RADON IMPLICIT FIELD TRANSFORM (RIFT): LEARNING SCENES FROM RADAR SIGNALS

Anonymous authors
Paper under double-blind review

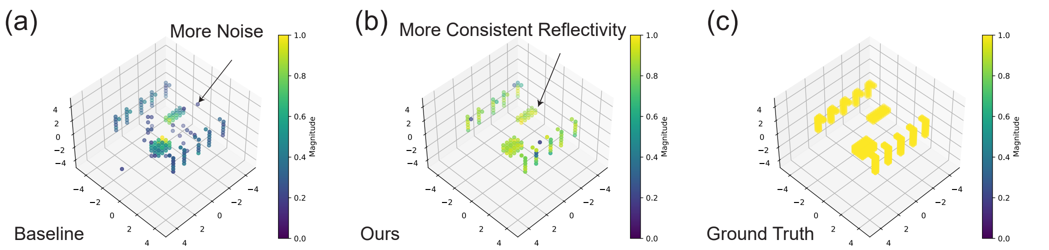

Figure 1: *Visualizations of the "mini parking lot" scene from Section 4: (a) Scene reconstructed by the baseline model. (b) Scene reconstructed by RIFT. (c) Ground truth scene visualized with the same granularity (defined in Section 2.2) as scene reconstruction. Under the same number of input, scene reconstruction by our RIFT model achieved 300% higher score in scene reconstruction than baseline by only using 40% of the data samples. The detailed data is in Table 2.*

ABSTRACT

Data acquisition in array signal processing (ASP) is costly because achieving high angular and range resolutions necessitates large antenna apertures and wide frequency bandwidths, respectively. The data requirements for ASP problems grow multiplicatively with the number of viewpoints and frequencies, significantly increasing the burden of data collection, even for simulation. Implicit Neural Representations (INRs) — neural network-based models of 3D objects and scenes — offer compact and continuous representations with minimal ground truth data. They can interpolate to unseen viewpoints and potentially address the sampling cost in ASP problems. In this work, we select Synthetic Aperture Radar (SAR) as a case from ASP and propose the ***R****adon* ***I****mplicit* ***F****ield* ***T****ransform* (RIFT). RIFT consists of two components: a classical forward model for radar (Generalized Radon Transform, GRT), and an INR based scene representation learned from radar signals. This method can be extended to other ASP problems by replacing the GRT with appropriate algorithms corresponding to different data modalities. In our experiments, we first synthesize radar data using the GRT. We then train the INR model on this synthetic data by minimizing the reconstruction error of the radar signal. After training, we render the scene using the trained INR and evaluate our scene representation against the ground truth. Due to the lack of existing benchmarks, we introduce two main new error metrics: ***phase-Root Mean Square Error*** (p-RMSE) for radar signal interpolation, and ***magnitude-Structural Similarity Index Measure*** (m-SSIM) for scene reconstruction. These metrics adapt traditional error measures to account for the complex nature of radar signals. Compared to traditional scene models in radar signal processing, with only 10% data footprint, our RIFT model achieves up to 188% improvement in scene reconstruction. Using the same amount of data, RIFT is $3\times$ better at reconstruction and shows a 10% improvement generalizing to unseen viewpoints as shown in Figure 3 and Table 1.

1 INTRODUCTION

Array signal processing (ASP) is a subdomain of digital signal processing (DSP) which involves multiple spatially distributed sensors (Swindlehurst et al., 2014). For some ASP problems – in particular, for imaging and detection problems – the cost of data acquisition is often high because of the relationship between resolution requirements, aperture size, and signal bandwidth. The angular resolution achieved by an array is a direct consequence of the aperture size, and the range resolution depends on the total bandwidth of the received signal (Richards, 2022; Moccia & Renga, 2011; Liu et al., 2021). The data acquisition cost grows linearly with each the number of samples, the aperture size (assuming antennas are Nyquist-spaced), and the number of frequency bins. In this work, we use radar imaging as a representative of ASP problems and address the cost of data acquisition with a deep learning model. Such combinations of machine learning and radar signal processing have been used in autonomous vehicles (Bilik et al., 2019), robotics (Ali et al., 2014), and geographic information systems (Javali et al., 2021).

Synthetic Aperture Radar (SAR) is a specialized radar imaging technique that synthesizes a large virtual antenna aperture by moving the sensor relative to the scene (Moreira et al., 2013). This process involves the coherent processing of successive radar echoes received at multiple points along the sensor path to reconstruct a high-resolution image. Most often, the radar samples are uniformly spaced throughout the synthetic aperture, so the size of the aperture determines the amount of viewpoint samples radar takes. In order to give a concrete example of the amount of data needed in SAR imaging, according to NASA, for a satellite with C-band radar, to get a spatial resolution of 10 m, the synthetic radar aperture size needs to be of the size of 47 soccer fields ($\sim$ 5km) (NASA, 2023). The resulting amount of data can be on the order of terabytes.

A potential remedy to the cost of data acquisition for SAR imaging is learning based reconstruction of objects and scenes. One example is Implicit Neural Representations (INR), which involves a neural network learning scene properties[1] (colors, opacity, and so on) through measurement signals like pictures. The prospect of INR interpolating between different view points, particularly for visual data, is accomplished with different scene representation mechanisms, e.g., voxels (Choy et al., 2016), point clouds (Achlioptas et al., 2018), meshes (Kanazawa et al., 2018) and especially the occupancy network by Mescheder et al. (2019). More recently, Neural Radiance Fields (NeRF) by Mildenhall et al. (2020) integrates a physical process called light field rendering from Levoy & Hanrahan (1996) to improve model performance. NeRF sparked works illustrate that the integration of underlying physical mechanisms enables better learning and scene representations.

In this study, we integrate deep learning methods with a traditional forward model for radar signals called the Generalized Radon Transform (GRT) (Nolan & Cheney, 2002; Monga et al., 2018). Analogous to the light marching in NeRF, the GRT is the physical mechanism integrated in the rendering process for radar, so the model can learn scene reconstruction directly from the observed radar signals. We denote our architecture as ***R****adon* ***I****mplicit* ***F****ield* ***T****ransform* (RIFT).

The main contributions of this work are as follows:

- We present the first method to learn implicit scene representations directly from radar signals.
- Using our method, we achieve better scene reconstruction and viewpoint interpolation with fewer measurements than traditional algorithms.
- We formulate the first joint benchmark for both radar scene reconstruction and signal interpolation which aligns with perceived quality.

2 BACKGROUND

2.1 ARRAY SIGNAL PROCESSING AND (SYNTHETIC APERTURE) RADAR

Array signal processing (ASP) generally refers to the use of two or more antennas for coherent processing. ASP is a fundamental technique and has diverse applications in radar, sonar, and communications. The use of multiple transmitters or receivers enhances overall system performance by

[1] In this work, the scene property of interest is complex reflectivity of the scene.

increasing gain, enabling beamforming, providing spatial filtering, and increasing signal-to-noise ratio (SNR)(Van Veen & Buckley, 1988). As wireless spectrum becomes more crowded and systems evolve to utilize higher frequency bands(Berger, 2014) – e.g., in 5G networks – the use of larger and more sophisticated antenna arrays has become crucial for achieving the precise beamforming necessary for efficient communication.

The basic principle of ASP in the narrowband setting involves adjusting the phase and amplitude of signals received by (or transmitted from) each element in the array. For signals that originate far from the antenna array, the spherical wavefront impinging on the antenna array appears locally as a plane wave. Coherently summing signals from any given direction can be accomplished by weighting the signal received at each element and adding up the signals over the array. The weights are simple phase shifts which depend on the array geometry (e.g., linear, planar, circular, etc.) and the direction-of-arrival (DoA) of the incoming signal. The phase adjustment allows for constructive interference in desired directions and destructive interference in others, effectively shaping the radiation pattern of the array. DoA estimation considers the signal angle as an unknown and tries to find the angle which best explains an observed signal.

Synthetic aperture radar (SAR) is related to DoA estimation in that goal is to sense an unknown environment. There are two important modifications, however. First, DoA estimation assumes that the signals exist in space and are being passively observed by the array. On the other hand, SAR techniques use a transmit antenna to excite the scene and observe reflections. A second modification that distinguishes SAR from conventional radar is the motion of the antenna array relative to the scene. As the array is moving, pulses are repeatedly transmitted, and reflections are stored at a variety of viewpoints. SAR processing takes these measurements and uses array position information to form a large “synthetic” aperture. Even with a small antenna array, the path traced by the antenna can be orders of magnitude larger, leading to greatly improved imaging capabilities (Moreira et al., 2013).

2.2 Generalized Radon Transform (GRT)

GRT is a standard forward model for a radar signal under the Born approximation and planar wave assumption Monga et al. (2018). In practice, we discretize the scene to voxel reflectors with position $\boldsymbol{x} \in \mathbb{R}^3$. For each voxel reflector, there is an associated complex-value reflectivity $\rho(\boldsymbol{x}) \in \mathbb{C}$. $\rho(\boldsymbol{x})$ is discretized to a look-up table for the INR to learn and interpolate.

Let s_{TX} and s_{RX} represent the slow-time variables corresponding to the positions of the TX and RX, respectively. Let $\gamma(s)$ denote the trajectory of the antennas, and let $R_b(\mathbf{x})$ be the range function under the bistatic configuration. Consider the fast-time temporal frequency ω within the range $[\omega_{\text{Lo}}, \omega_{\text{Hi}}]$, where ω_{Lo} and ω_{Hi} are the lowest and highest frequencies used by our radar system, respectively. Each frequency ω corresponds to a wave number $k(\omega)$ according to the standard definition. We define the GRT operator $\mathcal{F}$ such that the perceived radar signal $d(\omega, s)$ can be expressed as:

$$d(\omega, s) := \mathcal{F}[\rho] \approx \int_{\mathbb{R}^3} e^{j(k(\omega)R_b(\boldsymbol{x})} \boldsymbol{A}(\omega, s, \boldsymbol{x})\rho(\boldsymbol{x})e^{j\Phi(\rho(\boldsymbol{x})))}d\boldsymbol{x}$$

here the range function

$$R_b(\boldsymbol{x}) := ||\boldsymbol{x} - \gamma(s_{TX})|| + ||\boldsymbol{x} - \gamma(s_{RX})||,$$

and the function $\Phi(\rho(\boldsymbol{x}))$ stands for the phase of $\rho(\boldsymbol{x})$. The phase of reflector $\Phi(\rho(\boldsymbol{x}))$ corresponds to possible phase change takes place when the electromagnetic wave interacts with the scene.

2.3 Implicit Neural Representation

INR is a class of methods that learn a scene or an object through parameterized signals, providing a continuous interpolation that maps the signal to its domain. Compared to traditional grid-based representations, INR is more compact, as the spatial resolution in grid-based methods is inherently tied to the grid’s granularity.

Following the integration of light marching by NeRF (Mildenhall et al., 2020), there is a resurgence of research into INR trained on visual data to demonstrate concrete improvement of speed in training (Garbin et al., 2021), accuracy (Barron et al., 2021), and generalizability across viewpoints (Barron et al., 2022). The above works demonstrated data efficiency, training speed and reconstruction accuracy and set cornerstones to our work.

3 Methods and Experimental Setup

In this section, we present the relevant details of the GRT, the design of RIFT, and the error metrics we customize for radar data modality.

Overall, The INR from RIFT takes as input the location in the scene and returns the complex reflectivity of the scene at that point. Upon receiving the reflectivity estimate from the INR, the GRT directly produces the radar signal at different viewpoints. The predicted signal is compared to the ground truth, and automatic differentiation with gradient descent accumulates gradients to update the estimation of the scene.

3.1 Generalized Radon Transform (GRT) and Radar Signal Synthesis

Without loss of generality, in this research, we normalize all the magnitude of radar signal $\mathbf{S}$ to $[0, 1]$. To accurately simulate real-world scenarios, We use the bistatic radar configuration mentioned in 2.2 In the case, the transmitter (TX) and receiver (RX) are spatially separated at each time step. Further details of this setup are provided in Appendix B.1.

The combination of the INR learning $\rho(\boldsymbol{x})$ and the GRT generating the radar signal constitutes the inverse problem relative to the forward signal generation. We define the *granularity* of the forward and inverse problem pair as the distance between neighboring voxels along the coordinate axes. Unless otherwise noted, the scene extends a 3D cubical space with edges of 10m. The granularity of the forward problem to 0.2m, and that of the inverse problem to 0.4m. This twofold difference in granularity between the forward and inverse problems is designed to create a compact representation of the scene.

It is important to note that the dependence of the matrix $\mathbf{A}$ on $\mathbf{x}$ makes our assumption of no loss of generality nontrivial. To address this $\mathbf{x}$ dependence, we consider two distinct cases: the near-field approximation and the far-field approximation. However, since this pertains to an ASP problem and an extensive discussion would be tangential to the primary objective of developing a neural radar reconstruction algorithm, we defer the detailed analysis to Appendix B.3.

3.2 Implicit Neural Representation and the RIFT Workflow

In this study, we assume that the scene $\rho(\mathbf{x})$ remains constant over time. The INR we employ is a function $\hat{\rho}_\Theta(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{C}$, which parameterizes the scene's properties—specifically, the complex reflectivity for radar signals—using tunable parameters Θ. Consequently, we can formulate the optimization problem as follows[2]:

$$\arg\min_{\Theta} ||\mathbf{S} - \mathcal{F}[\hat{\rho}_\theta(\boldsymbol{x})]||_2$$

The approximation $\hat{\rho}$ utilized in this study is based on a multi-layer perceptron (MLP) model (Bishop, 1995), configured in two distinct ways: one incorporating layer normalization (Ba et al., 2016) and the other employing positional encoding, which is discussed in detail later. These configurations are designated as RIFT(N) and RIFT(S), respectively. Detailed descriptions of these configurations are provided in Appendix B.4. Both models are trained using standard backpropagation techniques (Rumelhart et al., 1986), with the exact architecture and training parameters outlined in Appendix B.

The positional encoding configuration for INR was first introduced in NeRF (Mildenhall et al., 2020) and subsequently analyzed by Tancik et al. (2020). In this study, we adopt a mathematically equivalent structure known as SIREN (Sitzmann et al., 2020) for positional encoding within the INR framework. This approach allows us to evaluate the effect of positional encoding on the learning process with radar signals. Further details regarding the neural network structures and their configurations are provided in Appendix B.4.

[2]In our experiments, we slightly modify the optimization process to enhance numerical convergence properties. Details are provided in Appendix B.5

It is crucial to note that during training, we employ a nonstandard approach of *accumulating gradients within an individual epoch* across different viewpoints. This methodology is essential for SAR systems, where the geometry of the scene is learned through the coherent addition of radar signals from various positions within the synthetic aperture. *This gradient accumulation is specifically designed to mimic the physical motion inherent in synthetic aperture radar systems*, ensuring that the neural network accurately captures the spatial relationships and scene geometry in the same way as SAR.

In Figure 2, we present a workflow chart for RIFT. The RIFT workflow models the physical process of radar sensing, where transmitted and received waves interact with the scene. The scene is discretized into a look-up table, serving as the ground truth for our Implicit Neural Representation (INR) to learn from. RIFT comprises two main components: a GRT Segment and an INR scene model.

The GRT segment transforms the learned scene representation into an approximation of the radar signal. It effectively bridges the gap between the continuous scene representation and the discrete radar measurements. The INR learns by backpropagating through the GRT segment and comparing the generated radar signals against the ground truth radar signals. This iterative process refines the scene representation to minimize discrepancies between the predicted and actual radar data.

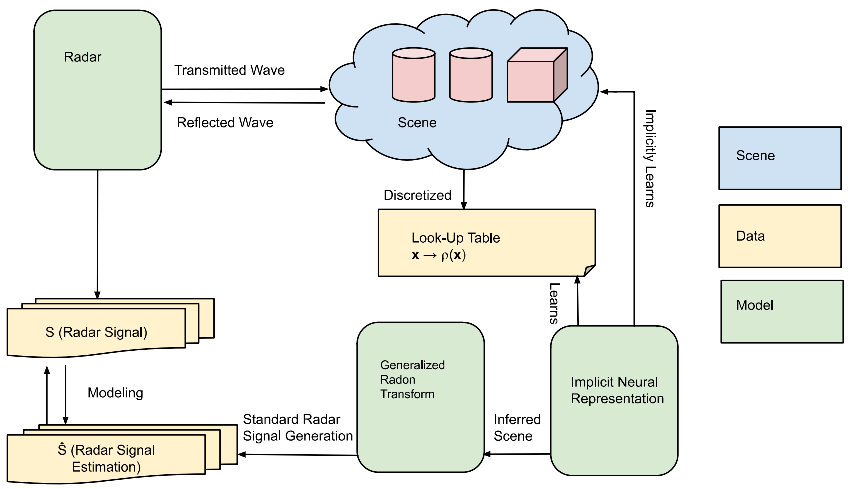

Figure 2: *Workflow chart of the RIFT architecture. The diagram illustrates how RIFT models physical radar sensing, transforms the learned scene into radar signals through the GRT segment, and iteratively refines the scene representation via backpropagation.*

3.3 Error Metrics and Benchmark

To our best knowledge, there are no existing benchmarks to gauge how well a neural net (NN) learns both the radar signal and the corresponding scene properties. In traditional SAR imaging algorithms, common error metrics include norm-based measures like Mean Square Error (MSE) (Gonzales & Woods, 2008), structural measures like the Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and probabilistic measures mainly used for classification tasks, such as Kullback-Leibler (KL) divergence (Gao, 2010).

However, although we used Mean Absolute Error (MAE) during the training stage of the INR (see Appendix B.5), we cannot always distinguish between good and bad reconstructions using MAE alone. This is because the coherent addition of phase information of the signal determines the geometry, as discussed in Franceschetti & Lanari (1999), but the scene is determined by the magnitude of reflectivity.

Therefore, we introduce two modified traditional error metrics: magnitude-SSIM (m-SSIM), magnitude-cosine similarity (m-COS), threshold Intersection-over-Union (tIoU), and phase-Root Mean Square Error (p-RMSE). The idea is splitting the metrics to two parts: magnitude and phase. The magnitude-based metrics are used for scene reconstruction, and phase-based metrics are used for radar signal interpolation for unseen viewpoints. The detailed definition are deferred to the Appendix B.6

As a benchmark for scene rendering, we use the inverse GRT operator $\mathcal{F}^{-1}$, since there is no existing machine learning algorithm that renders a reflective scene from radar signals. We select a block version of the Kaczmarz method (Kaczmarz, 1993) as our inversion technique because it allows us to compute the inversion using a least squares formulation with a reasonably fast convergence rate. Both RIFT and the Kaczmarz method utilize radar signals without downsampling in the transmitter/receiver (TX/RX) combinations or frequency bandwidth. Due to the differing granularity between the forward signal synthesis and the inverse scene reconstruction and unseen viewpoint interpolation, we resize the scene in the forward problem to match the granularity of the scene in the inverse problem using the scikit-image library (van der Walt et al., 2014) before calculating the benchmark.

4 Results and Discussion

The evaluation of RIFT's performance is twofold, addressing the core challenge of reducing data acquisition costs in SAR systems. RIFT offers two primary solutions. The first is that by reconstructing the scene using fewer radar measurements, RIFT lowers the demand for extensive data collection. This is particularly beneficial in scenarios where data acquisition is time-consuming or resource-intensive. The effectiveness of scene reconstruction is evaluated using error metrics detailed in Section 3.3, which compare the ground truth scene reflectivity with the reconstructed scene. Then, RIFT can interpolate radar signals between previously unseen viewpoints, effectively increasing the available data supply without additional measurements. This capability enhances the system's ability to generate comprehensive radar maps from limited viewpoints. The performance of unseen viewpoint interpolation is assessed by comparing the GRT of the learned scene with the ground truth radar signals in the test set.

4.1 Scene Reconstruction

For simplicity, in this section we assume that the materials of the scenes are perfect reflectors whose phase interactions are captured by $\mathcal{F}$; that is, $\Phi(\rho(\boldsymbol{x})) = 0$. We generate three simple objects and two complex scenes, and present our results and evaluation of scene reconstruction and unseen viewpoint interpolation. All data are generated from 51 uniformly sampled azimuth and elevation angles (within their respective domains) of the spherical coordinate system described in Appendix B.1. In total, there are 2,601 possible viewpoints in the dataset we generate. When presenting results in this section, the "viewpoints" are sampled from these 2,601 viewpoints.

We demonstrate that in all cases, RIFT models use at most 40% of the training data used by the baseline model, yet they perform better in scene reconstruction by at least 109.6% in the m-SSIM metric. In all but one case mentioned in Appendix A, RIFT models outperform the baseline model in unseen viewpoint interpolation, thus achieving better generalization with significantly less data in terms of p-RMSE. Empirically, the RIFT(N) models without SIREN-style positional encoding perform better in scene reconstruction, while the RIFT(S) models perform well in cases where RIFT(N) models converge to local minima and excel at viewpoint interpolation.

4.1.1 Simple Scene Reconstruction

The three simple objects are a sphere of radius 2 meters, a cube with an edge length of 2 meters, and a tetrahedron (denoted as "Pyramid") with a base measuring 2 meters by 2 meters and a height of 2 meters. Both the sphere and the cube are centered at the origin of the coordinate system described in Appendix B.1.

In Table 1[3], we compare the m-SSIM score and p-RMSE value, along with auxiliary metrics such as t-IoU and the cosine similarity of the reconstructed magnitude (denoted as m-COS). We reconstructed the scene using 100 and 1,000 viewpoints with the RIFT workflow; the corresponding datasets are denoted as RIFT(N or S)-(100 or 1000). Additionally, we applied the least squares

[3] In this work, we use bold font to highlight the best performances.

Table 1: Simple Scene Reconstruction Result for Cube corresponding to Figure 3

Model	m-SSIM	m-COS	t-IoU	p-RMSE
LS-100	0.0704	0.5698	0.0578	0.0155
LS-200	0.0908	0.6689	0.1663	0.0153
LS-500	0.1128	0.7822	0.1243	0.0151
LS-1000	0.1706	0.8511	0.2183	0.0150
RIFT(N)-100	**0.6395**	0.9792	0.3677	0.0147
RIFT(S)-100	0.1435	0.6957	0.3302	0.0152
RIFT(N)-1000	0.6298	**0.9833**	**0.3714**	**0.0145**
RIFT(S)-1000	0.6045	0.9606	0.3688	0.0147

reconstruction using 100, 200, 500, and 1,000 viewpoints; these datasets are denoted as LS-100, LS-200, LS-500, and LS-1000, respectively.

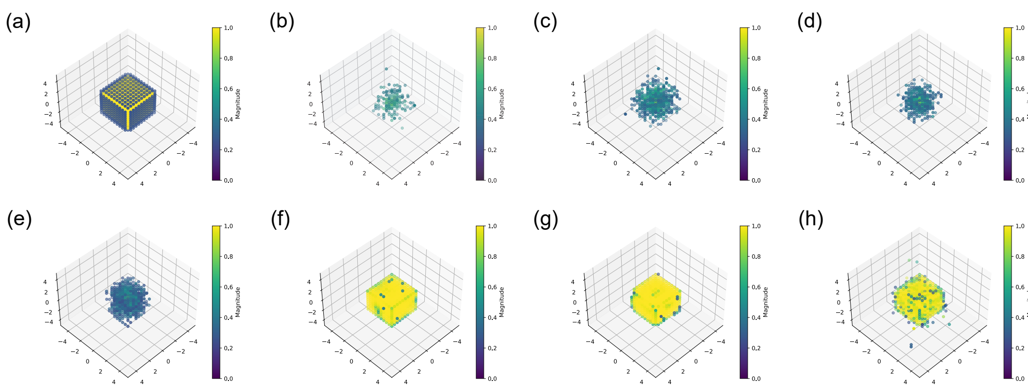

Figure 3: *Visualizations of the "cube" scene (a): Ground truth of a cube of edge of 2m. (b)-(e): Scene reconstruction by the baseline with 100, 200, 500, and 1000 viewpoints, respectively. (f): Scene reconstruction by RIFT(N) with 100 The m-SSIM score and p-RMSE of reconstruction in (f) is 0.6395 and 5.4986, 274% and 11% better than those of reconstruction in (e) while only using 10% of the viewpoints, respectively. (g), (h) Scene reconstruction by RIFT(N/S) with 1000 viewpoints as references.*

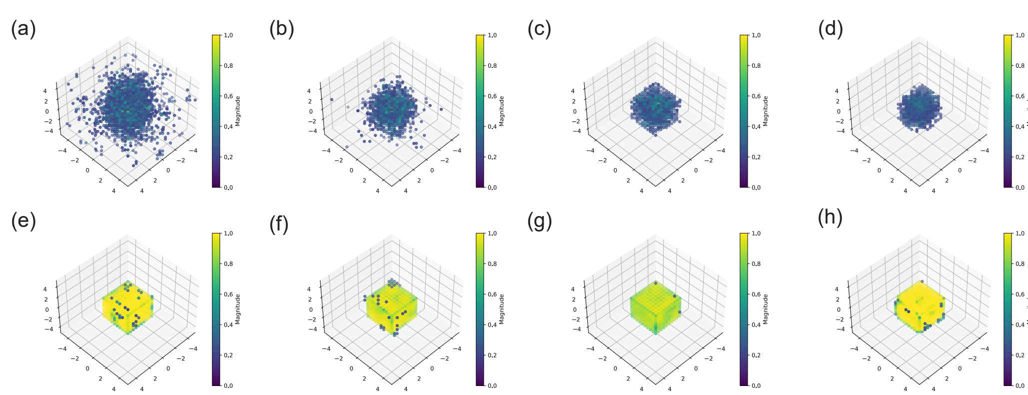

Figure 4: *Visualizations for presenting the need of data from different models. (a)-(d) Scene reconstruction by the baseline least square model with 100, 200, 500, and 1000 viewpoints. (e)-(h) Scene reconstruction by our RIFT(N) model with 100, 200, 500, and 1000 viewpoints. The data is*

In addition to comparing the RIFT models using small amounts of data against baseline models using more data, we present Figure 4, which visualizes a comparison of two reconstructions with the same number of inputs. To ensure a fair comparison, the assumed SNR for the visualization of all eight sub-graphs is set to 0.2. The data footprint of the baseline model demonstrates the necessity for more data to reconstruct the scene precisely, which aligns with the requirement for more samples in SAR and other active sensing ASP problems.

In all three simple scenes, using a tenth of the viewpoints, the RIFT-100 instance of our RIFT work flow scored up to 247.80% higher in m-SSIM score, up to 15.05% higher in m-COS, up to 68.44% higher in t-IoU, and up to 4.75% lower in p-RMSE across the three simple scene as compared to LS-1000. The detailed data for sphere and pyramid data and figures are available in Appendix A.

4.1.2 Complicated Scene Reconstruction

For more complex scenes, we constructed two scenarios: “mini parking lot” and “mini highway.” In the “mini parking lot” scene, we placed ten “street lights,” each 2.2 meters high, distributed evenly along a line 1.8 meters from the y-axis of the scene. There are also two “cars” of different sizes positioned on opposite sides of the road, with dimensions of 0.8 meters × 0.6 meters × 0.6 meters and 1.0 meters × 0.4 meters × 0.4 meters, respectively. We defer the figure and details of the “mini highway” scene to the additional results in Appendix A.

For complex scenes, we used 1,000 viewpoints for the RIFT workflow instances—denoted as RIFT(N)-1000 and RIFT(S)-1000 for the two configurations, respectively—and 2,500 viewpoints for the least squares baseline model, denoted as LS-2500. Using only 40% of the training data, our RIFT-1000 models achieved up to a threefold improvement in m-SSIM score, a 53.79% higher m-COS, and a 567.20% higher t-IoU, although the lead in p-RMSE is smaller. The instances in Sections 4.1.1 and 4.1.2 where the RIFT model does not perform as well in unseen viewpoint interpolation are likely due to the number of viewpoint samples used.

Table 2: Complicated Reconstruction Result for “Mini Parking Lot” Scene Corresponding to Figure 1

Model	m-SSIM	m-COS	t-IoU	p-RMSE
LS-2500	0.1662	0.5931	**0.0356**	0.0153
RIFT(N)-1000	0.5705	0.7833	0.0244	0.0152
RIFT(S)-1000	**0.6639**	**0.9120**	0.0345	**0.0149**

4.2 Case Study: Weak Target Detection in Far-Field

In this section, we investigate a real-world problem in radar signal processing to demonstrate the capabilities of the RIFT. Weak Target Detection (WTD) (Li et al., 2024; Bai et al., 2020) refers to scenarios where multiple objects in a scene have different reflectivities and are positioned close to each other, making it challenging for radar systems to distinguish them.

We use a far-field setting with a smaller scene extent ranging from -3 meters to 3 meters and the radar at a distance of 50 meters from the scene. The spatial granularity is set to 0.12 meters. In the scene, two rectangular reflectors are placed on opposite sides of the y-axis, separated by 1.2 meters. The dimensions of each bar are 2.4 meters in length, 0.72 meters in width, and 0.48 meters in height. In order to mimic the practical scenarios, instead of generating signal from a hypothetical sphere surrounding the scene, we limit the azimuth and elevation angle samples to 41 and 21 samples on $[0.1\pi, 0.3\pi]$, respectively.

The bar on the negative x-side is assigned a reflectivity of 1.0, while the bar on the positive x-side is assigned reflectivity of 1.0, 0.5, 0.333, and 0.25 in four separate experiments. Both the RIFT model and the least squares model use 500 input data points. All other training setups are identical to those in the experiments described in Section 4.1. Figure 5 presents the resulting reconstructions of the scene.

From Figure 5(a)–(d), we observe that in far-field simulations, the least squares baseline models are unable to resolve the two reflectors, regardless of differences in their reflectivities. This outcome corresponds to the Weak Target Detection (WTD) problem, where radar systems encountering

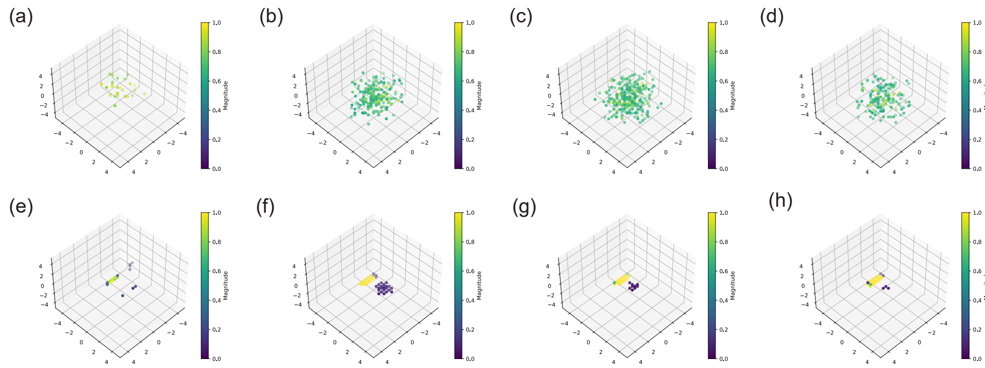

Figure 5: *Visualizations for Weak Target Detection: (a)-(d) Scene reconstruction by the baseline with no difference in reflectivity, 2×, 3× and 4× difference in reflectivity. (a)-(d) Scene reconstruction by the RIFT with no difference in reflectivity, 2 ×, 3× and 4× difference in reflectivity.*

such scenarios can only identify a general area of reflectivity or may even ignore the weaker object entirely.

In contrast, the RIFT model provides sufficient expressiveness to resolve the two reflectors, even when there is a fourfold difference in reflectivity between them. Although the reconstructed reflectivity of the weaker object is diminished, its accurate localization demonstrates the value of integrating the physical process into the Implicit Neural Representation (INR) in different modalities.

5 Conclusions and Future Works

In this paper, we introduced the Radon Implicit Field Transform (RIFT) workflow, which integrates an INR with a traditional forward model for radar signals to reconstruct scenes from radar data. Compared to traditional inverse models, RIFT achieves superior scene reconstruction across all experiments and enhances interpolation of unseen viewpoints in certain cases, all while utilizing significantly less data. These results indicate that RIFT effectively addresses the high cost of data acquisition in SAR problems by reconstructing scenes with reduced data requirements.

To assess the performance of RIFT-type models, we introduced customized error metrics for reconstruction and unseen viewpoint interpolation. The m-SSIM empirically aligns with our visual evaluations. However, since RIFT employs a neural network to model scene properties—in contrast to the Kaczmarz-based least square inversion of the forward radar model with well-established convergence properties—it may experience numerical stability issues. Consequently, there is one instance where the RIFT model underperforms in unseen viewpoint interpolation. As illustrated in Figure 6(g), the RIFT model occasionally fails due to convergence to local minima during optimization or vanishing gradients, highlighting the need for further investigation and customized optimization methods.

Beyond addressing sampling costs, we believe this work lays a cornerstone for research into the representation of INRs in less-explored data modalities. INRs show promise in reducing data acquisition costs for a wide range of active sensing problems (ASPs) with well-defined forward models.

To fully realize the potential of RIFT, we require datasets of real-world scenes and corresponding radar signals. We acknowledge the necessity of continued research into RIFT models to bridge them with real-world radar sensing applications, such as compact high-resolution mapping for autonomous vehicles or robotic navigation. The current problem we investigate involves a pair of forward and inverse problems. To further solidify the RIFT model, the immediate next step is learning through different radar forward models, such as those provided by commercial finite element simulators, may be necessary.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 40–49. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/achlioptas18a.html`.

Faiza Ali, Georg Bauer, and Martin Vossiek. A rotating synthetic aperture radar imaging concept for robot navigation. *IEEE Transactions on Microwave Theory and Techniques*, 62(7):1545–1553, 2014. doi: 10.1109/TMTT.2014.2323013.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In *Proceedings of the 2016 Conference on International Conference on Machine Learning*, pp. 477–485. JMLR.org, 2016.

Yang Bai, Tongyuan Zou, Shujia Ye, Zhenqiang Qin, Guoming Gao, and Yanfeng Gu. Weak target detection in high-resolution remote sensing images by combining super-resolution and deformable fpn. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 292–295, 2020. doi: 10.1109/IGARSS39084.2020.9323260. URL `https://ieeexplore.ieee.org/abstract/document/9323260`.

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. URL `https://arxiv.org/abs/2103.13415`.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. URL `https://arxiv.org/abs/2111.12077`.

Stephen Berger. Spectrum congestion - is it a technical problem? In *2014 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM)*, pp. 1–1, 2014. doi: 10.1109/USNC-URSI-NRSM.2014.6928004.

Igal Bilik, Oren Longman, Shahar Villeval, and Joseph Tabrikian. The rise of radar for autonomous vehicles: Signal processing solutions and future research directions. *IEEE Signal Processing Magazine*, 36(5):20–31, 2019. doi: 10.1109/MSP.2019.2926573. URL `https://ieeexplore.ieee.org/abstract/document/8828025`.

Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. URL `https://arxiv.org/abs/1604.00449`.

Giorgio Franceschetti and Riccardo Lanari. *Synthetic Aperture Radar Processing*, pp. 35–38. Springer US, Boston, MA, 1999. doi: 10.1007/0-306-47633-9. URL `https://link.springer.com/content/pdf/10.1007/0-306-47633-9.pdf`.

Gui Gao. Statistical modeling of sar images: A survey. *Sensors*, 10(1):775–795, 2010. ISSN 1424-8220. doi: 10.3390/s100100775. URL `https://www.mdpi.com/1424-8220/10/1/775`.

S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14326–14335, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.01408. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01408`.

R. C. Gonzales and R. E. Woods. *Digital Image Processing*, pp. 354. Prentice Hall, 3rd edition, 2008.

Biao Hou, Xingzhong Chen, and Licheng Jiao. Multilayer cfar detection of ship targets in very high resolution sar images. *IEEE Geoscience and Remote Sensing Letters*, 12(4): 811–815, 2015. doi: 10.1109/LGRS.2014.2362955. URL https://ieeexplore.ieee.org/abstract/document/6942175?casa_token=1HcehsC1EYwAAAAA:6Hx1LhfPW04Hjkp1a7HBphYN9TvcclzHlcwz24-4jdpymrgF8gdnPpzdjcNoKjR6Vnru1Zxk.

Abhishek Javali, Jagrati Gupta, and Anindita Sahoo. A review on synthetic aperture radar for earth remote sensing: Challenges and opportunities. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 596–601, 2021. doi: 10.1109/ICESC51422.2021.9532910. URL https://ieeexplore.ieee.org/document/9532910.

Stefan Kaczmarz. Approximate solution of systems of linear equations. *International Journal of Control*, 57(6):1269–1271, 1993. doi: 10.1080/00207179308934446. URL https://doi.org/10.1080/00207179308934446.

Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, pp. 386–402, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01266-3. doi: 10.1007/978-3-030-01267-0_23. URL https://doi.org/10.1007/978-3-030-01267-0_23.

Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pp. 31–42, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237199. URL https://doi.org/10.1145/237170.237199.

Chun Li, Kaina Zhang, Junzhou Zhang, and Ge Yang. A novel method for high-speed weak target detection and tracking. In *Proceedings of the 2024 8th International Conference on Digital Signal Processing*, ICDSP '24, pp. 61–66, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709029. doi: 10.1145/3653876.3653884. URL https://doi.org/10.1145/3653876.3653884.

Chao Liu, Mohammad Tayeb Al Qaseer, and Reza Zoughi. Influence of antenna pattern on synthetic aperture radar resolution for nde applications. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021. doi: 10.1109/TIM.2020.3026122. URL https://ieeexplore.ieee.org/document/9204828.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk9oneA5t7.

Andrew L. Maas, Andrea D. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 118–121. IEEE, 2013. doi: 10.1109/ICASSP.2013.6638947. URL https://doi.org/10.1109/ICASSP.2013.6638947.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. URL https://arxiv.org/abs/1812.03828.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020. URL https://arxiv.org/abs/2003.08934.

Antonio Moccia and Alfredo Renga. Spatial resolution of bistatic synthetic aperture radar: Impact of acquisition geometry on imaging performance. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3487–3503, 2011. doi: 10.1109/TGRS.2011.2115250. URL https://ieeexplore.ieee.org/abstract/document/5729807.

Vishal Monga, Eric Mason, Ilker Bayram, and Birsen Yazici. *Optimization Methods for Synthetic Aperture Radar Imaging*, pp. 71–98. Springer International, 2018. URL https://link.springer.com/chapter/10.1007/978-3-319-61609-4_4.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, 2013. doi: 10.1109/MGRS.2013.2248301.

NASA. What is sar. https://www.earthdata.nasa.gov/learn/backgrounders/what-is-sar, 2023. Accessed: 2024-09-27.

Clifford J Nolan and Margaret Cheney. Synthetic aperture inversion. *Inverse Problems*, 18(1): 221, jan 2002. doi: 10.1088/0266-5611/18/1/315. URL https://dx.doi.org/10.1088/0266-5611/18/1/315.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Mark A. Richards. *Fundamentals of Radar Signal Processing*. McGraw-Hill Education, New York, 3rd edition, 2022. ISBN 9781260468717. URL https://www.accessengineeringlibrary.com/content/book/9781260468717. Citing Chapter 1: Preliminaries (pp. 1–43), Chapter 3: Radar Waveforms (pp. 117–163), and Chapter 8: Space-Time Adaptive Processing (pp. 499–583).

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

J. Schou, H. Skriver, A.A. Nielsen, and K. Conradsen. Cfar edge detector for polarimetric sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 41(1):20–32, 2003. doi: 10.1109/TGRS.2002.808063.

Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

A. Lee Swindlehurst, Brian D. Jeffs, Gonzalo Seco-Granados, and Jian Li. Chapter 20 - applications of array signal processing. In Abdelhak M. Zoubir, Mats Viberg, Rama Chellappa, and Sergios Theodoridis (eds.), *Academic Press Library in Signal Processing: Volume 3*, volume 3 of *Academic Press Library in Signal Processing*, pp. 859–953. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-12-411597-2.00020-5. URL https://www.sciencedirect.com/science/article/pii/B9780124115972000205.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. ISSN 2167-8359. doi: 10.7717/peerj.453. URL https://doi.org/10.7717/peerj.453.

B.D. Van Veen and K.M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988. doi: 10.1109/53.665.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL `https://ieeexplore.ieee.org/document/1284395`.

Kaichao You, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. How does learning rate decay help modern neural networks? *arXiv preprint arXiv:2305.12345*, 2023. URL `https://arxiv.org/abs/2305.12345`.

A Additional Results

A.1 Additional Scene Reconstruction Results and Details

In this section, we provide the detail of the data of model performances summarized in Section 4. The Table 3 and 4 are for the sphere and pyramid scene discussed in Section 4.1.1.

Due to the nature of radar signal, there are cases when the signal viewing the same scene from a different angle can be apart by orders of magnitude. Hence, there are cases when RIFT training go into local minima or the gradient vanishes, for example, Figure 6 (g). The engineering detail for overcoming the numerical issues are in Appendix B.4 and B.5, but for the sake of completeness, we include the failed experiments in the tables[4].

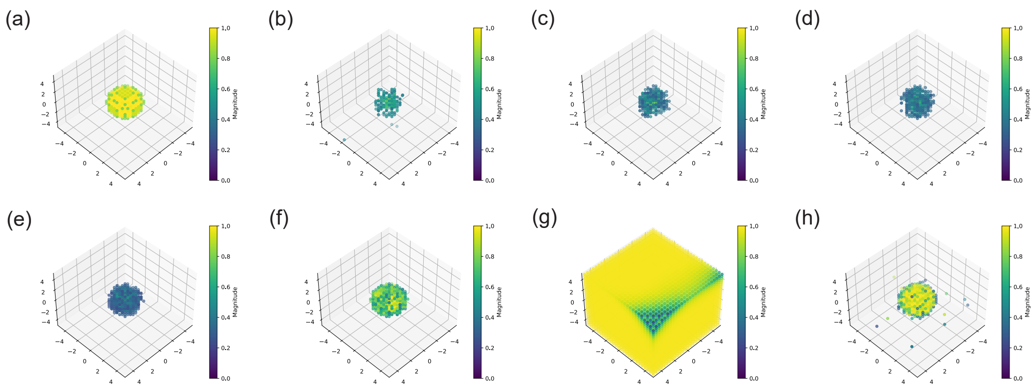

Figure 6: *Visualizations of the "sphere" scene (a): Ground truth of a sphere of radius 2m represented with a granularity of 0.2m. (b)-(e): Scene reconstruction by the baseline with 100, 200, 500, and 1000 viewpoints, respectively. (f): Scene reconstruction by RIFT(N) with 100 viewpoints, respectively. The RIFT-100 scores 188.5% higher in m-SSIM, 11.60% higher in m-COS, and 29.38% higher in t-IoU. However, the p-RMSE lags behind that of LS-1000 by 64.56%. (g),(h): Scene reconstructions by RIFT(N) and RIFT(S) with 1000 viewpoints. The detailed results are presented in Table 3.*

Table 3: Simple Scene Reconstruction Result for Sphere Data in Section 4.1

Model	m-SSIM	m-COS	t-IoU	p-RMSE
LS-100	0.0762	0.5986	0.0897	0.0151
LS-200	0.1007	0.7090	0.1741	0.0150
LS-500	0.1813	0.8412	0.2130	0.0148
LS-1000	0.2890	0.8886	0.2736	0.0146
RIFT-100(N)	**0.8343**	**0.9917**	0.3540	0.0187
RIFT-100(S)	0.2858	0.9473	0.3412	0.0145
RIFT-1000(N)†	0.0018	0.1893	0.0893	0.0214
RIFT-1000(S)	0.6002	0.9744	**0.3726**	**0.0145**

The "mini highway" scene is also in the comprises of a series of "streetlights" positioned 4.4 meters from the y-axis of the scene and uniformly spaced by 3.2 meters from one another. All "streetlights" are 1.8mmeters tall. There are "fences" placed 4m from the y-axis and right on the y-axis. The height is 2.0 meters. There is a 2.4 meters by 0.8 meters by 1.6meters "car" placed about 3.8 meters away from the origin of the scene. The Table 5 presents the comparison between the RIFT model and the baseline. This experiment is the only case we noticed a conspicuous disadvantage of the viewpoint interpolation by the RIFT model.

From the results above, we confirm that in all cases we presented, the RIFT models reconstructs the scene better with significantly less data. In most cases, the RIFT models interpolates the unseen

[4]In this work, we use †in the tables to denote the failed case we present.

Table 4: Simple Scene Reconstruction Result for Pyramid Data in Section 4.1

Model	m-SSIM	m-COS	t-IoU	p-RMSE
LS-100	0.1163	0.5584	0.0576	0.0151
LS-200	0.1444	0.7012	0.1740	0.0149
LS-500	0.3065	0.8158	0.1257	0.0147
LS-1000	0.4258	0.8689	0.1822	**0.0146**
RIFT(N)-100	0.8926	0.9840	**0.2549**	0.0185
RIFT(S)-100†	0.0266	0.1774	0.0581	0.0159
RIFT(N)-1000	**0.9513**	**0.9845**	0.2537	0.0186
RIFT(S)-1000	0.6385	0.9271	0.2730	0.0147

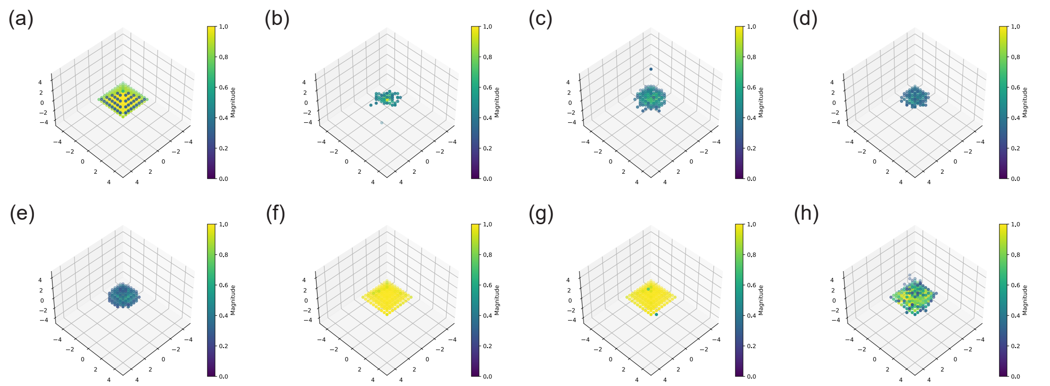

Figure 7: *Visualizations for the "Pyramid" scene(a): Ground truth of a pyramid of cubical base of 2m long and 2m tall represented with a granularity of 0.2m. (b)-(e): Scene reconstruction by the baseline with 100, 200, 500, and 1000 viewpoints, respectively. (f): Scene reconstruction by RIFT(N) with 100 viewpoints, respectively. The RIFT-100 scores 109.60% higher in m-SSIM, 13.25% higher in m-COS, and 39.41% higher in t-IoU. However, the p-RMSE lags behind that of LS-1000 by 59.77%. (g),(h): Scene reconstructions by RIFT(N) and RIFT(S) with 1000 viewpoints. The detailed results are presented in Table 4.*

viewpoints better with less data. Overall, we demonstrate the potential of the neural representations, with a relatively simple model configuration, in the under-researched field. Due to its distinctive nature, radar signal processing may need further investigation on improvement of optimization techniques to prevent the instability (like the one shown in Figure 6 (g)) caused by its wide distribution in magnitude.

B Technical Details

In this section, we specify the engineering details in the experiments including details in radar signal processing that are pertinent to this work, the structure of the RIFT models, and the optimization details.

B.1 Radar Setup

In general, we model the radar system as consisting of two components: transmitters (TX) and receivers (RX). We denote the number of TX and RX antennas as $|TX|$ and $|RX|$, respectively. In this study, we set $|TX| = |RX| = 16$. The radar operates over a band of angular frequencies ω, uniformly sampled from $[\omega_{\text{Lo}}, \omega_{\text{Hi}}]$. Here, ω follows the standard definition in radar signal processing, where $\omega = 2\pi f$ and f is the corresponding frequency. In our synthetic radar simulations, we use 100 frequencies uniformly sampled from the range $[95, 105]$ GHz.

In our problem setup, we define a main three-dimensional coordinate system with its origin at the geometric center of the scene. The radar trajectories for the transmitters and receivers, denoted as

Table 5: Complicated Reconstruction Result for "Mini Highway" Scene Corresponding to Figure 8

Model	m-SSIM	m-COS	t-IoU	p-RMSE
LS-2500	0.1914	0.5725	0.0128	**0.0154**
RIFT(N)-1000	**0.7354**	0.8778	0.0854	0.0160
RIFT(S)-1000	0.6897	**0.9116**	**0.0913**	0.0160

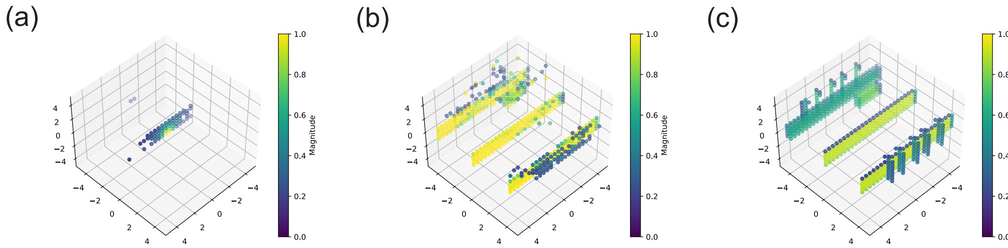

Figure 8: *Visualizations of the "mini highway" scene from Section 4: (a) Scene reconstructed by LS-2500 baseline model instance. (b) Scene reconstructed by RIFT-1000(N) instance. (c) Ground truth scene visualized with same granularity (defined in Section 2.2) as scene reconstruction. The scene reconstruction by RIFT model achieved 243.30% higher score in scene reconstruction and 1.50% better unseen viewpoint interpolation than baseline by only using 40% input data.*

$\gamma(s_{TX})$ and $\gamma(s_{RX})$, are determined by an auxiliary trajectory γ_{radar} that lies at a fixed distance $r_{\text{radar}} = 10$ meters from the origin. To ensure that the radar fully captures the synthetic scene, we assume that the scene's extent, $r_{\text{scene}} = 5$ meters, is smaller than r_{radar}. The TX and RX antennas are arranged such that the normal vector of the plane formed by the antennas always points toward the center of the main coordinate system.

The trajectory of γ_{radar} is defined as:

$$\gamma_{\text{radar}} := r_{\text{radar}} \left[\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta\right]^T$$

Here, θ and ϕ denote the azimuth and elevation angles relative to the scene center.

For the sake of resolution, we denote the speed of light as $c_0 = 299792458$ m/s and define the spacing between each individual antenna of the same kind $s = \frac{\lambda_{Max}}{2}$, where $\lambda_{max} = \frac{2\pi c_0}{\omega_{Lo}}$ is the longest wavelength which the radar system uses. In particular, the spacing we use for this work is 1.4276mm. Then for the m^{th} TX and the n^{th} RX, where $m \in [1, |Tx|]$ and $n \in [1, |Rx|]$, the trajectory $\gamma(s_{\text{TX}})$ and $\gamma(s_{\text{RX}})$ are:

$$\gamma(s_{\text{TX}}) = \gamma_{\text{radar}} s(m + \frac{1}{2}) \left[-\cos\theta\cos\phi, -\cos\theta\sin\phi, \sin\theta\right]^T$$

, and

$$\gamma(s_{\text{RX}}) = \gamma_{\text{radar}} s(n + \frac{1}{2}) \left[-\sin\phi, \cos\phi, 0\right]^T$$

respectively.

B.2 Radar Signal Formulation

In this work, the radar signal $\mathbf{S}$[5] is represented as a three-dimensional complex tensor $\mathbf{S} \in \mathbb{C}^{n_f \times |TX| \times |RX|}$. For each individual experiment, the dimensions of $\mathbf{S}$ are determined by the number of frequencies n_f, the number of transmitters $|TX|$, and the number of receivers $|RX|$. Based on the setup described in Appendix B.1, in this work we have $n_f = 100$, $|TX| = 16$, and $|RX| = 16$, so $\mathbf{S} \in \mathbb{C}^{100\times16\times16}$.

[5] In digital signal processing (DSP), $\mathbf{S}$ is often denoted as the S-parameter because it characterizes the scattering properties of a scene.

B.3 Scattering and Attenuation Factor

The definition of $\boldsymbol{A}$ comes from the work by Nolan & Cheney (2002):

$$\boldsymbol{A}(\omega, s, \boldsymbol{x}) = \frac{\omega^2 p(\omega) j_s(\omega(\widehat{x - \Gamma(s)}), \Gamma(s)) j_r(\omega(\widehat{x - \Gamma(s)}), \Gamma(s)) m(s)}{4\pi^2 |x - \Gamma(s)|^2}$$

The $\Gamma(s)$ term is the surface which the traces of radar antenna form. The j_s and j_r are Fourier transforms of current density of the radar, which is constant when we fix a radar pattern. The waveform $p(\omega)$ is waveform which we assume to be constant in Appendix B.1. The $m(s)$ term is a taper function which is also constant when we fix the radar. All terms are unity up to a normalization with no loss of information except $|x - \Gamma(s)|^2$.

There are two cases to discuss here: the first being the radar is far enough from the scene, which is often denoted as *far-field* in ASP, the the second being the radar is close to the scene. The difference between the two is that in the far-field case, the difference between the R_b of different combination of TX and RX pair is not significant as compared to the distance which the wave travels. The converse holds true for the *near-field* case.

Consequently, in order to not lose the information from R_b, in the case of near-field, the calculation $\frac{1}{R_b^2(s)}$ must be executed before normalization. For far-field, since the $\boldsymbol{x}$ dependence are all constant, the term $\boldsymbol{A}$ is absorbed by the normalization.

That is to say, the GRT of near-field case (which we denote as $\mathcal{F}_{NF}$ below) and far-field case (which we denote as $\mathcal{F}_{FF}$ below) are different where:

$$\mathcal{F}_{NF}[\rho] \approx \int_{\mathbb{R}^3} \frac{1}{R_b^2(s)} e^{j(k(\omega) R_b(\boldsymbol{x}))} \rho(\boldsymbol{x}) e^{\Phi(\rho(\boldsymbol{x}))} d\boldsymbol{x}$$

$$\mathcal{F}_{FF}[\rho] \approx \int_{\mathbb{R}^3} e^{j(k(\omega) R_b(\boldsymbol{x}))} \rho(\boldsymbol{x}) e^{\Phi(\rho(\boldsymbol{x}))} d\boldsymbol{x}$$

Note that in $\mathcal{F}_{FF}$, all $R_b^2 \approx ||\gamma_{radar}||$ for different combinations of TX/RX pairs, and the term is absorbed by normalization. From the radar setup in Appendix B.1, we use $\mathcal{F}$ as a shorthand $\mathcal{F}_{NF}$ for this study unless noted otherwise.

B.4 Model Structure

The two configurations of the RIFT models share the same structure shown in Figure 9. Both models have 3 dimensional input of the radar array center position γ_{radar}. The output of the models are 2 dimensional, which are real and imaginary part of learned scene reflectivity $\hat{\rho}(\boldsymbol{x})$. The hidden size of all hidden layers are 64.

The primary difference between the RIFT(N) and RIFT(S) models arises from the definitions of their units and nonlinearities. In the RIFT(N) models, each unit consists of a Linear layer, followed by a nonlinearity σ, and then a LayerNorm layer. The nonlinearities σ and τ are the LeakyReLU function (Maas et al., 2013) and the hyperbolic tangent function, respectively. In contrast, the RIFT(S) models have a simpler structure: each unit is a single Linear layer, and all nonlinearities σ and τ are sine functions.

As previously mentioned, the optimization process faces challenges due to the dynamic range of radar signals. Among all the different multilayer perceptron (MLP) structures we experimented with, the INR architectures used in RIFT(N) and RIFT(S) models were empirically found to perform the best. These structures optimize effectively despite frequent vanishing gradients and convergence to local minima.

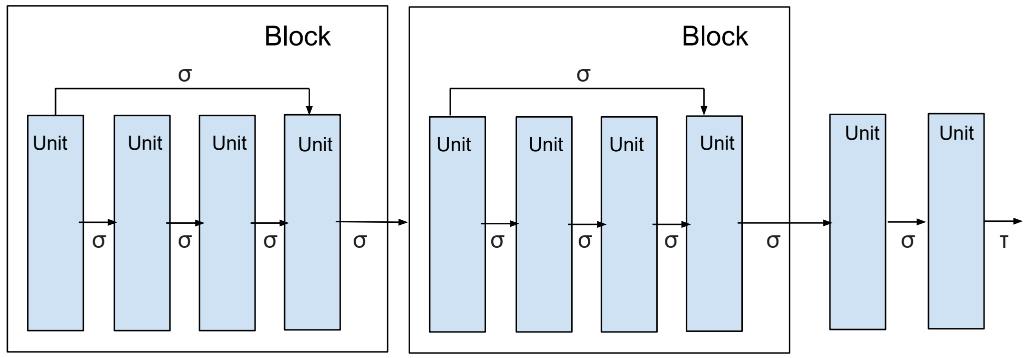

Figure 9: *The common architecture of two different configurations of RIFT models.*

B.5 Optimization Details

In practice, we observe that the optimization problem:

$$\arg\min_{\Theta} ||\mathbf{S} - \mathcal{F}[\hat{\rho}_\theta(\boldsymbol{x})]||_2$$

often faces challenges due to vanishing gradients. We speculate that since PyTorch's autodifferentiation Paszke et al. (2019) implements numerical calculations of differences in function values, the high dynamic range of radar signals can lead to computations involving very small numbers. Additionally, these small numbers are scattered across all frequencies and combinations of TX and RX, which further ill-conditions the loss landscape. To avoid multiplication of small numbers that could compromise numerical stability during training, we instead minimize the L_1 norm:

$$\arg\min_{\Theta} ||\mathbf{S} - \mathcal{F}[\hat{\rho}_\theta(\boldsymbol{x})]||_1$$

Our results support this approach.

For all experiments presented in this paper, we used the AdamW optimizer (Loshchilov & Hutter, 2018) with an initial learning rate of 10^{-2} and a weight decay rate of 10^{-2}. Optimization was set to cease after 500 epochs, with learning rate annealing (You et al., 2023) by halving every 100 epochs. Throughout this work, we handle the magnitude and phase of radar signals separately. We take the results with the lowest loss during the 500 epochs as the final result. During training, we process the real and imaginary parts of the loss function separately and assign them different weights because the magnitude of the radar signal spans the radar's dynamic range, whereas the phase is confined to $[0, 2\pi]$. In all experiments presented, the weight assigned to the phase term in the loss function is typically several thousand times greater than that of the magnitude term.

For the baseline models, all least square solution are solved with 100 iterations of block-KaczmarzKaczmarz (1993) algorithm.

B.6 Further Discussion is Metrics

The mathematical intuition behind separating SSIM for magnitude and phase stems from the formulation of the INR (see Section 3.2) and forward radar signal synthesis. In radar signal synthesis, we assign a reflectivity function $\rho(\mathbf{x})$ to the scene, as discussed in Section 2.2. For simplicity, when the scene is not occupied by a particular object, the reflectivity is set to zero. The subset $\mathbf{x}_0$, where $\rho(\mathbf{x}_0) = 0$, corresponds to regions of the scene without objects, while the subset $\mathbf{x} \setminus \mathbf{x}_0$ represents the parts of the scene occupied by objects.

By comparing the geometry of the occupied regions $\mathbf{x} \setminus \mathbf{x}_0$ with the ground truth geometry, we can assess how well the INR learns the scene, even at angles where reflections from the scene are weak. In all the experiments presented in this work, we select 100 unseen viewpoints and calculate the p-RMSE of all radar signals across these viewpoints.

The threshold Intersection-over-Union (tIoU) metric is inspired by the constant false alarm rate (CFAR) used in radar signal processing. In radar imaging, persistent background noise is present in radar signals, and various methods have been developed to reduce the influence of this constant false alarm (Schou et al., 2003; Hou et al., 2015). In this work, to address CFAR, we introduce a threshold to the magnitude of the learned reflectivity $\hat{\rho}(\mathbf{x})$ by assuming an apparent SNR, and discard all values below this threshold. We then calculate the Intersection-over-Union (IoU) to assess how much of the scene has been learned, especially in the reconstruction of a single object.

Due to the unavoidable noise level, tIoU values are often low because the presumed SNR cutoff must be held constant when comparing reconstructions from different methods, and it cannot be optimal for all methods simultaneously. However, tIoU can be considered a robustness measure for the model, since under the same presumed threshold, a higher tIoU value indicates better alignment between the ground truth scene and the reconstructed scene. The less noise present in the model's inference, the more of the inference surpasses the presumed SNR threshold, resulting in a higher tIoU.

For the calculation of tIoU and figure generation, we apply the thresholding of the presumed SNR at a fixed value for a fixed number of viewpoints. The thresholding only affects the results of tIoU and visualization. The m-SSIM calculations are conducted by slicing the 3D scene on a fixed 2D plane, and the result for a scene is the average m-SSIM of all slices.

For the unseen viewpoint interpolation, we measure the Mean Squared Error (MSE) of the phase of the radar signal, which we denote as p-RMSE. We discard the magnitude in the MSE calculation due to the nature of radar signals. Given a scene, reflections from certain viewpoints can differ by orders of magnitude if the reflectivity $\rho(\mathbf{x})$ is anisotropic. To test the generalizability of the model, we aim to reduce the impact of the signal magnitude, focusing instead on the phase information, which is necessary for the correct coherent addition of radar signals in SAR imaging.