CaRaFe: Camera-Radar Radiance Fields for Scene Reconstruction

DAVID BORTS, ETH Zürich, Switzerland
JULIAN OST, Princeton University, USA and Torc Robotics, USA
SHAMIK BASU, INRIA, France
TIM BRÖDERMANN, ETH Zürich, Switzerland
ANDREA RAMAZZINA, Mercedes Benz, Germany and Technical University of Munich, Germany
CHRISTOS SAKARIDIS, ETH Zürich, Switzerland
MARIO BIJELIC, Princeton University, USA and Torc Robotics, USA
FELIX HEIDE, Princeton University, USA and Torc Robotics, USA

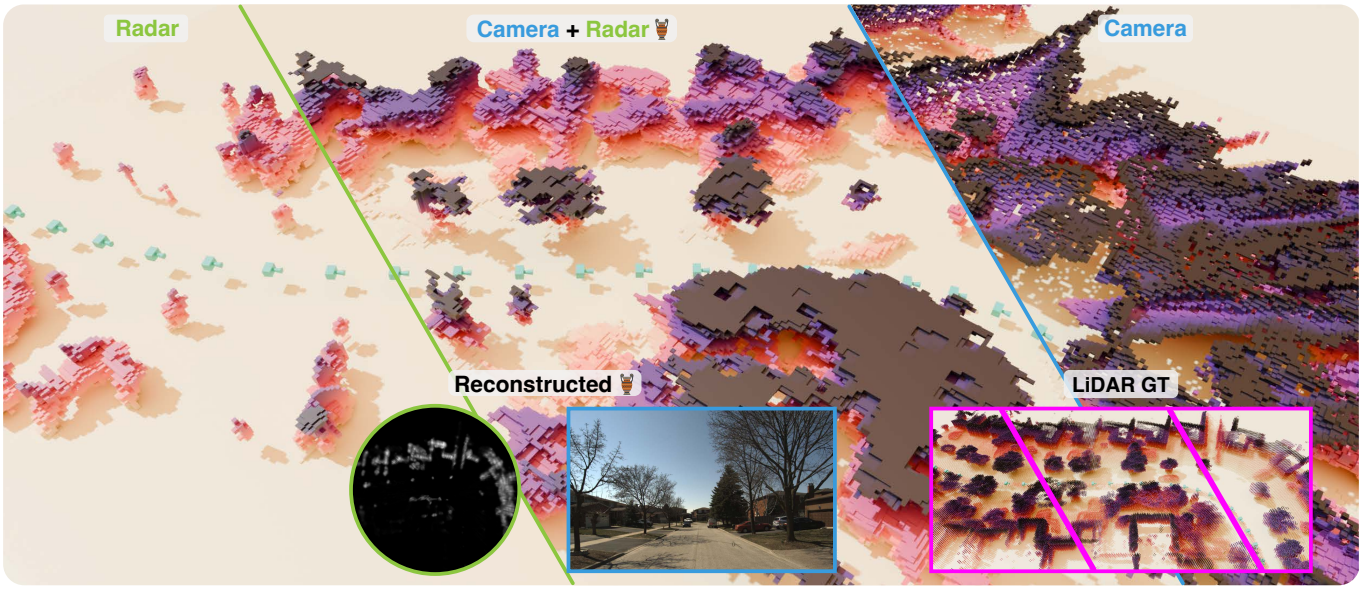


Fig. 1. For the same driving scene, we show a radar-only reconstruction (left), our joint camera–radar CaRaFe model (center), and a camera-only Gaussian-splatting reconstruction (right). The radar-only model fails to recover accurate elevation and misses many targets due to inconsistent radar reflectance. The camera-only model suffers from scale ambiguity, smearing occupancy radially and warping scene geometry. CaRaFe yields a sharper, more geometrically consistent 3D reconstruction by fusing both modalities: it inherits radar's metric depth accuracy while retaining the camera's high-frequency spatial detail.

Radar neural reconstruction methods have recently achieved robust 3D scene occupancy from radar measurements alone, as they provide metric depth and are insensitive to adverse weather and low light. However, while these methods can recover some 3D geometry, their input radar data mixes information across elevation into a 2D range-azimuth measurement. This fundamentally limits their elevation resolution, especially in automotive scenes with limited vertical baselines. Camera images offer the opposite tradeoff: they contain strong, high resolution cues for object elevation but struggle with accurate depth and in adverse conditions.

We propose CaRaFe, a method that leverages the complementary strengths of camera and radar for 3D reconstruction in challenging urban settings. CaRaFe employs a single multi-modal neural field, relying on conventional novel view synthesis as a supervision signal. Radar supervision provides a valuable geometric constraint to camera rendering that reduces shape-radiance ambiguity, while camera supervision allows for more accurate object elevation disambiguation and fewer missing structures in regions weakly observed by radar. We validate CaRaFe across diverse in-the-wild driving scenes, demonstrating favorable reconstruction quality over both radar and camera methods. Code for this paper is available at light.princeton.edu/carafe.

CCS Concepts: • **Computing methodologies** → **Active vision**.

Additional Key Words and Phrases: radar, neural fields, neural rendering

ACM Reference Format:
David Borts, Julian Ost, Shamik Basu, Tim Brödermann, Andrea Ramazzina, Christos Sakaridis, Mario Bijelic, and Felix Heide. 2026. CaRaFe: Camera-Radar Radiance Fields for Scene Reconstruction. *ACM Trans. Graph.* 1, 1 (May 2026), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM 1557-7368/2026/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large-scale outdoor scene reconstruction is essential for autonomous vehicles and robotics. Accurate scene geometry supports downstream understanding, localization, and navigation. The majority of work in this area [Chen et al. 2025; Guo et al. 2023; Rematas et al. 2022] relies on camera or LiDAR to recover scene geometry. However, recent advances in inverse rendering [Borts et al. 2024; Huang et al. 2024; Kung et al. 2025] have made radar a powerful alternative for reliable geometry reconstruction. These methods couple 3D data structures with an explicit model of radar sensing to recover dense 3D occupancy, even for large outdoor scenes. This breakthrough was enabled by training on full range-azimuth or range-Doppler radar measurements, leveraging the higher-resolution signal otherwise absent from conventional sparse thresholded point clouds.

Automotive radars typically operate at mm-wavelengths – three orders of magnitude larger than LiDAR – allowing for a robustness in their measurements not seen in optical modalities. This is especially true for adverse weather, like rain or fog, where the water particles are smaller than the wavelength of the radar signal and introduce minimal backscatter. Moreover, as an active sensor, radar directly provides metric depth and can function in low-light environments where passive sensors fail. Radar inverse rendering methods inherit these strengths, recovering consistent scene scales and maintaining performance in adverse weather and lighting.

However, most automotive radars are 2D frequency-modulated continuous-wave (FMCW) sensors, which mix signals across elevations into a flattened, range-azimuth scan. This limits the ability of current radar-based scene reconstruction methods to recover object elevation. The ambiguity is exacerbated in driving scenes, where elevation diversity across measurements is limited. Some approaches [Kung et al. 2025] prioritize robustness and novel view synthesis quality, resulting in often "collimated" reconstructions with limited elevation variance and tall objects. Another challenge in radar-based reconstruction is the prevalence of unique and powerful sensor noise and artifacts not seen in camera and LiDAR. FMCW radar measurements are prone to widespread speckle noise and ghost targets from strong reflections or multi-path effects [Kopp et al. 2021; Kraus et al. 2021; Kung et al. 2025], and objects are blurred by spectral leakage and bloom in the signal processing chain.

In this work, we propose a novel method for radar neural reconstruction that overcomes these limitations. We integrate camera into our radar neural representation and optimization, leveraging the rich spatial information and orthogonal plane of projection of camera images to constrain object elevations and improve geometric accuracy in our reconstructions. Informed by the insight that radar and optical volumetric extinction are comparable at centimeter scales and in urban environments, we introduce a novel joint camera-radar radiance field. Our model combines the complementary signals from both modalities into a shared geometric representation without compromising the adverse weather robustness of radar or high resolution of camera. Camera and radar share the same positional features, volume density, and geometric features, from which we perform both passive and active volume rendering to separately reconstruct input measurements from both sensors. We also introduce additional depth losses to supervise geometry with either dense depth cues from pre-trained camera monocular depth estimation models, or with reliable sparse depth cues from radar measurements. We make the following contributions:

- We propose a novel multi-modal inverse rendering method that recovers dense geometry from camera and radar measurements in unbounded outdoor scenes.
- We introduce a joint camera-radar radiance field with shared multi-modal volumetric density and geometric features, conditioning separate sensor-specific reflectance heads.
- We validate our method on the Boreas and Radar Fields datasets, and confirm across all tested scenes favorable performance over both radar and camera-only methods.

2 Related Work

Radar Sensing. has become a core cue for perception in robotics, maritime operations [Cheng et al. 2021], and autonomous vehicles [Bijelic et al. 2020; Hwang et al. 2022; Wang et al. 2023a]. Its millimeter-wave operation yields robustness to rain, fog, and other weather particles, complementing optical sensors [Bijelic et al. 2020; Hwang et al. 2022]. Recent work shows that radar can support dense scene understanding—depth estimation [Lin et al. 2020], semantic segmentation [Ouaknine et al. 2021; Zhang et al. 2023], scene flow [Ding et al. 2023], and reliable object detection [Bijelic et al. 2020; Hwang et al. 2022; Kim et al. 2023; Li et al. 2022]—and even enable non-line-of-sight perception in the wild [Scheiner et al. 2020]. Cross-modal fusion with cameras [Ding et al. 2023; Lin et al. 2020] and LiDAR [Hwang et al. 2022; Li et al. 2022] further mitigates radar sparsity and specularity, improving accuracy and range. Progress has been accelerated by public datasets that expose raw and processed radar signals with unified benchmarks [Brödermann et al. 2024; Caesar et al. 2020; Meyer and Kuschk 2019; Rebut et al. 2022].

Neural Rendering. has matured into a powerful tool for reconstructing 3D structure from posed observations, delivering both novel views and detailed geometry. Radiance-field models [Barron et al. 2021; Chen et al. 2022; Mildenhall et al. 2020; Müller et al. 2022] realize this by treating scenes as continuous volumetric functions and optimizing them via volumetric rendering. Encodings span implicit coordinates [Barron et al. 2021, 2022; Mildenhall et al. 2020; Zhang et al. 2021], dense grids [Chen et al. 2022; Fridovich-Keil et al. 2022; Yu et al. 2021], and hybrid data structures [Barron et al. 2023; Müller et al. 2022; Tancik et al. 2023], with numerous accelerations for training and inference [Barron et al. 2023; Chen et al. 2022; Müller et al. 2022; Yu et al. 2021]. Extensions to large, outdoor environments highlight persistent challenges under narrow-baseline, trajectory-aligned data common in automotive setups [Barron et al. 2022; Guo et al. 2023; Kundu et al. 2022; Liu et al. 2023; Ost et al. 2022; Ramazzina et al. 2023; Tancik et al. 2022; Turki et al. 2023; Wang et al. 2023b; Yang et al. 2023b; Zhang et al. 2020].

Scaling to street-scale environments introduces view-coverage and parallax limitations from single-trajectory captures [Guo et al. 2023; Kundu et al. 2022; Liu et al. 2023; Ost et al. 2022; Wang et al. 2023b; Yang et al. 2023b], often mitigated by auxiliary LiDAR [Guo et al. 2023; Hess et al. 2025; Ost et al. 2022; Rematas et al. 2022; Tonderski et al. 2024; Turki et al. 2023], predicted depth [Deng et al. 2022; Guo et al. 2023; Roessle et al. 2022], flow [Meuleman et al.

2023; Turki et al. 2023], or semantics [Kundu et al. 2022; Turki et al. 2023; Wang et al. 2023b]. While these methods utilize LiDAR and camera data, we explore radar-camera data as a robust alternative.

Neural Scene Reconstruction for Active Sensing. has emerged in parallel, proposing unique adaptations for sensors beyond conventional cameras. This interest has been especially present in specific domains, such as LiDAR [Huang et al. 2023; Zhou et al. 2025] or Radar [Borts et al. 2024; Kung et al. 2025] for automotive scene reconstruction, maritime sonar reconstruction [Qu et al. 2024; Sethuraman et al. 2025], or thermal imaging with infrared sensors [Ye et al. 2024]. Active sensor measurements are exposed to far stronger sensor and environment noise and influenced by complex multi-path effects, requiring high-fidelity sensor models coupled with physics-inspired rendering. LiDAR and Sonar have seen an increasing amount of work on multi-modal approaches that combine RGB and active sensor reconstructions, either through intense physics-inspired modeling efforts [Sethuraman et al. 2025; Turki et al. 2025] or deep sensor feature decoders [Hess et al. 2025; Yang et al. 2023b], while the integration of radar into a *multi-modal* reconstruction pipeline still remains underexplored. Note that existing neural reconstruction methods [Borts et al. 2024; Kung et al. 2025] use radar as the sole supervision modality.

3 Background

We review the principles of volume rendering for passive sensors (Sec. 3.1) and outline a signal formation model for FMCW radar (Sec.3.2). These form the basis of our proposed joint representation (Sec. 4.1) and active volume rendering for radar (Sec. 4.2).

3.1 Multi-view Volumetric Rendering

A large body of existing work in neural rendering [Barron et al. 2023; Müller et al. 2022; Rematas et al. 2022] reconstructs scene information by photometric reconstruction using volumetric rendering. The scene is represented as a volume with density $\sigma(\mathbf{x})$ that defines the differential density of reflective particles at any scene point $\mathbf{x} \in \mathbb{R}^3$. For any ray $\mathbf{r}(z) = \mathbf{o} + z\mathbf{d}$ with origin $\mathbf{o}$, direction $\mathbf{d}$, and bounds $\left[z_n, z_f\right)$, we can therefore define the accumulated transmittance $T(z)$ up to any distance z along that ray,

$$T(z) = \exp\left(-\int_{z_n}^{z} \sigma(\mathbf{r}(t))dt\right), \tag{1}$$

which is the proportion of radiation that propagates through the interval $[\mathbf{r}(z_n), \mathbf{r}(z))$ without hitting any particles. An appearance field $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ is a vector-valued function that defines the view-dependent camera-observable outgoing radiance, specific to a scene's illumination and camera. This allows us to compute the expected color of that same camera ray with the volume rendering integral

$$\mathbf{C}(\mathbf{r}) = \int_{z_n}^{z_f} T(z)\sigma(\mathbf{r}(z))\mathbf{c}(\mathbf{r}(z), \mathbf{d})dz. \tag{2}$$

Neural Radiance Fields [Mildenhall et al. 2020] use a discrete approximation of this integral by partitioning the ray into N piecewise-constant bins and employing quadrature. Now, any ray passing through bin i with density σ_i and width δ_i has probabilities $\alpha_i = 1 - \exp(-\sigma_i\delta_i)$ of terminating (opacity) and $T_i = 1 - \alpha_i$ of not terminating within that bin. This reduces to alpha compositing as

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} \mathbf{w}_i\mathbf{c}_i, \;\; \mathbf{w}_i = \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \tag{3}$$

and is broadly used to optimize radiance fields [Müller et al. 2022].

3.2 FMCW Radar Rendering

Frequency-modulated continuous-wave (FMCW) radars are active sensors that continuously emit radio waves and measure time-of-flight by correlating the original transmitted signal with the reflected signal [Jankiraman 2018; Richards et al. 2010; Skolnik 2001]. They modulate the frequency of the transmitted wave with a periodic function, such that phase shifts induced by travel time directly correspond to a frequency difference with the transmitted wave.

Correlating the spectrograms of the transmitted and received waves recovers the "beat frequency", or instantaneous frequency difference. Taking its fast Fourier transform (FFT) yields the final measurement: reflected power as a function of phase offset and therefore range. Any frequency bin in the FFT corresponds to a metric range bin along the emitted beam, and its power directly depends on the reflected power from that range bin. This measurement is a power-range profile that mixes multiple targets at different ranges, all irradiated by the same diverging beam. Recent work [Borts et al. 2024] leverages exactly these measurements for dense 3D scene reconstruction, as their cm-level range resolution offers more information than a conventional thresholded peak.

The received radar power $P(R, \boldsymbol{\omega})$ from a point target at range R with view direction $\boldsymbol{\omega}$ can be modeled using the radar equation [Jankiraman 2018; Richards et al. 2010; Skolnik 2001] as

$$P(R, \boldsymbol{\omega}) = \frac{P_t G^2(\boldsymbol{\omega})\lambda^2\Sigma(R, \boldsymbol{\omega})}{(4\pi)^3 R^4}, \tag{4}$$

where P_t is the transmitted power, $G(\boldsymbol{\omega})$ is the angle-dependent antenna gain, λ is the transmitted wavelength, and $\Sigma(R, \boldsymbol{\omega})$ is the radar cross-section (RCS) of the target.

RCS is a measure of how strongly a target scatters electromagnetic energy back toward a radar sensor [Knott et al. 2004; Skolnik 2001]. It is a compound function of target geometry, material properties, as well as incident and reflected angles. For a point target, RCS can be factorized as $\Sigma = \sigma\eta$, where σ is the volume density at that point as defined in Sec. 3.1, and η is its general radar scattering efficiency. Note η is the view-dependent product of target reflectivity and directivity, while σ is solely a geometric quantity.

While Eq. 4 pertains to point targets, radar beams have a non-negligible divergence. Therefore, to compute the total received radar power $P(R)$ at range R, we must integrate across the opening angle of the beam, Ω_b. The total RCS $\Sigma(R)$ at range R can be expressed as

$$\Sigma(R) = \int_{\Omega_b} \sigma(\mathbf{x})\eta(\mathbf{x}, \boldsymbol{\omega})R^2 d\boldsymbol{\omega}, \quad \mathbf{x} = \mathbf{r}(R), \tag{5}$$

where $\boldsymbol{\omega} = (\theta, \phi)$ is a ray direction within the beam, and $\mathbf{r}(R) = \mathbf{o} + R\boldsymbol{\omega}$ is a ray starting from the radar frame's origin $\mathbf{o}$, pointing in direction $\boldsymbol{\omega}$, and evaluated at range R. We combine Eq. 5 with Eq. 4 to derive

$$P(R) = \frac{P_t\lambda^2}{(4\pi)^3 R^2} \int_{\Omega_b} G^2(\boldsymbol{\omega})\sigma(\mathbf{x})\eta(\mathbf{x}, \boldsymbol{\omega})d\boldsymbol{\omega}, \tag{6}$$

where we weigh each ray's contribution to the received power by the antenna gain in the corresponding direction.

4 CaRaFe

We describe our joint camera-radar test-time optimization method for 3D scene reconstruction. Specifically, we propose a shared neural field representation (Sec. 4.1), active sensor volume rendering formulation for radar (Sec. 4.2), and depth supervision losses for both camera and radar-based depth signals (Sec. 4.3).

4.1 Joint Neural Scene Representation

Camera and radar measurements offer complementary geometric signals. Broadly deployed range-azimuth radars provide an accurate measure of depth and azimuth, but do not constrain the elevations of targets. Camera images have rich, high-resolution elevation information for localizing and classifying targets, but ambiguous metric depth for inferring their range. Existing radar reconstruction works [Borts et al. 2024; Huang et al. 2024; Kung et al. 2025] do not leverage camera data, while recent multi-modal reconstruction [Rafidashti et al. 2025] learn separate per-modality feature decoders and do not recover a single scene geometry. CaRaFe integrates both modalities into the same *joint radiance field* and recovers a single shared neural density that satisfies the geometric constraints of both sensors.

With separate density fields, structure can diverge into implausible geometries that overfit to their respective modalities. Instead, optimizing a single density field encourages more accurate scene reconstructions, as it must satisfy measurements from both sensors simultaneously. By supervising CaRaFe with both camera and radar measurement reconstruction losses, we preserve the robustness of both sensors; radar geometry persists even when camera measurements fail in fog or low light, while the high spatial resolution of camera images is not lost by the coarse radar depth signal.

We define an implicit shared neural field to reconstruct both input radar measurements and camera images of the same scene. The neural field $\mathbf{f} : (\mathbf{x}, \mathbf{d}) \mapsto (\sigma, \eta, \mathbf{c})$ is a function of scene position $\mathbf{x} \in \mathbb{R}^3$ and view direction $\mathbf{d} \in \mathbb{R}^3$, which maps them to shared volumetric density σ, radar-specific scattering efficiency η, and camera-specific color $\mathbf{c}$. Input positions $\mathbf{x}$ are first mapped to positional features $\chi \in \mathbb{R}^{32}$ with a multi-resolution hashgrid, while input directions $\mathbf{d}$ are projected onto the first four spherical harmonics coefficients $\mathcal{D} \in \mathbb{R}^4$. The neural field is parametrized with three MLPs as

$$\begin{aligned} f_\sigma &: \chi \mapsto (\sigma, \mathbf{f}_{\text{geo}}), \\ f_\eta &: (\mathcal{D}, \mathbf{f}_{\text{geo}}) \mapsto (\eta_0, \xi, \kappa), \\ f_{\mathbf{c}} &: (\mathcal{D}, \mathbf{f}_{\text{geo}}) \mapsto \mathbf{c}, \end{aligned}$$

which comprise a shared geometry network f_σ and separate radiance heads for camera, $f_{\mathbf{c}}$, and radar, f_η. Here, the field component f_σ maps positional features to density and extracts geometric features $\mathbf{f}_{\text{geo}} \in \mathbb{R}^{16}$ that condition both per-sensor view-dependent radiance heads. The field component $f_{\mathbf{c}}$ directly regresses color, while f_η parametrizes a von Mises Fisher distribution [Fisher 1953] on the sphere with amplitude η_0, sharpness factor κ, and mean scattering direction $\boldsymbol{\xi} \in \mathbb{R}^3$. From these parameters, we can recover η at point $\mathbf{x}$ as a function of $\mathbf{d}$ via

$$\eta(\mathbf{d}) = \frac{\eta_0 \kappa}{4\pi \sinh(\kappa)} \exp(\kappa(\boldsymbol{\xi} \cdot (-\mathbf{d}))). \tag{7}$$

We choose this parametrization for η because it aligns well with the specular lobes observed in radar sensing. Moreover, it constrains the angular frequency of f_η to mitigate shape-radiance ambiguity, preventing geometric errors from being absorbed into η.

4.2 Volume Rendering for FMCW Radar

The radar rendering Eq. (6) assumes that there are no other objects between the sensor and target. This results in inaccurate scene reconstructions and affects the model's ability to accurately resolve elevation, as it could "cheat" by placing density in front of a target and occlude it without reducing its reflected power. Therefore, we introduce a forward volumetric rendering model to derive the occlusion-aware received radar power, $P'(R)$, that is

$$P'(R) = \frac{P_t \lambda^2}{(4\pi)^3 R^2} \int_{\Omega_b} G^2(\boldsymbol{\omega}) T^2(R) \sigma(\mathbf{r}(R)) \eta(\mathbf{r}(R), \boldsymbol{\omega}) d\boldsymbol{\omega}, \tag{8}$$

where we scale the RCS by the squared transmittance $T^2(R)$, which is the proportion of transmitted power that reaches the target at distance z and returns back to the radar. Note that, unlike the camera volumetric rendering in Eq. (3), radar is an active sensor and radiation attenuates both during travel to and from a target.

If we discretize each radar ray into N_b piecewise-homogeneous range bins, then we can compute $T^2(R_i)\sigma_i$ at any bin i with center at range R_i with

$$T^2(R_i)\sigma_i = \mathbf{w}_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j), \tag{9}$$

where we define the two-way probability β_i of terminating within bin i as

$$\beta_i = 1 - \exp(-2\sigma_i \delta_i), \tag{10}$$

where δ_i is the bin width. Note that σ is the same shared volume density as in Sec. 3.1. We plug the discrete formulation of (9) into the continuous version of (8) and additionally approximate the beam integral by taking a discrete Monte Carlo estimate. This yields our occlusion-aware discrete radar power model

$$\hat{P}_{R_i} \propto \frac{\Omega_b}{R_i^2 S} \sum_{s=1}^{S} G^2(\omega_s) \mathbf{w}_{R_i} \eta(z_s, \omega_s), \tag{11}$$

where we sample S rays in the opening angle of the beam at ω_s.

Spectral Leakage Modeling. FMCW radars sample reflected beat signals over a finite time interval T, which multiplies the continuous-time signal $s(t)$ by a window $w(t)$ and thus convolves its spectrum with the window spectrum $W(f)$. This *spectral leakage* causes small, cm-scale objects to spread across multiple FFT range bins.

For the Navtech scanning radar used in Boreas [Burnett et al. 2023] and Radar Fields [Borts et al. 2024], the range FFT internally employs Hamming windowing, which we model as a broadening step in our radar volume rendering. Let $P_{\text{ideal}}(r, \mathbf{d})$ be the ideal power-range profile for beam direction $\mathbf{d}$. We approximate the combined effect of finite integration, windowing, and electronics by a one-dimensional radial blur kernel $h(r)$ and render

$$P_{\text{blur}}(r, \mathbf{d}) = \int h(r - r')\, P_{\text{ideal}}(r', \mathbf{d})\, \mathrm{d}r'. \tag{12}$$

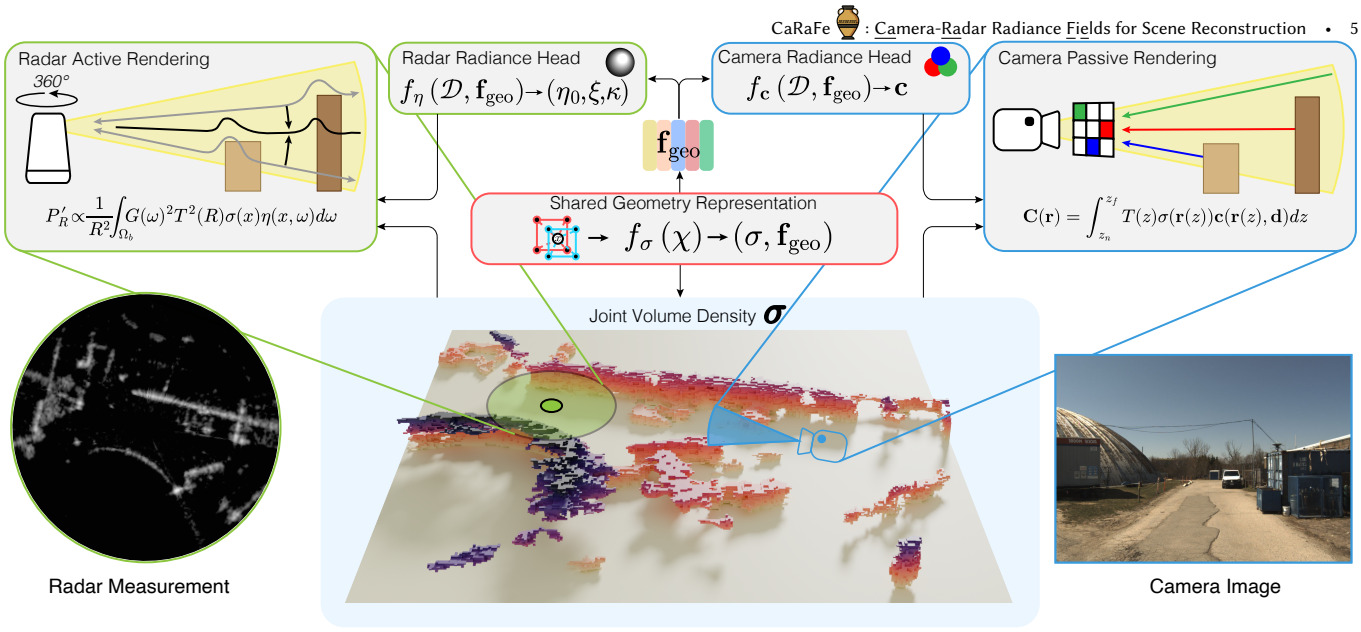


Fig. 2. CaRaFe learns a joint neural scene representation with a geometric embedding $\mathbf{f}_{\text{geo}}$. From this embedding and the viewing direction $\mathcal{D}$, a radar head $f_\eta : \mathbf{f}_{\text{geo}} \mapsto (\eta_0, \zeta, \kappa)$ predicts radar radiance parameters, while a camera head $f_c : \mathbf{f}_{\text{geo}} \mapsto \mathbf{c}$ predicts RGB colors. For rendering, we cast rays along radar beams and through camera pixels, sample points along each ray, and integrate the predicted scene response to reconstruct radar measurements (left) and camera images (right), which both serve as supervision signals during training.

Following [Kung et al. 2025], we parameterize h as a Gaussian and experiment with learning its scale parameter σ, but find that setting σ according to manufacturer specs performs equally well.

4.3 Training

We train CaRaFe by jointly reconstructing radar and camera measurements. We sample rays from both modalities, query the occupancy field f_σ at points along each ray, predict pixel colors with passive volume rendering according to Eq. 3, and calculate radar range bin power with active volume rendering via Eq. 11. Training minimizes the loss:

$$\mathcal{L} = \lambda_C \mathcal{L}_C + \lambda_R \mathcal{L}_R + \lambda_D \mathcal{L}_D + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{13}$$

which includes a camera reconstruction loss $\mathcal{L}_C$, radar reconstruction loss $\mathcal{L}_R$, depth supervision loss $\mathcal{L}_D$, and regularization $\mathcal{L}_{\text{reg}}$.

Reconstruction Losses. Our camera loss $\mathcal{L}_C$ is the vanilla mean squared error between predicted and ground-truth pixel colors

$$\mathcal{L}_C = \frac{1}{|C|} \sum_{\mathbf{r} \in C} \left\| \hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2, \tag{14}$$

where C is the set of all camera rays in a batch. The radar reconstruction loss penalizes deviations of predicted power across distinct range bins b of each beam through

$$\mathcal{L}_R = \frac{1}{B_R N_b} \sum_{i=1}^{B_R} \sum_{b=1}^{N_b} \left\| \hat{P}^i_{R_b} - P^i_{R_b} \right\|_2^2, \tag{15}$$

where B_R is the number of sampled radar beams per batch and N_b is the number of range bins per beam, each at range R_b.

Depth Supervision Losses. CaRaFe can be supervised with dense depth signals $\mathcal{L}_D = \mathcal{L}_{D_1} + \mathcal{L}_{D_2}$, such as from camera depth foundation models $\mathcal{L}_{D_1}$ and radar data $\mathcal{L}_{D_2}$. We draw inspiration from [Rematas et al. 2021] and leverage a free-space density loss

$$\mathcal{L}_{D_1} = \frac{1}{|C|N} \sum_{\mathbf{r} \in C} \mathcal{D}(\mathbf{r}) \sum_{i=1}^{N} \sigma_i \cdot \mathbb{1}\left(z_i \notin [\mathbf{z} - \epsilon, \mathbf{z} + \epsilon]\right), \tag{16}$$

where z_i is the depth of sample bin i, $\mathbf{z}$ is the estimated depth for that supervised ray, $\mathcal{D}(\mathbf{r})$ is a per-ray binary depth supervision mask, and ϵ is a tunable window hyperparameter.

Intuitively, $\mathcal{L}_{D_1}$ suppresses floaters and radar ghost targets by discouraging CaRaFe from placing density outside an ϵ-neighborhood of the input dense depth. We increase ϵ to accommodate the known noise and frame-to-frame inconsistencies of monocular depth predictions, which are not ground truth and exhibit local distortions. Crucially, we exploit radar's metric scale to make this supervision robust: before applying $\mathcal{L}_{D_1}$, we pre-compute a validity mask $\mathcal{D}(\mathbf{r})$ over monocular depth maps, filtering out and rescaling camera depth estimates that are warped or inconsistent in scale. Details of the mask computation are provided in the Supplementary Material.

For using 2D radar depth as a supervision signal in $\mathcal{L}_{D_2}$, defining a meaningful 3D depth loss is non-trivial. Standard 2D automotive radars integrate returns over elevation, i.e., a detection at range R and azimuth θ may originate from any elevation within the radar beam's vertical opening angle. If we naively enforce all rays whose azimuth is close to θ to terminate at R, the resulting gradients diffuse across the full vertical FOV, leading to columnar artifacts and reconstructions that collapse the third dimension. Moreover, due to occlusions and multi-path effects, many beams exhibit multiple peaks at different ranges, each corresponding to echos at different heights, which further exacerbates the depth-elevation ambiguity.

To address these challenges, we propose a soft-matching radar depth loss. We first extract a sparse set of candidate depth targets by peak-finding on each beam's range–power spectrum, allowing multiple peaks per beam. Given a radar training ray, we compute its one-way NeRF weights $\{w_i\}_{i=1}^{N}$ along the ray using Eq. (3). We score each of the P candidate depth targets by measuring how well the ray's weight profile aligns with a Gaussian kernel $\mathcal{K}(z) = \mathcal{N}(z; z_p, (\epsilon_k/3)^2)$ centered at the target depth z_p, yielding scores $S = \{s_1, \ldots, s_P\}$. These scores are converted into mixture weights $L = \{l_1, \ldots, l_P\}$ via a temperature-controlled softmax with temperature τ (annealed over training). Finally, we compute a free-space loss over each target and aggregate them using L as weights,

$$\mathcal{L}_{D_2} = \frac{1}{B_R} \sum^{B_R} \sum_{i=1}^{P} l_i \sum_j \sigma_j \cdot \mathbb{1}\left(z_j \notin [\mathbf{p}_i - \epsilon, \mathbf{p}_i + \epsilon]\right), \quad (17)$$

For each depth target $\mathbf{p}_i$, we penalize density outside an ϵ neighborhood, analogous to Eq. 16. But rather than enforcing a single termination depth per ray, we allow multiple candidate per-ray and compute a free-space loss for each. We weight these losses by how well each candidate aligns with the ray's current density profile, resulting in a softer and more stable radar depth supervision signal.

Density Regularization. Neural volume rendering can produce scattered or multi-modal density along a ray, especially in large-scale driving scenes with narrow baselines and underconstrained geometry. Such diffuse densities lead to ambiguous depth, floating structures, and blurred occupancy boundaries. To mitigate these artifacts, we introduce a regularization term $\mathcal{L}_{\text{reg}}$ that encourages each ray to form a unimodal density peak, yielding crisper geometry. We compute one-way volume rendering weights $\mathbf{w}_i$ (Eq. 3) for each sampled ray, and treat them like a probability distribution,

$$\mathcal{L}_{\text{reg}} = \frac{1}{|C| + |\mathcal{R}|} \sum_{\mathbf{r} \in C \cup \mathcal{R}} \mathcal{M}(\mathbf{r}) \sum_i -\mathbf{w}_i \log(\mathbf{w}_i), \quad (18)$$

where $\mathcal{M}(\mathbf{r})$ is a mask that is 1 only where $\sum_i \mathbf{w}_i > \epsilon_{\text{reg}}$ for some small ϵ_{reg} we choose. This mask prevents the regularization term from distorting empty scene rays.

4.4 Implementation Details

We optimize CaRaFe on outdoor automotive sequences of 50-70 radar frames captured at 4Hz while driving at speeds ranging from 10-15 km/h, for a total of about 34 to 73 meters of driving distance. For each sequence, we include the first 200 camera frames before the first radar frame, as well as the first 50 camera frames after the last radar frame; this helps minimize regions of our scene that receive only radar coverage, since our camera is forward-facing. For our radar measurements, we train on the first 1079 range bins, as the vast majority of targets lie within this radius and returned signals diminish strongly due to the quadratic attenuation. We omit the first 75 range bins, as these are usually dominated by reflections from the ego vehicle, leaving a total of 1024 supervised range bins per-beam. We optimize CaRaFe in 20k optimizer steps on a single A100 GPU, using the Adam optimizer. In camera space, we set $|C| = 8192$ rays per-batch, while for radar we set $S = 20$ super-sampled rays per-beam, and sample 60 beams per-batch, for a total $B_R = 1200$ rays. We use fully fused MLP kernels from [Müller et al. 2022] and train in half precision, allowing for faster training and inference.

5 Experiments

In this Section, we evaluate CaRaFe qualitatively in Fig. 4, 5, and quantitatively in Tab. 1, 2, 3 & Fig. 3. Specifically, we compare to two radar neural reconstruction methods [Borts et al. 2024; Kung et al. 2025] and four camera neural reconstruction methods [Chen et al. 2025; Yan et al. 2024; Yang et al. 2023a; Yu et al. 2024].

Experimental Protocol. To evaluate our approach, we follow the experimental protocol of [Borts et al. 2024; Kung et al. 2025] and select static scenes from both datasets. To assess geometry reconstruction, we measure the Relative Chamfer Distance (RCD) between the predicted occupancy representation and an accumulated LiDAR measurement from the dataset. The LiDAR point cloud is assembled by using the recorded poses to reproject all scans into a common reference frame. For novel view synthesis, we report PSNR and SSIM for both the reconstructed radar and camera measurements on a withheld split in Table 3. We withhold 10% of the samples per-scene.

5.1 Validation

Radar Fields [Borts et al. 2024] and RadarSplat [Kung et al. 2025] train only on radar, whereas the remaining baselines are camera-only. CaRaFe is the first inverse-rendering model to jointly leverage both modalities, which naturally creates a large performance gap to existing baselines since no fully comparable method yet exists. For camera novel view synthesis, CaRaFe surpasses all camera-only baselines, despite satisfying radar supervision. For radar novel view synthesis, CaRaFe outperforms the radar-only Radar Fields and RadarSplat, indicating that the camera branch provides a strong regularizing signal for reconstructing high-fidelity radar measurements. Overall, these results suggest the two modalities are complementary and interact constructively rather than interfering destructively.

In the boxed sections of Fig. 4 and 5, we highlight fine geometric details that are consistently missed by either camera or radar-only baselines. For example, the metallic nature of vehicles means that their radar reflectivity is extremely view-dependent - the radar baselines often fail to capture such objects. For example, rows 1 and 5 of Fig. 4 and row 1 in Fig. 5 show examples of radar baselines failing to reconstruct bollards, trees, and electrical transformers. Moreover, the radar baselines produce collimated scenes with inaccurate object elevation. This is especially true for RadarSplat [Kung et al. 2025], which trades elevation awareness for robustness, resulting in shorter objects like cars and shrubs being reconstructed at more than 4 meters above the ground. The camera baselines lose a robust notion of scale without active sensing, recovering dense scene geometries but with low certainty, as geometry is smeared out radially. This is most obvious in rows 1, 3 of Fig. 4 and 5 of Fig. 5, and is exacerbated in narrow baseline scenes where multiview consistency cues are weak. CaRaFe suffers minimally from both artifacts, maintaining metric scale while disambiguating elevation.

In Tables 1 & 2, CaRaFe consistently improves IoU and F-score over both radar-only and camera-only baselines, while still achieving the lowest RCD, indicating a closer match to the LiDAR geometry and recovering more targets with less spatial uncertainty.

Metric	Ours	EmerNeRF [Yang et al. 2023a]	Street Gaussians [Yan et al. 2024]	GOF [Yu et al. 2024]	PGSR [Chen et al. 2025]	Radar Fields [Borts et al. 2024]	RadarSplat [Kung et al. 2025]
Modality	*Both*	*Camera*	*Camera*	*Camera*	*Camera*	*Radar*	*Radar*
IoU (%) ↑	**19.15**	7.45	3.52	6.15	1.28	4.10	4.22
Precision (%) ↑	**24.53**	16.99	3.58	16.63	14.66	11.01	22.57
Recall (%) ↑	49.19	13.36	**66.68**	8.99	1.41	6.14	4.94
F-score ↑	**31.77**	13.66	6.80	11.47	2.52	7.88	8.11
RCD (full) ↓	**0.001**	0.002	0.004	0.010	0.012	0.003	0.003

Table 1. Geometric Occupancy Evaluation. We evaluate geometric accuracy and compare CaRaFe (Ours) to camera baselines EmerNeRF [Yang et al. 2023a], Street Gaussians [Yan et al. 2024], PGSR [Chen et al. 2025] and GOF [Yu et al. 2024], as well as to Radar baselines [Borts et al. 2024; Kung et al. 2025] on the Boreas Dataset [Burnett et al. 2023]. CaRaFe consistently outperforms all baselines by a substantial margin.

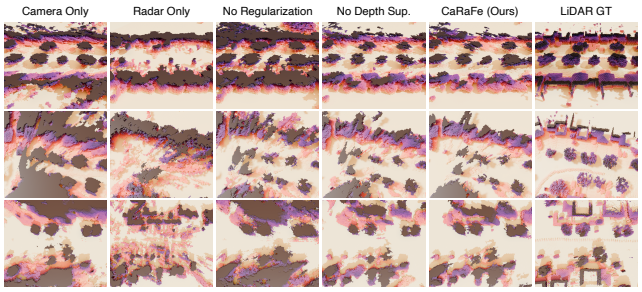
Metric	Ours	EmerNeRF [Yang et al. 2023a]	GOF [Yu et al. 2024]	PGSR [Chen et al. 2025]	RadarSplat [Kung et al. 2025]
IoU ↑	**14.13**	6.27	2.92	2.41	3.55
F-score ↑	**24.74**	5.66	4.71	11.76	6.87
RCD (full) ↓	**0.003**	0.004	0.005	0.006	0.006

Table 2. Additional Geometric Validation on the Radar Fields Dataset.

	Metric	Ours	GOF [Yu et al. 2024]	PGSR [Chen et al. 2025]	Radar Fields [Borts et al. 2024]	RadarSplat [Kung et al. 2025]
Camera	PSNR ↑	**33.33**	32.61	32.22	-	-
	SSIM ↑	**0.901**	0.885	0.875	-	-
Radar	PSNR ↑	**28.31**	-	-	27.53	28.05
	SSIM ↑	**0.956**	-	-	0.882	0.913

Table 3. Novel view synthesis results for all modalities. PGSR and GOF are camera-only methods; Radar Fields and RadarSplat are radar-only.

In Table 3, we validate that CaRaFe's improved geometry does not come at the expense of novel view synthesis quality. We report PSNR and SSIM for both camera and radar on held-out views. CaRaFe exceeds baseline methods for both camera and radar synthesis, while also outperforming them on all geometric metrics.



Variant	Radar	Camera	Reg.	Depth	IoU (%) ↑	F-score ↑	RCD ↓
Camera only	✗	✓	✗	✗	9.51	17.02	0.003
Radar only	✓	✗	✗	✗	10.10	18.12	0.002
No Regularization	✓	✓	✗	✓	18.83	31.24	0.002
No Depth Supervision	✓	✓	✓	✗	21.65	34.97	0.002
Full model (CaRaFe)	✓	✓	✓	✓	**23.02**	**36.71**	**0.001**

Fig. 3. Ablation study on the contributions of individual sensor inputs, regularization, and depth supervision for geometry reconstruction.

5.2 Ablation Study

To understand the contribution of each component of CaRaFe, we conduct an ablation study summarized in Table 3. We consider four variants: a radar-only model (Radar only), a camera-only model (Camera only), a joint model without density regularization (w/out regularization), and a joint model without any depth supervision (w/out depth sup.), and compare them to the full CaRaFe model.

We report three geometry-focused metrics: IoU, F-score, and Relative Chamfer Distance (RCD). Removing either modality (radar-only or camera-only) leads to a clear drop in IoU, F-score, and RCD, confirming that both sensors contribute complementary information: radar provides metrically accurate depth, while the camera enforces high-resolution structure. Disabling the density regularization term $\mathcal{L}_{\text{REG}}$ produces more diffuse density profiles and degrades all geometry metrics, highlighting the importance of encouraging single, sharper density peaks along each ray. Disabling the depth supervision likewise hurts geometric accuracy by increasing floaters.

Overall, the full CaRaFe configuration, combining radar and camera inputs with both regularization and depth supervision, achieves the best performance across all metrics. Notably, the gap between the full model and the depth/regularization ablations is smaller than the gap induced by removing an entire modality. This indicates that CaRaFe's gains are driven primarily by the core camera-radar fusion, and that the method still outperforms the same baselines even without depth supervision or density regularization.

6 Conclusion

We introduce CaRaFe, a novel joint camera–radar neural reconstruction method that fuses complementary camera and radar sensing cues for large-scale outdoor 3D scene recovery. By using radar as a depth prior and camera images as a high-resolution appearance and elevation cue, CaRaFe reduces shape–radiance ambiguities and overcomes the inherent elevation limitations of 2D radar-only reconstructions. Our proposed fusion outperforms *both* radar and RGB-based reconstruction methods in isolation. Specifically, the method achieves geometrically accurate reconstructions that closely match LiDAR ground truth, with errors on the order of 0.3m, enabling faithful recovery of static driving scenes at scale. In the future, we plan to generalize to dynamic scenes and self-supervise existing radar–camera perception models, such as [Singh et al. 2023] and [Hwang et al. 2022] with CaRaFe, allowing us to learn robust and fast feed-forward perception from multi-modal reconstruction.

References

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. In *ICCV*, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023.

Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

David Borts, Erich Liang, Tim Broedermann, Andrea Ramazzina, Stefanie Walz, Edoardo Palladin, Jipeng Sun, David Brueggemann, Christos Sakaridis, Luc Van Gool, Mario Bijelic, and Felix Heide. Radar fields: Frequency-space neural scene representations for fmcw radar. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery.

Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. MUSES: The Multi-Sensor Semantic Perception Dataset for driving under uncertainty. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Z Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. Boreas: A multi-season autonomous driving dataset. *The International Journal of Robotics Research*, 42(1-2):33–42, 2023.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022.

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6100–6111, 2025.

Yuwei Cheng, Hu Xu, and Yimin Liu. Robust small object detection on the water surface through fusion of camera and millimeter wave radar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15263–15272, 2021.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

Fangqiang Ding, Andras Palffy, Dariu M. Gavrila, and Chris Xiaoxuan Lu. Hidden gems: 4d radar scene flow learning using cross-modal supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9340–9349, 2023.

Ronald A. Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A*, 217(1130):295–305, 1953.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. StreetSurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023.

Georg Hess, Carl Lindström, Maryam Fatemi, Christoffer Petersson, and Lennart Svensson. SplatAD: Real-time lidar and camera rendering with 3D Gaussian splatting for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11982–11992, 2025.

Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural LiDAR fields for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Tianshu Huang, John Miller, Akarsh Prabhakara, Tao Jin, Tarana Laroia, Zico Kolter, and Anthony Rowe. DART: Implicit doppler tomography for radar novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty, Nicholas Armstrong-Crews, Tiffany Chen, and Dragomir Anguelov. CramNet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

M. Jankiraman. *FMCW Radar Design*. Artech House, 2018.

Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17615–17626, 2023.

E.F. Knott, J.F. Schaeffer, and M.T. Tulley. *Radar Cross Section*. Institution of Engineering and Technology, 2004.

Johannes Kopp, Dominik Kellner, Aldi Piroli, and Klaus Dietmayer. Fast rule-based clutter detection in automotive radar data, 2021.

Florian Kraus, Nicolas Scheiner, Werner Ritter, and Klaus Dietmayer. The radar ghost dataset – an evaluation of ghost objects in automotive radar data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 8570–8577. IEEE, 2021.

Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022.

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 (ICCV)*, 2025.

Yu-Jhe Li, Jinhyung Park, Matthew O'Toole, and Kris Kitani. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2022.

Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020.

Jeffrey Yunfan Liu, Yun Chen, Ze Yang, Jingkang Wang, Sivabalan Manivasagam, and Raquel Urtasun. Real-time neural rasterization for large scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8416–8427, 2023.

Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023.

Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th European Radar Conference (EuRAD)*, pages 129–132, 2019.

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. In *ECCV*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

Julian Ost, Issam Laradji, Alejandro Newell, Yuval Bahat, and Felix Heide. Neural point light fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15671–15680, 2021.

Ziyuan Qu, Omkar Vengurlekar, Mohamad Qadri, Kevin Zhang, Michael Kaess, Christopher Metzler, Suren Jayasuriya, and Adithya Pediredla. Z-splat: Z-axis gaussian splatting for camera-sonar fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Mahan Rafidashti, Ji Lan, Maryam Fatemi, Junsheng Fu, Lars Hammarstrand, and Lennart Svensson. Neuradar: Neural radiance fields for automotive radar point clouds, 2025.

Andrea Ramazzina, Mario Bijelic, Stefanie Walz, Alessandro Sanvito, Dominik Scheuble, and Felix Heide. Scatternerf: Seeing through fog with physically-based inverse neural rendering. *arXiv preprint arXiv:2305.02103*, 2023.

Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17000–17009, 2022.

Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields, 2021.

Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022.

Mark A. Richards, James A. Scheer, and William A. Holm. *Principles of Modern Radar: Basic principles*. The Institution of Engineering and Technology, 2010.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022.

Nicolas Scheiner, Florian Kraus, Fangyin Wei, Buu Phan, Fahim Mannan, Nils Appenrodt, Werner Ritter, Jurgen Dickmann, Klaus Dietmayer, Bernhard Sick, and Felix Heide. Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Advaith V Sethuraman, Max Rucker, Onur Bagoren, Pou-Chun Kung, Nibarkavi NB Amutha, and Katherine A Skinner. Sonarsplat: Novel view synthesis of imaging

sonar via gaussian splatting. *arXiv preprint arXiv:2504.00159*, 2025.

Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9275–9285, 2023.

M.I. Skolnik. *Introduction to Radar Systems*. McGraw-Hill, 2001.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14895–14904, 2024.

Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

Haithem Turki, Qi Wu, Xin Kang, Janick Martinez Esturo, Shengyu Huang, Ruilong Li, Zan Gojcic, and Riccardo de Lutio. Simuli: Real-time lidar and camera simulation with unscented transforms. *arXiv preprint arXiv:2510.12901*, 2025.

Yingjie Wang, Jiajun Deng, Yao Li, Jinshui Hu, Cong Liu, Yu Zhang, Jianmin Ji, Wanli Ouyang, and Yanyong Zhang. Bi-lrfusion: Bi-directional lidar-radar fusion for 3d dynamic object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13394–13403, 2023a.

Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. In *ECCV*, 2024.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision, 2023a.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023b.

Tianxiang Ye, Qi Wu, Junyuan Deng, Guoqing Liu, Liu Liu, Songpengcheng Xia, Liang Pang, Wenxian Yu, and Ling Pei. Thermal-nerf: Neural radiance fields from an infrared camera. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1046–1053. IEEE, 2024.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021.

Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Trans. Graph.*, 43(6), 2024.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

Liwen Zhang, Xinyan Zhang, Youcheng Zhang, Yufei Guo, Yuanpei Chen, Xuhui Huang, and Zhe Ma. Peakconv: Learning peak receptive field for radar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17577–17586, 2023.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021.

Chenxu Zhou, Lvchang Fu, Sida Peng, Yunzhi Yan, Zhanhua Zhang, Yong Chen, Jiazhi Xia, and Xiaowei Zhou. Lidar-rt: Gaussian-based ray tracing for dynamic lidar re-simulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1538–1548, 2025.

Fig. 4. Qualitative Assessment. We report qualitative results for CaRaFe and, from left to right, compare to the camera-only approaches GOF [Yu et al. 2024] and PGSR [Chen et al. 2025], as well as radar-only approaches [Borts et al. 2024] and [Kung et al. 2025] with sensor measurements on the left. In alternating rows we show reconstructed occupancy grids and rendered views from each method, for both modalities. We also show a camera-perspective render of the reconstructed CaRaFe grid. Boxes are used to highlight objects that are not recovered by either the camera or radar baselines, but are recovered by CaRaFe. Please rotate and zoom into digital version for details.

Fig. 5. Additional Qualitative Assessment. We report further qualitative results for CaRaFe and, from left to right, compare to the camera-only approaches GOF [Yu et al. 2024] and PGSR [Chen et al. 2025], as well as radar-only approaches [Borts et al. 2024] and [Kung et al. 2025] with sensor measurements on the left. In alternating rows we show reconstructed occupancy grids and rendered views from each method, for both modalities. We also show a camera-perspective render of the reconstructed CaRaFe grid. Boxes are used to highlight objects that are not recovered by either the camera or radar baselines, but are recovered by CaRaFe. Please rotate and zoom into digital version for details.