing direction produce strong coherent returns; off-angle regions are dominated by diffuse scattering with substantially weaker backscattered energy. These weak responses cause structural details to be blurred or missed.

To explicitly demonstrate this effect, we design a controlled experiment that isolates the contribution of aperture position. We separately reconstruct SAR images from the upper and lower six scans. As shown in Fig. 5f, radar positions are illustrated as black rectangles, and the subset highlighted in red corresponds to the aperture locations used to generate the image. The upper-aperture subset produces an image that clearly resolves the lower portion of the digit 9. The image generated from the lower-aperture subset in Fig. 5g reveals the upper portion of the digit. Although the reflected intensity is weaker in this case, the global shape remains partially recognizable. Together, these observations show that different aperture subsets contribute complementary measurements of the same object, while no individual subset provides the full support available from the dense aperture.

E Experimental Details

This appendix records the experimental settings used to reproduce the reported comparisons, including data splits, simulation assumptions, preprocessing, baseline definitions, optimization schedules, compute, and metrics.

E.1 Datasets

Synthetic corpus. After filtering incomplete object–measurement instances, the synthetic corpus contains 1,957 valid instances: 1,567 for training, 195 for validation, and 195 for testing. The partitions are object-disjoint at test time. Synthetic scenes cover heterogeneous object geometry, orientation, and surface structure.

MITO real measurements. We evaluate a 60-object subset of the MITO dataset Dodds et al. (2025). The real-data split is object-disjoint: 40 objects are used for Stage II residual adaptation, 10 for validation, and 10 for testing. Each object has two captures at horizontal orientations separated by 90°, and both captures of the same object remain in the same partition, yielding 80/20/20 train/validation/test captures. Stage I geometry learning uses no real MITO measurements; the real training split is used only to adapt the residual-refinement branch. Thus, the real test captures are held out from real-data adaptation, although the experiment is not presented as object-identity zero-shot with respect to the broader synthetic mesh corpus. The released complex-valued 77 GHz measurements and recorded radar poses are processed using the dataset trajectory; the physical measurements therefore include hardware response, positioning/calibration residuals, material-dependent scattering, and multipath effects not reproduced exactly by the synthetic generator or analytic renderer. Reference geometry is registered to the radar coordinate frame using the dataset calibration/object pose. The same aligned reference is used for all methods, with no method-specific post-hoc alignment before Chamfer Distance evaluation.

E.2 Radar Signal Simulation

Waveform Parameters. We simulate a 77 GHz FMCW radar, closely modeled after Texas Instruments mmWave sensors. The operating frequency range is 77.5–80.5 GHz, with center frequency $f_c = 79.0\,\text{GHz}$ and bandwidth $B = 3.0\,\text{GHz}$. The chirp slope is $S = 5.86 \times 10^{12}$ Hz/s, yielding a chirp duration $T_c = B/S \approx 512\,\mu\text{s}$. Each chirp contains $N_s = 512$ ADC samples, corresponding to a sampling rate of approximately 1 MHz. The theoretical range resolution is $\delta_r = c/(2B) \approx 5\,\text{cm}$.

Signal Processing. Raw ADC signals are converted to the frequency domain by applying a one-dimensional Fast Fourier Transform along the fast-time dimension. A Hanning window is applied

before the FFT to suppress sidelobes. All subsequent reconstruction, rendering, and learning-based processing operate on the resulting complex-valued range-frequency representation.

Forward Scattering Model. The synthetic measurement generator is distinct from the differentiable renderer optimized by SPIRIT. The generator operates on reference meshes with hard visibility/ray casting and prescribed scattering components, whereas SPIRIT renders from its predicted structured surface using soft visibility and learned reflectivity/roughness. Both are geometry-based first-order approximations and neither models all higher-order multi-bounce effects; the synthetic benchmark is therefore partially model-matched at the level of physical assumptions, but not an inversion of an identical renderer. For radar position $\mathbf{r}_m$ and wavenumber k, the simulated return is

$$y_m(k) = \sum_{i \in \mathcal{V}_{\text{vis}}} A(\|\mathbf{r}_m - \mathbf{v}_i\|)\, \rho_i \, \exp\big(-j\, 2k\, \|\mathbf{r}_m - \mathbf{v}_i\|\big), \tag{39}$$

where $\mathbf{v}_i$ denotes a visible mesh vertex, ρ_i is its reflectivity, and $A(\cdot)$ captures system scaling and distance-dependent attenuation.

Visibility Culling. Visible scatterers $\mathcal{V}_{\text{vis}}$ are determined via a two-stage culling process. First, backface culling removes vertices whose surface normals $\mathbf{n}_i$ satisfy $\mathbf{n}_i \cdot (\mathbf{p} - \mathbf{v}_i) \leq 0$. Second, exact ray-casting is performed to discard vertices occluded by other parts of the mesh. This ensures that only physically visible surface elements contribute to the simulated return.

E.3 Synthetic Aperture Geometry

Aperture Configuration. Radar measurements are collected over a planar rectangular synthetic aperture with dimensions approximately 200 mm $\times$ 100 mm. The aperture is sampled on a dense Cartesian grid, satisfying the spatial Nyquist criterion and preventing grating lobes.

E.4 Sparse Aperture Sampling

Measurement Representation. Full-aperture measurements are represented as a complex-valued tensor $\mathbf{Y}_{\text{full}} \in \mathbb{C}^{N_y \times N_x \times N_s}$, where N_x and N_y denote aperture positions and N_s denotes range-frequency bins. In our datasets, scanline sparsity corresponds to selecting a subset of N_y rows (vertical scan positions) while retaining all N_x columns within each selected row.

Sampling Strategy. We distinguish between training-time and evaluation-time sparsity patterns. During training, we apply random scanline masks as data augmentation, randomly selecting subsets of scan lines to improve robustness to sampling variations. For validation and all reported test results, we use fixed and deterministic center scanline masks, corresponding to the central 1, 3, or 6 scan positions described in Section 5 of the main paper. All baseline methods are evaluated under the identical sparsity masks to ensure a fair and controlled comparison. All quantitative results are computed using these fixed-center scanline masks.

Training and Evaluation Protocol. During training, a new sparse mask is sampled independently for each mini-batch to prevent overfitting to any fixed sampling pattern. During validation and testing, sparse masks are fixed to ensure reproducible and fair comparisons across methods. The Hadamard product gives the network input

$$\mathbf{Y}_{\text{in}} = \mathbf{Y}_{\text{full}} \odot \mathbf{M}, \tag{40}$$

with missing measurements zero-padded. The masked signal $\mathbf{Y}_{\text{in}}$ is used as input to the differentiable backprojection operator, whose output feature map serves as the direct input to the reconstruction backbone (Section 4.2). To preserve expected signal energy across sparsity levels, non-zero measurements are normalized by the inverse sparsity ratio. This normalization is applied consistently to all methods, including baselines.

E.5 Synthetic Scene Generation

Object Assets. Object meshes are sourced from the MuJoCo physics engine and the MITO object dataset. All meshes are preprocessed to ensure watertightness, metric scaling, and centered geometry.

Diversity and Randomization. Scenes include both simple geometric primitives and complex household objects. Each object undergoes random 3D rotation (roll, pitch, yaw uniformly sampled in $[0, 2\pi]$) and slight translational jitter of up to ± 2 cm. Test scenes contain object instances that were not seen during training.

E.6 Model Architecture

Geometry Network. The geometry reconstruction network follows an asymmetric ResNet–UNet architecture. The encoder comprises four stages with channel sizes $\{64, 128, 256, 512\}$, followed by a self-attention bottleneck. The decoder mirrors the encoder with skip connections. The output is a 128×128 structured surface grid of 3D vertex coordinates.

Parameterization. Geometry is parameterized using a global–local formulation. A global MLP predicts an axis-aligned bounding box, while convolutional heads predict normalized local offsets within this box.

Physics-Based Renderer. The VertexRenderer module is fully differentiable with respect to vertex positions. Visibility is modeled using a soft normal-based gating function $\sigma(\kappa\, \mathbf{n} \cdot \mathbf{v})$, which provides stable gradients for surface orientation during training.

E.7 Data Representation and Preprocessing

Input Signal Representation. The overall input to SPIRIT is the sparsely sampled complex-valued SAR signal. This signal is first transformed through a differentiable backprojection operator, which aggregates the sparse measurements into a spatial feature representation. The resulting BP feature map serves as the direct input to the UNet-style reconstruction backbone. The sparse complex signal is not fed directly into the UNet; instead, it is implicitly encoded by the BP operator. Complex signals are represented using real and imaginary components, stored as complex64 tensors in PyTorch.

Log-Magnitude Normalization. To stabilize training and reduce dynamic range, the input signal magnitude is normalized to the log-magnitude. Specifically, given a complex signal S:

$$m = 20 \log_{10}(|S| + \epsilon), \qquad m_{\text{clip}} = \text{clip}(m, -100\ \text{dB}, 0\ \text{dB}), \tag{41}$$

$$S_{\text{norm}} = \text{MinMax}(m_{\text{clip}}) \cdot e^{j \angle S}. \tag{42}$$

Unobserved measurements are set to zero, while the mask M restricts network attention and backprojection to valid samples.

Reconstruction Volume. The reconstruction volume is defined in the physical coordinate frame specified by the simulator configuration. Typical bounds are $[-0.15, 0.15]$ m per axis for validation crops, and up to $[-0.3, 0.3]$ m for full-field simulation. Voxel resolution is 128^3 or 256^3, depending on the experiment. These volumetric bounds and resolutions are used for simulation and volumetric baselines; SPIRIT itself predicts a structured surface grid.

Sparse Sampling Strategy. Sparsity is structured by scan lines rather than random point dropping. The dataset selects a subset of vertical azimuth scan lines and masks all remaining positions. This setting reflects realistic sparse-aperture collection and preserves angular coherence.

E.8 Baseline Implementations

All baselines operate on the exact sparse measurements and are evaluated under identical conditions. They are designed as matched controls under our acquisition model; the implicit and volumetric variants are not presented as faithful reimplementations of Radar Fields, GeRaF, SpINRv2, RadarSplat, or other systems whose inputs, forward operators, and scene regimes differ from ours.

Backprojection. The BP baseline implements a classical time-domain SAR backprojection using frequency-domain phase correction. Given a radar location $\mathbf{r}$ and spatial location $\mathbf{x}$, the contribution is:

$$I(\mathbf{x}) = \left| \sum_{m \in \mathcal{M}} \sum_{k} y_m(k) \exp\big(j\, 2k\|\mathbf{x} - \mathbf{r}_m\|\big) \right|, \tag{43}$$

where $y_m(k)$ denotes the measured frequency-domain signal. The output is the magnitude of the complex image.

Compressed Sensing (CS). The CS baseline solves a sparse inverse problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \tag{44}$$

where $\mathbf{x}$ is a volumetric reflectivity grid and $\mathbf{A}$ denotes the forward operator. Optimization is performed using FISTA for the fixed 20-iteration schedule reported in the main paper, with $\lambda = 0.01$ and step size 0.5. The same values are used for every test object and sparsity level, with no per-test-instance tuning. A common fixed threshold is applied to extract surfaces for geometric evaluation.

Image-Domain UNet. The UNet baseline takes a coarse BP image as input. Three channels are provided: real, imaginary, and magnitude components. The architecture matches the SPIRIT backbone but operates purely in the image domain.

Height Baseline. The height baseline uses the same backbone as SPIRIT but assumes a fixed planar surface at the mid-depth of the volume. Only reflectivity is predicted on this surface; no 3D deformation is allowed. This baseline isolates the benefit of explicit surface modeling.

Implicit and Volumetric Baselines. Implicit and volumetric neural-field controls optimize a representation independently for each test sample. They do not generalize across the dataset. Each sample uses the same fixed 100-iteration budget and minimizes the L2 discrepancy between predicted and observed signals. Appendix tables additionally retain a direct volumetric variant from the same matched benchmark pipeline. These controls are included to probe representation behavior under a common acquisition/evaluation protocol, not as faithful reproductions of named published neural-field systems.

E.9 Training Protocol

Stage 1 Training. Stage I focuses on stabilizing geometric reconstruction. During this stage, supervision is limited to geometry-related regularization and magnitude-based or image-domain losses derived from backprojected measurements. No complex-valued phase supervision or signal-level matching is enforced in Stage I. We use AdamW for 300 epochs with learning rate 10^{-4}, weight decay 10^{-4}, cosine annealing to 10^{-6}, gradient clipping at 0.5, batch size 8, and gradient accumulation over 8 steps.

Stage 2 Training. Stage 2 introduces complex-valued signal supervision and activates the residual refinement module. In the final SPIRIT configuration, the geometry backbone remains trainable but uses the reduced Stage II learning rate together with geometry-preserving regularization; material parameters and the refiner are optimized jointly. Frozen-backbone, partial-freeze, and no-refiner variants are used only as controlled ablations in Table 3. The optimizer and learning-rate scheduler are reinitialized at the beginning of Stage 2.

E.10 Computational Resources

We report the compute resources used for training and evaluation to improve reproducibility. All experiments were run on a machine with Intel(R) Xeon(R) Platinum 8268 CPUs and NVIDIA Quadro RTX 8000 GPUs with 48 GB VRAM. Unless otherwise noted, each experiment uses a single RTX 8000 GPU, although the full server includes multiple identical GPUs. The SPIRIT backbone contains approximately 122.2M parameters. Training on the 1,567-instance synthetic training split requires approximately 41 hours for Stage I and an additional 15 hours for Stage II. Inference takes approximately 80 ms per sample. We note that high-resolution rendering is memory-intensive, and GPU memory management becomes important in larger-resolution settings.

E.11 Evaluation Metrics

Signal Error. Signal error is computed in the frequency (phase-history) domain as the mean absolute complex residual over valid samples after the common preprocessing described above:

$$\mathcal{L}_{\mathrm{sig}} = \frac{1}{|\Omega|} \sum_{q \in \Omega} |S_{\mathrm{pred},q} - S_{\mathrm{gt},q}|\,, \tag{45}$$

where Ω indexes the valid phase-history entries used for evaluation. For image-domain baselines, the predicted image is first passed through the same forward operator before comparison.

Image Metrics. SSIM and PSNR are computed on 2D magnitude images obtained via standard backprojection:

$$I_{\mathrm{pred}} = \mathrm{BP}(S_{\mathrm{pred}}), \quad I_{\mathrm{gt}} = \mathrm{BP}(S_{\mathrm{gt}}). \tag{46}$$

SSIM is computed with data range normalized to $[0, 1]$. PSNR is defined as:

$$\mathrm{PSNR} = 20 \log_{10} \left(\frac{1}{\sqrt{\mathrm{MSE}}} \right). \tag{47}$$

Image-level metrics depend on the chosen image-formation operator and therefore complement rather than replace geometry- and signal-domain evaluation. The same metric definitions are used for the held-out MITO evaluation, with its common registered reference geometry and measured coherent data.

Chamfer Distance. Chamfer Distance is computed from metric scene coordinates and reported in millimeters throughout the paper. Synthetic coordinates stored in meters are converted to millimeters for reporting; MITO geometry is evaluated in its registered metric coordinate frame. For SPIRIT, predicted surface vertices are directly compared with the reference point set. For outputs represented on a reconstruction grid, a common fixed threshold and Marching Cubes surface extraction are used before geometric evaluation. CD is computed as the symmetric mean nearest-neighbor distance between the two point sets. No method-specific post-hoc alignment is applied in the controlled benchmark.

E.12 Reproducibility Notes

Data partitions and deterministic evaluation masks use random seed 42. The main SPIRIT comparison reports mean $\pm$ standard deviation over three independent training runs, whereas matched controls use the fixed schedules described above. All methods share the same evaluation pipeline and post-processing steps. Runtime comparisons are not emphasized, as this work focuses on reconstruction fidelity under extreme sparsity rather than real-time performance.

F Extended Experiments

This appendix provides a comprehensive experimental analysis of SPIRIT beyond the main quantitative results. We focus on understanding performance under practical sensing conditions and isolating the effect of key design choices. Specifically, we examine (i) the interaction between Stage 1 geometry learning and Stage 2 refinement, (ii) the effect of aperture coverage on SAR image formation, (iii) held-out evaluation on real MITO measurements, (iv) robustness to measurement noise, and (v) the role of constrained signal refinement. Together, these studies characterize when the proposed design is effective, how its components interact, and where the evidence remains limited.

F.1 Two-Stage Ablation: Geometry Learning and Signal Refinement

We examine the interaction between Stage 1 geometry learning and Stage 2 signal-driven refinement under extreme sparsity. Table 3 summarizes the main ablation results.

Table 3 in the main paper reports the complete stage-wise comparison. The qualitative examples below visualize the geometric drift observed for unconstrained Stage II optimization.

Unconstrained refinement improves image-level metrics but can induce geometric drift, as illustrated in Fig. 6, motivating the constrained refinement strategy adopted in Stage 2.

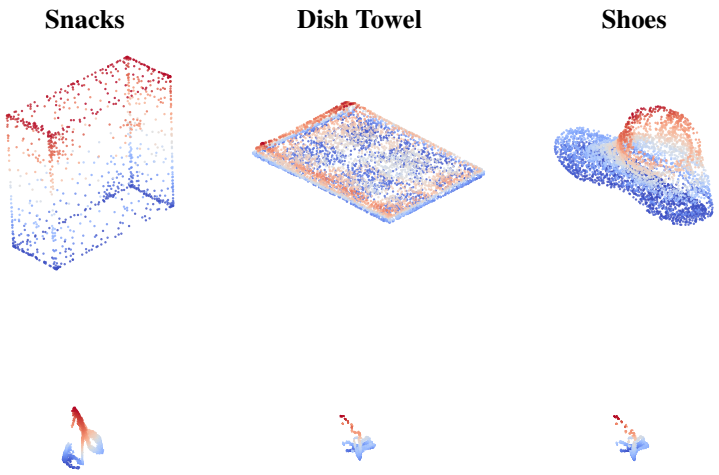


Figure 6: Geometry collapse under unconstrained Stage 2 fine-tuning. The top row shows ground truth, while the bottom row shows Stage 2 full fine-tuning without geometric constraints. Despite improved image appearance, unconstrained optimization leads to geometric drift and collapse.

F.2 Aperture Coverage and SAR Image Formation

Figure 7 visualizes the effect of retaining different amounts of aperture support. A single scanline provides only a restricted subset of the spatial-frequency and angular observations available in the full aperture. Adding scanlines introduces additional viewpoint-dependent measurements and reduces, but does not eliminate, the ambiguity of the inverse problem. The experiment therefore illustrates complementarity among retained aperture positions; it should not be interpreted as evidence that sparse sampling preserves all scene information.

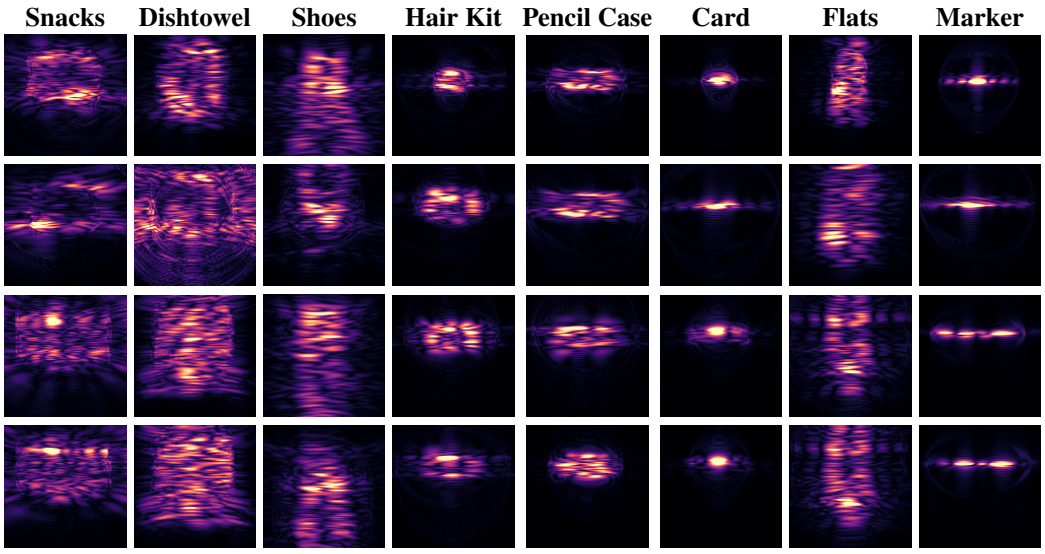


Figure 7: Reconstruction under increasing aperture coverage. Rows correspond to 1-, 3-, and 6-scan inputs. Additional scanlines provide complementary angular and spatial-frequency measurements, reducing ambiguity relative to the 1-scan setting. Sparse sampling nevertheless leaves unobserved support and does not uniquely determine the scene.

F.3 Real-World Evaluation on MITO

We evaluate the final SPIRIT system on held-out physical measurements from the MITO dataset Dodds et al. (2025). The real-data split is object-disjoint: 40 physical objects are used to adapt the Stage II residual-refinement branch, 10 for validation, and 10 for testing; the two $90°$-separated captures of each object remain in the same partition. Stage I geometry learning uses only synthetic measurements, and no real test capture is used during residual adaptation. Because the full pipeline has access to a disjoint real training partition, we describe this as synthetic geometry learning plus real residual adaptation rather than zero-shot synthetic-to-real transfer.

The real measurements contain hardware response, trajectory/calibration residuals, material-dependent scattering, occlusion, and multipath effects that are not reproduced exactly by either the synthetic generator or SPIRIT's single-scattering renderer. For geometric evaluation, we use the dataset-provided reference meshes registered to the radar coordinate frame using the released calibration and object pose. The same reference is used for every method and no method-specific post-hoc alignment is applied. We report Chamfer Distance in millimeters, phase-history ℓ_1 error, and PSNR/SSIM on backprojected magnitude images using the same metric definitions as the controlled benchmark.

Table 4: Quantitative comparison on the object-disjoint MITO real-world test split after residual-only Stage II adaptation on the disjoint real training split. The listed methods are matched controls under the common evaluation pipeline rather than claimed faithful reproductions of every recent radar system. CD is computed against the common registered reference geometry; signal and image metrics use the corresponding measured/reference radar products. Lower is better for CD and signal error; higher is better for PSNR/SSIM.

Method	SSIM ↑	PSNR (dB) ↑	CD (mm) ↓	Signal ℓ_1 ↓
Backprojection	0.452	18.15	18.52	0.851
Compressed Sensing	0.484	10.42	6.15	0.621
Implicit Neural Field	0.621	12.10	4.85	0.554
UNet (Image Domain)	0.655	13.50	4.20	0.482
Volumetric Neural Field	0.710	14.85	6.50	**0.353**
Direct Volumetric Field	0.685	10.10	5.85	0.401
SPIRIT (Ours)	**0.805**	**20.45**	**3.15**	0.362

SPIRIT obtains the lowest Chamfer Distance and the highest PSNR/SSIM among the evaluated methods on the held-out physical test objects. Its signal error is competitive but not the lowest: the volumetric neural-field control reaches 0.353 versus SPIRIT's 0.362. This mirrors the controlled ablations and reinforces the distinction between signal residual and geometric correctness. Because the experiment remains a controlled, object-scale near-field setting, we do not extrapolate these results to cluttered, multi-object, airborne, or strongly multipath SAR scenes.

Auxiliary-sweep scope. The following robustness and regularization sweeps are auxiliary controlled analyses used to study within-table trends. Their absolute CD values are not used to support the primary benchmark claims in Table 2; those claims are based on the object-disjoint benchmark and MITO protocols stated explicitly above.

F.4 Noise Robustness

To assess robustness to measurement noise, we inject additive complex Gaussian noise at different signal-to-noise ratios (SNRs). For each noise level, we report Chamfer Distance (mm) and signal ℓ_1 error. Results are summarized in Table 5.

SPIRIT has the lowest Chamfer Distance at every tested noise level. Signal error is not uniformly lowest, again showing that signal residual and geometric fidelity need not rank methods identically.

Table 5: Noise robustness analysis under varying SNR levels. Each entry reports Chamfer Distance (mm) / Signal ℓ_1 error. Lower values indicate better geometric fidelity and signal consistency.

Method	∞ dB	30 dB	20 dB	10 dB (High Noise)
Backprojection	16.8 / 0.61	22.5 / 0.79	26.4 / 0.78	55.3 / 0.82
Compressed Sensing	8.9 / 0.24	12.1 / 0.43	17.6 / 0.52	36.2 / 0.72
Implicit Neural Field	5.6 / 0.32	8.1 / **0.06**	9.4 / **0.11**	19.7 / 0.86
UNet (Image Domain)	4.1 / 0.64	5.0 / 0.72	8.8 / 0.87	15.6 / 0.91
Volumetric Neural Field	7.4 / 0.37	15.2 / 0.45	19.6 / 0.51	48.9 / 0.74
Direct Volumetric Field	4.0 / 0.40	5.8 / 0.71	10.5 / 0.13	22.4 / 0.58
SPIRIT (Ours)	**2.9 / 0.20**	**3.2** / 0.28	**4.6** / 0.48	**7.9 / 0.57**

F.5 Effect of Identity Regularization

We vary the identity-regularization coefficient λ_{id} while keeping the residual-refiner architecture fixed. Table 6 reports the resulting geometry–signal trade-off.

Table 6: Sensitivity to the identity-regularization coefficient λ_{id}.

λ_{id}	CD (mm) $\downarrow$	Signal $\ell_1 \downarrow$
0.0 (No identity reg.)	4.8	**0.22**
0.01	3.5	0.24
0.05	**2.1**	0.28
0.1	2.8	0.35
0.5	5.5	0.52
No Refiner	6.2	0.58

The sweep exposes a Pareto trade-off rather than a single metric optimum: weak identity regularization gives the smallest signal residual, while moderate regularization gives the lowest geometric error. This behavior reinforces our decision to interpret signal and geometry metrics jointly.

G Discussion, Limitations, and Societal Impact

This section discusses the applicability, limitations, and broader implications of SPIRIT. Rather than repeating experimental results, we aim to clarify when the method is effective, where it may fail, and how its design choices shape these behaviors.

G.1 Applicability and Failure Modes

Scope of empirical evidence. Our primary mechanism and attribution conclusions are established on the controlled synthetic benchmark, where the generating geometry and measurement process are fully specified. Appendix F.3 additionally evaluates held-out physical MITO measurements after residual-only adaptation on a disjoint real training split. These measurements provide evidence under model–reality mismatch, but the experiment remains controlled, object-scale, and near-field; it does not establish zero-shot generalization or performance in operational multi-object or remote-sensing SAR scenes.

Single-Object Assumption. SPIRIT is primarily designed for scenes dominated by a single main object. The controlled geometry experiments in this work use centered, isolated objects. The current structured $N \times N$ surface grid imposes a strong inductive bias toward a single coherent surface and is not designed to robustly represent arbitrary numbers of disconnected components or volumetric scatterers. We therefore restrict the geometric conclusions to this single-object regime; multi-object reconstruction requires a more flexible scene representation and dedicated evaluation. In particular, the present experiments do not cover clutter, layover, ground-object double bounce, vegetation volume scattering, or operational airborne/spaceborne SAR geometries.

G.2 Attribution Scope

The main experiments separate two sources of improvement only to the extent supported by the available controls: Table 2 compares matched representation/reconstruction controls, while Table 3 isolates the Stage I/Stage II optimization choices within SPIRIT. This is not a fully factorial comparison in which every representation is trained with identical parameter count, optimizer, and staged schedule. We therefore attribute the reported performance to the combined geometry-first system and claim only that the structured surface prior is effective in the studied regime, not that fixed topology is uniquely necessary or universally optimal.

G.3 Limitations of Structured Surface Representation

Surface Parameterization and Self-Intersection. The structured surface representation used in SPIRIT does not enforce topological constraints. Each point in the $N \times N$ grid predicts its (x, y, z) coordinates independently, allowing self-intersection and folding. This flexibility enables the representation of complex shapes, such as concavities or overlapping structures, but does not guarantee manifold consistency.

Holes and Thin Structures. Failures on holes and thin structures arise from both measurement ambiguity and representation limits. Sparse aperture coverage may provide insufficient support for weak, shadowed, or self-occluded structures, in which case reconstruction is necessarily prior-dependent. The fixed-resolution surface grid additionally limits features below its point spacing and favors coherent topology. Increasing aperture density can improve conditioning, while wider or more diverse trajectories may introduce new lines of sight; neither change removes the representation limits of the current fixed-topology surface. A visually plausible completion in a poorly observed region should therefore not be interpreted as evidence that the corresponding geometry was uniquely determined by the measurement.

Future Directions. Addressing multi-object scenes may require multi-grid or instance-wise surface representations combined with explicit segmentation. Recovering thin structures or complex topology may benefit from hybrid explicit–implicit surface formulations or adaptive resolution strategies. These extensions are left for future work.

G.4 Physics-Constrained Modeling

Role of the Signal Refiner. The refiner is a statistical model-mismatch compensator rather than an electromagnetic simulator. Identity regularization and a small residual weight constrain its contribution, but do not guarantee that every correction corresponds to a physically modeled scattering mechanism. Under strong multipath, calibration error, or other unmodeled effects, the refiner may absorb signal structure that is not explained by the recovered surface. For this reason, improvements in signal error are interpreted jointly with geometric accuracy rather than as independent evidence of geometry recovery.

G.5 Risks and Trade-offs in Signal Refinement

Geometry Collapse. Unconstrained signal-driven optimization can lead to geometric collapse, in which the reconstructed surface deforms to minimize signal loss rather than accurately representing the actual structure. Our ablation studies show that this behavior is configuration-dependent but poses a high risk without proper constraints.

H Potential Applications and Broader Scope

The sparse-aperture inverse-rendering formulation may be relevant to object-scale millimeter-wave applications in which dense spatial scanning is costly, including security sensing, human-centric radar perception, and close-range material or agricultural sensing. These domains share a collection-efficiency challenge, but they also introduce scene statistics and scattering effects not evaluated in this work.

SPIRIT has not been validated here for concealed-object detection, face authentication, vegetation reconstruction, airborne SAR, or other downstream tasks. The current evidence combines controlled synthetic attribution with held-out physical-object MITO evaluation; application-specific deployment would require dedicated measurements, forward-model validation, and task-level evaluation.

I AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT, OpenAI Codex, and Google Gemini, in several stages of this work.

For writing support, these tools were used to edit and polish author-provided text, with the goal of improving clarity, readability, organization, grammar, and presentation without altering the intended scientific meaning.

For retrieval and discovery, generative AI was used to assist in identifying potentially relevant literature, locating related research directions, and organizing references for the related-work discussion. All cited references were independently checked by the authors against their original sources before inclusion.

For research ideation and execution, generative AI was used to help structure research questions and technical ideas, examine alternative system designs, provide feedback on experimental plans, and assist with code development, debugging, and implementation. The authors determined the research objectives, methodological choices, experimental protocols, and final system design.

Generative AI was also used to help draft and revise portions of the manuscript based on outlines, technical descriptions, experimental results, and scientific content provided and verified by the authors. All generated or revised text was reviewed, edited, and approved by the authors before inclusion in the final manuscript.

Generative AI was not used to fabricate experimental data, numerical results, or evaluation outcomes. All reported results were produced by executing the implemented experimental pipeline on the datasets described in the paper. AI-assisted code, experimental configurations, citations, technical statements, results, interpretations, and scientific claims were inspected and verified by the authors.

The authors retain full responsibility for the research design, implementation, analysis, interpretation, claims, conclusions, and final content of the paper.