Shoot-Bounce-3D: Single-Shot Occlusion-Aware 3D from Lidar by Decomposing Two-Bounce Light

TZOFI KLINGHOFFER, Massachusetts Institute of Technology, USA
SIDDHARTH SOMASUNDARAM*, Massachusetts Institute of Technology, USA
XIAOYU XIANG*, Meta, USA
YUCHEN FAN, Meta, USA
CHRISTIAN RICHARDT, Meta, Switzerland
AKSHAT DAVE, Massachusetts Institute of Technology, USA
RAMESH RASKAR, Massachusetts Institute of Technology, USA
RAKESH RANJAN, Meta, USA

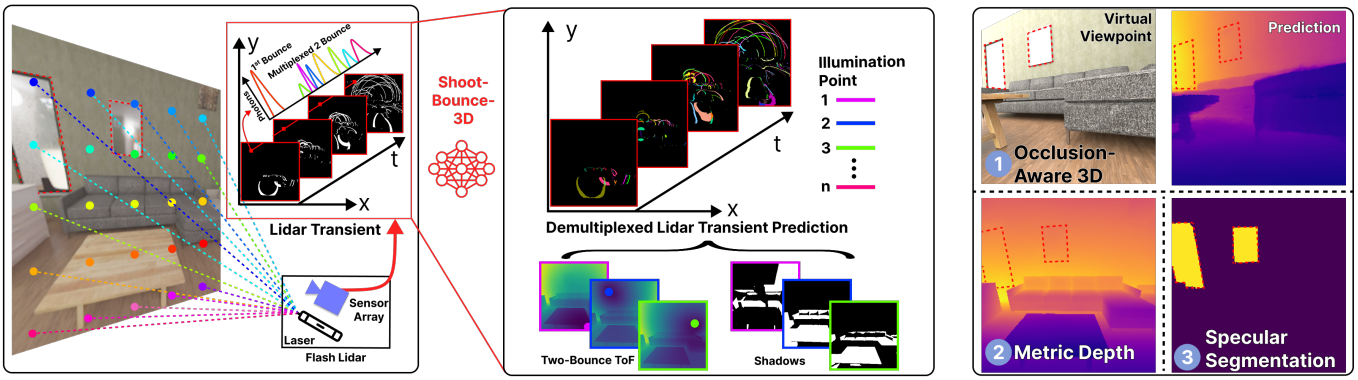

Fig. 1. **Overview.** We introduce *Shoot-Bounce-3D (SB3D)*: a method to decompose temporal light transport in a scene from a single-view, single-shot capture, enabling recovery of 3D geometry, despite specular surfaces and occlusions. **(a)** A single-photon lidar ***shoots*** light into the scene at multiple points at once, referred to as *multiplexed illumination*. Some light reflects directly back to the sensor, while other light ***bounces*** multiple times first. The lidar captures histograms containing photon intensity over time – known as *transients*. The multiplexed light mixes together in the transients. **(b)** We create the first-of-its-kind simulated dataset of multiplexed lidar transients from ~100k scenes and use it to train a model to *demultiplex* two-bounce light. **(c)** Our model enables single-shot ***3D***, including both dense metric depth and occluded geometry, in the presence of specular surfaces.

3D scene reconstruction from a single measurement is challenging, especially in the presence of occluded regions and specular materials, such as mirrors. We address these challenges by leveraging single-photon lidars. These lidars estimate depth from light that is emitted into the scene and reflected directly back to the sensor. However, they can also measure light that bounces multiple times in the scene before reaching the sensor. This *multi-bounce light* contains additional information that can be used to recover dense depth, occluded geometry, and material properties. Prior work with single-photon lidar, however, has only demonstrated these use cases when a laser sequentially illuminates one scene point at a time. We instead focus on the more practical – and challenging – scenario of illuminating multiple scene points simultaneously. The complexity of light transport due to the combined effects of multiplexed illumination, two-bounce light, shadows, and specular reflections is challenging to invert analytically. Instead, we propose a data-driven method to invert light transport in single-photon lidar. To enable this approach, we create the first large-scale simulated dataset of ~100k lidar transients for indoor scenes. We use this dataset to learn a prior on complex light transport, enabling measured two-bounce light to be decomposed into the constituent contributions from each laser spot. Finally, we experimentally demonstrate how this decomposed light can be used to infer 3D geometry in scenes with occlusions and mirrors from a single measurement. Our code and dataset are released on our project webpage.

CCS Concepts: • **Computing methodologies → 3D imaging**.

ACM Reference Format:
Tzofi Klinghoffer, Siddharth Somasundaram, Xiaoyu Xiang, Yuchen Fan, Christian Richardt, Akshat Dave, Ramesh Raskar, and Rakesh Ranjan. 2025.

*Both authors contributed equally to this research.

Authors' Contact Information: Tzofi Klinghoffer, tzofi@mit.edu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Siddharth Somasundaram, sidsoma@mit.edu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Xiaoyu Xiang, Meta, Menlo Park, California, USA, xiangxiaoyu@meta.com; Yuchen Fan, Meta, Menlo Park, California, USA, ycfan@meta.com; Christian Richardt, Meta, Zurich, Switzerland, crichardt@meta.com; Akshat Dave, ad74@mit.edu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Ramesh Raskar, raskar@mit.edu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; Rakesh Ranjan, Meta, Menlo Park, California, USA, rakeshr@meta.com.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763945

Shoot-Bounce-3D: Single-Shot Occlusion-Aware 3D from Lidar by Decomposing Two-Bounce Light. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757377.3763945

1 Introduction

Single-shot 3D scene understanding is a long-standing problem in computer vision and graphics – critical to applications ranging from autonomous vehicles to extended reality. However, recovering 3D information from a single RGB image is ambiguous: lack of multiview correspondences makes metric depth estimation ill-posed, occlusions must be hallucinated, and specular surfaces can be mistaken as holes or "portals" in the scene. We present a machine learning (ML) approach to leverage *single-photon lidar* for single-shot 3D reconstruction in scenes with occlusions and specular surfaces.

Single-photon lidars – composed of a pulsed laser and a single-photon avalanche diode (SPAD) sensor – shoot light pulses into the scene and measure the time light takes to return to the sensor. Similar to traditional lidar, time taken by the light *directly* reflecting back from the scene – called time of flight (ToF) – encodes the depth of illuminated points. However, single-photon lidars can also capture the time taken by light that *indirectly* reflects – or "bounces" – to other parts of the scene before hitting the sensor. In particular, single-photon lidars measure time-resolved histograms, called *transients*, in which multiple bounces of light appear as multiple peaks (Fig. 1a).

In our work, we focus on two-bounce light: light that has reflected up to two times in the scene. Prior works use two-bounce light in lidar transients to recover dense depth [Henley et al. 2022], occluded geometry [Klinghoffer et al. 2024], and material properties [Henley et al. 2023]. However, these works rely on lidars that scan the laser *sequentially* over the scene, one point at a time. Instead, we consider multiplexed illumination – meaning the scene is illuminated at multiple points *simultaneously*. As a result, measured transients have multiple peaks corresponding to multi-bounce light from *all* illuminated points – causing prior work to fail. Our investigation of multiplexed illumination is motivated by its use on high-resolution SPADs [Henderson et al. 2019] found on consumer devices, such as mobile phones, tablets, and headsets [4sense 2021; Allain 2022].

Extracting 3D information from transients with multiplexed illumination is challenging due to the ambiguity in mapping *peaks* in the transient to corresponding illumination *points*. In this work, we demonstrate the potential of ML to address this challenge by *demultiplexing* captured transients using data priors (Fig. 1b). While ML's application to RGB images has transformed the field of computer vision, single-photon lidar has only recently emerged as a common sensor on consumer devices – meaning large-scale datasets and ML approaches do not yet exist. Yet, single-photon lidars capture a rich set of features that RGB cameras cannot. We posit that by harnessing these features, ML could enable a new set of abilities in computer vision. Our work is intended as an initial step towards this vision by (1) building the first-of-its-kind simulated multi-bounce transient dataset on ~100k scenes, and (2) applying it to train our proposed approach, Shoot-Bounce-3D (SB3D), that recovers metric 3D reconstructions, including in occluded areas and scenes with specular surfaces – from a single-view, single-shot capture (Fig. 1c).

SB3D consists of three steps. First, we estimate the two-bounce time of flight (ToF) of each illumination point. To do this, we train a model to estimate dense depth, which can be directly used to compute the two-bounce ToF for each illumination point. However, not all scene points are illuminated by each illumination point. Scene points that are not illuminated are in shadow; thus, we next train our model to estimate shadow maps for each illumination point. Our model uses the earlier predicted two-bounce ToF for this step. Finally, once both two-bounce ToF and shadows have been predicted for each illumination point, we train an existing method for neural reconstruction to learn 3D scene geometry, including in occluded regions. Because our dataset contains specular surfaces, such as mirrors and windows, our method is robust to these everyday objects. Interestingly, we find that the features learned to demultiplex two-bounce ToF in the first step can also be used to accurately predict specular segmentations. We posit this is a sign that the features may be a generalizable representation for single-photon lidar. Dataset and code will be released upon acceptance.

1. Data-Driven Demultiplexing: From a single-photon lidar measurement of a scene illuminated at multiple points at once, we propose a data-driven method to decompose the two-bounce signal, enabling separation of two-bounce time of flight and shadows.

2. Occlusion-Aware 3D: We show that the demultiplexed two-bounce ToF and shadows can be used for 3D reconstruction, enabling occluded areas to be revealed, despite the presence of specularities.

3. Large-Scale Multi-Bounce Lidar Dataset: To enable the above contributions, we introduce a dataset of ~100k simulated multi-bounce transient measurements of indoor scenes. This dataset can be used to drive future work in ML for single-photon lidar.

4. Generalizable Multi-Bounce Lidar Features: We find that our demultiplexing model learns features that can be transferred to other tasks, such as specular object segmentation, in simulation. This is a step towards a generalizable multi-bounce transient representation.

Scope of this Work. While our work is motivated by the use of high-resolution SPADs [Henderson et al. 2019; Kumagai et al. 2021] used with multiplexed point illumination on consumer devices [4sense 2021; Allain 2022], we acknowledge that these sensors also introduce a variety of practical challenges that are beyond the scope of this work, such as cross talk, hot pixels, blooming, and dead time. Instead, the purpose of our work is to (a) introduce a data-driven approach to demultiplexing illumination in multi-bounce flash lidars, and (b) show proof-of-concept results. We do not consider low-resolution SPADs with diffuse illumination [AMS OSRAM 2023; Jungerman et al. 2022]. We assume objects are purely specular or diffuse. While high-resolution consumer SPADs continue to be developed, they are not yet widely available off the shelf. As this technology continues to mature, we expect our work to become increasingly relevant.

2 Related Work

3D from RGB. Recovering 3D information from a single RGB image is ambiguous due to occluded geometry, specular surfaces, and lack of correspondences. Recent foundation models for depth leverage large datasets, learning statistical correlations to address the correspondence ambiguity [Bochkovskii et al. 2024; Guo et al. 2025;

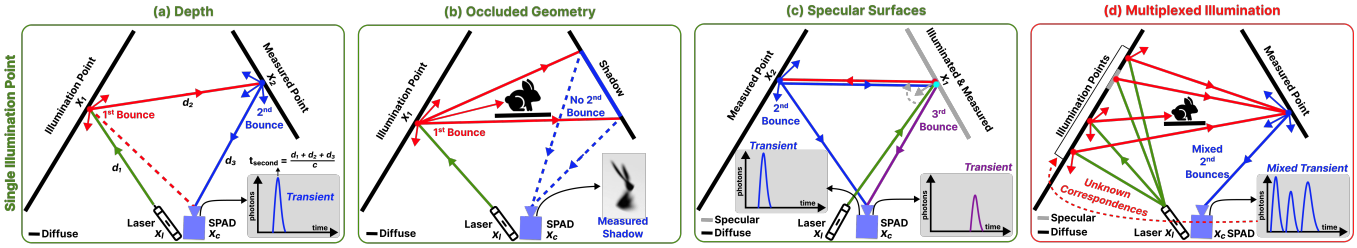

Fig. 2. **Multi-Bounce Signals.** Shoot-Bounce-3D leverages multi-bounce signals measured from single-photon lidar. Multi-bounce light encodes **(a)** dense depth (from geometric constraints), **(b)** occluded geometry (from shadows), and **(c)** specular surfaces (from two- and three-bounce pairs), but existing techniques assume a single scene point is illuminated at a time, scanning a laser over the scene. However, multi-bounce lidars on consumer devices instead use **(d)** *multiplexed* illumination, meaning multiple points are illuminated at once – causing existing methods to fail due to (1) lack of correspondence between two-bounce peaks and illumination points, and (2) mixing of signals from (a), (b), and (c). To resolve these ambiguities, we employ a learning-based technique.

Ke et al. 2024; Yang et al. 2024a,b]. Diffusion models are widely used to generate 3D from a single image, including occluded geometry. Several methods generate novel views from an RGB image [Liu et al. 2023; Qian et al. 2024; Sargent et al. 2024; Tewari et al. 2023; Yu et al. 2024], though they often focus on objects and struggle in scenes. Starting with Yu et al. [2021], many techniques incorporate data priors in neural radiance fields (NeRF) [Mildenhall et al. 2021] to learn 3D geometry from single or few images [Gao et al. 2024; Xu et al. 2022]. However, NeRF often struggles with the challenge of specular surface materials [Ma et al. 2024; Tiwary et al. 2023; Verbin et al. 2024]. More broadly, specular surfaces cause ambiguities, leading to "portals" being hallucinated. Work by He et al. [2021] and Yang et al. [2019] tries to address this by learning mirror and glass segmentation from RGB. Despite the progress, each of these challenges remains an open problem; rather than relying on RGB, we explore using single-photon lidar for scene-level 3D.

3D from Single-Photon Lidar. Single-photon lidars offer additional cues for 3D understanding by capturing the distribution of light intensity with travel time, called *transients*. Direct reflections encoded in transients enable photon-efficient depth imaging [Gupta et al. 2019; Heide et al. 2018; Shin et al. 2016]. Recent works also explore multi-view lidar measurements for neural 3D reconstruction [Behari et al. 2024; Malik et al. 2024; Mu et al. 2024]. Three-bounce information in transients has been extensively used for looking around corners using non-line-of-sight imaging [Ahn et al. 2019; Kirmani et al. 2009; Liu et al. 2019; Maeda et al. 2019; O'Toole et al. 2018; Pediredla et al. 2019; Shen et al. 2024; Velten et al. 2012]. In this work, we focus on two-bounce light – which has higher signal quality than three-bounce light due to one less scattering attenuation. Prior works leverage two-bounce light for dense depth from sparse illumination [Henley et al. 2022], seeing behind occluders [Henley et al. 2020] and specular surface mapping [Henley et al. 2023] – but require sequentially illuminating the scene one point at a time. Single-photon lidars on consumer devices have multiplexed illumination [4sense 2021] – captured transients with mixed light contributions complicate 3D understanding. Lin et al. [2024] explore specular surface mapping with multiplexed illumination, but require multiple measurements from different viewpoints. Somasundaram et al. [2023] explore occluded object imaging from multiplexed transients using an analytical approach. We leverage learning to *demultiplex* the captured two-bounce transient to recover depth and occluded 3D geometry in scenes with both diffuse and specular objects in a single shot.

Data-Driven Methods for Single-Photon Lidar. There is a growing interest in applying deep learning methods to single-photon lidar data. Prior works leverage direct bounce information in transients for data-driven depth estimation [Lindell et al. 2018; Nishimura et al. 2020; Peng et al. 2020; Plosz et al. 2023; Sun et al. 2020; Yang et al. 2022; Zang et al. 2021], human pose [Ruget et al. 2022] and activity recognition [Mora-Martín et al. 2024]. There are also works on non-line-of-sight imaging from three-bounce transients that develop simulated datasets [Chen et al. 2020] and novel convolutional neural network-based [Chen et al. 2020; Cho et al. 2024; Mu et al. 2022; Sun et al. 2024, 2020; Zhu et al. 2023], transformer-based [Li et al. 2023; Yu et al. 2023] and motion-aware [Chopite et al. 2025; Isogawa et al. 2020; Ye et al. 2024] models. To the best of our knowledge, we are the first to bring data-driven methods to two-bounce transients.

3 Single-Photon Lidar Image Formation Model

The most common measurement model for lidar only accounts for *direct illumination*: emitted light that directly reflects back to the sensor. Any other returning light is due to *indirect illumination* – light that interacted with or originated from other parts of the scene – and is treated as ambient noise. However, the secondary light paths from indirect illumination encode useful scene properties, and can be more informative than direct paths. In this section, we show how complex indirect illumination captured by lidar serves as a useful cue for 3D scene geometry, occlusions, and specularity. The ToF dimension is particularly valuable for light transport decomposition when multiple scene points are illuminated simultaneously.

3.1 Multi-Bounce Light Transport

Indirect illumination in single-photon lidar arises due to *multi-bounce* light paths. Multi-bounce paths occur when the illuminated scene point becomes a *virtual light source*. A virtual source acts as a light source by reflecting incident light from another source towards other scene points [Henley et al. 2022]. In this work, a virtual light source can either have isotropic or directional radiance, depending on whether the scene point is diffuse or specular, respectively.

An n-bounce path is a light path that consists of exactly n surface reflections before returning to the camera. An example of a multi-bounce path is shown in Fig. 2a. A scene point $\mathbf{x}_1$ is illuminated by the laser located at $\mathbf{x}_l$ and imaged by the sensor located at $\mathbf{x}_c$. The light that travels along the path $\mathbf{x}_l \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_c$ is referred to as *1-bounce light*. Similarly, light that travels along the path $\mathbf{x}_l \rightarrow \mathbf{x}_1 \rightarrow$

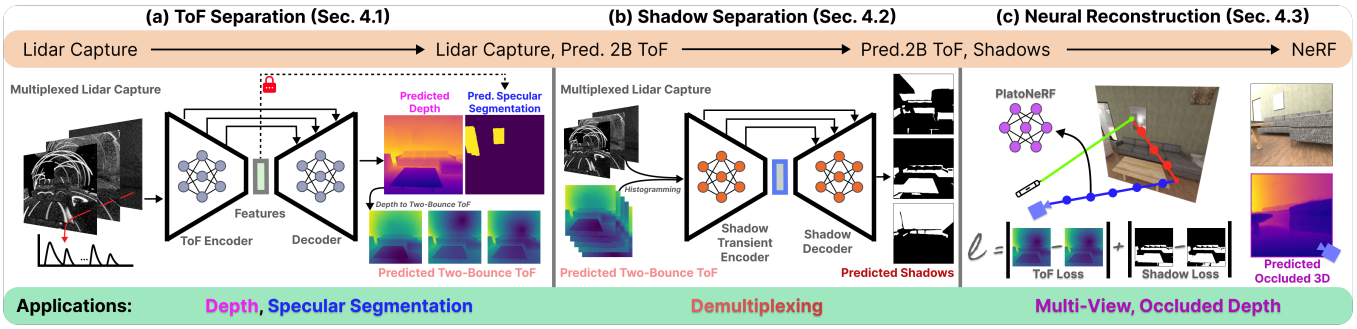

Fig. 3. **Method overview.** Shoot-Bounce-3D (SB3D) performs 3D reconstruction from a single lidar measurement. The pipeline consists of three steps, each with its own output. **(a)** First, from a measurement taken with multiplexed illumination (meaning multiple points in the scene are illuminated at once), SB3D is trained to learn to predict depth – allowing the 2-bounce time-of-flight (ToF) for each illumination point to be separated using ray geometries. Because our scenes contain specular objects, we find the ToF encoder used for this step also learns features that enable specular object segmentation. **(b)** The predicted 2-bounce ToF is unprojected into histograms and used with the lidar measurement to estimate shadows. Using the 2-bounce ToF allows the network to learn *shadow transients*, improving performance. **(c)** Finally, usin g the predicted 2-bounce ToF and shadows, PlatoNeRF can be trained for 3D reconstruction.

$\mathbf{x}_2 \rightarrow \mathbf{x}_c$ is referred to as *2-bounce* light. The presence, pathlength, and bounce order (i.e., 1-, 2-, or 3-bounce) of multi-bounce light is indicative of the geometry and materials present in a scene.

Depth. A key benefit of multi-bounce light is that a scene point doesn't have to be directly illuminated to infer its properties [Henley et al. 2022; Klinghoffer et al. 2024]. Consider the illuminated scene point $\mathbf{x}_1$ and non-illuminated scene point $\mathbf{x}_2$ shown in Fig. 2a. The pathlength of the 1-bounce light can be used to infer the 3D position of $\mathbf{x}_1$ using conventional time-of-flight techniques [Charbon 2014]. Once the location of $\mathbf{x}_1$ is recovered, the location of $\mathbf{x}_2$ can be computed. The light travels a distance of $d_2 + d_3$ along the path $\mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \mathbf{x}_c$. This distance constrains the possible locations for $\mathbf{x}_2$ to be on the surface of an ellipsoid, with foci at $\mathbf{x}_1$ and $\mathbf{x}_c$ and major axis length $d_1 + d_2$. We also know that $\mathbf{x}_2$ must lie along the pixel viewing direction. Taken together, these two constraints uniquely determine the location of $\mathbf{x}_2$. In this way, 2-bounce light can provide a physical cue for depth even if the scene point isn't directly illuminated.

Occlusions. Multi-bounce light can also probe parts of the scene that aren't directly visible to the camera. Consider the example in Fig. 2b, where a bunny is behind an obstacle and therefore outside the camera's line of sight. Here, the presence or absence of two-bounce light measured at $\mathbf{x}_2$ is an indication of the presence or absence of an object behind the occluder. If 2-bounce light is absent at $\mathbf{x}_2$ (i.e., $\mathbf{x}_2$ is in shadow), then an object lies along the ray connecting $\mathbf{x}_1$ and $\mathbf{x}_2$. By analyzing the two-bounce intensities (i.e., shadows) along the entire surface on the right wall, the shape of the occluded object can be inferred [Henley et al. 2020].

Specular Surfaces. A useful cue to identify specular surfaces is that light returning to the sensor from the specular surface will always arrive after (i.e., have a longer pathlength than) light returning to the sensor from a diffuse surface. For example, consider the case where the virtual source at $\mathbf{x}_1$ is specular, as shown in Fig. 2c. In this case, 2-bounce and 3-bounce light will be observed. The 2-bounce light will travel along the path $\mathbf{x}_l \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \mathbf{x}_c$, and corresponds to light returning from the diffuse surface. The observed 3-bounce light will travel along the path $\mathbf{x}_l \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_c$, due to the geometry of specular geometry, and corresponds to light returning from the specular surface. The three-bounce pathlength will be longer than the two-bounce pathlength due to the triangle inequality theorem. A similar argument can be made for the case that $\mathbf{x}_1$ is diffuse and $\mathbf{x}_2$ is specular, in which case 1-bounce light would be observed from the diffuse surface and 2-bounce light would be observed from the specular surface. The resulting property provides a natural cue for detecting the presence of mirrors in a scene.

3.2 ToF-Based Multi-Bounce Path Separation

ToF cameras, such as lidars, can measure the presence and pathlength of multi-bounce light, and separate different light bounces at a pixel due to their high timing precision (picosecond scale). These lidars are now widely available on consumer devices, making them a promising sensing modality for occlusion-aware 3D.

Single-Photon Lidar. A single-photon lidar system consists of a pulsed illumination source and a 2D SPAD array. The SPAD array consists of $n_\mathrm{x} \times n_\mathrm{y}$ pixels, and each pixel captures a temporal histogram of n_t bins. The kth bin of the histogram contains the number of detected photons in the time interval $[k\Delta, (k+1)\Delta]$, where Δ is the temporal bin width of the sensor. The resulting SPAD measurement $\mathbf{i} \in \mathbb{R}^{n_\mathrm{x} \times n_\mathrm{y} \times n_\mathrm{t}}$ is a 3D data cube. There is also a $n_{\mathrm{x_l}} \times n_{\mathrm{y_l}}$ laser spot illumination grid within the field of view of the camera, where $n_{\mathrm{x_l}} \ll n_\mathrm{x}$ and $n_{\mathrm{y_l}} \ll n_\mathrm{y}$. Prior works have shown that 3D geometry and specular surfaces can be recovered from $\mathbf{i}$ when each laser spot is illuminated sequentially [Henley et al. 2022, 2023].

Multiplexed Illumination. In practice, single-photon lidars found on consumer devices typically do not illuminate one laser source at a time: they emit all laser spots simultaneously as shown in Fig. 2d. This multiplexed illumination produces an ambiguous signal that integrates the contributions of all laser spots into a single measurement. To recover scene properties, as discussed in Sec. 3.1, we must "unmix" the signal contributions from each light source.

Practical Challenges. Demultiplexing the contributions of each virtual source is challenging because many of the multi-bounce paths will produce similar pathlengths, making inversion highly ill-posed. Analytical methods for demultiplexing based on linear

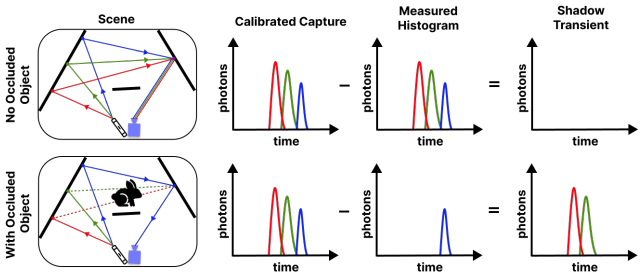

Fig. 4. **Shadow Transients.** We leverage the idea of shadow transients to improve network training for shadow demultiplexing. The key idea is to estimate the light that never reached the sensor due to the object casting a shadow. In the top row, there's no object, so the shadow transient is empty. In the bottom row, two of the three light paths are blocked, so only one peak shows up in the measurement. The shadow transient, on the other hand, measures the two light sources that were blocked by the occluded object. In practice, the calibrated capture is estimated from the data, not measured. As a result, the calibrated capture is input to the network with the measured histogram to prevent errors due to inaccurate shadow transient estimation.

models [Somasundaram et al. 2023] and heuristic algorithms [Henley et al. 2023; Lin et al. 2024] are based on simplifications of the underlying light transport that do not generalize to natural scenes. Our key insight is to use a data-driven approach that can learn features directly from the physically-constrained, but difficult to model, cues present in transient measurements.

4 Data-Driven Demultiplexing from Two-Bounce Lidar

The key challenge that we must solve to leverage multi-bounce light transport in single-photon lidar is demultiplexing illumination. Each laser spot will result in a complex mixture of shadows, specular reflections, and multi-bounce reflections. The presence of multiple laser spots during illumination will further complicate the light transport during capture. The demultiplexing problem entails determining the light transport caused by each individual laser spot.

In practice, directly reconstructing the per-laser-spot transient is challenging. Instead, we break the problem into several substeps based on the intuitions in Sec. 3.1. First, we predict the depth of the visible scene using a neural network (Sec. 4.1). From predicted depth, we estimate separate 2-bounce ToF for each laser spot by tracing the distances $\mathbf{x}_l \rightarrow \mathbf{x}_i \rightarrow \mathbf{x}_{u,v} \rightarrow \mathbf{x}_c$ for all laser spots $\mathbf{x}_i$ and scene points $\mathbf{x}_{u,v}$. For this step, we assume $\mathbf{x}_i$ is known from 1-bounce light. Second, we combine the predicted 2-bounce ToF with the multiplexed measurement to estimate the individual shadows caused by each laser spot (Sec. 4.1). Finally, we combine the 2-bounce ToF and shadows to reconstruct the 3D scene. Demultiplexing is performed in a data-driven manner, and final reconstruction is performed via neural rendering. The pipeline overview is in Fig. 3 and implementation details, such as architecture, are in the supplement.

4.1 Demultiplexing Two-Bounce Time-of-Flight

The 2-bounce ToF of light from individual laser spots is useful for obtaining the visible scene geometry. Predicting the ToF of each laser spot individually, however, is challenging due to the high dimensionality of the output data structure ($n_\mathrm{x} \times n_\mathrm{y} \times n_{\mathrm{x_l}} n_{\mathrm{y_l}}$). Instead, we use depth estimation as a proxy task. The depth is a lower-dimensional quantity ($n_\mathrm{x} \times n_\mathrm{y}$) and can subsequently be used to compute the 2-bounce ToF for each laser spot. We train an encoder-decoder to directly learn depth from $\mathbf{i}$ in a supervised fashion. The loss for the depth network consists of a data fidelity and an edge-aware smoothness regularization term:

$$\mathcal{L}_\mathrm{depth} = \mathcal{L}_\mathrm{data} + \mathcal{L}_\mathrm{smooth}. \tag{1}$$

Following prior work in depth estimation [Godard et al. 2017, 2019], the data fidelity term consists of a weighted combination of SSIM [Wang et al. 2004] and an L1 loss

$$\mathcal{L}_\mathrm{data} = \alpha(1 - \mathrm{SSIM}(d, \hat{d})) + (1 - \alpha)|d - \hat{d}|_1, \tag{2}$$

where d is the ground-truth depth, $\hat{d}$ is the predicted depth, and α is a weight hyperparameter. We set α=0.15 as in past work [Godard et al. 2017]. We also find that an edge-aware smoothness loss provides a slight improvement in accuracy. This loss encourages the predicted depth to be smooth without blurring edges [Godard et al. 2017]. We use the time-integrated transient measurement $I_{u,v} = \sum_t \mathbf{i}(u, v, t)$ to obtain an estimate of the edges because we don't have access to an RGB image. The resulting loss is

$$\mathcal{L}_\mathrm{smooth} = \frac{\beta}{N} \sum_{u,v} |\partial_x \hat{d}_{u,v}| \mathrm{e}^{-|\partial_x I_{u,v}|} + |\partial_y \hat{d}_{u,v}| \mathrm{e}^{-|\partial_y I_{u,v}|}, \tag{3}$$

where $\hat{d}_{u,v} = \hat{d}(u, v)$, ∂_x and ∂_y are the gradients of the depth and intensity images, and β is a hyperparameter which we set to 10^{-3}.

From the predicted depth, we compute the 2-bounce time-of-flight for each illumination spot, assuming *no occlusions*

$$t_{2B}(\mathbf{x}_i, \mathbf{x}_{u,v}; \mathbf{x}_l, \mathbf{x}_c) = \frac{|\mathbf{x}_i - \mathbf{x}_l| + |\mathbf{x}_{u,v} - \mathbf{x}_i| + |\mathbf{x}_{u,v} - \mathbf{x}_c|}{\mathrm{c}}, \tag{4}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ is the location of the ith virtual source (known from 1-bounce) and $\mathbf{x}_{u,v} \in \mathbb{R}^3$ is the location of the scene point imaged by pixel (u, v), which can be computed via depth unprojection.

4.2 Demultiplexing Shadows

We use the shadows cast by the hidden objects from each illumination spot as the cue for occlusions. Therefore, we aim to recover a set of binary shadow masks $\{\mathbf{s}_1, ..., \mathbf{s}_{n_{\mathrm{l_x}} n_{\mathrm{l_y}}}\}$, where $\mathbf{s}_j \in \mathbb{R}^{n_\mathrm{x} \times n_\mathrm{y}}$, *for each illumination spot* from the multiplexed measurement $\mathbf{i}$.

We find that directly training a model to learn $n_{\mathrm{l_x}} n_{\mathrm{l_y}}$ shadow masks from the multiplexed measurements, similar to the approach in Sec. 4.1, does not work well. Conditioning the input on the laser spot index similarly does not work well. One possible explanation for poor performance with this approach is due to the inconsistency between the network input and output. The input multiplexed measurements measures the *net amount of light* arriving to the sensor from all illumination spots. The output shadow masks, on the other hand, predicts the *absence of light* from each illumination spot.

To handle this misalignment, we modify our input based on the concept of *shadow transients* [Somasundaram et al. 2023]. The shadow transient $\mathbf{i}_\mathrm{shadow}$ was used in prior work to linearize the forward model for 2-bounce occluded imaging [Somasundaram et al. 2023], and helps align the network input and output in our case. Shadow transients measure the multi-bounce light that *doesn't reach* the sensor due to obstructions from the occluded objects. The

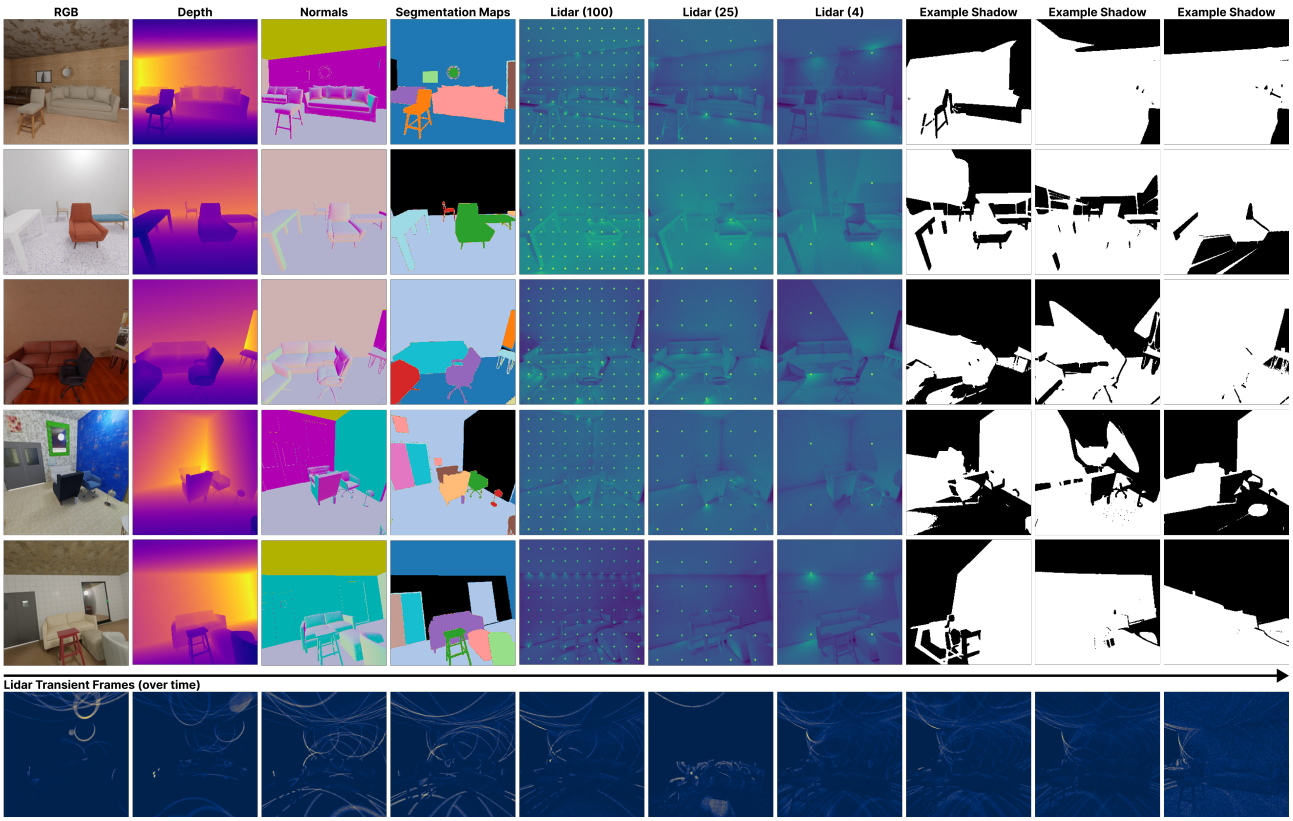

Fig. 5. **Proposed Dataset.** Samples from our simulated dataset of multi-bounce transients for ~100k scenes. Our dataset also contains RGB, depth, normals, and segmentation maps for each scene. Transients are simulated with varying amounts of multiplexed illumination – shown as the intensity maps. Binary shadow maps are provided for each illumination point. The last row shows frames of the simulated lidar transient for an example scene.

shadow transient is $\mathbf{i}_{\text{shadow}} = \mathbf{i}_{\text{calib}} - \mathbf{i}$, where $\mathbf{i}_{\text{calib}}$ is a calibrated capture. This calibrated capture is performed by removing all occluded objects in a scene (leaving only objects that are directly visible to the camera), then capturing the transients, as shown in Fig. 4. Rather than removing all occluded objects, $\mathbf{i}_{\text{calib}}$ can be estimated as

$$\hat{\mathbf{i}}_{\text{calib}}(u, v, t) = \sum_{i=1}^{n_{x_l} n_{y_l}} A_i(u, v) \cdot \delta(t - t_{2B}(\mathbf{x}_i, \mathbf{x}_{u,v}; \mathbf{x}_l, \mathbf{x}_c)), \quad (5)$$

where $A_i(u, v)$ is the intensity of two-bounce light returning to pixel (u, v) from virtual source i and δ is the Dirac delta. We estimate $\hat{\mathbf{i}}_{\text{calib}}$ using the predicted two-bounce ToF from Sec. 4.1. However, A_i is unknown, resulting in inaccurate estimates of the shadow transients when subtracting $\mathbf{i}$ from $\hat{\mathbf{i}}_{\text{calib}}$. While past work required hyperparameter tuning for A_i, we instead set $A_i = 1$ and input both $\mathbf{i}$ and $\hat{\mathbf{i}}_{\text{calib}}$ to the network. We find that concatenating $\hat{\mathbf{i}}_{\text{calib}}$ to the input $\mathbf{i}$ significantly improves predicted shadow mask quality. We use a binary cross entropy loss to train the network.

4.3 Single-Shot 3D Reconstruction

We use the predicted 2-bounce ToF from Sec. 4.1 and the predicted shadows from Sec. 4.2 to train PlatoNeRF [Klinghoffer et al. 2024], a neural reconstruction model to learn 3D geometry. PlatoNeRF requires a separate 2-bounce ToF map and shadow mask for each laser spot. By using the output of our SB3D, PlatoNeRF can be trained without modification, yielding the ability to render depth from extreme novel views that reveal occluded geometry. Because the contributions from all laser spots were captured simultaneously through multiplexed illumination, SB3D enables single-shot capture of the light transport needed for 3D reconstruction. Please refer to the supplement for a comprehensive review of PlatoNeRF.

4.4 Single-Photon Lidar Feature Generalization

We take inspiration from existing work in represention learning for RGB images [Kolesnikov et al. 2019; Tian et al. 2020] and apply it to the context of single-photon lidar. We observe that the features learned from demultiplexing ToF in Sec. 4.1 can also be used to perform other tasks, such as specular surface segmentation. To do so, we freeze the pre-trained encoder and train a randomly initialized decoder to predict binary segmentation masks from the learned features. The specular segmentation decoder is supervised with ground-truth segmentation masks and a binary cross-entropy loss.

5 Shoot-Bounce-3D Transient Dataset

One of our contributions is a large-scale dataset of simulated lidar transients (Fig. 5), built on top of the Aria Synthetic Environments

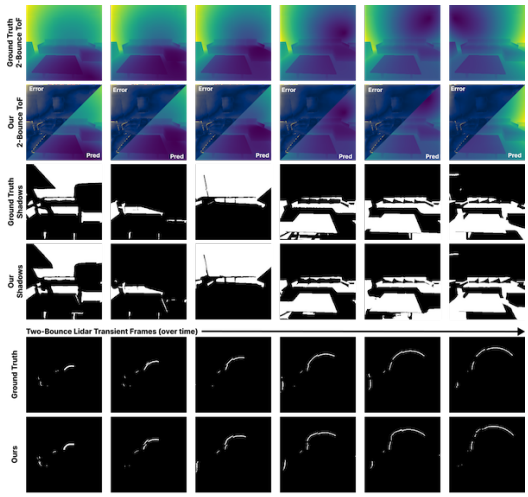

Fig. 6. **Demultiplexing Results.** We show qualitative results for demultiplexing both time of flight (ToF) and shadows. Each column denotes a different illumination source – our method extracts the two-bounce ToF and shadow maps for each from the multiplexed lidar measurement. The last row shows frames from the predicted "light in flight" video (video provided in the supplement), which is rendered by combining the predicted two-bounce ToF and shadows into a transient measurement, allowing visualization of two-bounce light propagation per illumination point.

(ASE) dataset [Avetisyan et al. 2024]. While there are many point-cloud lidar datasets, only limited single-photon lidar datasets exist. Our dataset contains transients capturing multi-bounce light transport (1st, 2nd, 3rd, etc bounces) that can be exploited for a diverse range of tasks. To the best of our knowledge, this is the first large-scale transient dataset, with past datasets containing ~5,000 simulated transients [Gutierrez-Barragan et al. 2021]. We render one 256×256 transient at 128 ps temporal resolution for each of the 97,432 ASE scenes (assembled from ~8,000 unique objects) we used. These ASE scenes were procedurally created with SceneScript [Avetisyan et al. 2024], which was shown to produce scene geometry sufficiently realistic for real-world generalization. Renderings include single-photon lidar (4, 25, and 100 illumination points), RGB, depth, normals, specular segmentation, instance segmentation, and binary shadow maps. In addition, data is rendered at the same poses as in the ASE dataset, allowing both datasets to be used in conjunction. We leverage the lidar transients, ground-truth depth, specular segmentation masks, and shadow masks in our work. More details are available in the supplement.

6 Experiments

We present simulated results for demultiplexing, depth estimation, specular segmentation, and occlusion-aware 3D using data containing 25 illumination points. We use ~87k samples for training and 6k for test metrics. We end with proof-of-concept real-world results.

6.1 Demultiplexing Results

First, we investigate the ability of our model to decompose two-bounce light into separated two-bounce ToF and shadows per illumination point. Qualitative results are shown in Fig. 6 for (a) demultiplexing ToF, (b) demultiplexing shadows, and (c) re-rendering two-bounce transients, which can be visualized as "light in flight" videos [Velten et al. 2013]. The transient video for the ith laser spot can be rendered by applying the summand in Eq. (5) and setting $A_i(u, v) = \mathbf{s}_i(u, v)$. The last result enables visualization of light propagation per illumination point, as shown in the supp. video. The mean absolute test error for two-bounce pathlength was 0.2736 m. While correlated to depth error, this error is higher due to the longer paths of two-bounce light. The pixel mean absolute error (MAE) and IoU for predicted shadow maps was 0.0214 and 95.3%, respectively.

Table 1. **Qualitative Results.** We report metrics for each task. For depth and specular segmentation, metrics are computed over 6k test samples. For 3D reconstruction, metrics are computed for predicted multi-view depth, averaged over four scenes (shown in Fig. 7) with 80 novel test views each.

(a) Depth Estimation

Approach	MAE ↓	F1 Boundary ↑
Bounce Flash Lidar	0.4922	0.0138
CompletionFormer	0.4394	0.0066
Depth Anything V2	0.1640	0.1999
Depth Pro	0.1089	0.2930
Shoot-Bounce-3D	**0.0228**	**0.6238**

(b) Specular Segmentation

Approach	Pixel MAE ↓	IoU (%) ↑
EBLNet	0.0117	81.21
Shoot-Bounce-3D	**0.0010**	**86.52**

(c) Occlusion-Aware 3D Reconstruction

Approach	MAE ↓	F1 Boundary ↑
ZeroNVS	0.5619	0.0090
Shoot-Bounce-3D	**0.0983**	**0.2725**
PlatoNeRF Oracle	0.0950	0.3317

6.2 Depth Results

Our method is able to predict accurate dense depth from a single image. While dense depth can also be recovered by using a lidar with diffuse illumination and dense pixel array, existing consumer devices such as the iPhone use point illumination [4sense 2021; Allain 2022], which results in sparse depth but better range. Our method, which also uses point illumination, also recovers dense depth due to its use of both one-bounce and two-bounce paths.

Baselines. We compare to a physics-based lidar approach, a learned RGB/lidar fusion approach, and an RGB depth foundation model:

1. Bounce-Flash (BF) Lidar [Henley et al. 2022] uses geometric constraints to estimate depth from two-bounce lidar. Because BF lidar assumes scanned illumination, we adapt it to the multiplexed setting by using it to compute depth candidates for all (two-bounce peak, illumination point) pairs, via peak finding. Intuitively, if a scene point is not in shadow for two illumination points, then BF Lidar will yield approximately the same depth for both. We compute the mode of the discretized depth candidates to compute depth.

2. CompletionFormer (CF) [Zhang et al. 2023] recovers dense depth from monocular RGB and sparse lidar. We use RGB images rendered at the same view as our lidar transients and depth from our sparse illumination points (computed from first bounce) as input.

3 & 4. Depth Anything V2 & Depth Pro [Bochkovskii et al. 2024; Yang et al. 2024b] are RGB foundation models for monocular depth,

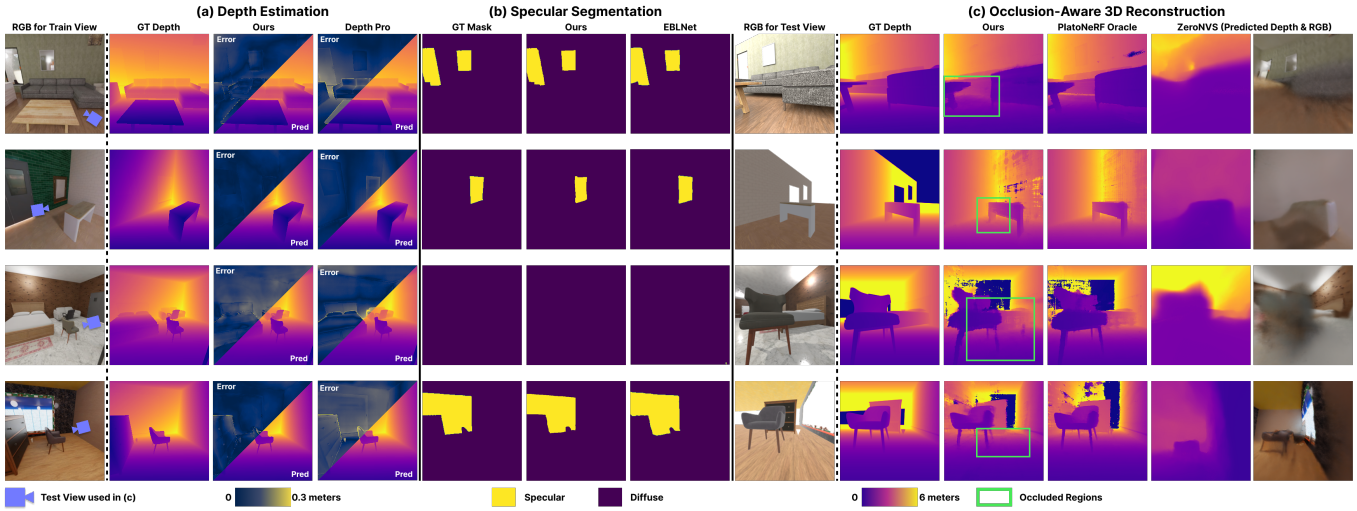

Fig. 7. **Qualitative Results.** We show qualitative results for the tasks of **(a)** depth estimation, **(b)** specular surface segmentation, and **(c)** occlusion-aware 3D reconstruction – both for our method and the best baseline per task (as well as PlatoNeRF oracle for 3D reconstruction). Our method consistently generates interpretable and accurate results across each of the scenes. For 3D reconstruction, we show novel views that lie in regions occluded from the training view, demonstrating our method's ability to predict demultiplexed shadows that enable inference of hidden geometry.

trained on large sets of real and simulated RGB. Although both models can compute metric depth, we found this inaccurate. We rescaled predictions via a least squares regression on 25 anchor points, acquired from the one-bounce returns in our lidar measurements.

Metrics. We compute metric depth mean absolute error and boundary F1 (a scale-invariant metric defined by Bochkovskii et al. [2024]).

Results. We find that our method significantly outperforms the baselines. Quantitative results are provided in Tab. 1a and qualitative results and error maps for our method and Depth Pro are shown in Fig. 7. Depth Pro produces qualitatively similar depths as the ground truth, but struggles to preserve scale, even after rescaling with anchor points. In addition, we notice higher depth error around edges with Depth Pro compared to our method. Because BF Lidar has no learnable mechanism for denoising or infilling, it produces noisy depth maps and is unable to resolve depth for shadowed pixels. CF struggles when provided depth from only 25 points. Even when 100 points are provided, it still achieves only 0.149 m MAE.

6.3 Specular Surface Segmentation Results

Baselines. We compare our method for specular segmentation to EBLNet [He et al. 2021], which learns to segment glass and mirrors from an RGB image. For fair comparison, we retrain EBLNet on our dataset. We considered lidar comparisons, but found that existing methods either do not focus on multiplexed illumination [Henley et al. 2023] or use a different hardware setup [Lin et al. 2024].

Metrics. Specular surface segmentation is a binary segmentation task where values of one indicate a specular surface and values of zero indicate a diffuse surface. We report pixel mean absolute error ($\sum_{u,v} |s_{u,v} - \hat{s}_{u,v}|$, where s and $\hat{s}$ are ground-truth and predicted segmentation values per pixel), and intersection over union (IoU).

Results. We find that our method is able to detect mirrors with high accuracy, outperforming EBLNet. We posit the increase in performance is due the availability of physical cues correlated to specular surfaces (Fig. 2c), whereas, in RGB images, detecting specular surfaces is inherently ambiguous (a specular surface and a "portal" often look identical). While our work focuses on two-bounce signals, since we use the full transient as input to our model, it may also use three-bounce signals. We ablate this further in the supplement.

6.4 3D Reconstruction Results

Finally, we evaluate SB3D's 3D reconstruction quality. Recall from Fig. 3 that 3D reconstruction is done per-scene by supervising PlatoNeRF with SB3D's predicted two-bounce ToF and shadows.

Baselines. We compare our method to the following approaches. We also attempted a comparison with Somasundaram et al. [2023], but the method assumes that a calibrated capture $\mathbf{i}_{\text{calib}}$ is available, as described in Sec. 4.2, which is not the case in practice.

1. ZeroNVS [Sargent et al. 2024] trains a NeRF from a single RGB image via score distillation sampling of a diffusion model.

2. PlatoNeRF Oracle [Klinghoffer et al. 2024]: Our method is built atop PlatoNeRF – a recent method for 3D reconstruction from two-bounce lidar. Therefore, we train an "oracle" PlatoNeRF model from ground truth two-bounce ToF and shadows for each scene to disentangle performance of our method and performance of PlatoNeRF.

Metrics. We compute MAE and boundary F1 on multi-view depth rendered from NeRF as a proxy for 3D reconstruction.

Results. Qualitative results for our method and ZeroNVS are shown in Fig. 7c. We find that our method is able to accurately reconstruct not only visible regions – but also occluded regions, yielding interpretable and detailed geometry. ZeroNVS is able to discern coarse

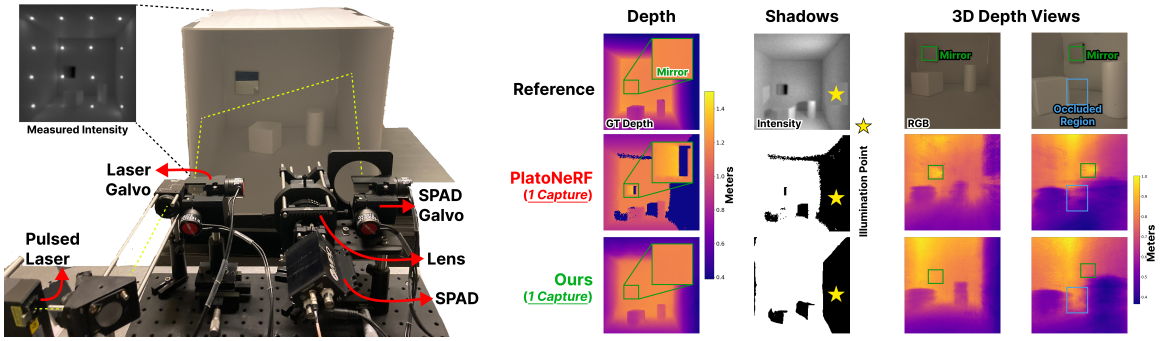

Fig. 8. **Real-World Results.** We provide proof-of-concept real-world results on a new dataset that we capture with multiplexed illumination (16 laser spots). We compare our method to PlatoNeRF for depth, shadows, and 3D depth views. Since our goal is single-shot 3D, we restrict PlatoNeRF to a single capture, which, since it is unable to handle multiplexing, means it is trained with 1 illumination point. Our method, on the other hand, learns to demultiplex 16 illumination points, leading to better performance, especially in specular regions. A comparison to PlatoNeRF with more captures is provided in the supplement.

structure – and in some cases, such as the last scene, carve out empty space in occluded regions, but fails to recover detailed geometry. Lack of detailed geometry may be because of geometric inconsistencies that emerge from training with a diffusion model. In addition, ZeroNVS relies entirely on hallucination in regions that are fully occluded, whereas SB3D relies on a physically meaningful quantity – demultiplexed shadows. Although ZeroNVS struggles to perform accurate novel view synthesis in the extreme views and occluded areas emphasized in this work, we note that it performs better with small view changes without occlusion, as shown in Fig. 9, yielding better, albeit still less accurate depth. Figure 9 also highlights the importance of understanding specular surfaces during 3D reconstruction, which our method is able to do due to earlier steps implicitly learning about them when demultiplexing.

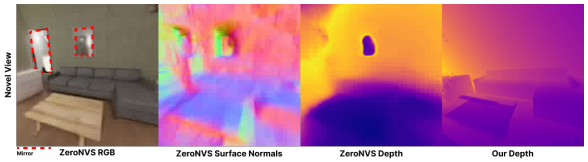

Fig. 9. **Portals in Mirrors.** RGB-based methods – such as ZeroNVS (shown above) – are prone to hallucinating "portals" inside mirrors due to the lack of cues to distinguish the mirror from a physical space. By leveraging multi-bounce transients, our method can handle specular surfaces, such as mirrors, enabling accurate 3D reconstruction even in the presence of mirrors.

6.5 Real-World Results

We provide an overview of our real-world results and refer the reader to the supplement for more details and discussion.

Dataset. We collect a real-world dataset by scanning a single-pixel SPAD (MPD PDM Series) over the field of view of the scene (containing a cube and cylinder inside a room with a mirror on the wall) using a two-axis scanning galvanometer (ThorLabs GVS412). This process is repeated for 16 illumination points and the transients are summed to create a multiplexed measurement for testing. The real-world data is 256×256 with a temporal resolution of 32 ps.

Results. To validate our method, we retrain our models with a simulated dataset of 10k scenes containing a cube, cylinder, and mirror randomly placed inside a room of varying scale. During training, we add Poisson noise and Gaussian timing jitter to the transients. We then test the models on the real-world dataset and find that they are able to recover accurate depth and shadow masks, enabling 3D reconstruction and outperforming PlatoNeRF in the single-shot setting (Fig. 8). Our reconstruction is created by manually selecting the 4 best shadow masks, since, as shown in the supplement, some predictions contain artifacts. These artifacts can be mitigated in the future by improving SNR of test data, incorporating more realistic noise in simulated training data, and introducing real-world data in training. Our depth error is 0.028 m and boundary F1 is 0.556. We find that as the number of captures used to train PlatoNeRF increases, so does its performance, but SB3D remains competitive, as shown in the supplement. These results demonstrate that the proposed method can be successfully extended to real-world data.

6.6 Ablations

We find our method continues to work when retrained on separate datasets with 4 or 100 illumination points (Fig. 11). We also ablate adding realistic pulse shapes, noise, and timing jitter on a simplified dataset in Fig. 10. More details on these ablations, as well as additional ablations, including on shadow estimation, training on 1- and 2-bounce light only, amount of training data, out-of-distribution geometry, and temporal resolution, are in the supplement.

7 Conclusion

We present a method for single-shot estimation of depth and 3D scene geometry in the presence of specularities using single-photon lidars with *multiplexed* illumination. At the heart of our method is the first large-scale simulated dataset of ~100k lidar transients, which enables our learned technique for *demultplexing* ToF and shadows from a single lidar measurement. Not only does demultiplexing enable 3D reconstruction, but we also find that the learned features could generalize to other tasks, such as specular segmentation.

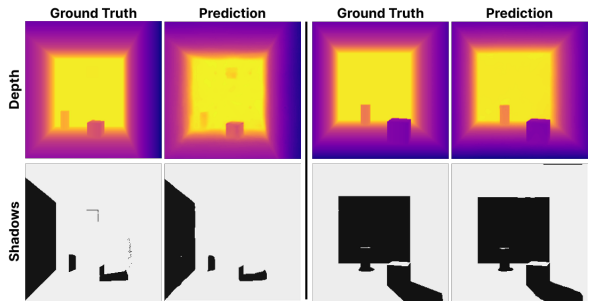

Fig. 10. **Pulse Shape, Noise, & Timing Jitter.** To demonstrate that the ideas introduced in this work can generalize to realistic pulse shapes, noise, and timing jitter, we (a) convolve our simulations with real-world sensor pulses, (b) add Poisson noise to the histogram intensities (such that 2-bounce peak photon counts range from 10 to 400), and (c) add Gaussian timing jitter to the histogram peaks (50 ps full width at half max). Accurate depth and shadow estimation show robustness to some practical challenges with lidar.

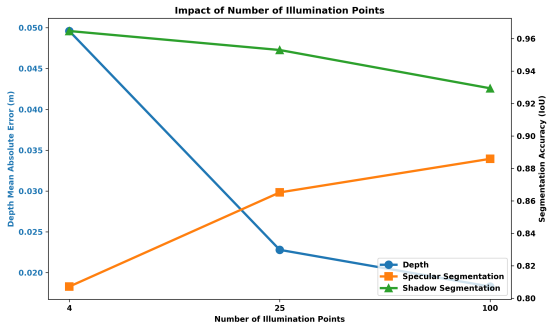

Fig. 11. **Ablation: Number of Illumination Points.** We show the trade-off in performance as the number of illumination points is varied from 4 to 25 to 100. While depth and specular surface estimation improve with more illumination points – despite increased ambiguities from multiplexing – shadow mapping becomes less accurate. Therefore, a Pareto optimum exists that balances the accuracy of all three tasks.

Limitations. Due to our pipeline design, errors from ToF separation propagate to shadow separation. In addition, this work does not consider practical challenges that arise on consumer SPADs, such as cross talk, hot pixels, dead time, and blooming. Our work is a step towards enabling single-shot 3D on these sensors.

Future Work. Our dataset and method open opportunities for future work in single-photon lidar foundation models and fusion with RGB images. In addition, future work is needed to automate the selection of top shadow mask predictions and to reduce the propagation of errors, potentially through end-to-end methods.

Acknowledgments. We thank Suvam Patra and Armen Avetisyan for support with Aria Synthetic Environments. Tzofi Klinghoffer is supported by the Department of Defense (DoD) National Defense Science and Engineering Graduate (NDSEG) Fellowship Program. Siddharth Somasundaram is funded by the National Science Foundation (NSF) Graduate Research Fellowship Program (Grant #2141064).

References

4sense. 2021. Apple LIDAR Demystified: SPAD, VCSEL, and Fusion. https://4sense.medium.com/apple-lidar-demystified-spad-vcsel-and-fusion-aa9c3519d4cb Online; posted 1 March 2021.

Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C Sankaranarayanan. 2019. Convolutional approximations to the general non-line-of-sight imaging operator. In *International Conference on Computer Vision (ICCV)*. 7889–7899.

Rhett Allain. 2022. What an iPhone Lidar Can Show About the Speed of Light. https://www.wired.com/story/what-an-iphone-lidar-can-show-about-the-speed-of-light/ Online; posted 12 August 2022.

AMS OSRAM. 2023. TMF8820 Datasheet. https://look.ams-osram.com/m/52236c476132a095/original/TMF8820-21-28-Multizone-Time-of-Flight-Sensor.pdf.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. 2024. SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model. In *European Conference on Computer Vision (ECCV)*.

Nikhil Behari, Aaron Young, Siddharth Somasundaram, Tzofi Klinghoffer, Akshat Dave, and Ramesh Raskar. 2024. Blurred LiDAR for Sharper 3D: Robust Handheld 3D Scanning with Diffuse LiDAR and RGB. (2024). arXiv:2411.19474.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. 2024. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. (2024). arXiv:2410.02073.

Edoardo Charbon. 2014. Introduction to time-of-flight imaging. In *IEEE Sensors*. 610–613.

Wenzheng Chen, Fangyin Wei, Kiriakos N Kutulakos, Szymon Rusinkiewicz, and Felix Heide. 2020. Learned feature embeddings for non-line-of-sight imaging and recognition. *ACM Transactions on Graphics (ToG)* 39, 6 (2020).

In Cho, Hyunbo Shim, and Seon Joo Kim. 2024. Learning to Enhance Aperture Phasor Field for Non-Line-of-Sight Imaging. In *European Conference on Computer Vision (ECCV)*. 72–89.

Javier Grau Chopite, Patrick Haehn, and Matthias Hullin. 2025. Non-Line-of-Sight Estimation of Fast Human Motion with Slow Scanning Imagers. In *European Conference on Computer Vision*. 176–194.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024. CAT3D: Create anything in 3D with multi-view diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 270–279.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *International Conference on Computer Vision (ICCV)*. 3828–3838.

Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. 2025. Depth Any Camera: Zero-Shot Metric Depth Estimation from Any Camera. In *CVPR*.

Anant Gupta, Atul Ingle, and Mohit Gupta. 2019. Asynchronous single-photon 3D imaging. In *International Conference on Computer Vision (ICCV)*. 7909–7918.

Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu. 2021. itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging* 7 (2021), 1205–1214.

Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. 2021. Enhanced boundary learning for glass-like object segmentation. In *International Conference on Computer Vision (ICCV)*. 15859–15868.

Felix Heide, Steven Diamond, David B Lindell, and Gordon Wetzstein. 2018. Sub-picosecond photon-efficient 3D imaging using single-photon sensors. *Scientific reports* 8, 1 (2018), 17726.

Robert K Henderson, Nick Johnston, Sam W Hutchings, Istvan Gyongy, Tarek Al Abbas, Neale Dutton, Max Tyler, Susan Chan, and Jonathan Leach. 2019. 5.7 A 256× 256 40nm/90nm CMOS 3D-stacked 120dB dynamic-range reconfigurable time-resolved SPAD imager. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. 106–108.

Connor Henley, Joseph Hollmann, and Ramesh Raskar. 2022. Bounce-flash lidar. *IEEE Transactions on Computational Imaging* 8 (2022), 411–424.

Connor Henley, Tomohiro Maeda, Tristan Swedish, and Ramesh Raskar. 2020. Imaging behind occluders using two-bounce light. In *European Conference on Computer Vision (ECCV)*. 573–588.

Connor Henley, Siddharth Somasundaram, Joseph Hollmann, and Ramesh Raskar. 2023. Detection and mapping of specular surfaces using multibounce lidar returns. *Optics Express* 31, 4 (2023), 6370–6388.

Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris M Kitani. 2020. Optical non-line-of-sight physics-based 3D human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 7013–7022.

Sacha Jungerman, Atul Ingle, Yin Li, and Mohit Gupta. 2022. 3D scene inference from transient histograms. In *European Conference on Computer Vision*. Springer, 401–417.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9492–9502.

Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. 2009. Looking around the corner using transient imaging. In *International Conference on Computer Vision (ICCV)*. 159–166.

Tzofi Klinghoffer, Xiaoyu Xiang, Siddharth Somasundaram, Yuchen Fan, Christian Richardt, Ramesh Raskar, and Rakesh Ranjan. 2024. PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 14565–14574.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. 2019. Revisiting self-supervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1920–1929.

Oichi Kumagai, Junichi Ohmachi, Masao Matsumura, Shinichiro Yagi, Kenichi Tayu, Keitaro Amagawa, Tomohiro Matsukawa, Osamu Ozawa, Daisuke Hirono, Yasuhiro Shinozuka, et al. 2021. 7.3 A 189× 600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 64. 110–112.

Yue Li, Jiayong Peng, Juntian Ye, Yueyi Zhang, Feihu Xu, and Zhiwei Xiong. 2023. Nlost: Non-line-of-sight imaging with transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 13313–13322.

Tsung-Han Lin, Connor Henley, Siddharth Somasundaram, Akshat Dave, Moshe Laifenfeld, and Ramesh Raskar. 2024. Handheld Mapping of Specular Surfaces Using Consumer-Grade Flash LiDAR. In *International Conference on Computational Photography (ICCP)*.

David B Lindell, Matthew O'Toole, and Gordon Wetzstein. 2018. Single-photon 3D imaging with deep sensor fusion. *ACM Trans. Graph.* 37, 4 (2018), 113.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3D object. In *International Conference on Computer Vision (ICCV)*. 9298–9309.

Xiaochun Liu, Ibón Guillén, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. 2019. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature* 572, 7771 (2019), 620–623.

Li Ma, Vasu Agrawal, Haithem Turki, Changil Kim, Chen Gao, Pedro Sander, Michael Zollhöfer, and Christian Richardt. 2024. SpecNeRF: Gaussian directional encoding for specular reflections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 21188–21198.

Tomohiro Maeda, Guy Satat, Tristan Swedish, Lagnojita Sinha, and Ramesh Raskar. 2019. Recent advances in imaging around corners. (2019). arXiv:1910.05613.

Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kyros Kutulakos, and David Lindell. 2024. Transient neural radiance fields for lidar view synthesis and 3D reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Germán Mora-Martín, Stirling Scholes, Robert K Henderson, Jonathan Leach, and Istvan Gyongy. 2024. Human activity recognition using a single-photon direct time-of-flight sensor. *Optics Express* 32, 10 (2024), 16645–16656.

Fangzhou Mu, Sicheng Mo, Jiayong Peng, Xiaochun Liu, Ji Hyun Nam, Siddeshwar Raghavan, Andreas Velten, and Yin Li. 2022. Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* preprints (2022).

Fangzhou Mu, Carter Sifferman, Sacha Jungerman, Yiquan Li, Mark Han, Michael Gleicher, Mohit Gupta, and Yin Li. 2024. Towards 3D Vision with Low-Cost Single-Photon Cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5302–5311.

Mark Nishimura, David B Lindell, Christopher Metzler, and Gordon Wetzstein. 2020. Disambiguating monocular depth estimation with a single transient. In *European Conference on Computer Vision*. 139–155.

Matthew O'Toole, David B Lindell, and Gordon Wetzstein. 2018. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature* 555, 7696 (2018), 338–341.

Adithya Pediredla, Akshat Dave, and Ashok Veeraraghavan. 2019. Snlos: Non-line-of-sight scanning through temporal focusing. In *International Conference on Computational Photography (ICCP)*.

Jiayong Peng, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. 2020. Photon-efficient 3D imaging with a non-local neural network. In *European Conference on Computer Vision (ECCV)*. 225–241.

Sandor Plosz, Aurora Maccarone, Stephen McLaughlin, Gerald S Buller, and Abderrahim Halimi. 2023. Real-time reconstruction of 3D videos from single-photon LiDaR data in the presence of obscurants. *IEEE Transactions on Computational Imaging* 9 (2023), 106–119.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. 2024. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *ICLR*.

Alice Ruget, Max Tyler, Germán Mora Martín, Stirling Scholes, Feng Zhu, Istvan Gyongy, Brent Hearn, Steve McLaughlin, Abderrahim Halimi, and Jonathan Leach. 2022. Pixels2pose: Super-resolution time-of-flight imaging for 3D pose estimation. *Science Advances* 8, 48 (2022), eade0123.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. 2024. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*.

Siyuan Shen, Suan Xia, Xingyue Peng, Ziyu Wang, Yingsheng Zhu, Shiying Li, and Jingyi Yu. 2024. HOLI-1-to-3: Transient-Enhanced Holistic Image-to-3D Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* preprints (2024).

Dongeek Shin, Feihu Xu, Dheera Venkatraman, Rudi Lussana, Federica Villa, Franco Zappa, Vivek K Goyal, Franco NC Wong, and Jeffrey H Shapiro. 2016. Photon-efficient imaging with a single-photon camera. *Nature Communications* 7, 1 (2016), 12046.

Siddharth Somasundaram, Akshat Dave, Connor Henley, Ashok Veeraraghavan, and Ramesh Raskar. 2023. Role of Transients in Two-Bounce Non-Line-of-Sight Imaging. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9192–9201.

Shida Sun, Yue Li, Yueyi Zhang, and Zhiwei Xiong. 2024. Generalizable Non-Line-of-Sight Imaging with Learnable Physical Priors. (2024). arXiv:2409.14011.

Zhanghao Sun, David B Lindell, Olav Solgaard, and Gordon Wetzstein. 2020. SPADnet: deep RGB-SPAD sensor fusion assisted by monocular depth estimation. *Optics Express* 28, 10 (2020), 14948–14962.

Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. 2023. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 12349–12362.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*. 776–794.

Kushagra Tiwary, Akshat Dave, Nikhil Behari, Tzofi Klinghoffer, Ashok Veeraraghavan, and Ramesh Raskar. 2023. Orca: Glossy objects as radiance-field cameras. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 20773–20782.

Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. 2012. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications* 3, 1 (2012), 745.

Andreas Velten, Di Wu, Adrian Jarabo, Belen Masia, Christopher Barsi, Chinmaya Joshi, Everett Lawson, Moungi Bawendi, Diego Gutierrez, and Ramesh Raskar. 2013. Femto-photography: capturing and visualizing the propagation of light. *ACM Transactions on Graphics (ToG)* 32, 4 (2013), 1–8.

Dor Verbin, Pratul P Srinivasan, Peter Hedman, Ben Mildenhall, Benjamin Attal, Richard Szeliski, and Jonathan T Barron. 2024. Nerf-casting: Improved view-dependent appearance with consistent reflections. In *SIGGRAPH Asia 2024 Conference Papers*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*. 736–753.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 10371–10381.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth anything v2. *Advances in Neural Information Processing Systems* 37 (2024), 21875–21911.

Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. 2019. Where is my mirror?. In *International Conference on Computer Vision (ICCV)*. 8809–8818.

Xu Yang, ZiYi Tong, PengFei Jiang, Lu Xu, Long Wu, Jiemin Hu, Chenghua Yang, Wei Zhang, Yong Zhang, and Jianlong Zhang. 2022. Deep-learning based photon-efficient 3D and reflectivity imaging with a 64× 64 single-photon avalanche detector array. *Optics express* 30, 18 (2022), 32948–32964.

Juntian Ye, Yu Hong, Xiongfei Su, Xin Yuan, and Feihu Xu. 2024. Plug-and-Play Algorithms for Dynamic Non-line-of-sight Imaging. *ACM Transactions on Graphics* 43, 5 (2024).

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4578–4587.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. (2024). arXiv:2409.02048.

Yanhua Yu, Siyuan Shen, Zi Wang, Binbin Huang, Yuehan Wang, Xingyue Peng, Suan Xia, Ping Liu, Ruiqian Li, and Shiying Li. 2023. Enhancing Non-line-of-sight Imaging via Learnable Inverse Kernel and Attention Mechanisms. In *International Conference on Computer Vision (ICCV)*. 10563–10573.

Zhenya Zang, Dong Xiao, and David Day-Uei Li. 2021. Non-fusion time-resolved depth image reconstruction using a highly efficient neural network architecture. *Optics express* 29, 13 (2021), 19278–19291.

Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. 2023. Completionformer: Depth completion with convolutions and vision transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 18527–18536.

Shenyu Zhu, Yong Meng Sua, Ting Bu, and Yu-Ping Huang. 2023. Compressive non-line-of-sight imaging with deep learning. *Physical Review Applied* 19, 3 (2023), 034090.