Gated Fields: Learning Scene Reconstruction from Gated Videos

Andrea Ramazzina[1*] Stefanie Walz[1,2*] Pragyan Dahal[3] Mario Bijelic[4,5] Felix Heide[4,5] [1]M

Abstract

Reconstructing outdoor 3D scenes from temporal observations is a challenge that recent work on neural fields has offered a new avenue for. However, existing methods that recover scene properties, such as geometry, appearance, or radiance, solely from RGB captures often fail when handling poorly-lit or texture-deficient regions. Similarly, recovering scenes with scanning LiDAR sensors is also difficult due to their low angular sampling rate which makes recovering expansive real-world scenes difficult. Tackling these gaps, we introduce Gated Fields – a neural scene reconstruction method that utilizes active gated video sequences. To this end, we propose a neural rendering approach that seamlessly incorporates time-gated capture and illumination. Our method exploits the intrinsic depth cues in the gated videos, achieving precise and dense geometry reconstruction irrespective of ambient illumination conditions. We validate the method across day and night scenarios and find that Gated Fields compares favorably to RGB and LiDAR reconstruction methods. Our code and datasets are available here[1].

1. Introduction

Large-scale outdoor scene reconstruction is essential for advancing autonomous robotics, drones, and driver-assistance systems, serving as the foundation for scene understanding, safe navigation, dataset generation and validation. Existing works in this domain [68, 76, 96] have typically adopted a two-step approach. Initially, they infer depth maps from different poses, utilizing time-of-flight sensors or RGB captures. Subsequently, these depth estimates are fused to produce a coherent 3D representation, using either classical methodologies [25] or learned-based representations [72]. In contrast, more recent studies [16, 65, 79] have proposed end-to-end strategies that bypass the estimation of local depth maps as an intermediate representation. Instead, they directly regress a Truncated Signed Distance Function (TSDF) [65, 79] or an occupancy volume [16]. A rapidly growing body of work on neural rendering [62, 92] offers not only geometrically-accurate scene reconstruction from posed RGB images but also to generate novel perspectives from unobserved angles. Hinging on implicit coordinate-based neural representations, RGB-based methods [7, 8, 55, 100] have been adapted to large open outdoor environments. A recent line of work [37, 71, 86, 93] includes LiDAR scans for auxiliary depth supervision and to improve scene reconstruction for urban environments. However, recovery based on the RGB images exhibit is fundamentally limited in the presence of low light [63, 91] or in the presence of scattering such as fog [49, 69].

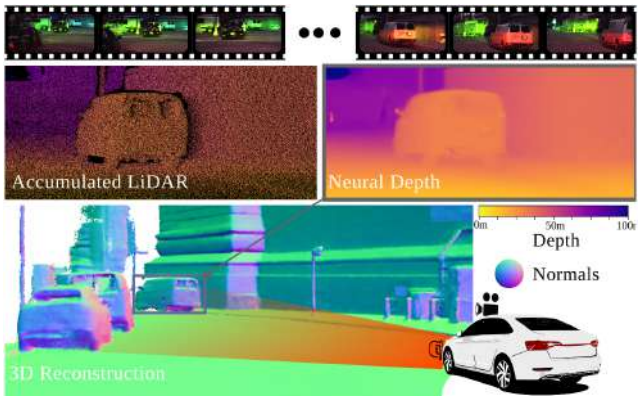

Figure 1. From a single video of gated captures (top-row), we reconstruct an accurate scene representation and render depth projections (mid-row, right) as accurate as LiDAR scans (mid-row, left), and we recover 3D geometry and normals (bottom-row).

A parallel direction of research investigates neural rendering techniques tailored to Time-of-Flight (ToF) sensors as opposed to the conventional RGB cameras. Existing methods [41, 84] tackle scene understanding with posed LiDAR scans, and have modeled the raw output from a single-photon LiDAR system [59] or continuous-wave ToF sensor [4] as additional depth supervision. However, all of these existing methods struggle with recovering large unbounded outdoor scenes: while continuous-wave ToF sensors [4] offer signal only in room-sized scenes, methods based on scanning LiDAR suffer from low angular sampling which mandates temporal aggregation. Specifically, even today's LiDAR sensors boasting 200 scan lines, lag drastically in resolution compared to current HDR cameras that offer two orders of magnitude higher vertical pixel counts nearing 10k and three orders of magnitude higher total resolution.

Addressing these challenges, our work explores scene reconstruction using active gated imaging. Gated imaging functions by integrating the transient response from a scene that has been flash-illuminated by a synchronized

*These authors contributed equally to this work.

[1] https://light.princeton.edu/gatedfields/

light source [19]. This imaging technique is robust to adverse weather conditions – temporal gating allows us to filter out backscatter – and provides signal in poorly-lit scenes [13]. Existing work has exploited this sensing modality to achieve state-of-the-art depth estimation [89, 90] as well as object detection [44]. In our approach, we train a neural field-based representation of the scene, concurrently learning its geometry, illumination, and material properties. This is accomplished by integrating the gated imaging formation model with a neural rendering framework, that jointly learns the associated gating parameters along with the scene reconstruction. By leveraging the implicit depth cues present in gated video captures, we are able to reconstruct a detailed 3D geometric model of the scene, as shown in Fig. 1. Compared to LiDAR-based approaches, our method offers distinct advantages, as LiDAR systems are inherently constrained by their resolution, necessitating additional time-multiplexed scene captures to aggregate points. This results in an extended acquisition process for LiDAR-based methods or, conversely, compromises the geometric detail of the final estimate. Specifically, for a fixed time acquisition budget the scene reconstruction from a LiDAR sensor is less supervised, although offering highly accurate depth information, the sensor yields data at a volume one order of magnitude less than that of a camera stereo pair. This disparity in data quantity means that while the LiDAR provides precise depth points, it is unable to provide dense and fine detailed predictions.

To assess our method, we captured a dataset of varied scenes at day and night conditions, using a vehicle test setup comprising of LiDAR, RGB and Gated sensors. We compare our method with feed-forward and 3D reconstruction methods, and assess its superiority in novel depth synthesis beating the next best method by 21.87% MAE, 3D reconstruction improving on the baseline by 11% IoU, and performs novel view synthesis with a PSNR of 32.28 dB.

In our work, we make the following contributions:

- We propose a novel neural rendering method and scene representation that is capable of reconstructing scene geometry and radiance from active gated camera videos.
- By modeling the gated image formation process and integrating into differentiable volume rendering, our approach is able to reconstruct and decompose both passive and active light transport components conditioned on the scene parameters in a physically accurate way.
- We validate our method on large outdoor scenes, captured across different scenarios in both day and night, achieving a reduction of MAE error in depth precision of 59.8% to the next best RGB+LiDAR method and 31.7% to the next best methods using gated images.

2. Related Work

Monocular and Stereo Depth Estimation Depth estimation tasks, from a single image [32, 36, 51, 54], from stereo image pairs [5, 22, 52, 95], or single/stereo images with a LiDAR scan [24, 40, 67, 82, 83, 94, 102] have been at the center of a large body of work. Single CMOS sensor-based depth estimation from RGB color images is inherently limited by scale ambiguity. Additional measurements from LiDAR [9] or ego-vehicle speed [36] can resolve this ambiguity at the cost of an additional sensor. Similarly, stereo methods rely on an additional camera sensor to resolve ambiguity by triangulating between two camera views [22]. Training approaches for learned depth estimation methods using intensity images cover both unsupervised methods [30, 32, 33, 36, 103], which harnesses multi-view geometry consistency, and supervised techniques [22, 27, 42, 45, 51, 54, 58, 60] relying primarily on multi-view datasets [45, 51, 60] or time-of-flight captures [22, 27, 42, 58]. In particular, LiDAR measurements have been proven as a ground-truth signal for depth supervision [2, 22, 27, 42, 54, 58]. Several methods [31, 87] have mitigated the sparsity and range limitations of scanning LiDAR by accumulating scans. However, adverse weather [14] can make LiDAR ground truth unreliable, and methods that rely on consistency between camera and LiDAR [24, 40, 67, 82, 83, 94, 102] suffer from degradations, including scan pattern artifacts and temporal distortions.

Time-gated Depth Sensing Time-of-Flight (ToF) sensors determine depth by emitting light into a scene and calculating the distance based on the round trip time of the light. Successful sensing methods can be categorized into three classes: correlation ToF cameras [38, 46, 48], pulsed ToF sensors [77], and gated imaging [34, 39]. Correlation ToF sensors[38, 46, 48] estimate depth by continuously flood-illuminating the scene and assessing the phase shift between the emitted and received light. While these sensors can provide high-resolution depth information, their use is primarily confined to indoor settings due to their susceptibility to external light interference. Pulsed ToF sensors [77] function by emitting light pulses toward specific scene points and measure the total travel time to estimate depth. Emitting collimated light, this approach is robust against ambient illumination and allows for outdoor depth measurements. However, its spatial resolution is fundamentally limited owing to its scanning illumination technique, and its efficacy can be compromised in adverse weather due to backscatter [11, 21, 43]. In contrast, gated cameras [12, 34, 39] capture light from a scene over brief intervals, essentially constraining the observable depth to specific range segments. The inherent gating mechanism of these cameras offers resistance to backscattering, and they allow for the recovery of detailed depth maps when using a large number of short

gates [3, 17, 18] Subsequent works have improved gated depth estimation with few gates by adopting Bayesian approaches [1, 75] or deep neural networks [35, 89, 90], and they achieve accurate gated depth estimation for dynamic outdoor scenes, even under challenging conditions. Recently, Gated Stereo [90] reached state-of-the-art results using a stereo-gated setup and self-supervision [89].

Neural Scene Reconstruction Recent research has amalgamated sets of single-sensor measurements to recover comprehensive scene representations. This synthesis has led to advancements in both novel-view generation [6, 23, 62, 64] and depth estimation [85] , with neural radiance field methods [6, 23, 62, 64] emerging as a pivotal approach for representing scenes as continuous volumetric fields of radiance. These methods combine this representation with volumetric rendering as a forward model in a test-time optimization approach. Specific representations that these methods explore include coordinate-based networks [6, 7, 62, 101], 3D voxel-grid representation [23, 28, 98], or hybrid approaches [8, 64, 81]. Subsequent works have extended this representation to large outdoor scenes [7, 100], and increased the efficiency at training and test time [8, 23, 64, 98]. Other works departed from radiance-based representation and explicitly learned the scene illumination, geometry, and material proprieties [15, 74, 101]. A particular challenge within the field is reconstructing large urban terrains based on imagery captured from vehicles [37, 47, 56, 66, 71, 80, 86, 93, 97], given that a significant portion of the scene is seen from only a narrow range of viewpoints. This problem is tackled by additional supervision cues from sparse LiDAR [37, 66, 71, 86], pre-estimated depths [26, 37, 73], optical flow [61, 86] and semantic segmentation [47, 86, 93]. Departing from RGB-based approaches, recent works have investigated neural reconstruction methods using ToF sensors [41, 59, 84]. Existing methods [41, 84, 99] learn a neural field from posed LiDAR scans, allowing for the synthesis of realistic LiDAR scans from novel views. Recently, Malik et al. [59] represent the time-resolved photon count acquired by a single-photon LiDAR system with a neural reconstruction method.

3. Gated Imaging Model

In this section, we provide a brief background on gated imaging as presented in [35] and introduce the proposed gated image formation model. Unlike prior work, this model incorporates shadow effects and features self-calibrated parameter learning.

A gated imaging system, illustrated in Fig. 2, utilizes a pulsed flood-light illumination source p with a synchronized imager that operates with a nanosecond (ns) gated exposure g that is delayed by ξ compared to the pulse. This allows us to capture only photons with round-trip times inside the gates, hence specific distance segments in the scene.

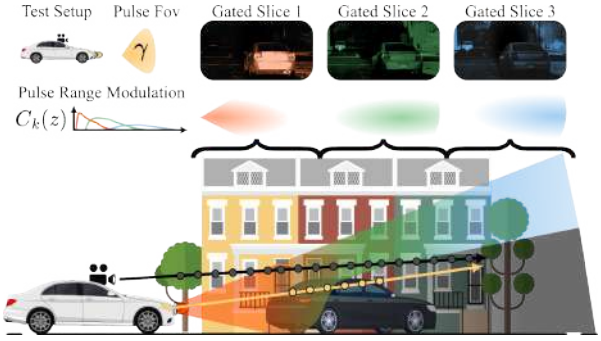

Figure 2. Gated Image Formation and Bi-Directional Sampling. Top-row: Our test vehicle is equipped with a synchronized stereo camera setup and illuminator that flood-lits the scene with a light pulse and FoV γ. Using different gating profiles $C(z)$, we capture three slices with intensity visualised here in red, green and blue. Illustrated in the middle row, the gating profiles describe pixel intensity for a point at sensor distance z. The first slice (red) accounts for close ranges, the green for mid-ranges, and the blue for far ranges. Bottom-row: we show the ray sampling employed in our method, based on a bidirectional sampling strategy. We cast the rays from the illuminator view to explore the occluded areas, while the rays casted from the camera integrate the reflected scene response. The shadowed areas are marked in gray.

We formalize this using so-called range intensity profiles $C_k(2z_c)$ given distance z_c from the camera, time t, and a parameter set k [35], that is

$$C_k(2z_c) = \int_{-\infty}^{\infty} g_k(t-\xi) p_k \left(t - \frac{2z_c}{c} \right) \beta(2z_c) dt, \quad (1)$$

where c is the speed of light and $\beta(\cdot)$ models the distance-dependent decay of the reflected light pulse. The resulting gated pixel value is

$$I_k(z_c) = \alpha \iota C_k(2z_c) + \Lambda + \mathcal{D}_k, \quad (2)$$

where Λ represents the passive ambient contribution, α is the scene reflection, ι the laser illumination, and $\mathcal{D}_k$ is an additive noise term.

This model assumes camera position $\mathbf{o}_c$ and the illuminator position $\mathbf{o}_i$ are collocated. To allow for non-collocated positions, we express the travel time as $z = z_c + z_i$, where z_i denotes the distance between the illuminator and the point on the surface impacted by the light beam, represented as $z_i = |\mathbf{x} - \mathbf{o}_i|_2$. Additionally, there may be areas visible to the camera that remain dark due to potential occlusions. Modeling shadow effects and attenuation due to incident angle ω results in the following image formation

$$I_k(z) = \alpha \iota \psi |\mathbf{n} \cdot \omega| C_k(z) + \Lambda + \mathcal{D}_k. \quad (3)$$

Here, $\psi \in [0, 1]$ serves as a shadow indicator for the pixel, and ω is the direction of the incident light at that point.

We extend this model to fit the range intensity profiles during optimization, thereby eliminating the need for their direct measurement. This approach overcomes potential calibration inaccuracies encountered in previous approaches [89, 90]. We model both the laser pulse p_k and the gate g_k as rectangular functions with durations $t_{l,k}$ and $t_{g,k}$, respectively. This simplification permits the analytical computation of the integral in Eq. (1), that is

$$\tilde{C}_k = \begin{cases} \frac{2z}{c} - \xi_k + t_{l,k} & \text{if } \xi_k - t_{l,k} < \frac{2z}{c} < \xi_k \\ t_{l,k} & \text{if } \xi_k < \frac{2z}{c} < \xi_k + t_{g,k} - t_{l,k} \\ -\frac{2z}{c} + \xi_k + t_{g,k} & \text{if } \xi_k + l_{g,k} - t_{l,k} < \frac{2z}{c} < \xi_k + t_{g,k} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

4. Gated Field

We reconstruct a scene by fitting a neural field representation to gated videos. We collect videos of three active gated slices $I_{k\in\{1,2,3\}}$ with different gating parameters and one passive slice I_P. We model active illumination by jointly estimating light and material properties, and separately represent the ambient light as a radiance field. The proposed reconstruction method relies on both photometric reconstruction cues and scene priors and is described in the following.

4.1. Neural Gated Fields

We describe the scene proprieties using two neural fields f_{Gp} and $f_{G\alpha}$, respectively representing the ambient light scattered in the scene and the reflection of the scene surfaces, conditioned on a spatial embedding χ. Moreover, the laser illumination contribution is represented by a physics-based model, while shadow effects are simulated through ray-tracing using the volumetric density field f_{Gd}.

Neural Ambient and Reflection Field We represent a scene as a neural field $f_G : \{\mathbf{x}, \mathbf{d}, \omega\} \rightarrow \{\sigma, \alpha, \Lambda, \mathbf{n}\}$ mapping each point in space $\mathbf{x}$ viewed from a direction $\mathbf{d}$ and laser direction ω to its volumetric density σ, normal vector $\mathbf{n}$, scene reflection α and the passive component Λ, that is

$$\begin{aligned} f_{Gp} &: \{\mathbf{d}, \chi\} \longrightarrow \{\Lambda\} && \text{Ambient Component} \\ f_{Gn} &: \{\mathbf{x}, \chi\} \longrightarrow \{\mathbf{n}\} && \text{Surface Normal} \\ f_{G\alpha} &: \{\mathbf{d}, \omega, \chi\} \rightarrow \{\alpha\} && \text{Surface Reflection} \\ \text{with } f_{Gd} &: \{\mathbf{x}\} \longrightarrow \{\sigma, \chi\} && \text{Vol. Density and Embedding} \end{aligned}$$

We condition here normal, ambient and reflectance on a volumetric embedding via the field $f_{Gd} : \{\mathbf{x}\} \rightarrow \{\sigma, \chi\}$ estimating the density σ and embedding χ. This embedding is being shared by the network branches to estimate the normal with $f_{Gn} : \{\mathbf{x}, \chi\} \rightarrow \{\mathbf{n}\}$, scene reflection $f_{G\alpha} : \{\mathbf{d}, \omega, \chi\} \rightarrow \{\alpha\}$, and ambient light component with $f_{Gp} : \{\mathbf{d}, \chi\} \rightarrow \{\Lambda\}$. An overview of the overall Neural Gated Fields is shown in Fig. 3. We also use a proposal sampler f_P as in [7] for efficiency. Both f_G and f_P are MLPs (of different size) with multi-resolution hash encoding [64].

Shadow and Illumination Model As the light pulse emitted by the illuminator is a diverging light beam, we model it as cone of light with irradiance maximum at the cross-section center and exponentially decreasing as it diverges from the center by the angles γ. As such, we express the illumination intensity as a 2D higher-order Gaussian $\mathcal{G}$ with mean $\boldsymbol{\Xi}$, standard deviation $\boldsymbol{\Omega}$ and power $\boldsymbol{\Theta}$

$$\iota = \eta \mathcal{G}_i(\gamma; \boldsymbol{\Xi}, \boldsymbol{\Omega}, \boldsymbol{\Theta}), \tag{5}$$

where η is a scaling parameter.

Instead of predicting the shadow indicator $\psi(\mathbf{x})$, we can directly estimate it using the density field, by computing the accumulated transmittance along the ray from the pixel to the point $\mathbf{r}_{ill}(l) = \mathbf{o}_i + \omega l$

$$\psi(\mathbf{x}) = \exp\left(-\int_0^{l_\mathbf{x}} \sigma(\mathbf{r}_{ill}(l))\mathrm{d}l\right) \tag{6}$$

The illuminator origin and direction $\mathbf{o}_i, \omega$ are obtained from the camera as $[\mathbf{o}_i, \omega] = \mathbf{R}[\mathbf{o}_c, \mathbf{d}_c] + \mathbf{T}$. During training, we jointly fine-tune the translation and rotation matrices $\mathbf{T}$, $\mathbf{R}$, as well as $\mathbf{o}_c$. We also treat illuminator profile proprieties as learnable parameters, namely η, $\boldsymbol{\Xi}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Theta}$ and the gating parameters, i.e. number of accumulated laser pulses m_k before read-out, laser pulse duration $t_{l,k}$, camera exposure $t_{g,k}$, and delay ξ_k between laser pulse emission and gated exposure for all three slices $k \in \{0, 1, 2\}$. In addition, we optimize a general distance offset d_0 for the range intensity profiles to compensate for internal signal processing delays.

4.2. Gated Field Learning

We learn to acquire a gated capture with camera origin $\mathbf{o}_c$ and direction $\mathbf{d}$ by casting a ray $\mathbf{r}(l) = \mathbf{o_c} + l\mathbf{d}$ for each pixel into the scene and computing the intensity $\tilde{I}_k(\mathbf{r})$ through volume rendering. Using the gated imaging formation model from Sec. 3, we define volume rendering as

$$\begin{aligned} \tilde{I}_k(\mathbf{r}) = \int_0^\infty T(l)\sigma(\mathbf{x}) \int_{-\infty}^{\infty} g_k(t - \xi_k)\beta(l) \\ \cdot \left(\alpha\iota\psi|\mathbf{n}\cdot\omega|p_k\left(t - \frac{l + l_i}{c}\right) + \kappa\right) \mathrm{d}t\,\mathrm{d}l + \mathcal{D}_k, \end{aligned} \tag{7}$$

where κ is the ambient level, $T(l) = \exp(-\int_0^l \sigma(u)du)$ is the accumulated transmittance along the ray and l_i is the distance of $\mathbf{x}(l)$ from the illuminator origin $\mathbf{o}_i$. As illustrated in Fig. 2, in our volume rendering formulation the pixel intensity contribution of a point along the ray depends not only on its accumulated transmittance and volumetric

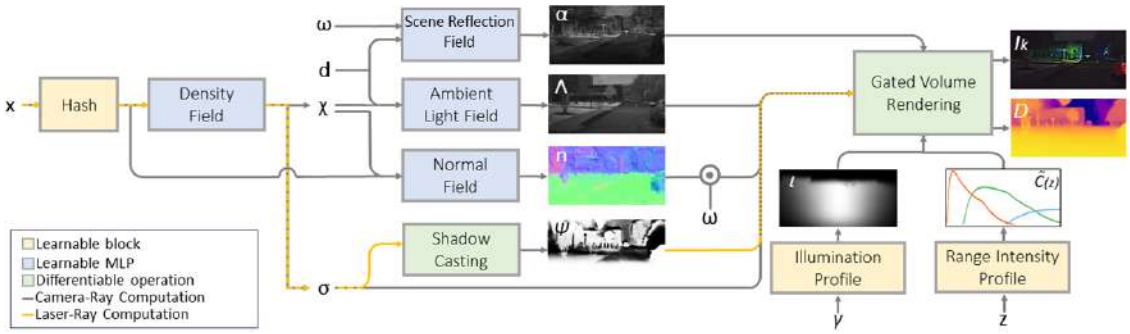

Figure 3. Neural Gated Fields. For any point in space $\mathbf{x}$, we learn its volumetric density σ, normal $\mathbf{n}$, reflectance α and ambient lighting Λ through four neural fields, conditioned on direction $\mathbf{d}$, incident laser light direction ω and spatial embedding χ. The illuminator light ι is represented by a physics-based model dependent on the displacement angle γ, while the gating imaging process is described by the range intensity profiles $\tilde{C}(z)$ using as input the camera-point-laser distance z, as explained in Sec. 3. With this information, we reconstruct a gated image I_k through the gated volume rendering formulation introduced in Sec. 4.2. As this process is fully differentiable, we simultaneously fit neural fields and physical parameters through image reconstruction together with other regularization losses discussed in Sec. 4.3.

density, but also on its distance from the illuminator and camera origins through $C(z)$, as well as on its relative position to the illuminator source via ι and ψ.

We simplify the time-dependent integral using Eq. (4) as

$$\tilde{I}_k(\mathbf{r}) = \int_0^\infty T(l)\sigma(\mathbf{x})\big(\alpha(\mathbf{x},\mathbf{d},\omega)\tilde{C}_k(l+l_i)\psi(\mathbf{x}) \cdot|\mathbf{n}\cdot\omega|\iota(\gamma)+\Lambda(\mathbf{x},\mathbf{d})\big)\mathrm{d}l + \mathcal{D}_k. \tag{8}$$

We numerically estimate this spatial integral by numerical quadrature [62, 88], approximating it with a set of points $\mathbf{X}_{ray}$. Specifically, for each point $\mathbf{x}_j \in \mathbf{X}_{ray}$, we query the neural field f_G to infer the normal vector $\mathbf{n}_j$, scene reflection α_j, volumetric density σ_j and ambient light component Λ_j. The laser illumination intensity ι_j is instead computed following the physics-based model defined in Eq. (5). The gated intensity $\tilde{I}_k$ is then expressed as

$$\tilde{I}_k(\mathbf{r}) = \sum_{j=0}^{N} w_j \Big(\underbrace{\alpha_j \tilde{C}_j \psi_j |\mathbf{n}_j\cdot\omega_j|\iota_j}_{\text{Active Component}} + \underbrace{\Lambda_j}_{\text{Passive Component}} \Big) + \mathcal{D}_k, \tag{9}$$

$$w_j = \exp\Big(-\sum_{k=1}^{j-1}\sigma_k\delta_k\Big)\big(1-\exp(-\sigma_j\delta_j)\big). \tag{10}$$

The shadow indicator ψ_j from Eq. (6) is similarly approximated by sampling on $\mathbf{r}_{ill} = \mathbf{o}_i + \omega_j l$ a set of points $\mathbf{X}_{ill}$ bounded between $\mathbf{o}_i$ and $\mathbf{x}_j$

$$\psi_j = \exp\Big(-\sum_{k}^{\mathbf{X}_{ill}}\sigma_k\delta_k\Big). \tag{11}$$

For the passive slice, the active component is null, further simplifying to

$$\tilde{I}_P(\mathbf{r}) = \sum_{j=0}^{N} w_j\Lambda_j + \mathcal{D}_P. \tag{12}$$

Both $\mathbf{X}_{ray}$ and $\mathbf{X}_{ill}$ are sampled using a proposal network [7] f_P that, analogously to f_{Gd}, predicts point-wise densities converted with Eq. (10) to proposal weights $\hat{w}$ for sampling with piece-wise-constant probabilities.

4.3. Training Supervision

We supervise the predicted passive and active gated frames applying a photometric loss $\mathcal{L}_c$, regularize the volumetric density with a depth loss $\mathcal{L}_d$ and by supervising the shadow estimate with $\mathcal{L}_s$. We regularize normal and reflectance estimates through $\mathcal{L}_{nc}$ and $\mathcal{L}_\alpha$, respectively.

Photometric Loss We supervise with ground truth captures for active and passive gated slice reconstruction as

$$\mathcal{L}_c = \sum_{k,r}\|\tilde{I}_k(\mathbf{r}) - I_k(\mathbf{r})\|_2 + \sum_{r}\|\tilde{I}_P(\mathbf{r}) - I_P(\mathbf{r})\|_2 \tag{13}$$

Volume Density Regularization As additional training supervision, we use the depth estimate $\hat{D}(\mathbf{r})$ of a pretrained stereo depth estimation algorithm [90] as pseudo ground-truth to regularize the ray termination distribution [26]

$$\mathcal{L}_d = \sum_{\mathbf{r}}\sum_{j} \log w_j \exp\left(-\frac{(l_j-\hat{D}(\mathbf{r}))^2}{2s^2}\right)\delta_i \tag{14}$$

We regularize the density field by partially supervising the shadow indicator ψ. Each pixel whose active intensity $I_{kA} = I_k - I_P$ in any of the three gated slices is above a certain threshold ϵ_i is considered as visible from the illuminator. We hence supervise the expected shadow value for such rays $\mathbf{r}_v \in \{\mathbf{r}|\forall k \in \{1,2,3\} : I_{kA}(\mathbf{r}) > \epsilon_i\}$ as

$$\mathcal{L}_s = \sum_{\mathbf{r}_v}\|1 - \int T(l)\sigma(\mathbf{x})\psi(\mathbf{x})dl\|_2 \tag{15}$$

Normals Consistency Following [88], for each sampled point $\mathbf{x}$ we enforce a consistency between the predicted normal $\mathbf{n}$ and the density gradient $\hat{\mathbf{n}}(\mathbf{x}) = -\frac{\nabla\mathbf{x}}{||\nabla\mathbf{x}||}$, and we penalize normals which are back-facing the camera as

$$\mathcal{L}_{nc} = \sum_{\mathbf{x}} w(\mathbf{x})\big(\|\mathbf{n}(\mathbf{x}) - \hat{\mathbf{n}}(\mathbf{x})\|_2 + \max(0, \hat{\mathbf{n}}(\mathbf{x})\cdot\mathbf{d})^2\big) \tag{16}$$

Reflectance Regularization We enforce the predicted scene reflection α to be spatially consistent within $\epsilon_{\mathbf{x}}$, i.e.

$$\mathcal{L}_\alpha = \sum_{\mathbf{x}} w(\mathbf{x})\left(\|\alpha(\mathbf{x},\mathbf{d}) - \alpha(\mathbf{x}+\epsilon_{\mathbf{x}}, \mathbf{d}+\epsilon_{\mathbf{d}})\|_2\right). \tag{17}$$

Here the ω dependency of α is omitted here for brevity. We also include an angular noise $\epsilon_{\mathbf{d}}$ that we set high at the beginning of the training and then decrease it exponentially. By forcing the reflectance to behave as fully diffuse at the beginning of the training, we disincentive it to bake in the effects from lighting or shadowing, hence improving the disjoint learning of the scene components.

Total Training Loss Combining the different losses we obtain the following loss formulation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_d + \lambda_3 \mathcal{L}_s + \lambda_4 \mathcal{L}_{nc} + \lambda_5 \mathcal{L}_\alpha, \qquad (18)$$

see $\lambda_{1,\dots,5}$ hyperparameters in the Supplementary Material.

5. Implementation Details

We train for 35,000 steps and a batch size of 4096 rays. As optimizer we use ADAMW [57] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate 10^{-2} for f_P and f_G, 10^{-4} for the camera poses optimization, 10^{-4} for the laser profile and gated parameters. We train on two NVIDIA V100 GPUs, for approximately 3 hours. The proposal network f_P is comprised of two MLPs and trained following [7]. Additional architecture details, training procedures, and hyper-parameters are found in the Supplementary Material.

6. Dataset

To conduct this work, we have collected a diverse set of 10 static sequences, recorded in both day and night conditions across North America. To this goal, we equipped a test vehicle with a NIR gated stereo camera setup (BrightWay Vision), an automotive RGB stereo camera (OnSemi AR0230), a LiDAR sensor (Velodyne VLS128) and a GNSS with IMU (Xsens MTi-7), as shown in Fig. 4. Each gated camera has a resolution of 1280x720 pixels, 10 bit depth and runs at 120 Hz, split up to collect the three active and one passive slice. The illuminator source consists of two vertical-cavity surface-emitting laser (VCSEL) modules, which illuminate the scene with a laser pulse with duration of 240-370 ns and a wavelength of 808 nm. The RGB cameras provide 12 bit HDR images with resolution of 1920x1080 pixels and 30 Hz frame-rate. The LiDAR has a vertical resolution of 128 lines and 10 Hz framerate, while the GNSS sensor runs at 4 Hz. Example captures from the dataset are being visualized in Fig. 4. In total, we collect 2650 samples, captured in both day (1223 samples) and night (1427 samples). We divide it in training, validation and test splits with a 50-25-25 split.

As ground-truth, we construct a large-scale ground-truth pointcloud by aggregating LiDAR scans with LIO-SAM [78] and removing noisy points. Additional details on the accumulation are provided in the Supplementary Material.

7. Assessment

In this section, we validate the proposed method quantitatively and qualitatively. Specifically, we investigate scene reconstruction at both day and night, using novel depth and view synthesis for 2D evaluation, and surface reconstruction for the 3D evaluation. To this end, we compare our approach to state-of-the-art feed-forward depth estimation algorithms and neural scene reconstruction methods. We also conduct ablation experiments to validate our design choices.

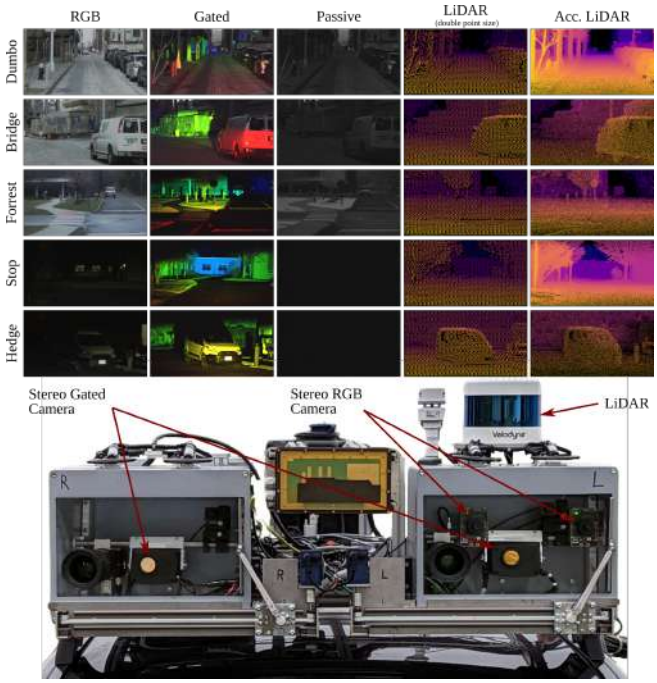

Figure 4. Top: Example captures from our collected dataset across different urban and suburban areas in North America. From left to right: RGB image, active gated slices (with red for slice 1, green for slice 2 and blue for slice 3), passive slice, projected LiDAR scan, accumulated LiDAR. Bottom: Sensors setup with LiDAR, stereo Gated camera, stereo RGB camera, IMU and GNSS.

Depth Reconstruction We assess the quality of depth synthesis of Gated Fields for camera poses unseen during training. We use as ground truth the accumulated and filtered LiDAR pointcloud. Unlike previous works relying on single LiDAR scans for evaluation [35, 89, 90], we use as ground truth an accumulated LiDAR pointcloud as described in Sec. 6. This allows us to evaluate the depth reconstruction up to 160 m accurately and without bias. We follow previous works [89, 90] and use as depth evaluation metrics RMSE, MAE, ARD, $\sigma_i < 1.25^i, i \in \{1, 2, 3\}$. We compare our method against 9 feed-forward depth estimation methods, namely SimIPU [53], AdaBins [10], DPT [70], DepthFormer [54] and CREStereo [50] for monocular and stereo RGB methods, Gated2Gated [89] and GatedStereo [90] for monocular and stereo gated estimation methods. We also compare our method against depths rendered with other neural reconstruction algorithms, using RGB images [7],[63],[29],[26] LiDAR [84], RGB+LiDAR [86], [37], and a varying-appearance method [29] for gated

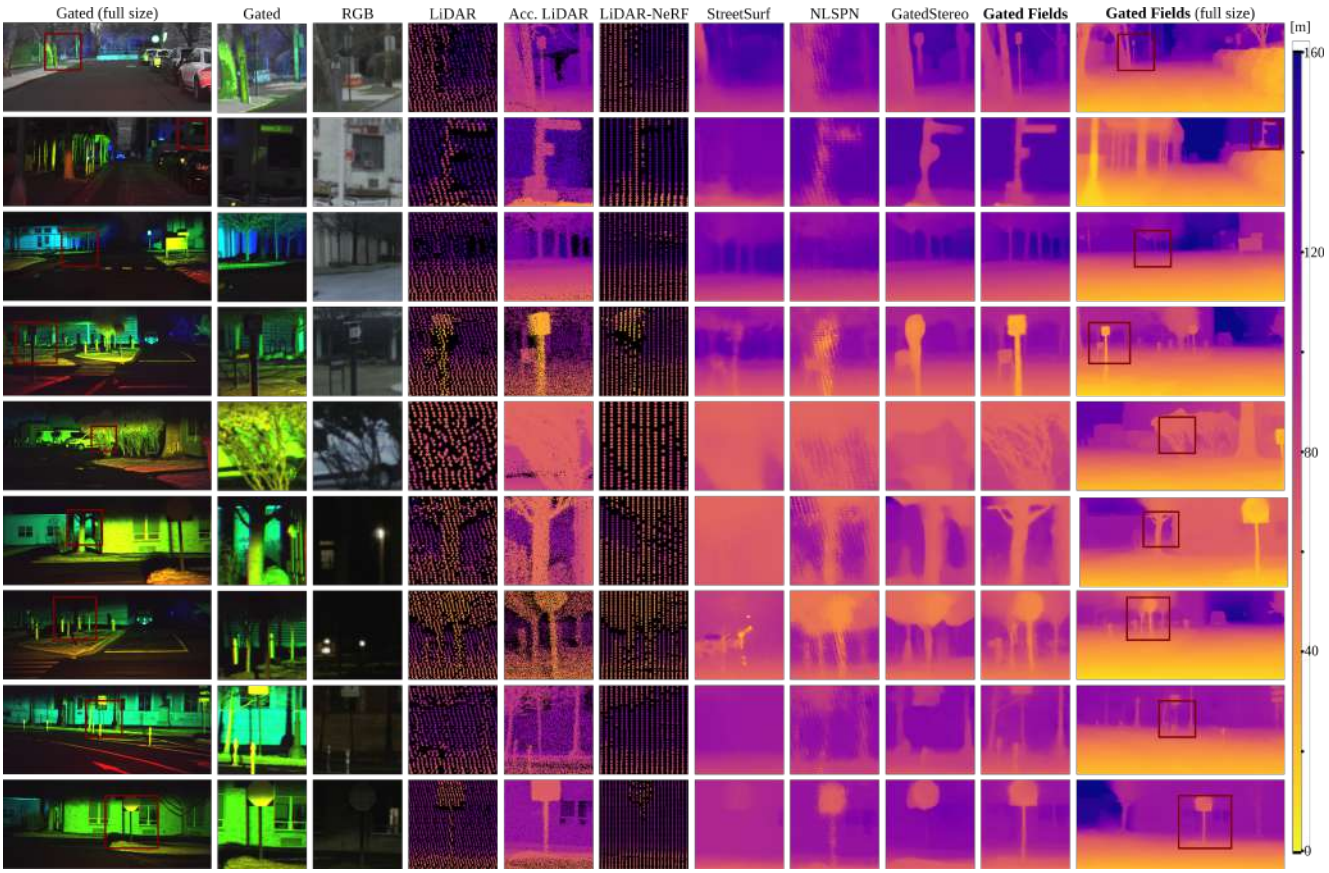

Figure 5. Qualitative comparison of the proposed **Gated Fields** and state-of-the-art depth estimation approaches, including LiDAR-NeRF [84], StreetSurf [37], NLSPN [67], and Gated Stereo [90]. Compared to baseline methods, we are able to reconstruct fine geometry details like branches or poles, also for far distances. Unlike RGB methods, Gated Fields is unaffected by poor ambient lighting, and unlike LiDAR-based methods it is able to reconstruct sharp object discontinuities. The active gated slices are visualized in red for slice 1, green for slice 2 and blue for slice 3.

captures. Results for day and night sequences are presented in Tab. 1. We outperform the next best neural field method by 21.87% MAE and 30.35% in RMSE. For night sequences the performance difference sharpens, with Gated Fields outperforming the best RGB-based method [50] by 3.14 m MAE. This performance decline is to be attributed to the limited pixel information present in RGB captures taken at night time, making impossible to learn a meaningful 3D representation of the scene, as shown qualitatively in Fig. 5. LiDAR-based methods are unaffected by the change in illumination, but suffer from the limited sensor resolution. On the other hand, gated cameras retrieve information-rich captures both day and night, which Gated Fields can explicitly leverage during training. We confirm that employing state-of-the-art neural field methods on gated captures does not yield accurate results, as they are not able to model the gated imaging formation and can only fit the ambient light component.

3D Reconstruction We evaluate the 3D scene reconstruction capabilities of Gated Fields using the accumulated LiDAR pointcloud as ground truth. We follow [20] and extract for both the 3D ground truth pointcloud and different neural field-based methods a voxelized occupancy grid of the scene, and compute intersection over union (IoU), Precision and Recall between ground truth voxels and estimated ones. Quantitative results are shown in Tab. 3. Our method outperforms RGB baselines [7, 29] by an average of 15% IoU. RGB baselines using additional LiDAR sensor [37, 86] data partially improve the results, but such methods are still unable to reconstruct finer surfaces details and struggle at night. On the other hand, Gated Fields is able to recover finer geometries using gated, illumination and depth cues, and the quality does not degrade with diminishing ambient light, as shown in Fig. 5. See details on evaluation and further qualitative results in the Supplementary Material.

Novel View Synthesis For novel view synthesis, we compare our method with Mip-NeRF360 [7], a state-of-the-art neural radiance field-based method, and K-Planes [29], to implicitly model the time-varying appearance of the static scene. Mip-NeRF [7] struggles to reconstruct novel views due to the inherent difficulty of modeling the gating imaging effects, resulting in a PSNR of 17.16dB. By learning

	Method	Modality	RMSE [m]	ARD	MAE [m]	δ_1 [%]	δ_2 [%]	δ_3 [%]
Test Data – Night (Evaluated on Accumulated LiDAR Ground-Truth Points)								
2D Depth Comparison	Gated2Gated [89]	Gated	13.33	0.28	8.57	61.78	90.30	94.90
	GatedStereo [90]	Stereo-Gated	10.10	0.20	5.97	82.86	93.35	96.52
	SimIPU [53]	RGB	19.33	0.44	14.21	40.67	77.99	89.80
	AdaBins [10]	RGB	21.14	0.38	14.28	51.72	80.00	90.64
	DPT [70]	RGB	14.17	0.28	9.90	62.82	88.32	94.42
	DepthFormer [54]	RGB	14.32	0.29	10.08	61.19	87.72	94.57
	CREStereo [50]	Stereo-RGB	14.22	0.13	7.27	82.21	90.10	94.32
	NLSPN [67]	RGB+LiDAR	11.12	0.18	6.48	77.15	90.81	96.00
	MipNeRF360 [7]	RGB	23.80	0.51	16.43	41.60	61.75	76.16
	K-Planes [29]	RGB	19.70	0.41	13.66	44.39	66.20	81.46
	RawNeRF [63]	RGB	27.46	0.64	19.75	34.32	54.33	69.41
	Depth-NeRF [26]	RGB	15.23	0.26	10.04	61.67	86.23	93.87
	SUDS [86]	RGB+LiDAR	11.07	0.17	6.17	79.49	88.41	95.31
	StreetSurf [37]	RGB+LiDAR	10.86	0.15	5.90	82.16	93.57	96.96
	LiDAR-NeRF [50]	LiDAR	10.21	0.12	4.72	87.71	95.05	97.78
	Gated Fields [50]	Gated	**7.92**	**0.12**	**4.13**	**90.61**	**95.76**	**97.90**
Test Data – Day (Evaluated on Accumulated LiDAR Ground-Truth Points)								
2D Depth Comparison	Gated2Gated [89]	Gated	9.26	0.22	6.69	58.46	93.70	97.38
	GatedStereo [90]	Stereo-Gated	6.32	0.09	3.30	92.86	97.31	98.53
	SimIPU [53]	RGB	13.54	0.31	10.08	52.90	86.49	95.65
	AdaBins [10]	RGB	12.74	0.25	8.39	69.84	89.13	95.52
	DPT [70]	RGB	10.34	0.21	7.08	77.61	94.29	97.24
	DepthFormer [54]	RGB	9.06	0.19	6.09	81.19	94.14	97.42
	CREStereo [50]	Stereo-RGB	7.35	0.09	3.45	**94.48**	97.23	98.50
	NLSPN [67]	RGB+LiDAR	10.34	0.17	5.97	77.83	91.16	96.11
	MipNeRF360 [7]	RGB	16.91	0.31	9.53	70.89	84.43	90.92
	K-Planes [29]	RGB	12.37	0.24	8.55	63.16	79.68	91.70
	RawNeRF [63]	RGB	15.10	0.23	9.38	65.90	84.90	92.43
	Depth-NeRF [26]	RGB	10.34	0.17	6.07	78.97	90.75	96.53
	SUDS [86]	RGB+LiDAR	9.11	0.17	5.84	80.96	95.08	98.33
	StreetSurf [37]	RGB+LiDAR	9.60	0.13	5.35	83.94	95.32	98.56
	LiDAR-NeRF [50]	LiDAR	8.13	0.10	3.86	88.99	95.49	98.06
	Gated Fields [50]	Gated	**6.15**	**0.09**	**2.91**	93.88	**97.32**	**98.75**

Table 1. Comparison of our proposed approach and state-of-the-art approaches on depth synthesis. Best results in each category are in **bold** and second best are underlined.

	Im. Formation Model	Depth Sup.	Normal	Active Illum.	Shadow	RMSE [m]	MAE [m]	PSNR [dB]	SSIM
Test Data – Night									
Ablation	End2End [62]	✗	✗	✗	✗	27.34	19.95	17.43	0.633
	Gated	✗	✗	✗	✗	9.34	8.35	23.47	0.889
	Gated	✓	✗	✗	✗	6.72	5.43	25.84	0.908
	Gated	✓	✓	✓	✗	3.78	2.82	30.66	0.946
	Gated	✓	✓	✓	✓	**3.39**	**2.51**	**30.91**	**0.95**
Test Data – Day									
Ablation	End2End [62]	✗	✗	✗	✗	30.03	19.22	16.77	0.66
	Gated	✗	✗	✗	✗	11.64	6.34	26.77	0.915
	Gated	✓	✗	✗	✗	9.20	4.22	26.88	0.922
	Gated	✓	✓	✓	✗	9.45	4.31	32.15	0.946
	Gated	✓	✓	✓	✓	**8.88**	**4.12**	**32.28**	**0.948**

Table 2. Ablation studies of the Gated Fields contributions, on a subset of the test dataset. We investigate different image formation models, neural fields components and supervision losses.

a time-varying appearance, K-Planes improves the quality reaching 27.42dB PSNR for day but only 19.35dB for night, as the model fails to learn an accurate scene geometry representation without ambient light information. Gated Fields outperforms these baselines in both day and night, reaching a PSNR of 32.28dB.

Ablation Experiments To assess the role and contribution of the different components of our method, we conduct an ablation study in Tab. 2. In particular, we consider as starting point a single neural field directly inferring one intensity value for each of the four slices (3 active + 1 passive), and obtain an average MAE of 19.58 m. By separately predicting ambient light and reflectance, and reconstructing the gated image as in Eq. (2), we significantly improve the MAE to 7.34 m . However, this approach still performs poorly on flat-color areas during the day and in unilluminated areas during the night due to lack of any depth cue. By adding the depth supervision, we are able to supervise also such areas and the PSNR improves by 4.83dB. By adding the angular-dependent attenuation and regularizing the reflectance in Eq. (17), the model is able to disentangle the material proprieties from other spurious effects. Finally, by explicitly modeling the shadow, casted by the illuminator, we improve final depth reconstruction to 3.32 m MAE.

	Method	Modality	IoU [%]	Precision [%]	Recall [%]
3D Rec.	MipNeRF360 [7]	RGB	6.32	7.34	31.21
	StreetSurf [37]	RGB+LiDAR	5.41	6.31	27.35
	SUDS [86]	RGB+LiDAR	8.96	9.88	49.09
	LiDAR-NeRF [84]	LiDAR	20.03	**32.38**	34.44
	Gated Fields (ours)	Gated	**22.25**	25.01	**66.51**

Table 3. Comparison of Gated Fields and state-of-the-art scene reconstruction methods. We evaluate over 3D occupancy reconstruction, using as ground truth the voxelized accumulated LiDAR pointcloud. Best results in each category are in **bold** and second best are underlined.

8. Conclusion

We introduce Gated Fields, a neural rendering method capable of reconstructing scene geometry from video captures of active time-gated cameras. The method hinges on a differentiable gated image formation as part of the rendering formulation, and it jointly learns geometry, ambient light and surface proprieties, represented implicitly as neural field components, alongside illumination and gating parameters, represented with physics-based models. Extensive experiments on real-world large-scale scenes validate that our method is able to precisely reconstruct a 3D scene both in day and night-time conditions. Our approach outperforms existing RGB and LiDAR methods by 21.87% on MAE, as well as baseline methods using gated captures by 31.67%. In the future, we hope to extend the proposed approach by "closing the loop" and providing dynamic feedback to the gated acquisition, allowing for adaptive gated scene reconstructions.

Acknowledgments This work was supported by the AI-SEE project with funding from the FFG, BMBF, and NRC-IRA. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

References

[1] Amit Adam, Christoph Dann, Omer Yair, Shai Mazor, and Sebastian Nowozin. Bayesian time-of-flight for realtime shape, illumination and albedo. 39(5):851–864, 2017. 3

[2] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5861–5870, 2023. 2

[3] Pierre Andersson. Long-range three-dimensional imaging using range-gated laser radar images. 45(3):034301, 2006. 3

[4] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in neural information processing systems*, 34:26289–26301, 2021. 1

[5] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3D: Stereo depth estimation via binary classifications. In *arXiv preprint arXiv:2005.07274*, 2020. 2

[6] 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. 3

[7] 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. 1, 3, 4, 5, 6, 7, 8

[8] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 1, 3

[9] Florent Bartoccioni, Éloi Zablocki, Patrick Pérez, Matthieu Cord, and Karteek Alahari. Lidartouch: Monocular metric depth estimation with a few-beam lidar. *Computer Vision and Image Understanding*, 227:103601, 2023. 2

[10] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 6, 8

[11] Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for lidar sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 760–767, 2018. 2

[12] Mario Bijelic, Tobias Gruber, and Werner Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. 2018. 2

[13] Mario Bijelic, Tobias Gruber, and Werner Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1773–1779. IEEE, 2018. 2

[14] 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. 2

[15] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022. 3

[16] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 1

[17] Jens Busck. Underwater 3-D optical imaging with a gated viewing laser radar. 2005. 3

[18] Jens Busck and Henning Heiselberg. Gated viewing and high-accuracy three-dimensional laser radar. 43(24):4705–10, 2004. 3

[19] Jens Busck and Henning Heiselberg. Gated viewing and high-accuracy three-dimensional laser radar. *Applied optics*, 43(24):4705–4710, 2004. 2

[20] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023. 7

[21] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda. Libre: The multiple 3d lidar dataset. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[22] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2

[23] 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. 3

[24] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021. 2

[25] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 1

[26] 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. 3, 5, 6, 8

[27] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. 2

[28] 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. 3

[29] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 6, 7, 8

[30] Ravi Garg, B.G. Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016. 2

[31] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 2

[32] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2

[33] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 2

[34] Yoav Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. 2

[35] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 6

[36] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 2

[37] 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. 1, 3, 6, 7, 8

[38] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 2

[39] Paul Heckman and Robert T. Hodgson. Underwater optical range gating. 3(11):445–448, 1967. 2

[40] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. 2021. 2

[41] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. 2023. 1, 3

[42] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. pages 52–60, 2018. 2

[43] Maria Jokela, Matti Kutila, and Pasi Pyykönen. Testing and validation of automotive point-cloud sensors in adverse weather conditions. *Applied Sciences*, 9, 2019. 2

[44] Frank Julca-Aguilar, Jason Taylor, Mario Bijelic, Fahim Mannan, Ethan Tseng, and Felix Heide. Gated3d: Monocular 3d object detection from temporal illumination cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2938–2948, 2021. 2

[45] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017. 2

[46] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, pages 141–159. Wiley Online Library, 2010. 2

[47] 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. 3

[48] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000. 2

[49] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. Seathru-nerf: Neural radiance fields in scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 56–65, 2023. 1

[50] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 6, 7, 8

[51] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241, 2021. 2

[52] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6186, 2021. 2

[53] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1500–1508, 2022. 6, 8

[54] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Depthformer: Exploiting long-range corre-

lation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2, 6, 8

[55] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[56] 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. 3

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2018. 6

[58] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. pages 1–8, 2018. 2

[59] Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kiriakos N Kutulakos, and David B Lindell. Transient neural radiance fields for lidar view synthesis and 3d reconstruction. *arXiv preprint arXiv:2307.09555*, 2023. 1, 3

[60] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2

[61] 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. 3

[62] 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. 1, 3, 5, 8

[63] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 1, 6, 8

[64] 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. 3, 4

[65] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 1

[66] 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. 3

[67] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 2, 7, 8

[68] Vivek Pradeep, Christoph Rhemann, Shahram Izadi, Christopher Zach, Michael Bleyer, and Steven Bathiche. Monofusion: Real-time 3d reconstruction of small scenes with a single web camera. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 83–88. IEEE, 2013. 1

[69] 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. 1

[70] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 6, 8

[71] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. *CVPR*, 2022. 1, 3

[72] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *2017 International Conference on 3D Vision (3DV)*, pages 57–66. IEEE, 2017. 1

[73] 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. 3

[74] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022. 3

[75] Michael Schober, Amit Adam, Omer Yair, Shai Mazor, and Sebastian Nowozin. Dynamic time-of-flight. In *CVPR*, pages 6109–6118, 2017. 3

[76] Thomas Schöps, Torsten Sattler, Christian Häne, and Marc Pollefeys. 3d modeling on the go: Interactive 3d reconstruction of large-scale scenes on mobile devices. In *2015 International Conference on 3D Vision*, pages 291–299. IEEE, 2015. 1

[77] Brent Schwarz. Lidar: Mapping the world in 3D. *Nature Photonics*, 4(7):429, 2010. 2

[78] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5135–5142. IEEE, 2020. 6

[79] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1

[80] 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. 3

[81] 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. 3

[82] Jiexiong Tang, John Folkesson, and Patric Jensfelt. Sparse2dense: From direct sparse odometry to dense 3-d reconstruction. *IEEE Robotics and Automation Letters*, 4 (2):530–537, 2019. 2

[83] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 2

[84] Tang Tao, Longfei Gao, Guangrun Wang, Peng Chen, Dayang Hao, Xiaodan Liang, Mathieu Salzmann, and Kaicheng Yu. Lidar-nerf: Novel lidar view synthesis via neural radiance fields. *arXiv preprint arXiv:2304.10406*, 2023. 1, 3, 6, 7, 8

[85] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 855–866, 2023. 3

[86] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 6, 7, 8

[87] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 2

[88] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 5

[89] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images. 2022. 2, 3, 4, 6, 8

[90] Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13252–13262, 2023. 2, 3, 4, 5, 6, 7, 8

[91] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson WH Lau. Lighting up nerf via unsupervised decomposition and enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2023. 1

[92] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[93] 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)*, 2023. 1, 3

[94] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12747–12756, 2021. 2

[95] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[96] Zhenfei Yang, Fei Gao, and Shaojie Shen. Real-time monocular dense mapping on aerial robots using visual-inertial fusion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4552–4559. IEEE, 2017. 1

[97] 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, 2023. 3

[98] 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. 3

[99] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. *arXiv preprint arXiv:2304.14811*, 2023. 3

[100] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 3

[101] 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. 3

[102] Yongjian Zhang, Longguang Wang, Kunhong Li, Zhiheng Fu, and Yulan Guo. Slfnet: A stereo and lidar fusion network for depth completion. *IEEE Robotics and Automation Letters*, 7(4):10605–10612, 2022. 2

[103] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2