Learning to Reconstruct Scenes with Single-Photon Cameras

by

Fangzhou Mu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: 12/07/2023

The dissertation is approved by the following members of the Final Oral Committee:

Yin Li, Assistant Professor, Biostatistics and Medical Informatics
Mohit Gupta, Associate Professor, Computer Sciences
Yong Jae Lee, Associate Professor, Computer Sciences
Yixuan Li, Assistant Professor, Computer Sciences
Andreas Velten, Associate Professor, Biostatistics and Medical Informatics

To my parents.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Yin Li, for his unwavering support throughout my doctoral journey. His expertise, insightful feedback, and commitment to academic excellence have not only shaped my research but also cultivated a passion for intellectual pursuit.

I would like to thank the rest of my dissertation committee: Dr. Mohit Gupta, Dr. Yong Jae Lee, Dr. Sharon Li, and Dr. Andreas Velten. Their invaluable comments have improved the quality and depth of my research.

I am extremely fortunate to have worked with many amazing colleagues: Carter Sifferman, Sacha Jungerman, Sicheng Mo, Dr. Xiaochun Liu, and Dr. Ji Hyun Nam. Many aspects of my dissertation would not have been possible without their expertise, dedication, and hard work.

I am also indebted to the following talented individuals whom I collaborated with on projects not included here: Dr. Felipe Gutierrez-Barragan, Yuheng Li, Dr. Ran Xu, and Dr. Chen-Lin Zhang. Their thoughtful discussions have been a constant source of inspiration for my research.

I am immensely grateful to my managers and mentors for the guidance I received during my internships: Dr. Jian Wang, Dr. Yicheng Wu, Dr. Shree Nayar, Guru Krishnan, Dr. Yanli Liu, and Dr. Bochen Guan. Their extensive knowledge and helpful advice have not only enhanced my skills but also broadened my perspective on the industry.

Special thanks to my labmates and friends: Dr. Yiwu Zhong, Zixuan Huang, Abrar Majeedi, Zhuoyan Xu, Dr. Sizhuo Ma, and Dr. Zijian Du. I appreciate the time and effort you spent helping me navigate the challenges, and your support has been a vital source of strength and joy.

Finally, my heartfelt thanks go to my parents for their unparalleled support, understanding, and encouragement throughout my PhD study. Their love and reassurance have provided me with the strength to overcome challenges and pursue my academic goals with determination.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Time-of-flight imaging with single-photon cameras has recently gained popularity in 3D vision. In this dissertation, I study an emerging paradigm of single-photon imaging for 3D scene reconstruction, where a spatially distributed set of transient histograms is captured by SPAD sensors with diffuse lighting and a wide-field-of-view detector. It is indicative of real-world sensor characteristics and novel imaging applications, yet presents major algorithmic challenges to 3D reconstruction due to the sophisticated transient formation model and unstructured sensor positioning. To overcome these challenges, I develop a new class of reconstruction algorithms based on the analysis-by-synthesis principle. My dissertation work combines expressive neural scene representations with differentiable transient volume rendering for the reconstruction of complex scene geometry with flexible sensor placement. It further incorporates careful sensor modeling and physics priors to account for non-idealities of real-world imaging hardware. I demonstrate the effectiveness of this reconstruction approach with two single-photon 3D vision systems, one for direct line-of-sight reconstruction using low-cost SPAD sensors, the other for high-speed non-line-of-sight imaging. My work provides strong evidence that 3D reconstruction of real-world objects can be achieved using spatially distributed single-photon cameras, and thus represents a solid step toward general and practical single-photon 3D vision.

1 INTRODUCTION

Reconstructing the 3D shape of real-world objects remains a central problem in computer vision. Solutions to this 3D reconstruction problem have evolved into two parallel branches. *Image-based modeling* leverages a plethora of visual cues from multiple photographs (*e.g.*, correspondence, shading, focus/defocus), leading to well-known approaches including multi-view stereo (Seitz et al., 2006), photometric stereo (Ackermann et al., 2015), shape-from-X (Grossmann, 1987; Subbarao and Surya, 1994; Zhang et al., 1999), and the more recent neural radiance fields (NeRF) (Mildenhall et al., 2021). Conversely, *active range scanning* combines an active light source with an imaging sensor, giving rise to widely-adopted computational imaging techniques such as structured light (Geng, 2011) and time-of-flight (Hansard et al., 2012). Conventional wisdom suggests that range scanning yields more precise 3D geometry as compared to image-based modeling at the cost of using specialized, expensive hardware.

An increasingly popular approach for range scanning is time-of-flight imaging with active *single-photon cameras*, a form of time-resolved sensors based on the single-photon avalanche diode (SPAD) technology. This approach couples a pico-to-nanosecond detector with a fast coherent light source, illuminates the scene with a very short pulse of light, and measures the intensity of the light over time as it reflects back from the scene. The resulting incident wavefront is quantized and recorded, forming a *transient histogram*. A typical use case of this approach is single-photon LiDAR, in which the light source (laser) is highly focused, the detector with a narrow field of view (FoV) finds the peak in the histogram, and the sensor reports a single distance value per detector pixel. Accurate reconstruction with LiDAR requires capturing a dense set of pixels. This can be achieved by either using raster scanning, which prolongs the imaging time, or employing a high-resolution 2D detector array, which increases

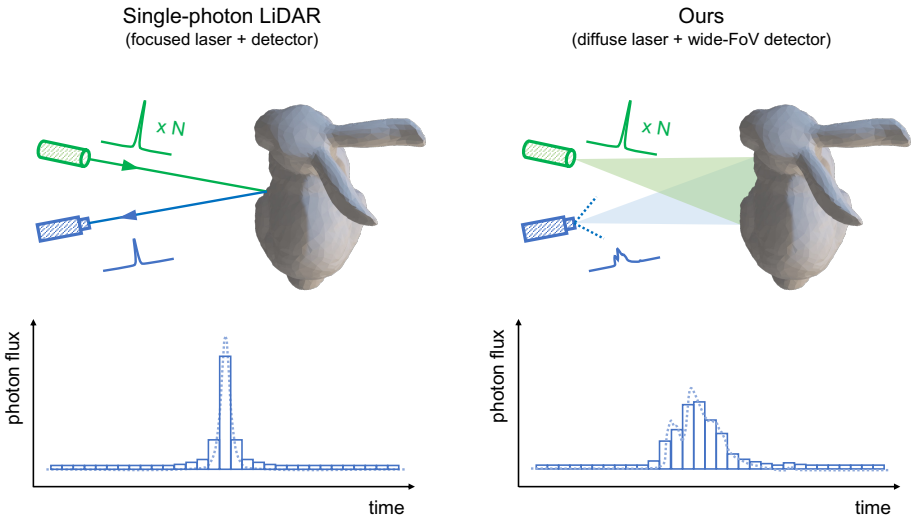


Figure 1.1: **Comparison of single-photon imaging paradigms.** Traditional single-photon LiDAR for depth ranging (*left*) consists of highly focused laser and a detector. The detector records histograms with a dominant peak from which the depth of a single point can be extracted. By contrast, our imaging paradigm (*right*) employs diffuse laser and a wide-FoV detector. The measured histograms encode richer information about the imaged scene patch.

the hardware cost.

In my dissertation, I investigate a less known yet emerging paradigm of single-photon 3D imaging, where a sensor consists of a *diffuse* light source and a SPAD detector with *wide FoV*. These sensors scatter light towards a potentially large scene patch, and record distributions of times-of-flight from a continuous range of incident angles. In doing so, rich information about scene geometry and reflectance is encoded in the *entirety* of their transient histograms, as opposed to the per-point distance extracted from a single dominant peak in LiDAR. Figure 1.1 illustrates the difference between the two imaging paradigms. We hypothesize that

this new paradigm will likely support sparser and less structured scan patterns, and reduce the cost of imaging hardware, thereby unlocking new capabilities of single-photon 3D imaging.

This imaging paradigm has been previously explored in applications such as non-line-of-sight imaging (Velten et al., 2012), fluorescence lifetime imaging (Lagarto et al., 2020), and most recently for direct line-of-sight depth estimation (Jungerman et al., 2022; Sifferman et al., 2023). However, a key bottleneck that prevents its wide adoption is the lack of an effective reconstruction framework. This difficulty stems from the sophisticated imaging model under diffuse lighting as compared to single-photon LiDAR. Direct inversion of the imaging model is only possible under highly restrictive sensor placements, whereas alternative approaches have thus far only been able to recover simple parametric shapes (*i.e.*, planes) and coarse depth maps using a single pixel or a small 2D pixel array. Further, the imaging systems built around this paradigm have been proof-of-concept in nature, and thus lag far behind image-based 3D vision systems in terms of generality and practicality.

Towards general and practical 3D vision with single-photon cameras, my dissertation is structured around the following statement:

Dissertation Statement

Transient histograms captured by a distributed set of single-photon cameras under diffuse lighting can be harnessed to reconstruct complex 3D scene geometry, thereby supporting new applications with practical imaging systems.

Overview

Towards a *general* algorithmic framework, I draw insights from the latest development in image-based modeling, and approach 3D reconstruc-

tion through the lens of analysis by synthesis. In particular, I propose to use neural implicit functions (*i.e.*, NeRF (Mildenhall et al., 2021) and NeuS (Wang et al., 2021)) as scene representation, and develop volume renderers tailored for our imaging paradigm to simulate the transient histogram formation process. Importantly, the neural representations support the modeling of arbitrary object shapes, whereas volume rendering enables flexible positioning of sensors. The resulting algorithms thus provide a general approach to 3D vision using single-photon cameras.

Towards *practical* single-photon imaging systems, I showcase the success of the reconstruction approach with two hardware prototypes. The first one is built using commodity SPAD sensors available at low cost and with low power consumption (Chapter 3). The second one operates at a real-time rate for high-speed imaging through clever approximation of optics (Chapter 4). In both cases, the algorithmic framework is customized to accommodate the non-idealities of the imaging system, either by careful calibration and modeling of sensors, or by incorporating physics priors as strong optimization constraints. These prototype systems shed light on the design of practical 3D vision systems with single-photon cameras.

Organization of the Dissertation

The rest of my dissertation is organized as follows:

Chapter 2 provides the essential background for the development of imaging and reconstruction approach in subsequent chapters. It starts with a brief introduction of single-photon cameras, followed by an overview of the imaging framework of interest and its associated transient formation model. It then reviews the basics of neural implicit functions as scene representation, with a focus on NeRF (Mildenhall et al., 2021) and NeuS (Wang et al., 2021), which I later adapt for 3D reconstruction from

transients. These preliminaries prepare us for the two practical use cases of our imaging framework.

Chapter 3 describes an imaging system with low-cost SPAD sensors, and a learning-based algorithm for the direct line-of-sight reconstruction of complex scenes. These sensors naturally exhibit diffuse lighting and wide FoV due to their imprecise optics. I introduce an imaging approach that leverages this property and reconstructs neural implicit surfaces using transient volume rendering given sparsely distributed transients. This approach demonstrates strong qualitative and quantitative results on both simulated and real-world data, and showcases the potential of low-cost SPAD sensors for practical 3D vision and imaging.

Chapter 4 presents a physics-inspired deep model for non-line-of-sight (NLOS) reconstruction with a high-speed imaging system. I first show that NLOS imaging can be formulated as a special case of the imaging framework, with the sensors placed on sparse scanlines of a relay wall. I then highlight the challenges of NLOS reconstruction using a high-speed imaging system, which motivates a robust learning-based reconstruction approach that combines a physical model with a volume renderer. The model learns entirely from simulated data, while generalizing well on both out-of-distribution simulations and real-world captures.

Chapter 5 concludes the dissertation and discusses several promising directions for future research.

Key Contributions

The key contributions of my dissertation are summarized as follows:

- I propose a general approach for 3D reconstruction using a distributed set of single-photon cameras with diffuse lighting and wide-FoV detectors. It is capable of reconstructing complex scenes with flexible positioning of sensors under various imaging conditions.

- I demonstrate encouraging reconstruction results with two real-world imaging systems designed for line-of-sight and non-line-of-sight imaging. My work thus represents a solid step towards practical single-photon 3D vision.

2.1 Single-Photon Cameras

A single-photon camera consists of an active light source (*e.g.*, continuous-wave or pulsed laser) and a time-resolved detector based on the single-photon avalanche diode (SPAD) technology. These cameras operate under the time-of-flight principle, and have recently gained popularity in 3D vision and imaging due to their single-photon sensitivity, extreme time resolution, and low power consumption.

In this dissertation, we study single-photon cameras with *pulsed* laser and *single-pixel* SPAD detector. They are often found in SPAD-based LiDAR systems for depth ranging, with broad applications in autonomous driving, robotics, wearable sensing, and virtual reality. In this regime, the laser source emits periodic pulses of light towards the scene. The SPAD in synchronization with the laser detects at most one returning photon within each pulse cycle, and records the time of arrival. Following a photon detection event, the SPAD enters a short dead time to reset itself, during which it will not be able to detect any photons. Photon timestamps over many pulse cycles are aggregated and binned into a temporal histogram (*i.e.*, *transient histogram*). This histogram counts the number of photons over discretized time intervals in the range of picoseconds, and can be thought of as a one-dimensional "image" of the scene.

In the following, we introduce a general form of our imaging framework with single-photon cameras. Two different instantiations of this conceptual framework for scene reconstruction are described in subsequent chapters.

2.2 Imaging Framework

Our single-photon imaging framework is uniquely characterized by *diffuse* laser and *wide-FoV* SPAD detector, and captures a *spatially distributed* set of transient histograms for scene reconstruction.

Diffuse Lighting

Our imaging system flash illuminates the scene with a periodic train of laser pulses. Importantly, the laser power is spread over an illumination field of view (FoV), allowing the outgoing light to interact with an extended surface area of the scene. This diffuse lighting condition differentiates our imaging setup from conventional SPAD-based depth ranging, where the laser is directed to a single scene point. Diffuse laser is often seen in low-cost proximity sensors as a consequence of imprecise optics, or can be generated by diverging collimated laser using a diffuser. It also arises in emerging applications such as fluorescence lifetime imaging (Lagarto et al., 2020) and non-line-of-sight imaging (Velten et al., 2012).

Wide-FoV Detector

In our framework, the diffuse laser is synchronized with a single-pixel, wide-FoV SPAD detector, possibly placed at a different location. A wide FoV can be easily achieved with a lensless or poorly focused detector. The detector accumulates returning photons from a range of incident angles within its FoV. The resulting histogram thus reflects the complex interplay between the diffuse lighting and the geometry and reflectance of a local scene patch visible from the detector. This is again different from conventional depth ranging, where a detector co-located with the laser source is focused at the illuminated scene point, and the depth can be directly read off the histogram via peak finding. In our case, it is not immediately

clear how to make sense of the histogram for scene reconstruction as it represents a superposition of signal from numerous scene points.

Distributed Sensing

We capture a scene with our imaging setup by placing the laser and detector at many different locations and collecting a distributed set of transient histograms. The laser-detector pairs have known poses, and are activated one at a time to avoid mutual interference. While each histogram has a limited FoV, we hypothesize that the full set of measurements will provide sufficient coverage of the scene needed for detailed reconstruction.

In the following, we describe the transient formation model for our imaging framework, which lays the foundation for the reconstruction algorithms we present in subsequent chapters.

2.3 Transient Formation Model

We model the formation of a transient histogram in two steps. We first derive the *transient waveform* of a scene by modeling the interaction of light with scene geometry and reflectance. This waveform is an idealized version of what we can expect to measure using real hardware. In the next step, we derive a *sensor model* that accounts for non-linearities that occur when capturing transient histograms with real hardware.

Transient Waveform

Given a diffuse laser source at location $\mathbf{l}$, a wide-FoV SPAD detector at location $\mathbf{s}$, and a scene with surface $\mathcal{S}$, the transient waveform $\tau_{\mathbf{l},\mathbf{s}}(t)$ can

be written in path integral form as

$$\begin{aligned}
\tau_{\mathbf{l},\mathbf{s}}(t) &= \int_{\mathcal{S}} R(\mathbf{x},\mathbf{l},\mathbf{s}) G(\mathbf{x},\mathbf{l}) G(\mathbf{x},\mathbf{s}) \delta\left(\|\mathbf{x}-\mathbf{l}\|_2 + \|\mathbf{x}-\mathbf{s}\|_2 - ct\right) dA(\mathbf{x}), \\
R(\mathbf{x},\mathbf{l},\mathbf{s}) &= \rho(\mathbf{x}) f_r(\mathbf{x}, \mathbf{n}_\mathbf{x}, \boldsymbol{\omega}_{\mathbf{x}\to\mathbf{l}}, \boldsymbol{\omega}_{\mathbf{x}\to\mathbf{s}}), \\
G(\mathbf{x},\mathbf{l}) &= \frac{\langle \mathbf{n}_\mathbf{x}, \boldsymbol{\omega}_{\mathbf{x}\to\mathbf{l}} \rangle}{\|\mathbf{x}-\mathbf{l}\|_2^2} V(\mathbf{x},\mathbf{l}), \\
G(\mathbf{x},\mathbf{s}) &= \frac{\langle \mathbf{n}_\mathbf{x}, \boldsymbol{\omega}_{\mathbf{x}\to\mathbf{s}} \rangle}{\|\mathbf{x}-\mathbf{s}\|_2^2} V(\mathbf{x},\mathbf{s}).
\end{aligned} \tag{2.1}$$

In Equation 2.1, $\mathbf{x}$ is a scene point on $\mathcal{S}$, and $dA(\mathbf{x})$ an infinitesimal area around $\mathbf{x}$. The time Dirac delta function δ compares the light traveling distance ct to the length of light path $\mathbf{l} \to \mathbf{x} \to \mathbf{s}$, with c the speed of light. The reflectance term R models the refletance property of $\mathcal{S}$, with ρ the albedo, f_r the bidirectional reflectance distribution function (BRDF), $\mathbf{n}_\mathbf{x}$ the surface normal at $\mathbf{x}$, and $\boldsymbol{\omega}_{\mathbf{x}\to\cdot}$ a unit directional vector originating from $\mathbf{x}$. The geometry term G, governed by scene geometry, illumination and detector FoV, captures the quadratic intensity fall-off, foreshortening effect, and visibility V. $V(\cdot,\cdot)$ is an indicator function that evaluates to 1 when the two points are visible to each other.

Equation 2.1 may be interpreted as the *impulse response* of the imaging system with respect to a scene $\mathcal{S}$. Note that we ignore high-order light paths and only model direct reflection from the scene surface. Previous works (Jungerman et al., 2022; Sifferman et al., 2023) have demonstrated that indirect reflection makes insignificant contribution to the transients.

Sensor Model

The sensor model accounts for laser and detector characteristics when converting a transient waveform into a transient histogram (Hernandez et al., 2017). Our sensor model considers the laser pulse, laser power,

detector quantum efficiency, ambient photon flux, internal detector noise, pile-up effect, and time jitter.

In practice, the laser pulse is not a perfect impulse and, despite band-pass filters, the measured transient also captures some constant ambient light. To model this, we convolve $\tau(t)$[1] with the laser's impulse response $g(t)$, scaled by ϕ^{scale} which absorbs laser power and quantum efficiency of the detector, and then offset its intensity by ϕ^{bkgd} which encapsulates ambient photon flux and internal detector noise:

$$\tilde{\tau}(t) = \phi^{scale}(\tau * g)(t) + \phi^{bkgd}. \tag{2.2}$$

$\tilde{\tau}(t)$ is subsequently discretized into a histogram of Poisson rates $\mathbf{r} = [r_1, ..., r_B]$ with B bins. The probability q_i of at least one photon falling inside the i^{th} bin is given by (Coates, 1968)

$$q_i = 1 - \exp(-r_i). \tag{2.3}$$

SPADs aggregate photon counts over C laser cycles, with only the first incident photon being detected in each cycle. This results in pile-up, or nonlinear distortion of transients, leaving photons arriving at a later timestamp less likely to be detected (Pediredla et al., 2018). Specifically, the probability p_i of detecting a photon in the i^{th} bin in a cycle is given by (Pediredla et al., 2018)

$$p_i = q_i \prod_{k=1}^{i-1} (1 - q_k). \tag{2.4}$$

The photon counts $[h_1, ..., h_B]$ in a transient histogram $\tilde{\mathbf{h}}$ follow a multino-

[1] We drop the subscript $\mathbf{l}, \mathbf{s}$ hereafter for clarity.

mial distribution (Gupta et al., 2019b):

$$[h_1, ..., h_{B+1}] \sim \text{Multinomial}(C, (p_1, ..., p_{B+1})), \tag{2.5}$$

where $p_{B+1} = 1 - \sum_{i=1}^{B} p_i$, and h_{B+1} counts the number of cycles without detected photons. $\tilde{\mathbf{h}}$ is subsequently convolved with a discretized time jitter kernel $\mathbf{s}$ to account for the temporal uncertainty of photon detection events, yielding the final histogram $\mathbf{h}$ measured by a SPAD detector:

$$\mathbf{h}[b] = \sum_{k} \mathbf{h}[k]\mathbf{s}[b - k]. \tag{2.6}$$

Practical Considerations

Our goal in scene reconstruction is to invert the transient formation process and recover an appropriate representation of the scene from the measurements. In particular, this representation is expected to reproduce the measured transients when rendered with the transient formation model. Solving this challenging inverse problem often requires knowing intermediate quantities in the forward model including the albedo and BRDF, which in practice are often not available alongside scene geometry. Likewise, hardware vendors may not disclose certain laser and detector characteristics needed for accurate sensor modeling. Finally, building an imaging system in the real world inevitably introduces approximations and errors to the optics due to practical constraints. In Chapter 3 and 4, we discuss how we overcome these challenges in the context of direct line-of-sight and non-line-of-sight scene reconstruction.

2.4 Neural Implicit Representation

Neural representations, as popularized by NeRF (Mildenhall et al., 2021), enable novel view synthesis and 3D reconstruction by representing the

scene using a neural network. While the original NeRF representation encoded view-dependent volumetric effects, alternative encodings have been proposed to better model geometry and reconstruct surfaces. In particular, NeuS (Wang et al., 2021) represents the scene as a level set, allowing for better modeling of surfaces at the expense of not being able to represent volumetric effects.

Many works extend these ideas to work with different sensing modalities and external supervision, such as depth queues from structure-from-motion (Deng et al., 2022), RGB images plus continuous-wave time-of-flight sensors (Attal et al., 2021), only depth information (Ortiz et al., 2022; Liu et al., 2023), or more recently using only transients of depth-ranging LiDAR systems (Huang et al., 2023; Malik et al., 2023).

In this work, we adapt NeRF (Mildenhall et al., 2021) and NeuS (Wang et al., 2021) for scene reconstruction using transients of our imaging framework. In the following, we provide a concise review of their basics.

Neural Radiance Fields

A neural radiance field or NeRF (Mildenhall et al., 2021) represents a scene as a continuous volume of color and density values. Specifically, a function $f_\theta : \mathbb{R}^6 \to \mathbb{R}^4$ maps a point $\mathbf{x} \in \mathbb{R}^3$ in space and a viewing direction $\mathbf{d} \in \mathbb{R}^3$ to RGB color $\mathbf{c} \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}$:

$$(\sigma, \mathbf{c}) = f_\theta(\mathbf{x}, \mathbf{d}). \tag{2.7}$$

f_θ is often realized as a multi-layer perceptron (MLP) with learnable weights θ, and can be rendered into pixel value $\hat{\mathbf{C}} \in \mathbb{R}^3$ in an RGB image $\hat{\mathbf{I}}$

using the volume rendering equation (Max, 1995):

$$\begin{aligned} \hat{\mathbf{C}}(\mathbf{r}) &= \int_0^\infty T(\mathbf{r}, u)\sigma(\mathbf{r}(u))\mathbf{c}(\mathbf{r}(u))du, \\ T(\mathbf{r}, u) &= exp\left(-\int_0^u \sigma(\mathbf{r}(s))ds\right), \\ \mathbf{r}(u) &= \mathbf{o} + u\mathbf{d}. \end{aligned} \tag{2.8}$$

In Equation 2.8, the transmittance T models occlusion along a ray $\mathbf{r}$ originating from the camera center $\mathbf{o} \in \mathbb{R}^3$ and in the direction $\mathbf{d}$. In practice, the integrals are numerically estimated using quadrature, with N importance-sampled points within pre-defined depth bounds along $\mathbf{r}$:

$$\begin{aligned} \hat{\mathbf{C}}(\mathbf{r}) &\approx \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i, \\ T_i &= \exp\left(-\sum_{j=1}^{i=1} \sigma_j\delta_j\right). \end{aligned} \tag{2.9}$$

In Equation 2.9, T_i is the piecewise transmittance over a ray segment δ_i between adjacent samples, and σ_i and $\mathbf{c}_i$ are evaluated at the mid-point of each segment.

The rendered image $\hat{\mathbf{I}}$ can then be compared with the ground-truth image $\mathbf{I}$, and θ is learned by minimizing the reconstruction loss

$$\mathcal{L} = \|\hat{\mathbf{I}} - \mathbf{I}\|_2^2. \tag{2.10}$$

In Chapter 4, we adapt NeRF for non-line-of-sight reconstruction with a focus on view synthesis. In particular, we propose a transient volume rendering framework in place of Equation 2.8 to approximate the transient waveform formation process of our imaging system (Equation 2.1).

Neural Implicit Surfaces

A neural implicit surface or NeuS (Wang et al., 2021) is a variant of NeRF tailored for surface reconstruction. It learns an MLP $f_\theta : \mathbb{R}^6 \to \mathbb{R}^4$ that jointly encodes the signed distance field and radiance field of a scene in its weights θ:

$$(d_s, \mathbf{c}) = f_\theta(\mathbf{x}, \mathbf{d}). \tag{2.11}$$

While locating the precise boundary of an object can be difficult with the density-based geometry representation of NeRF, the surface $\mathcal{S}$ of a scene is explicitly defined in NeuS as the zero-level set of the learned SDF [2]:

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | d_s(\mathbf{x}) = 0\}. \tag{2.12}$$

Similar to NeRF, f_θ can be volume-rendered into RGB images using Equation 2.8 and 2.9, yet the key difference lies in the estimation of volume density σ. Unlike in NeRF where σ is directly output from f_θ, NeuS models σ as a function of the signed distance d_s at each point $\mathbf{x}$. We provide a quick derivation below for completeness.

Derivation of σ

From Equation 2.8, we observe that the weights $w(u) := T(u)\sigma(\mathbf{r}(u))$ characterize the contribution of each point along $\mathbf{r}$ to the rendered color value $\hat{\mathbf{C}}$. For unbiased surface estimation, it is natural to enforce that w peaks at the point $\mathbf{r}(u^*)$ where $\mathbf{r}$ and $\mathcal{S}$ first intersects. We thus set

$$\frac{dw(u^*)}{du} = \frac{d(T(u^*)\sigma(\mathbf{r}(u^*)))}{du} = 0. \tag{2.13}$$

[2] We slightly abuse the notation and denote the learned SDF as d_s.

In the meantime, note that

$$\frac{dT(u)}{du} = -T(u)\sigma(\mathbf{r}(u)) \;\Rightarrow\; \sigma(\mathbf{r}(u)) = -\frac{T'(u)}{T(u)}. \tag{2.14}$$

By plugging Equation 2.14 into Equation 2.13, we arrive at

$$T''(u^*) = 0. \tag{2.15}$$

Since $d_s(\mathbf{r}(u^*)) = 0$, a simple choice of T that satisfies Equation 2.15 is

$$T(u) = \phi(d_s(\mathbf{r}(u))), \tag{2.16}$$

where ϕ is the sigmoid function[3] with

$$\phi'(u) = \phi(u)(1 - \phi(u)). \tag{2.17}$$

Plugging Equation 2.16 and 2.17 into Equation 2.14, we obtain

$$\sigma(\mathbf{r}(u)) = \underbrace{(\phi(d_s(\mathbf{r}(u))) - 1)}_{<0}\underbrace{d_s'(\mathbf{r}(u))}_{\mathbf{n}_{\mathbf{r}(u)}}\underbrace{\mathbf{r}'(u)}_{\mathbf{d}}, \tag{2.18}$$

To ensure that σ is always non-negative, we modify Equation 2.18 by clamping σ above zero:

$$\sigma(\mathbf{r}(u)) = \max((\phi(d_s(\mathbf{r}(u))) - 1)\mathbf{r}'(u)\mathbf{d}, 0). \tag{2.19}$$

In Chapter 3, we adapt NeuS for direct line-of-sight reconstruction with a focus on shape recovery. Our method renders transients as opposed to RGB images, and our rendering equation models surface reflectance as opposed to radiance.

[3]Without loss of generality, we omit the temperature α for clarity.

3 DIRECT LINE-OF-SIGHT RECONSTRUCTION WITH LOW-COST SPADS

In this chapter, we describe the first use case of our imaging framework for direct line-of-sight scene reconstruction using *low-cost* SPAD sensors[1].

3.1 Introduction

Low-cost single-photon cameras have recently become available as commercial products. They include one or more SPADs paired with an eye-safe diffuse light source (*e.g.*, an infrared VCSEL laser), are very small (<20 mm^3), inexpensive (<$5 USD), and power efficient (<10 milliwatts per measurement). Notable examples include the TMF8820 from AMS[2], and the VL53L8CH from ST Microelectronics[3]. They are sold as commodity proximity sensors, and some can be configured to report transient histograms. These sensors have proven successful for material classification (Becker and Koerner, 2023), human pose recognition (Ruget et al., 2022), and simple shape recovery (*i.e.*, a planar surface) (Jungerman et al., 2022; Sifferman et al., 2023).

Compared to laboratory-grade single-photon cameras, however, these low-cost SPADs lack precise optics, calibration, and timing characteristics and have an order of magnitude lower temporal resolution. In particular, they exhibit large illumination and detector FoVs with the low-quality

[1] This is joint work with Carter Sifferman and Sacha Jungerman, under the supervision of Michael Gleicher, Mohit Gupta and Yin Li. Fangzhou led the project, developed the algorithm, performed data simulation, conducted most experiments, and prepared the results. Carter co-led the project, built the hardware prototype, captured real-world data, analyzed the baselines, and prepared the figures. Sacha captured real-world data and prepared the figures. All participants designed the study, interpreted the results, and drafted the paper that is in submission at the time of writing.

[2] `https://ams.com/documents/20143/6015057/TMF882X_DS000693_8-00.pdf`

[3] `https://www.st.com/resource/en/datasheet/vl53l8ch.pdf`

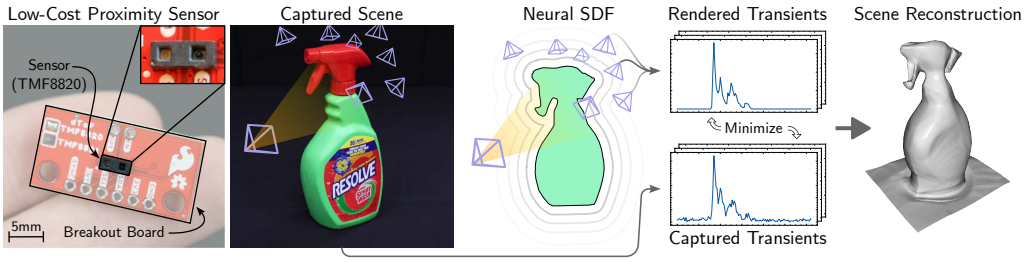


Figure 3.1: **Low-cost single-photon imaging.** We demonstrate that measurements from spatially distributed low-cost single-photon proximity sensors (left) can be used to reconstruct 3D shape of real world objects (right). Our method combines a differentiable image formation model and neural rendering to recover 3D geometry based on measurements (transient histograms) from sensors with known poses. This is done by minimizing the difference between the observed and rendered sensor measurements. For clarity, a subset of sensor poses and measurements are shown.

laser and lens, and the sensor model and post-processing on-board are near black-box. These non-idealities break the conventional depth ranging model and have hindered the wide adoption of these sensors for 3D imaging in the wild.

Contributions

In this work, we address the problem of reconstructing 3D shape of arbitrary Lambertian objects from a collection of spatially distributed, low-cost single-photon cameras with known poses. In particular, this capture setup represents a special case our proposed imaging framework with co-located laser and detector for each transient, and is reminiscent of multi-view stereo and NeRF-like systems for conventional RGB cameras. A key distinction between our system and multi-view image or scanning LiDAR based systems, however, is that our measurements consist merely of a few hundred pixels sparsely distributed in space, as opposed to capturing

dozens of RGB or transient images, each with millions of pixels. This is a major overhaul of imaging paradigms for 3D reconstruction and opens up new opportunities for single-photon 3D vision.

To support our novel 3D imaging system, we present an effective reconstruction algorithm that combines a neural signed-distance-field scene representation, a differentiable transient formation model for practical single-photon cameras, and an optimization scheme following the analysis-by-synthesis principle. Figure 3.1 illustrates our sensor, imaging setup, and reconstruction approach. We show that our approach can successfully recover complex 3D shapes with simulated data. We further demonstrate 3D object reconstruction from real-world captures, utilizing measurements from a low-cost, off-the-shelf proximity sensor.

Our setup extends recent low-cost SPAD imaging systems (Jungerman et al., 2022; Sifferman et al., 2023) in two directions. First, our system consists of multiple posed sensors as opposed to a single sensor. Second, it is capable of capturing and reconstructing complex, non-parametric scenes as opposed to the parametric geometry of a single plane. With these upgrades, our system marks a substantial step towards practical 3D vision using commodity proximity sensors.

3.2 Related Work

SPAD-based LiDAR Systems

SPAD-based LiDAR systems capture a spatiotemporal volume of the scene using a single pixel with a galvo for raster scanning, or using a 2D pixel array. These systems have found wide applications, ranging from autonomous driving [4] to depth sensing on smartphones [5]. Recent work has

[4] `https://www.sony-semicon.com/en/products/is/automotive/tof.html`

[5] `https://developer.apple.com/documentation/avfoundation/additional_data_capture/capturing_depth_using_the_lidar_camera`

focused on the robust reconstruction of depth and albedo (O'Toole et al., 2017; Heide et al., 2018), the analysis of optimal incident flux (Gupta et al., 2019b) and mitigation of non-linear distortion due to pile-up (Pediredla et al., 2018; Gupta et al., 2019a), denoising and super-resolution of depth reconstructions (Lindell et al., 2018; Peng et al., 2020; Mora-Martín et al., 2023), and the compression of transients for efficient in-sensor processing and data transmission off sensor (Gutierrez-Barragan et al., 2022, 2023). Our imaging model generalizes the conventional depth-ranging model (focused light and detector) and investigates a new imaging framework (diffuse light and wide-FoV detector), and further demonstrates the feasibility of 3D reconstruction with low-cost SPAD sensors.

3D Imaging with Low-Cost SPADs

With cheap SPAD sensors becoming commonplace, recent works have investigated their use for 3D imaging. Callenberg et al. (2021) demonstrate that high-resolution depth imaging using low-cost SPADs is possible with some additional hardware. A low-cost SPAD has also been used to augment an RGB SLAM system (Liu et al., 2023). Other works learn to recover geometric information from low-cost SPAD measurements, such as 3D human pose (Ruget et al., 2022).

Most relevant to our work, Jungerman et al. (2022) use differentiable rendering to recover two degrees of freedom of a planar surface from a single low-cost SPAD transient histogram. Sifferman et al. (2023) extend this method to fully recover a planar surface and its albedo, from a single measurement comprised of 9 histograms from a multi-zone low-cost SPAD. Our method extends this line of work by learning neural representations to represent arbitrary scene geometry.

3.3 Method

The reconstruction problem we attempt to solve is extremely challenging. Unlike in conventional LiDAR where the depth of a scene point can be directly read off the histogram via peak finding, a transient from our system represents the superposition of light reflected from numerous scene points, as illuminated by a diffuse laser, and is further contaminated by non-idealities of the detector (*e.g.*, pile-up). The direct inversion of the signal is thus a highly ill-posed problem. Further, we cannot adapt methods from the non-line-of-sight imaging literature (O'Toole et al., 2018; Lindell et al., 2019; Liu et al., 2019b) as they only support dense 2D scans, whereas out system uses a distributed, sparse and unstructured set of measurements.

To overcome these challenges, we resort to an analysis-by-synthesis approach based on differentiable rendering. Our approach allows flexible positioning of sensors and accurate modeling of histogram formation, thereby enabling high-quality reconstruction of scene geometry. We now describe our reconstruction algorithm in detail.

Neural Scene Representation

Following NeuS (Wang et al., 2021), we represent the scene geometry as a signed distance function (SDF), parameterized as a multi-layer perceptron (MLP) $f_\theta : \mathbb{R}^3 \to \mathbb{R}$, with θ as its weights. f_θ maps the position-encoded (PE) xyz-coordinates of a point $\mathbf{x}$ to its signed distance d_s:

$$d_s = f_\theta(\mathrm{PE}(\mathbf{x})). \tag{3.1}$$

Compared to Jungerman et al. (2022) and Sifferman et al. (2023), this neural SDF allows our method to represent scene geometry beyond simple parametric shapes as the level set $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | f_\theta(\mathbf{x}) = 0\}$.

Transient Volume Rendering

The key idea behind our analysis-by-synthesis approach is to render f_θ into transients and compare them with those captured by our system. To adapt Equation 2.1 for the rendering of f_θ, we first rewrite it in angular integral form as

$$\tau(t) = \int_\Omega \frac{\rho}{\pi} \frac{V(\mathbf{x})\langle -\boldsymbol{\omega}, \mathbf{n}_\mathbf{x}\rangle}{\|\mathbf{x}\|^2} \delta\left(\|\mathbf{x}\|_2 - \frac{ct}{2}\right) d\boldsymbol{\omega}, \tag{3.2}$$

where $\boldsymbol{\omega}$ are ray directions in the sensor FoV Ω, and $\mathbf{x}$ the point where $\boldsymbol{\omega}$ intersects with the object surface $\mathcal{S}$ (∞ if no intersection). For simplicity, we assume a learned spatially uniform albedo ρ and Lambertian BRDF $f_r = 1/\pi$ [6].

Inspired by NeRF (Mildenhall et al., 2021) and NeuS (Wang et al., 2021), we approximate Equation 3.2 via volume rendering to resolve surface discontinuities, enabling the optimization of θ via gradient descent:

$$\hat{\tau}(t) = \int_\Omega \frac{\rho}{\pi} \frac{T^2(t)\sigma(\mathbf{p}(\boldsymbol{\omega}, t))\langle -\boldsymbol{\omega}, \mathbf{n}_\mathbf{p}\rangle}{\|\mathbf{p}(\boldsymbol{\omega}, t)\|^2} d\boldsymbol{\omega}. \tag{3.3}$$

Here, $\mathbf{p}(\boldsymbol{\omega}, t) = \frac{ct}{2}\boldsymbol{\omega}$ are points along $\boldsymbol{\omega}$, the volume density σ is a function of f_θ as in Wang et al. (2021), and the transmittance T is given by

$$T(t) = \exp\left(-\int_0^t \sigma(\mathbf{p}(u))du\right). \tag{3.4}$$

In practice, we discretize $\hat{\tau}(t)$ over the transient bin intervals $\{[t_i, t_{i+1})\}_{i=1}^B$ and work with the histogram $\hat{\boldsymbol{\tau}} = [\hat{\tau}_1, ...\hat{\tau}_B]$, where

$$\hat{\tau}_i = \int_\Omega \frac{\rho}{\pi} \int_{t_i}^{t_{i+1}} \frac{T^2(t)\sigma(\mathbf{p}(\boldsymbol{\omega}, t))\langle -\boldsymbol{\omega}, \mathbf{n}_\mathbf{p}\rangle}{\|\mathbf{p}(\boldsymbol{\omega}, t)\|^2} dt\, d\boldsymbol{\omega}. \tag{3.5}$$

[6] This assumption may be relaxed to allow more expressive BRDF models such as the Phong reflection model (Phong, 1975).

We estimate the intractable Equation 3.5 via Monte Carlo sampling of $\boldsymbol{\omega}$ and subsequently of $\mathbf{p}(\boldsymbol{\omega}, t)$.

Similar to Mildenhall et al. (2021) and Wang et al. (2021), the sampling of $\mathbf{p}(\boldsymbol{\omega}, t)$ is weighted by a probability density function (PDF) over the equally sized bin intervals. This PDF is proportional to the per-bin weights $w_i(\boldsymbol{\omega})$ given by

$$w_i(\boldsymbol{\omega}) = \exp\left(-\sum_{j=1}^{i-1} \sigma_j(\boldsymbol{\omega})\Delta\right)\alpha_i(\boldsymbol{\omega}), \tag{3.6}$$

where $\sigma_i(\boldsymbol{\omega})$ is evaluated at the mid-point of the i^{th} bin along $\boldsymbol{\omega}$, Δ is the bin size in distance, and

$$\alpha_i(\boldsymbol{\omega}) = (1 - \exp(-\sigma_i(\boldsymbol{\omega})\Delta)) \tag{3.7}$$

is the opacity along $\boldsymbol{\omega}$ for the i^{th} bin.

We extend this idea to the importance sampling of $\boldsymbol{\omega}$. Specifically, the sampling PDF over a uniform partitioning of FoV Ω is proportional to the cumulative weights $w^{(k)}(\boldsymbol{\omega})$ over rays $\boldsymbol{\omega}^{(k)}$ drawn from each partition k:

$$w^{(k)}(\boldsymbol{\omega}) = \sum_{i=1}^{B} w_i^{(k)}(\boldsymbol{\omega}). \tag{3.8}$$

Intuitively, this allows us to point more rays at high-density regions occupied by the object surface.

Differentiable Sensor Modeling

Modeling sensor behavior is particularly important for our analysis-by-synthesis approach. This is because the synthesis targets $\boldsymbol{\tau}$ are not determined by the scene geometry alone but reflect the complex interplay of geometry with sensor non-idealities including pulse shape, pile-up and

time jitter. To this end, we cascade $\hat{\boldsymbol{\tau}}$ to a *differentiable sensor model* Γ to simulate the transformation applied by the sensor to raw waveforms.

Specifically, Γ closely follows the sensor model in Section 2.3; Equations 2.2-2.4 are differentiable and applied sequentially on $\hat{\boldsymbol{\tau}}$, yielding per-bin photon detection probabilities $\hat{\mathbf{p}} = [\hat{p}_1, ..., \hat{p}_B]$. Instead of sampling photon counts using Equation 2.5, we directly convolve $\hat{\mathbf{p}}$ with the jitter kernel as in Equation 2.6. This allows us to sidestep the non-differentiable sampling step while producing an unbiased estimate of the transient $\hat{\mathbf{h}}$ for loss evaluation:

$$\hat{\mathbf{h}} = \Gamma(\hat{\boldsymbol{\tau}}). \tag{3.9}$$

Optimization Objectives

The optimization of θ is driven by three loss terms. First, the histogram reconstruction loss $\mathcal{L}_{\mathtt{hist}}$ minimizes the L1 distance between $\hat{\mathbf{h}}$ and $\mathbf{h}$:

$$\mathcal{L}_{\mathtt{hist}} = \sum_{i=1}^{B} |\hat{\mathbf{h}}_i - \mathbf{h}_i|. \tag{3.10}$$

Second, the Eikonal loss (Gropp et al., 2020) $\mathcal{L}_{\mathtt{Eikonal}}$ encourages f_θ to approximate an SDF:

$$\mathcal{L}_{\mathtt{Eikonal}} = \mathbb{E}_{\mathbf{x}}(\|\nabla_{\mathbf{x}}(f_\theta(\mathbf{x}))\| - 1)^2. \tag{3.11}$$

Finally, the total variation regularizer (Mu et al., 2022) $\mathcal{L}_{TV}$ penalizes floaters in empty space:

$$\mathcal{L}_{TV} = \mathbb{E}_{\boldsymbol{\omega}} \sum_{i=1}^{B-1} |\log \alpha_{i+1}(\boldsymbol{\omega}) - \log \alpha_i(\boldsymbol{\omega})|. \tag{3.12}$$

The combined loss function $\mathcal{L}$ is thus given by

$$\mathcal{L} = \mathcal{L}_{\mathtt{hist}} + \lambda_{\mathtt{Eikonal}}\mathcal{L}_{\mathtt{Eikonal}} + \lambda_{\mathtt{TV}}\mathcal{L}_{\mathtt{TV}}, \tag{3.13}$$

where $\lambda_{\mathtt{Eikonal}}$ and $\lambda_{\mathtt{TV}}$ are the respective loss weights.

3.4 Experiments

We demonstrate the effectiveness of our method for 3D geometric reconstruction of various objects in simulation, and in the real world with a low-cost SPAD on scenes of varying geometry and texture. We provide qualitative and quantitative results for both settings.

The rest of this section is organized a follows. We first describe the implementation details of our method and discuss the baselines and evaluation metrics. We then present our simulated experiments, followed by a sensitivity analysis to understand the impact of imaging conditions. Finally, we discuss our hardware prototype and provide results on real-world experiments.

Implementation Details

We use an 8-layer MLP with 256 hidden units as our SDF, f_θ, and initialize it as a sphere, centered at the origin with radius 0.3m, using geometric initialization (Atzmon and Lipman, 2020). For each transient, we sample 256 rays $\boldsymbol{\omega}$ over Ω and sample 256 points per ray. We set $\lambda_{\mathtt{Eikonal}}$ to 0.1 across all experiments and set $\lambda_{\mathtt{TV}}$ to 0 and 0.01 respectively for the simulated and real-world experiments. We train f_θ for $300,000$ steps using Adam (Kingma and Ba, 2014) with a mini-batch size 2, a learning rate 0.0005, and cosine decay. The learned SDFs are converted to meshes using Marching Cubes (Lorensen and Cline, 1987).

Baselines

We compare our method to two baselines: reprojection (Heide et al., 2018; Gupta et al., 2019b) and space carving (Kutulakos and Seitz, 2000; Tsai et al., 2017). We briefly describe these baselines.

Reprojection, also known as back-projection, reconstructs a scene as a point cloud and is the *de facto* standard for depth ranging. We compare to two forms of reprojection. The peak method finds the distance d corresponding to the histogram bin with the highest intensity. For a sensor at position $\mathbf{s}$ with an outwards pointing optical axis $\mathbf{u}$, a point is placed in the scene at position $\mathbf{s} + d\mathbf{u}$. The threshold method works in the same way but finds d by locating the lowest-index bin with intensity above a threshold t_p. If no bin passes the threshold, no point is projected.

Space carving reconstructs a scene as a voxel grid. Like thresholded reprojection, it finds the distance d corresponding to the lowest-index bin with intensity above a threshold t_s. All voxels in the sensor's FoV and nearer than d are marked empty, along with voxels outside the FoV. Voxels in the FoV and further than d are marked as occupied. The carved scene is the union of the occupied set for all sensors.

In our simulated experiments, t_p and t_s are scaled alongside relevant sensor parameters (bin count, FoV, power, and number of cycles) to remain consistent. To ensure strong baselines for real-world experiments, we perform a brute-force search over t_p and t_s and choose values that minimize Chamfer distance over the entire real-world dataset. Space carving voxel size was set to 1.0cm.

Evaluation Protocol

Following NeuS (Wang et al., 2021), we evaluate all methods using Chamfer distance $d_{Chamfer}$ between two point clouds X and Y:

$$d_{Chamfer}(X, Y) = \underbrace{\sum_{x \in X} \min_{y \in Y} \|x - y\|_2}_{X \to Y} + \underbrace{\sum_{y \in Y} \min_{x \in X} \|x - y\|_2}_{Y \to X}. \tag{3.14}$$

We report standard (two-way) Chamfer on simulated data. For real-world captures, we report Chamfer in both directions to evaluate the quality of reconstruction. Prior to Chamfer calculation, we convert ground-truth meshes and reconstructions from our method to point clouds by drawing 5 million points uniformly at random on the mesh surface. For space carving, occupied voxels are converted to points if they touch unoccupied space, excluding the edge of the grid.

Simulated Experiments

Experiment Setup

We simulate transients for eight scenes of varying complexity using the image formation model in Section 2.3. The objects are centered on the ground plane ($z = 0$) with the largest dimension ≈ 0.3m. Sensors with a conical FoV are uniformly distributed on a hemisphere at the origin with a radius of 0.5m, and are all pointed at the origin. In our simulation, $N = 256$, $B = 256$, $\Delta = 5$mm, FoV $= 30^\circ$, $\phi^{scale} = 1$, $\phi^{bkgd} = 0.001$, $C = 5000$ and $\rho = 0.8$. The laser pulse, g, has a full-width-at-half-maximum (FWHM) of 50ps, and $\mathbf{s}$ is a tabulated PDF obtained from experiments Hernandez et al. (2017). The sensor parameters are deliberately chosen to reflect the characteristics of low-cost sensors.

Method	Chamfer Distance (mm) ↓							
	Armadillo	Bear	Bunny	Digit	Einstein	Skull	Soap	Sphere
Reprojection (Peak)	54.29	40.36	34.95	55.85	43.25	48.90	51.71	51.07
Reprojection (Threshold)	65.43	60.72	54.05	60.64	61.31	65.14	68.74	63.16
Space Carving	34.78	24.53	22.29	45.44	26.60	25.49	21.44	25.47
Ours	**3.93**	**5.95**	**3.84**	**3.27**	**3.51**	**3.22**	**3.23**	**3.77**

Table 3.1: **Quantitative results on simulated data.** Our method more accurately recovers 3D shapes than baselines across 8 objects.

Reconstruction Results

Table 3.1 summarizes the quantitative results of all methods. Our method achieves an average Chamfer distance of < 5mm, an order of magnitude lower than all baselines. A key reason is that the baselines only use depth information from a single histogram bin, whereas our method makes effective use of the entire waveform, which contains rich geometry cues about a large scene patch.

We provide visualizations of our results in Figure 3.2. Our method recovers global scene structure as well as local geometry details. In contrast, reprojection yields sparse point clouds without sufficient coverage of the scene. While space carving produces dense reconstructions, the occupancy grid only represents an envelope of the scene, leaving it difficult to recognize the precise shape of an object.

Surface Normal Estimation

Moving beyond the 3D shapes, we further examine the surface normal of our reconstructed 3D objects. The surface normal of a point $\mathbf{x}$ on the reconstructed mesh is estimated as

$$\tilde{\mathbf{n}}_{\mathbf{x}} = \frac{\nabla_{\mathbf{x}}(f_\theta(\mathrm{PE}(\mathbf{x})))}{\|\nabla_{\mathbf{x}}(f_\theta(\mathrm{PE}(\mathbf{x})))\|}, \tag{3.15}$$

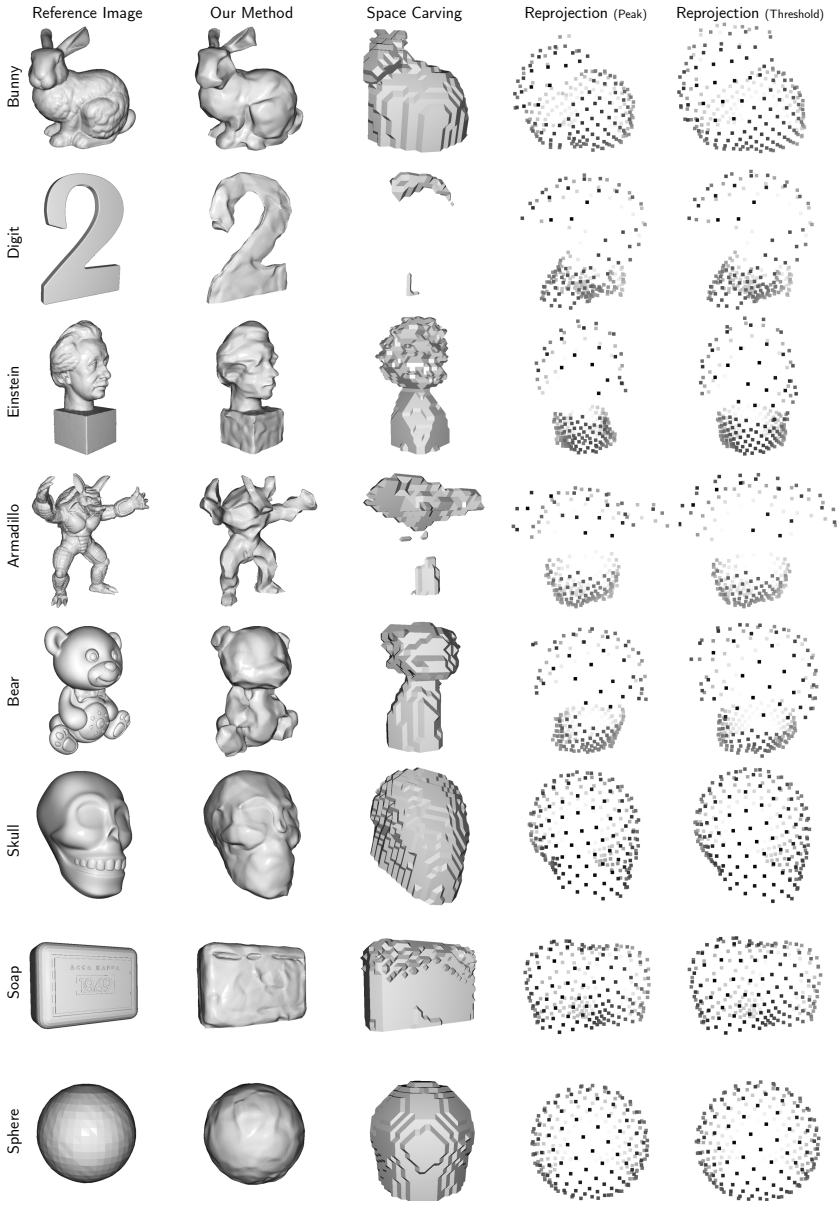


Figure 3.2: **Qualitative results on simulated data.** Our method reconstructs dense and detailed 3D shapes. Space carving provides only hulls of a target shape, and is prone to carving away extra space when thin structures are present. Reprojection yields sparse points.

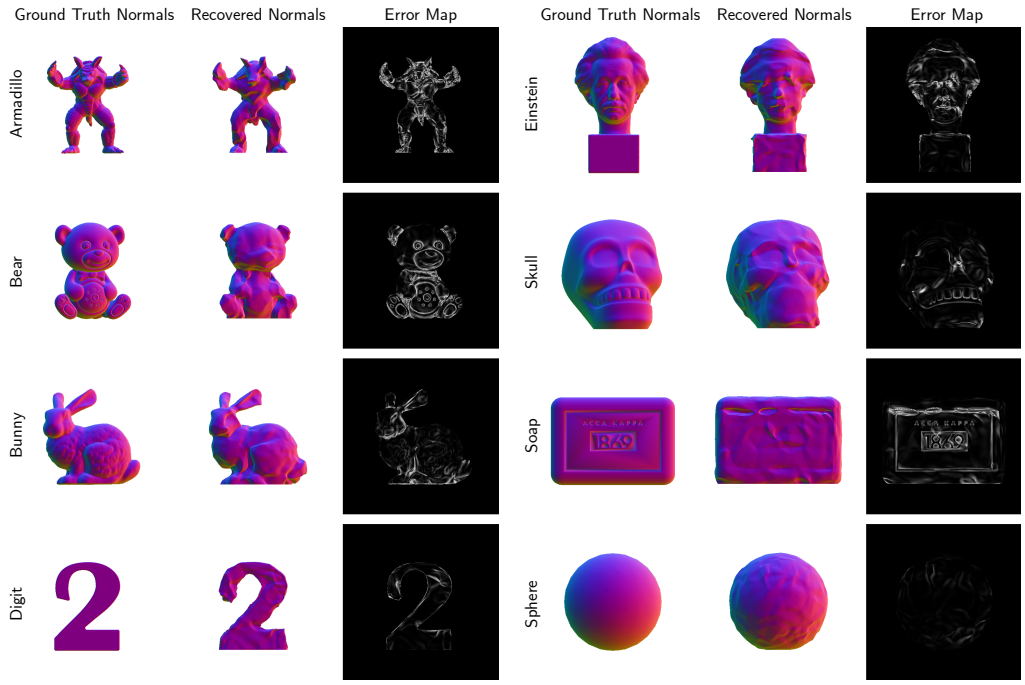


Figure 3.3: **Visualizations of surface normals for simulated data.** Our method correctly estimates surface normals in flat regions. Error mainly occurs at edges and depth discontinuities. We hypothesize that sensors with higher temporal and spatial resolution are needed to detect rapid changes in surface normals.

where f_θ is the learned SDF and PE denotes the positional encoding function. The error e_x w.r.t. the ground-truth normal $\mathbf{n}_x$ is given by

$$e_x = |\langle \mathbf{n}_x, \tilde{\mathbf{n}}_x \rangle|. \tag{3.16}$$

We provide visualizations of surface normals for simulated data in Figure 3.3. Our method can successfully recover smoothly varying normals. Error typically occurs at edges and depth discontinuities with fast-changing normals. We hypothesize that sensors with higher temporal and spatial resolution are needed for more accurate surface normal estimation.

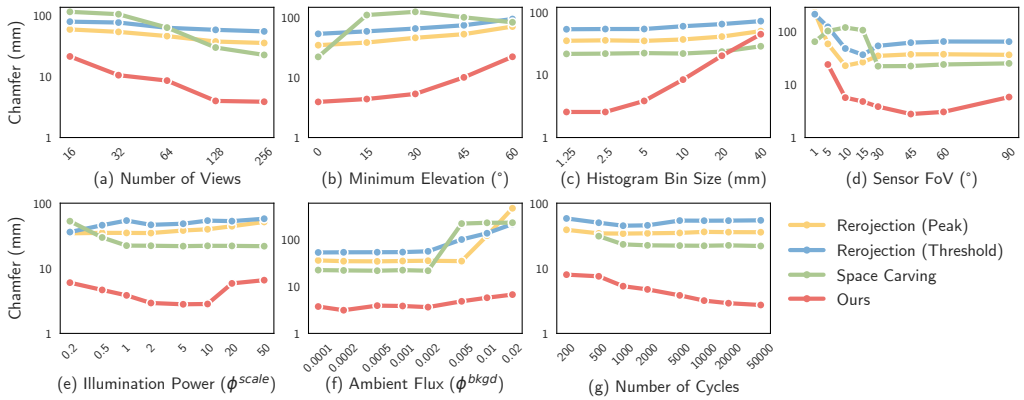


Figure 3.4: **Sensitivity analysis of our method compared to baselines across a range of imaging parameters.** In almost every case, our method outperforms baseline methods on Chamfer distance. Missing datapoint in (d) indicates that our method failed to converge. Illumination power (e) is unit-less as it also absorbs factors like quantum efficiency and does not map directly to any real world parameter.

Sensitivity Analysis

We perform extensive experiments to understand the robustness of our method in comparison to baselines under varying sensor parameters in simulation. All experiments are based on the Bunny scene and the parameters are varied one at a time while other parameters remain fixed at the base condition as described in Section 3.4. To ensure strong baselines for every sensor configuration, we calibrate the thresholds t_p and t_s for the projection (threshold) and space carving baselines respectively per reconstruction. We perform a brute force search over possible thresholds and report the best Chamfer achieved. As this amounts to calibrating on the test set, the numbers reported represent the best possible performance of the baseline methods on the given data. The results of this sensitivity analysis are presented in Figure 3.4. In what follows we discuss some of the main findings.

Sensor Placement

We study two key parameters that control sensor placement: the number of views and the minimum elevation angle at which the sensors are placed. Our method consistently outperforms all baselines in Chamfer distance by an order of magnitude across a broad range of parameter choices. In particular, our method readily supports as few as 128 views above a considerably large elevation angle of 30° without harming reconstruction quality. This robust gain in performance confirms that our method takes advantage of broad-band signal in transients not exploited by the baseline methods.

Temporal Resolution

Our system takes advantage of the temporal information in transient histograms, and therefore benefits when that information is present at a high resolution. Because of this, our method outperforms baselines by a very wide margin at a small bin size, but the margin vanishes as bins become wider than 2cm (equivalently 66ps), because decomposing the temporal signal becomes impractical beyond this limit. Fortunately, today's commodity SPADs operate at a smaller bin size ($\sim$ 40ps). Baseline methods show no performance gain at small bin sizes, as they do not take advantage of the temporal resolution.

Angular Resolution

Our system resolves spatial resolution from wide-FoV sensors by taking advantage of the time dimension. In this regime, the optimal sensor field-of-view size is not obvious: a smaller FoV means more highly constrained geometry, as each histogram images a smaller region, but too small of a field-of-view means a lack of coverage and under-constrained geometry. We find that an angular resolution in the 30° to 60° range is optimal for

reconstructing 3D geometry with our method on the bunny scene. Reprojection based methods benefit more from a smaller field-of-view, while space carving performs best with a wider field-of-view so that space is sufficiently carved away. In every case, our method outperforms baselines by a wide margin.

Signal-to-noise ratio (SNR)

We consider three parameters that jointly impact SNR: illumination power, ambient flux, and number of illumination cycles. Our method again outperforms all baselines by a significant margin across all test conditions. Notably, the baselines fail or perform considerably worse under high ambient flux, as signal photons are blocked by background photons due to pile-up. By contrast, our method is robust against a broad range of ambient flux levels, as we model the effects of ambient flux directly.

Hardware Prototype

We use the SPAD-based AMS TMF8820 proximity sensor (AG), which retails for $10 USD. We connect the sensor to a microcontroller via I^2C and use the AMS-provided driver to extract transient histograms.

The sensor contains a total of 216 SPADs, which are pooled onboard the sensor into 3×3 zones, each of which images a different FoV. The sensor captures one transient histogram for each zone. We pool histograms from all zones, which is equivalent to capturing one wide-FoV histogram per-measurement as SPADs do not suffer from readout noise (Zappa et al., 2007). In doing so, we avoid inter-histogram interference previously observed by Sifferman et al. (2023) and avoid the need to model individual fields-of-view of the sensor, which we empirically observed to have soft and poorly specified boundaries. We slightly modify our method to accommodate the AMS TMF8820 sensor used in real-world experiments.

Laser Impulse

The laser impulse response of the TMF8820 is not Gaussian and varies slightly between measurements. Fortunately, the sensor captures the shape of its laser impulse for each measurement in a "reference histogram". We record this histogram for each measurement and incorporate it into our forward model by cross-correlating the idealized scene response with this recorded reference histogram. We observe that the bin size Δ_r of the reference histogram is smaller than the bin size Δ of the transient histograms captured by the sensor. To account for this, we scale the reference histogram in the temporal dimension by a factor Δ_r/Δ before cross-correlation. Further, we find that it is necessary to temporally shift the reference histogram by a fixed amount ϕ^{delay} before correlation.

To calibrate the parameters Δ, Δ_r, and ϕ^{delay}, we perform the one-off intrinsic calibration procedure separately introduced by Sifferman et al. (2023). The TMF8820 sensor is pointed at a planar surface from a range of known distances and angles-of-incidence. A differentiable render-and-compare method is used to optimize for the unknown sensor intrinsic parameters given known planar geometry.

Pile-up Correction

While our forward model assumes that the target transients exhibit non-linear distortion due to pile-up, the TMF8820 sensor performs pile-up correction on-sensor, and it cannot be disabled. To accommodate this, we incorporate the differentiable Coates' correction (Coates, 1968) as a final step in the forward model.

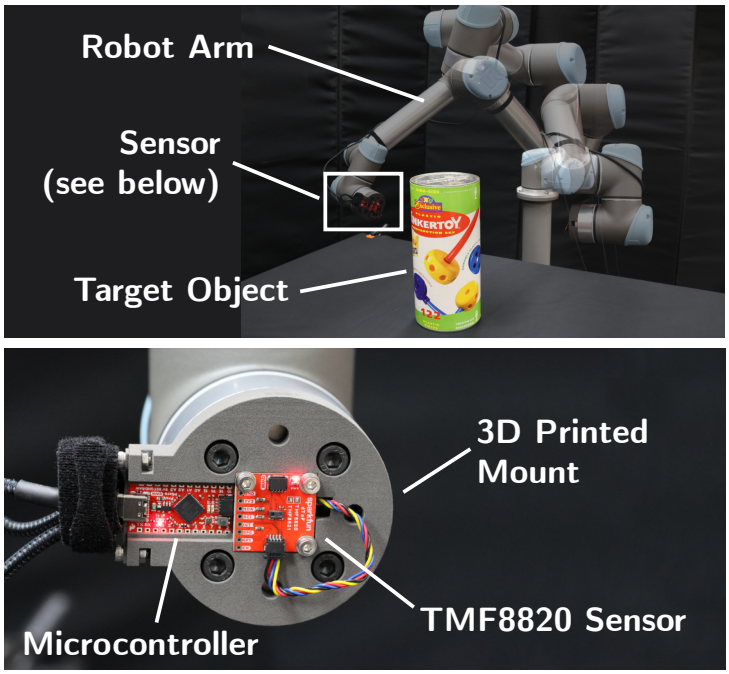


Figure 3.5: **Hardware prototype.** To capture real-world data from a wide set of viewpoints, we mount the TMF8820 proximity sensor to a robot arm. Forward kinematics of the robot are used to gather sensor pose.

Real-World Experiments

Experiment Setup

We capture a real-world tabletop dataset of eight objects of varying geometry and texture. To capture many posed views of the target object, we attach the sensor to a Universal Robots UR5 robot arm (Figure 3.5). We program the arm to automatically move to a set of poses and record sensor measurements at each pose. To obtain sensor poses, we use the forward kinematics of the robot, which are accurate to ± 0.5mm (Pollák et al., 2020). Each object is captured from between 128 and 240 viewpoints. Five of the objects are simple geometric primitives, for which we manually generate ground-truth meshes based on the dimensions of the target object and

Method	Chamfer Distance (mm) ↓ Rec → GT / GT → Rec									
	Big Box*		Block*		Pyramid*		Toy Container†		Cereal Box†	
Reproj. (Peak)	77.4	24.9/52.5	51.8	12.8/39.0	94.7	17.5/77.1	71.0	24.9/46.0	49.3	17.3/31.9
Reproj. (Thresh.)	67.5	14.8/52.7	52.3	8.5/43.8	75.4	**5.9**/69.5	52.4	8.9/43.5	51.6	19.2/32.4
Space Carving	67.9	35.1/32.8	69.2	33.4/35.8	80.1	39.5/40.6	98.9	52.8/46.0	44.1	24.4/19.6
Ours	**12.5**	**6.1**/ **6.4**	**9.8**	**5.6**/ **4.2**	**18.4**	9.0/ **9.3**	**11.5**	**5.8**/ **5.6**	**16.3**	**8.3**/ **8.0**

Table 3.2: **Quantitative results on real-world captures**. Our method more accurately reconstructs real-world objects with homogeneous (*) and rich (†) texture. Reprojection yields sparse and unevenly distributed points, harming one-way Chamfer from GT to reconstruction.

measurements of its position from the robot's forward kinematics. Meshes are trimmed to an axis-aligned bounding box 16cm larger than the target object in each dimension before the Chamfer distance calculation.

Reconstruction Results

As seen in Table 3.2, our method outperforms all baselines by a wide margin as measured by two-way Chamfer distance. While reprojection is at times competitive in one-way distance from reconstruction to ground truth, it performs substantially worse in the opposite direction owing to the sparse and unevenly distributed point cloud generated, as visualized in Figure 3.6. While space carving outperforms reprojection on simulated data under a highly structured sensor pose distribution (*i.e.*, all sensors are facing the center of the object), it yields poor results on real-world scenes, in which we vary sensor orientation by $\pm 10^\circ$ to emulate real-world capture conditions and increase coverage. By contrast, our method benefits from the more varied sensor poses as is shown in Figure 3.7.

Further, our method is surprisingly robust to violation of assumptions made about surface reflectance; it successfully reconstructs non-Lambertian objects with rich texture despite assuming Lambertian BRDF with a spatially uniform albedo. These include both simple shapes (Toy Container in Figure 3.6 and Cereal Box) and challenging objects with

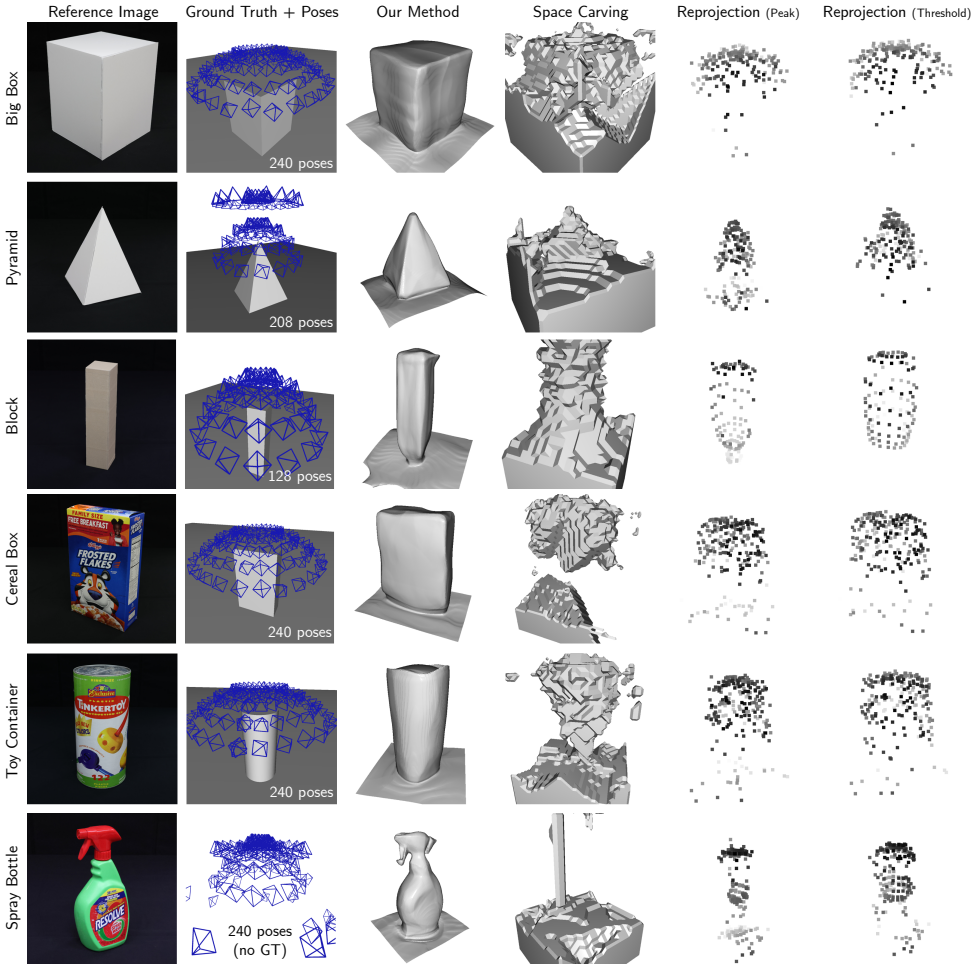


Figure 3.6: **Qualitative results on real-world captures.** Our method again attains the highest reconstruction quality. Poses in column two are subsampled by a factor of two for clarity.

complex geometry (Spray Bottle in Figure 3.6). We hypothesize that the overlapping FoVs of distributed sensors help constrain the optimization of our model and encourage a plausible reconstruction that best explains all transients.

Our method falls short on highly specular scenes, such as the glossy

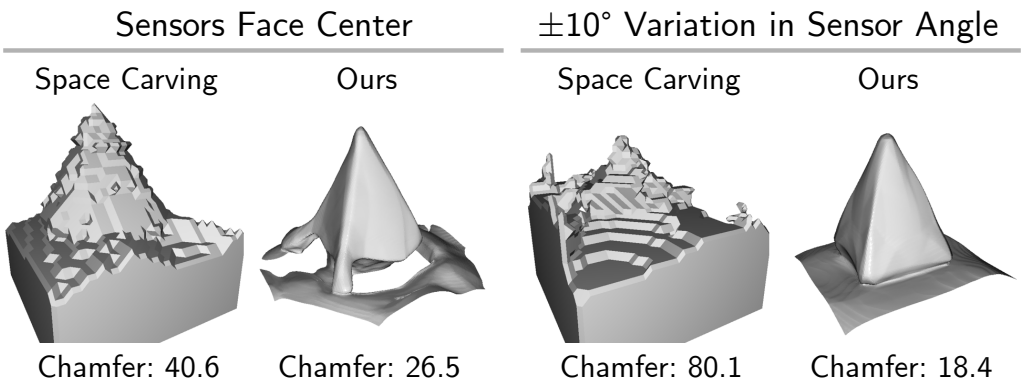


Figure 3.7: **Variation in sensor pose facilitates accurate reconstruction.** Space carving performs more poorly when sensor poses include some variation in target point, while our system takes advantage of the increased view diversity and coverage.

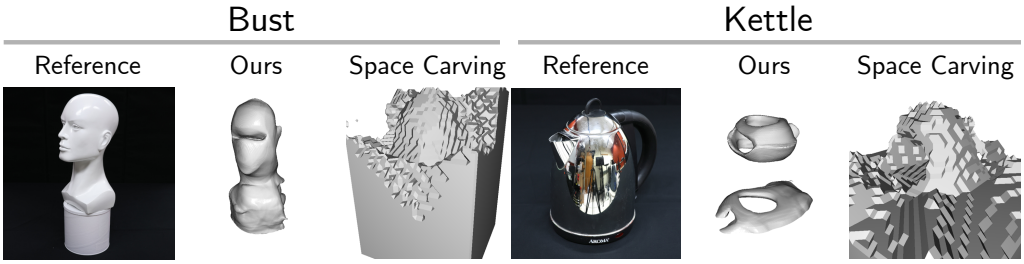


Figure 3.8: **Failure cases.** Because our reconstruction method assumes a Lambertian surface, it fails to reconstruct highly specular scenes, such as a glossy white bust (left) or mirror-finish kettle (right).

white bust and mirror-finish kettle in Figure 3.8. Nevertheless, it compares favorably to baseline methods and often yield meaningful shapes that may serve as an initial guess. We conjecture that proper lighting models and carefully distributed sensor poses are essential to the reconstruction of these challenging scenes.

Overall, our strong results on both simulated and real-world datasets validate our modeling approach and demonstrate a promising single-

photon 3D vision system for real-world scene reconstruction.

3.5 Conclusion and Discussion

We presented a method for recovering 3D geometry based on measurements from distributed single-photon cameras realized using low-cost proximity sensors, unlocking a new avenue of granular sensing on existing commodity hardware. Despite assuming spatial uniform reflectance and albedo, our method is robust to rich textures (Fig. 3.6) and compares favorably to baseline methods for reconstructing challenging scenes with high specularities (Fig. 3.8). Future work will investigate recovering spatially varying reflectance or incremental learning of geometry as new measurements become available (Sucar et al., 2021; Ortiz et al., 2022), enabling applications like real-time mapping and SLAM. Our method may be particularly relevant in applications such as robotics and wearable computing, where the small size, low power requirements, and robust hardware of proximity sensors are very valuable.

Beyond Lambertian Assumption

Our method assumes a spatially uniform Lambertian BRDF, but in practice can effectively reconstruct objects with spatially varying albedo and slightly glossy appearance (*e.g.*, the spray bottle). In theory, our method can easily be adapted to incorporate a parametric lighting model. Recovery of the parameters of such a model are likely possible because, by sharing information among many observations, the BRDF is effectively sampled at many incident and exitant angles. An intriguing direction for future work is investigating which BRDF parameterizations can be recovered with our imaging setup, and the effect of the reflectance model on reconstruction quality. We suspect that a non-parametric NeRF-like BRDF would not be suitable as it does not sufficiently constrain the optimization. A parametric

lighting model, *e.g.*, Phong (Phong, 1975) or Oren-Nayar (Oren and Nayar, 1994) may appropriately constrain the optimization while allowing the model to learn a more accurate scene representation.

Runtime Efficiency

Our method takes on the order of hours to reconstruct a scene, making it unsuitable for real-time applications in its current state. Future work should investigate ways to speed up forward rendering and model training, such as using plenoxels (Fridovich-Keil et al., 2022), multi-resolution hash encodings and custom CUDA kernels (Wang et al., 2023). Improved importance sampling and better initialization schemes would likely yield modest improvements in convergence time. Another option is to render only summary statistics of the histogram (*e.g.*, mean, peak locations or widths) rather than the entire histogram, which would likely be faster to render at the expense of yielding a lower-quality reconstruction.

Calibration of Sensor Pose

In this work, we used an industrial robot arm to gather posed sensor measurements. We chose this modality as it is guaranteed to provide highly accurate sensor poses, and allows control over precise sensor placement. For applications like wearable computing, camera poses might be pre-calibrated. Alternatively, the low-cost single-photon camera could be combined with a sensor-based localization system (*e.g.*, an IMU based (Yi et al., 2007) or a camera based (Mur-Artal et al., 2015) system) to recover camera pose, a setup which is standard in related works (Mildenhall et al., 2021; Ortiz et al., 2022). Such a capture setup would allow capture of more organic and large scale scenes, which more closely mimic the potential use cases of the sensor (*e.g.*, on mobile robots and drones).

Comparison to Other 3D Imaging Modalities

Our work provides a low-cost 3D imaging system using single-photon cameras. We provide detailed comparisons between our method and baseline methods, but do not compare our reconstructions to those gathered from other 3D modalities, such as continuous wave time-of-flight (Attal et al., 2021), LiDAR (Huang et al., 2023) or visual SLAM (Macario Barros et al., 2022) and multi-view stereo (Seitz et al., 2006) with conventional RGB images. Future work should provide a comparison to these other modalities to provide insights into the niche (in terms of accuracy, size, power, etc.) filled by each.

Commodity Sensors

One challenge for future work is a lack of hardware support for measurement and use of transient histograms. Very few low-cost sensors allow access to transient histograms, and those that do often perform preprocessing that is proprietary or undocumented. Our work has demonstrated the significance of sensor modeling for accurate 3D reconstruction. We hope that manufacturers will see value in users having access to transient histogram data and support the use of this data with documentation and low-level access in the future.

4 NON-LINE-OF-SIGHT RECONSTRUCTION FOR HIGH-SPEED IMAGING

In this chapter, we present a second use case of our imaging framework for non-line-of-sight (NLOS) scene reconstruction with a high-speed imaging system[1].

4.1 Introduction

Time-resolved NLOS imaging recovers information about hidden scenes based on indirect reflectance scattered by the surrounding environment (*e.g.*, a relay wall) (Velten et al., 2012). It has the potential to revolutionize many critical applications such as medicine, robotics, military and law enforcement operations, and scientific imaging.

Imaging Setup

A typical *three-bounce* model for NLOS imaging is illustrated in Figure 4.1. To image a hidden scene $\mathcal{S}$ behind an occluder, the observer sends a periodic train of collimated laser pulses from $\mathbf{l}_0$ to illuminate a point $\mathbf{l}$ on a diffuse relay wall, which scatters the light towards the hidden scene (1^{st} bounce). The portion of light reaching the scene is further reflected (2^{nd} bounce), and a fraction of it hits the wall again and heads back to the

[1]This is joint work with Sicheng Mo, Jiayong Peng, Xiaochun Liu, Ji Hyun Nam and Sid Raghavan, under the supervision of Andreas Velten and Yin Li. Fangzhou led the project, designed the study, developed the algorithm, performed data simulation, conducted most experiments, interpreted the results, and prepared the figures. Sicheng conducted experiments and prepared the figures. Jiayong analyzed the baselines. Xiaochun contributed to algorithm design. Ji Hyun built the hardware prototype and helped collect real-world data. Sid helped collect real-world data. All participants are involved in paper writing(Mu et al., 2022).

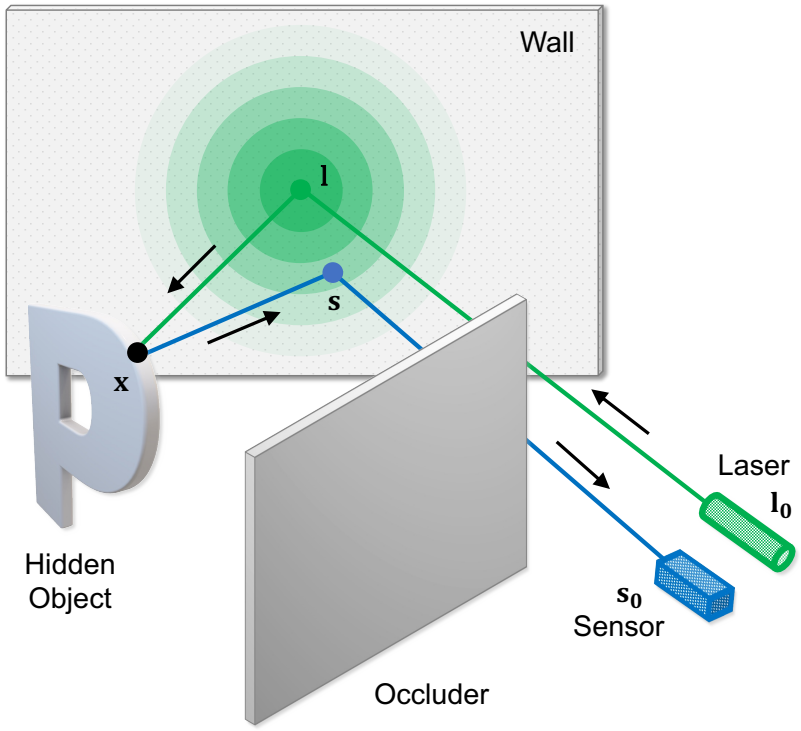
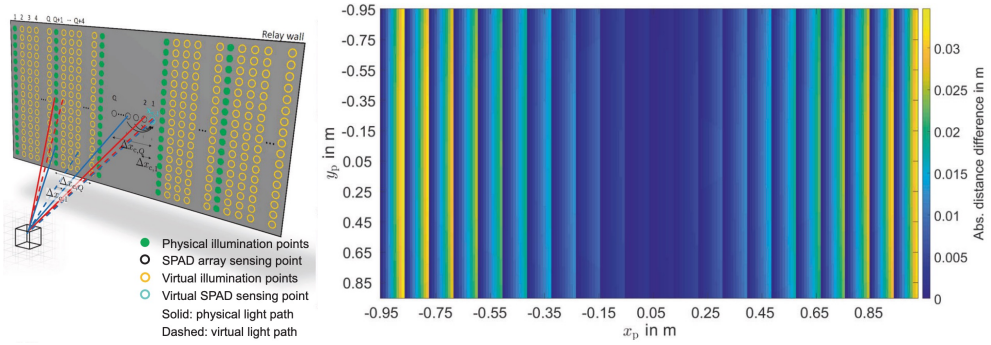


Figure 4.1: **NLOS imaging setup.** A pulsed laser at $\mathbf{l}_0$ illuminates a relay wall. The light bounces off the wall, interacts with the occluded scene, scatters to the wall again, and is finally captured by a time-resolved detector at $\mathbf{s}_0$. Figure adapted from Mu et al. (2022).

observer (3^{rd} bounce). A SPAD detector at $\mathbf{s}_0$ records photons returning from a point $\mathbf{s}$ on the wall.

One way to simplify this three-bounce model is to convert it into an equivalent, direct line-of-sight model with a *single* bounce. This is possible because $\mathbf{l}$ can be thought of as a virtual diffuse light source, and $\mathbf{s}$ can be similarly considered as a virtual wide-FoV detector. With this simplification, NLOS imaging becomes a special case of our imaging framework, with the laser and detector both virtually placed on a 2D plane (*i.e.*, the relay wall).



(a) Pixel Remapping (b) Remapping Error Map

Figure 4.2: **Pixel remapping.** (a) A sparse scan pattern with pixel remapping. Significant speedup is achieved by skipping scanlines and fill in the gaps later with pixels from slightly misaligned positions in a 1D detector array. (b) Pixel remapping results in non-uniform geometry error, visualized as a heap map. Figure adapted from Nam et al. (2021).

In practice, a full NLOS measurement contains a 2D array of transient histograms $\mathbf{H} = \{\mathbf{h}_{(\mathbf{l}_{ij},\mathbf{s}_{ij})}\}_{i=1\ldots H,j=1\ldots W}$ sampled at a dense $H \times W$ grid of virtual detector locations $\mathbf{s}_{ij}$ on the wall. The placement of virtual light sources $\mathbf{l}_{ij}$ differentiates *confocal* and *non-confocal* NLOS imaging.

Confocal *vs.* Non-confocal NLOS

In the traditional confocal setting, the virtual detector is co-located with the virtual light source, that is, $\forall i, j, \mathbf{l}_{ij} = \mathbf{s}_{ij}$. By contrast, non-confocal NLOS assumes a fixed virtual laser located at the center of the wall $\mathbf{l}$, that is, $\forall i, j, \mathbf{l}_{ij} = \mathbf{l}$. Most research to date (O'Toole et al., 2018; Lindell et al., 2019; Chen et al., 2020) studies confocal reconstruction due to its simpler physics, leaving the more challenging non-confocal reconstruction problem less explored.

High-Speed NLOS Imaging

One key advantage of the non-confocal setup is its rapid data acquisition speed. A confocal measurement is acquired by a dense raster-scan of the wall, which takes tens of seconds with the existing hardware (Lindell et al., 2019). By contrast, the state-of-the-art non-confocal imaging system runs at 5 FPS (Nam et al., 2021), enabling numerous applications that require high-speed imaging.

The fast acquisition speed of Nam et al. (2021) is made possible with two key ideas: (1) the Helmholtz reciprocity of light transport; (2) a sparse scan pattern with local pixel remapping. In particular, pixel remapping, as illustrated in Figure 4.2, enables the collection of *dense* measurements with a *sparse* raster scan that skips $> 90\%$ of the scanlines for a significant speedup. This is achieved by capturing multiple transients simultaneously using a one-dimensional SPAD array and mapping pixels to skipped scanlines at the expense of some approximation errors in geometry.

In this work, we study this non-confocal imaging approach (Nam et al., 2021), which we believe represents a more promising direction for future development of NLOS imaging systems.

Challenges

Despite favorable date acquisition rate, non-confocal NLOS imaging brings approximation errors to the lighting model, most notably through pixel remapping, thereby posing new challenges for hidden scene reconstruction. Previous methods that rely on the precise modeling of light paths (Tsai et al., 2019; Shen et al., 2021) will inevitably fall short due to the inaccurate physics in this setup. Recent development has thus turned to the modeling of wave propagation (Liu et al., 2019b, 2020; Jiang et al., 2021), resulting in methods that can tolerate modest errors in the lighting model. Despite their robustness, these methods oftentimes yield reconstructions that lack

fine details (*e.g.*, textures and edges), and do not incorporate any prior about the hidden scenes.

Contributions

To bridge this gap, we propose to embed physical models, consisting of an wave propagation module and a volume renderer, into a deep neural network for non-confocal NLOS reconstruction. Specifically, the wave propagation module adapts the Rayleigh-Sommerfeld diffraction (RSD) operator for feature propagation, and the volume renderer is inspired by neural transient field (Shen et al., 2021). Our key intuition is that using wave propagation helps regularize the solution space of the volume renderer, alleviating dependency on the accurate modeling of light transport, and thus leading to robust generalization beyond an idealized transient formation model.

Further, we devise a unified learning framework that enables flexible training of our model using diverse supervision signals, including intensity images and transient histograms. Once trained, our model renders both intensity and depth images at inference time in a single forward pass at an interactive rate.

Finally, we showcase several benefits of our method through extensive experiments. First, our model, despite being trained on simulated data, generalizes well on real-world captures. Second, our method, when implemented on a high-end GPU, processes 11.8 frames per second (FPS), thus paving the ways for fast NLOS imaging. Finally, our method supports key functionalities beyond NLOS reconstruction; it can synthesize images from a non-frontal view (*i.e.*, novel view synthesis), and the learned features can facilitate accurate NLOS object recognition.

4.2 Related Work

NLOS imaging has recently gained popularity with many applications including object detection (Scheiner et al., 2020; Chen et al., 2020), tracking (Scheiner et al., 2020; Smith et al., 2018), and human pose estimation (Isogawa et al., 2020). Several imaging systems (Velten et al., 2012; Lindell et al., 2019; Musarra et al., 2019; Nam et al., 2021) have been developed to capture multi-bounce indirect reflections scattered by the surrounding environment of a hidden scene. These systems have enabled NLOS reconstruction methods that recover an "image" of the hidden scene. Most relevant to our work are methods that reconstruct the appearance and/or geometry of a hidden scene using active illumination and time-resolved sensors. Faccio et al. (2020) provides a comprehensive review of NLOS imaging.

Physics-based NLOS Reconstruction

Since the seminal work of Kirmani et al. (2009), physics-based NLOS reconstruction has seen rapid progress, with methods falling into one of the four categories, namely back-projection methods (Velten et al., 2012; O'Toole et al., 2018), wave propagation methods (Lindell et al., 2019; Liu et al., 2019b, 2020; Nam et al., 2021), iterative optimization methods (La Manna et al., 2018; Tsai et al., 2019; Iseringhausen and Hullin, 2020) and geometry-based methods (Tsai et al., 2017; Xin et al., 2019). Our model draws insight from the wave propagation method of Liu et al. (2019b), and shares learning objectives with the iterative optimization method of Shen et al. (2021).

Learning-based NLOS Reconstruction

Learning-based methods have recently emerged for NLOS reconstruction. Grau Chopite et al. (2020) represents the first deep model for NLOS reconstruction. It trains a U-Net (Ronneberger et al., 2015) on simulated

depth maps for depth estimation of hidden scenes. More recently, Chen et al. (2020) proposed to learn feature embeddings from simulated data for NLOS reconstruction and recognition. Shen et al. (2021) introduced neural transient field (NeTF) for the implicit modeling of hidden scenes. Zhu and Cai (2022) developed a deep generative model for NLOS imaging using inexpensive commercial LiDAR.

Real-Time NLOS Imaging

Efficient hardware (Nam et al., 2021; Liao et al., 2021) and software implementations (Arellano et al., 2017; Liu et al., 2020; Jiang et al., 2021) have been developed for real-time NLOS imaging and reconstruction. Our method is tailored for the low-latency hardware prototype of Nam et al. (2021), and incorporates the fast RSD implementation of Liu et al. (2020).

Theoretical Analysis of NLOS Visibility

Liu et al. (2019a) demonstrates that hidden scenes positioned in certain poses are inherently not recoverable with a physics-based approach. Our approach goes beyond this theoretical limit by learning statistical priors from large datasets.

4.3 Method

Figure 4.3 provides an overview of our method. Our deep model consists of three key components. An physics-inspired encoder first extracts features from the time-domain histograms and transforms the features via through feature-space wave propagation. A neural radiance field (NeRF) is subsequently conditioned on the projected features to represent the shape and appearance of the hidden scene. This neural representation is subsequently volume-rendered into the desired target (*i.e.*, intensity

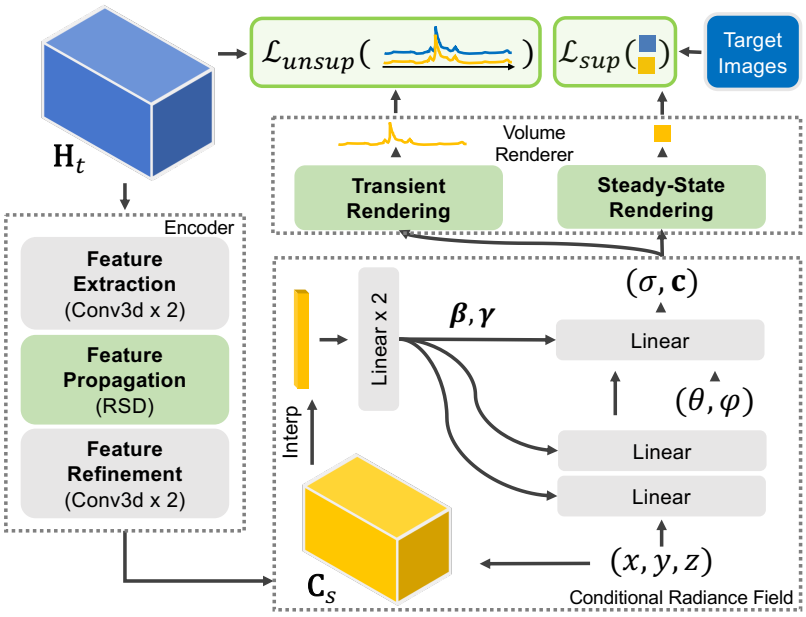


Figure 4.3: **Method Overview.** Our model consists of an encoder, a conditional radiance field, and a volume renderer. Our volume renderer can synthesize 2D intensity images using a steady-state forward model for supervised training, or transient histograms using a transient forward model for unsupervised training. At inference time, our method renders 2D images in a feed-forward manner for NLOS reconstruction. Modules in green are physics-based and parameter-free. Modules in gray have learnable parameters.

images and/or transients) for training and inference. We now describe each component in detail.

Physics-Inspired Encoder

Feature Extraction

Given a 2D array of histograms $\mathbf{H}$ of size $H \times W \times T$, a strided 3D convolution with a kernel size of 3 immediately down-samples the input along all spatiotemporal axes. This yields a feature cube $\mathbf{C}_t$ of size $H/2 \times W/2 \times$

$T/2 \times C$, where C is the number of feature channels. $\mathbf{C}_t$ is processed by another 3D convolutional layer with a kernel size of 3 before undergoing RSD wave propagation as detailed next. Compared to the ResNet-like design of LFE (He et al., 2016; Chen et al., 2020), our feature extraction module is more parameter-efficient and facilitates stable training.

Feature-Space Wave Propagation

A non-confocal NLOS measurement represents the impulse response of the hidden scene with the light source at $\mathbf{l}$ and detector at $\mathbf{s}$. Liu et al. (2019b, 2020) previously showed that one may convolve the measurement with any pulse waveform $\mathcal{P}_l(t)$ to simulate a *virtual* line-of-sight transient camera from behind the wall. Importantly, this provides a means to use the Rayleigh-Sommerfeld diffraction (RSD) theory for NLOS reconstruction. Further, Nam et al. (2021) observed that RSD is resilient to pixel remapping, a key approximation that enables fast NLOS imaging yet breaks existing reconstruction methods.

Inspired by this observation, we interpret $\mathbf{C}_t$ as the *featurized* impulse response and leverage RSD as a robust physical model for the transformation of $\mathbf{C}_t$

$$\mathbf{C}_s = \mathrm{RSD}(\mathcal{P}_l(t) * \mathbf{C}_t), \tag{4.1}$$

where $*$ is discrete 1D convolution and $\mathrm{RSD}(\cdot)$ is applied independently on each feature channel. The transformed feature cube $\mathbf{C}_s$ has a size of $H/2 \times W/2 \times D \times C$, where D is the number of depth planes. Note that the RSD operator is fully differentiable, thus enabling the end-to-end training of our model. We adopt the efficient RSD implementation of Liu et al. (2020) and refer readers to Liu et al. (2020) for the derivation and implementation details.

Feature Refinement

The output of RSD at a reduced resolution is prone to artifacts due to aliasing. As the last step, the encoder refines $\mathbf{C}_s$ with two 3D convolutional layers with a kernel size of 3. The refined features are then fed into a conditional neural radiance field.

Conditional Neural Radiance Field

Central to our deep model is a radiance field $f_\theta : (\mathbf{x}, \mathbf{d}; \mathbf{C}_s) \rightarrow (\mathbf{c}, \sigma)$ that is *conditioned* on the features $\mathbf{C}_s$. Similar to NeRF (Mildenhall et al., 2021), f_θ is realized as an MLP with learnable weights θ. It maps spatial locations $\mathbf{x}$ and viewing directions $\mathbf{d}$ to intensity or color $\mathbf{c}$ and volume density σ. Unlike NeRF and its transient variant NeTF (Shen et al., 2021) that learn separate θ for each scene through lengthy iterative optimization, our scene representation shares the same θ across all scenes. The network activations of f_θ are dynamically modulated given the conditioning features $\mathbf{C}_s$ as shown in Figure 4.3.

Our key intuition is in three-folds: (1) learning a shared θ facilitates the distillation of scene priors from diverse training data; (2) conditioning on scene-dependent features enables fast *feed-forward* reconstruction; and most importantly, (3) f_θ seamlessly bridges the physics-inspired encoder and volume renderer, thereby placing strong constraints on their respective learning. In doing so, our model, despite being trained exclusively on simulated data, generalizes well on real-world captures as we demonstrate in our experiments.

Conditioning Mechanism

Our conditioning mechanism goes as follows. We sample $\mathbf{C}_s$ at $\mathbf{x}$ via tri-linear interpolation and feed it into a small MLP to predict the affine weight $\boldsymbol{\gamma}_i$ and bias $\boldsymbol{\beta}_i$ for the activations $\mathbf{h}_i$ of the i^{th} layer. The activations

are then transformed as

$$\mathbf{h}_i' = \boldsymbol{\gamma}_i \odot \mathbf{h}_i + \boldsymbol{\beta}_i, \tag{4.2}$$

where $\odot$ stands for element-wise multiplication. Note that $\mathbf{C}_s$ injects scene-dependent information into the activations $\mathbf{h}_i'$ through the affine parameters. This approach is inspired by π-GAN (Chan et al., 2021), ECRF (Liu et al., 2021) and pixelNeRF (Yu et al., 2021), which similarly condition a radiance field on latent vectors for image generation, object editing, and novel view synthesis.

Volume Rendering

One key strength of our volume rendering framework is its ability to render the conditional radiance field f_θ into measurements of any sensor type given an appropriate forward model. We explore two such models for training and inference, namely the *steady-state rendering* of 2D intensity images and *transient rendering* of transient histograms. Importantly, these models are fully differentiable, thus enabling the end-to-end training of our deep model.

Steady-State Rendering

We adopt the volume rendering equation of NeRF (Mildenhall et al., 2021) (Equation 2.8) for the steady-state rendering of f_θ. We further calculate a depth value $\hat{\mathbf{D}}$ in a 2D depth image as follows:

$$\hat{\mathbf{D}}(\mathbf{r}) = \int_0^\infty T(u)\sigma(\mathbf{r}(u))u\,du. \tag{4.3}$$

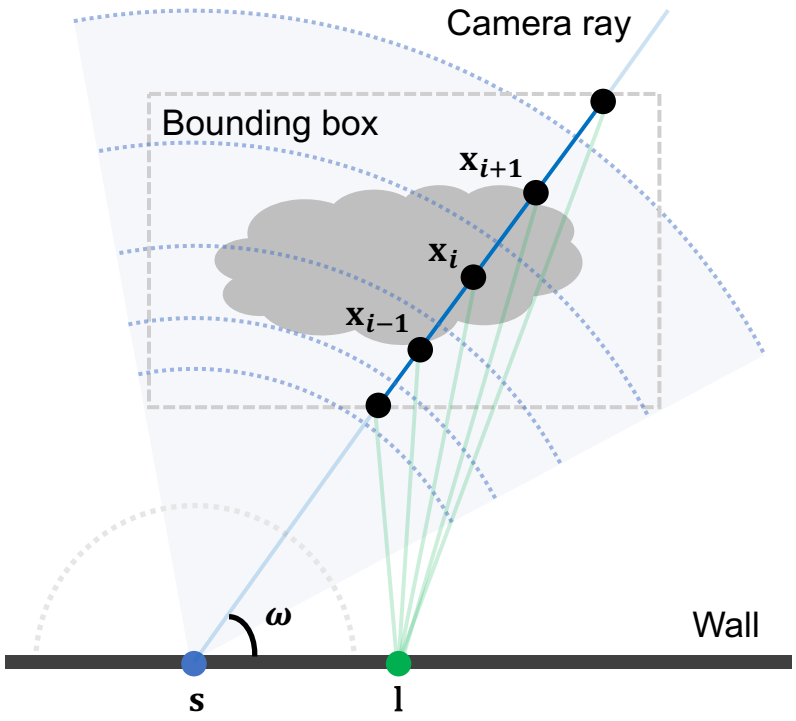


Figure 4.4: **Ray and point sampling for transient rendering.** Rays originating from a virtual sensor $\mathbf{s}$ are uniformly drawn from within the cone shaded in blue so that they always intersect the bounding volume (gray dotted box). Points $\mathbf{x_i}$ are sampled along a ray such that the length of path $\mathbf{l} \rightarrow \mathbf{x_i} \rightarrow \mathbf{s}$ is uniformly distributed. Note that the length of camera subpath $\mathbf{s} \rightarrow \mathbf{x_i}$ is *not* uniformly distributed (blue dotted arcs), and no point is drawn from light subpath $\mathbf{l} \rightarrow \mathbf{x_i}$ (green solid lines) as we only model outgoing radiance.

Transient Rendering

We approximate the surface integral in the transiant formation model (Equation 2.1) with a volume integral. Concretely, we calculate the flux $\hat{\tau}_i$ of the i^{th} bin by summing the radiance of all points $\mathbf{x}$ from within a hemi-ellipsoidal shell with the light travel time along the (reversed) path

$\mathbf{s} \to \mathbf{x} \to \mathbf{l}$ falling inside $[i\Delta t, (i+1)\Delta t)$:

$$\begin{aligned} \hat{\tau}_i &= \int_{i\Delta t}^{(i+1)\Delta t} \hat{L}(t)dt, \\ \hat{L}(t) &= \int_{\phi}\int_{\gamma} T(\mathbf{r}, u(t,\gamma))\sigma(\mathbf{r}(u(t,\gamma)))\mathbf{c}(\mathbf{r}(u(t,\gamma)), \mathbf{d})d\gamma d\phi, \\ u(t,\gamma) &= \frac{c^2t^2 - \|\mathbf{s}-\mathbf{l}\|_2^2}{2ct - 2\cos\gamma}. \end{aligned} \tag{4.4}$$

In Equation 4.4, u is the distance from $\mathbf{s}$ to $\mathbf{x}$ in the ray direction $\mathbf{d} = (\gamma, \phi)$, with the path $\mathbf{s} \to \mathbf{x} \to \mathbf{l}$ having a length of ct, and c the speed of light.

Practical Considerations

In practice, we interchange the order of integrals in Equation 4.4, and apply the line-to-point sampling strategy of Jarabo et al. (2014) for the *unbiased* estimation of $\hat{\tau}_i$. Further, we assume an axis-aligned bounding box (AABB) around the hidden scene, and follow Ureña et al. (2013) to draw ray directions from within the spherical projection of the AABB's face facing the wall. As illustrated in Figure 4.4, this sampling approach encourages the rays and in turn the sampled points along a ray to concentrate around the hidden scene.

Comparison to NeTF

Our transient rendering recipe differs from NeTF (Shen et al., 2021) in three aspects. First, we estimate the outgoing radiance at a scene point without factoring it into irradiance and BRDF. This simplification, in line with NeRF (Mildenhall et al., 2021), allows the implicit modeling of occlusion and multi-bounce lighting, and better supports feed-forward rendering. Second, we present a more principled and efficient sampling framework for flux estimation. Finally, we empirically found that our conditional parametrization bootstraps the learning of f_θ, and thus none of

the advanced training techniques (*e.g.*, two-stage training and hierarchical sampling) from Shen et al. (2021) is needed. Our experiments show that our recipe can be trivially adapted for iterative optimization and yields better reconstruction than NeTF.

Model Training and Inference

In the absence of a large dataset of real measurements, we train our model on simulated data. Our model naturally supports two training strategies based on the analysis-by-synthesis principle: (1) *supervised learning* by comparing rendered and ground-truth 2D intensity images; (2) *unsupervised learning* by matching rendered and ground-truth transients. At inference time, we render 2D intensity and depth images, and interpret them as reconstructions of the hidden scene.

Supervised Training

When multi-view instensity images are available in the training data, we minimize the mean squared error (MSE) between the rendered and ground-truth pixel values

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} |\hat{\mathbf{I}}(\mathbf{r}) - \mathbf{I}(\mathbf{r})\|_2^2, \tag{4.5}$$

where $\mathcal{R}$ is the number of pixels.

Following Lombardi et al. (2019), we add a Beta distribution prior on the cumulative transmittance $\mathsf{T}_{\mathbf{r}}$:

$$\mathcal{L}_{\text{Beta}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \log(\mathsf{T}_{\mathbf{r}}) + \log(1 - \mathsf{T}_{\mathbf{r}}), \tag{4.6}$$

and a total variation (TV) prior on log opacities:

$$\begin{aligned} \mathcal{L}_{TV} &= \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \sum_{i} \|\Delta \log \alpha(\mathbf{r})_i\|_1, \\ \alpha(\mathbf{r})_i &= 1 - \exp\left(\sigma(\mathbf{r}(s_i))\Delta s_i\right) \end{aligned} \tag{4.7}$$

where i indexes over the sampled steps along $\mathbf{r}$, and $\alpha(\mathbf{r})_i$ is the opacity at $\mathbf{r}(s_i)$ over the discretized step Δs_i.

The full training objective is a weighted combination of the three terms

$$\mathcal{L}_{sup} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{Beta}\mathcal{L}_{Beta} + \lambda_{TV}\mathcal{L}_{TV}, \tag{4.8}$$

where λ_{MSE}, λ_{Beta} and λ_{TV} are the respective loss weights.

Unsupervised Training

In the absence of target images, we minimize the Poisson negative log-likelihood[2] of the rendered transients:

$$\mathcal{L}_{Poisson} = \frac{1}{|\mathcal{S}||\mathcal{B}|} \sum_{\mathcal{S}} \sum_{\mathcal{B}} \hat{\tau}_{\mathbf{s},b} - n_{\mathbf{s},b} \log(\hat{\tau}_{\mathbf{s},b}), \tag{4.9}$$

where $\hat{\tau}_{\mathbf{s},b}$ is the rendered flux of the b^{th} bin at virtual sensor location $\mathbf{s}$, $n_{\mathbf{s},b}$ the photon counts in the same bin of the target histogram, $\mathcal{B}$ the set of rendered bins, and $\mathcal{S}$ the set of sampled sensor locations. The full objective is

$$\mathcal{L}_{unsup} = \lambda_{Poisson}\mathcal{L}_{Poisson} + \lambda_{Beta}\mathcal{L}_{Beta} + \lambda_{TV}\mathcal{L}_{TV}, \tag{4.10}$$

where $\lambda_{Poisson}$ is the weight for $\mathcal{L}_{Poisson}$.

[2]We omit the constant term $\log(n_{\mathbf{s},b}!)$ for conciseness.

Joint Training

One may further combine the supervised and unsupervised learning objectives for the joint training of our model:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{sup}} + \mathcal{L}_{\mathrm{unsup}}. \tag{4.11}$$

Inference

At inference time, we equip the volume renderer with the steady-state forward model (Equation 2.8 and 4.3) to synthesize intensity and depth images of the hidden scene in a single forward pass. Our model runs at *11.8 FPS* on an NVIDIA V100 GPU. This exceeds the data acquisition rate (5 FPS) of the imaging system and potentially enables real-time NLOS reconstruction. RSD (Liu et al., 2020) and LFE (Chen et al., 2020) run at 380 FPS and 30.8 FPS on the same GPU, whereas NeTF (Shen et al., 2021) requires 6 hours of training on each scene.

4.4 Experiments

We demonstrate the effectiveness of our method for non-confocal NLOS reconstruction in simulation and on real-world captures with the high-speed imaging system of Nam et al. (2021). We report qualitative and quantitative results for both settings. Further, we present object recognition results with the learned features to showcase the strength of our method for representation learning.

Implementation Details

For supervised training, We sample $4,096$ rays uniformly at random from all target views, and use λ_{MSE}=1, λ_{Beta}=0.0001 and λ_{TV}=0.01 in our experiments. For unsupervised training, we use $\lambda_{\mathrm{Poisson}}$=1, λ_{Beta}=0.0001

and λ_{TV}=0.01. We sample one sensor location at a time and draw 4,096 rays originating from the sensor as discussed in Section 4.3. We combine the two sets of ray samples in joint training. The models are trained for 50 epochs using the Adam optimizer (Kingma and Ba, 2014) with a mini-batch size of 2 and a learning rate of 0.0001. Training on measurements of size $128 \times 128 \times 512$ takes 3 hours on a single NVIDIA V100 GPU and requires 8 GB of memory.

Baselines

We compare our method to three baselines: RSD (Liu et al., 2020), LFE (Chen et al., 2020) and NeTF (Shen et al., 2021).

RSD is the state-of-the-art physics-based method for non-confocal NLOS reconstruction based on phasor-field virtual wave optics. Using the phasor-field method, NLOS imaging can be interpreted as line-of-sight diffractive wave propagation, and thus can be solved using the Rayleigh-Sommerfeld diffraction theory for conventional line-of-sight imaging (Sommerfeld, 1964) . Our encoder subsumes RSD as its feature propagation module.

LFE is a learning-based method with an encoder-decoder architecture. It learns scene priors from large-scale datasets for improved reconstruction quality in comparison to physics-based methods. Similar to our method, LFE's encoder comprises a physical operator for feature propagation. Both models are trained end-to-end in simulation and output hidden view reconstructions in a single forward pass at real-time rates. Different than our method, LFE's convolutional decoder design lacks physical constraints, and the model only supports synthetic RGB or intensity images as supervision target.

NeTF is an iterative optimization method for NLOS reconstruction. Similar to NeRF (Mildenhall et al., 2021), NeTF represents the hidden scene as a neural implicit function, and learns a separate model for each

scene. At each iteration, it renders the scene into transients using volume rendering and minimizes their difference with respect to the ground-truth measurements. While our method has a similar volume rendering component, it learns a single model to reconstruct arbitrary scenes and solves reconstruction in a single forward pass. We further introduce *NeTF++*, a variant of NeTF with our volume renderer for an apple-to-apple comparison of the transient rendering recipes.

In our experiments, we adopt the official code release for the baselines and use their default hyper-parameters. For completeness, we additionally compare with three physics-based methods originally for confocal NLOS reconstruction, namely fitlered back-projection (FBP) (Velten et al., 2012), light-cone transform (LCT) (O'Toole et al., 2018) and f-k migration (Lindell et al., 2019). These methods can be adapted for non-confocal reconstruction through interpolation, yet at the expense of geometric distortion, reduced FoV, and loss in image resolution.

Evaluation Protocol

Following LFE (Chen et al., 2020), we evaluate all methods using root mean squared error (RMSE), peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) in our simulated experiments.

Simulated Experiments

Experiment Setup

We simulate two large datasets with the rasterizer from Chen et al. (2020). The first **alphanumerics** dataset contains 2,775 samples of 111 objects, including all lower and upper case letters from the English and Greek alphabets as well as digits 0 to 9. The transient measurements are $128 \times 128 \times 512$ (height $\times$ width $\times$ time) in size. The target images have a resolution of 256×256, and include 25 randomly posed views in addition

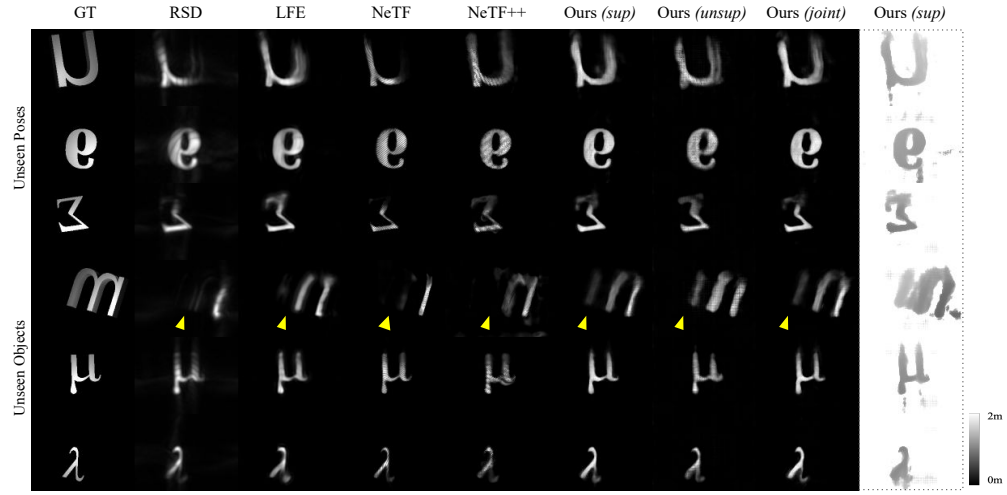


Figure 4.5: **Reconstruction results on the *alphanumerics* dataset.** Compared to the baselines, our method produces sharper reconstructions with finer details. Thanks to the learned scene priors, our method can infer missing scene content (yellow arrows). The rightmost column presents depth estimation of our supervised model.

to the canonical frontal view as in Chen et al. (2020). They both have a single brightness channel to match the real captures. We create a training split and two test splits. The training split has 2,000 samples of 100 objects, each with 20 poses. The "*Unseen Poses*" test split has the remaining 5 poses for each training object. The "*Unseen Objects*" test split has the remaining 11 objects and their full set of poses. We report results on both test splits to evaluate the generalizability of all methods.

The second **motorbikes** dataset contains 6,925 samples of 277 motorbikes from ShapeNet (Chang et al., 2015). Each object is again rendered in 25 random poses in addition to the canonical pose. The NLOS measurements are $256 \times 256 \times 512 \times 3$ (height $\times$ width $\times$ time $\times$ color) with RGB color channels for fair comparison to Chen et al. (2020). We generate training and test splits using the same protocol as before, with 5,000 samples of 250 motorbikes in the training split.

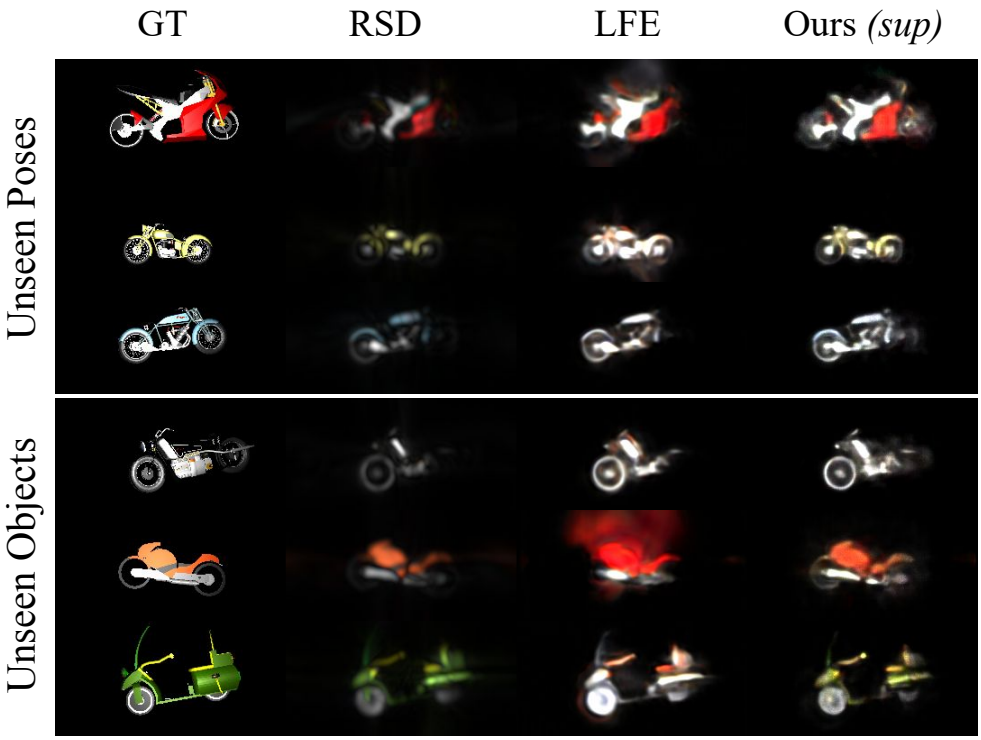


Figure 4.6: **Reconstruction results on the *motorbikes* dataset.** Reconstructions from our method achieve better color balance and contain geometry details (e.g., wheels) missed by RSD and LFE.

Finally, we simulate two small datasets for testing. The first **CMU** dataset contains six objects with complex geometry (Tsai et al., 2019). The second has two objects from the **Z-NLOS** dataset Galindo et al. (2019) and is previously used by NeTF (Shen et al., 2021). We use it in an ablation study to evaluate our transient rendering recipe.

Reconstruction Results

We report quantitative results on the test splits of **alphanumerics** and **motorbikes** in Table 4.1 and 4.2. Our supervised model consistently outperforms RSD and LFE by a wide margin. Our unsupervised model, which

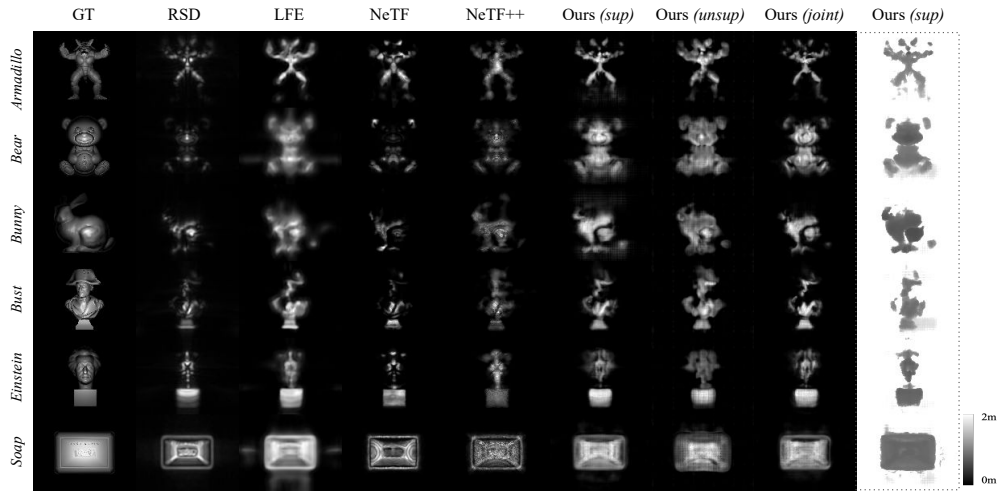


Figure 4.7: **Reconstruction results on the *CMU* dataset.** Our models generalize well on complex out-of-distribution shapes thanks to the strong regularization effect of the physics priors. The rightmost column presents depth estimation of our supervised model.

Methods	Unseen Poses			Unseen Objects		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
RSD	0.095	20.83	0.395	0.094	20.92	0.407
LFE	0.079	22.42	0.652	0.077	22.51	0.656
Ours (*sup*)	0.062	24.60	0.816	0.060	24.80	0.821
Ours (*unsup*)	0.093	20.97	0.709	0.092	21.02	0.726
Ours (*joint*)	**0.060**	**25.05**	**0.861**	**0.059**	**25.02**	**0.868**

Table 4.1: **Quantitative results on the full *alphanumerics* test sets.** Our method outperformsbaselines on both unseen poses and unseen objects.

is solely trained to enforce cycle consistency, compares favorably against RSD. Importantly, our model achieves the best results with the joint training objective, highlighting the benefit of our unified modeling approach. The qualitative results in Figure 4.5 and 4.6 demonstrate that our method reconstructs sharper contours and finer details, and in particular, is able to infer content that a physics-based method cannot recover thanks to the

Method	Unseen Poses			Unseen Objects		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
RSD	0.087	21.87	0.632	0.079	22.65	0.648
LFE	0.092	21.09	0.854	0.088	21.41	0.854
Ours (*sup*)	**0.065**	**24.05**	**0.881**	**0.062**	**24.59**	**0.886**

Table 4.2: **Quantitative results on the full *motorbikes* test sets.** Our supervised model outperforms baselines on reconstructing RGB images.

Method	Alphanumerics			CMU		
	RMSE	PSNR	SSIM	RMSE	PSNR	SSIM
RSD	0.084	22.02	0.395	0.086	21.54	0.456
LFE	0.064	24.34	0.886	0.082	21.84	0.700
NeTF	0.087	21.99	0.893	0.100	20.31	0.795
NeTF++	0.083	21.84	0.895	0.071	**23.26**	**0.815**
Ours (*sup*)	0.059	24.94	0.905	0.076	22.46	0.799
Ours (*unsup*)	0.073	23.12	0.833	**0.070**	23.18	0.775
Ours (*joint*)	**0.057**	**25.15**	**0.896**	0.079	22.11	0.798

Table 4.3: **Quantitative results on selected *alphanumerics* and *CMU* test samples.** Our method generalizes well on both in-distribution (**alphanumerics**) and out-of-distribution (**CMU**) samples in comparison to all baselines.

learned scene priors.

Moving forward, we investigate how models trained on **alphanumerics** generalize on the challenging out-of-distribution scenes from the **CMU** dataset. We present quantitative results averaged over the six available scenes in Table 4.3 and qualitative results in Figure 4.7. Our method outperforms RSD and LFE despite the increasing scene complexity. Notably, our unsupervised model performs even better than the supervised LFE model thanks to the strong regularization effect of the physics priors.

Our method outperforms NeTF and NeTF++ on **alphanumerics** [3]

[3]It is infeasible to train NeTF and NeTF++ on all test samples. We hence report results averaged over six random test scenes.

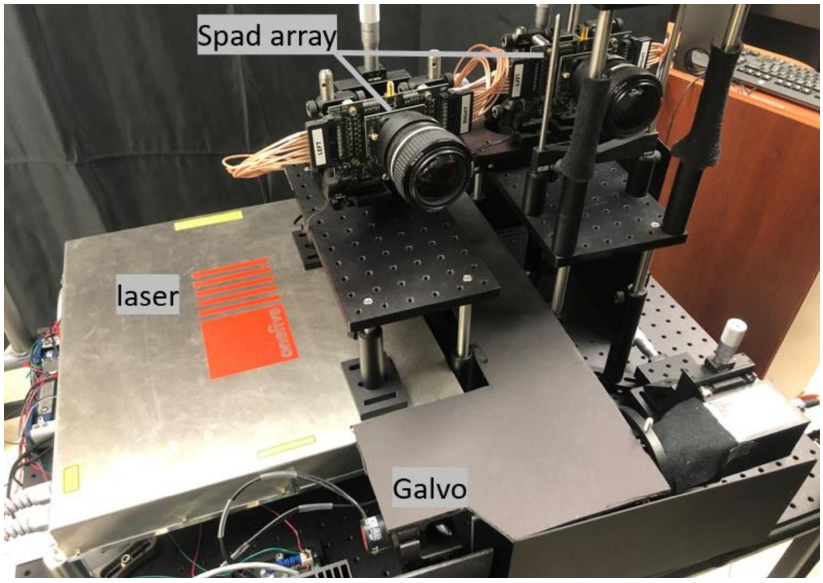


Figure 4.8: **Hardware prototype.** Our prototype includes an ultra-fast pulsed laser, two 1D SPAD arrays and a galvo for laser redirection.

(Table 4.3 and Figure 4.5) while supporting fast feed-forward inference. We attribute the improved reconstruction quality to the learned scene priors, which enable our model to reason beyond individual measurements. However, the scene priors inevitably fall short in the presence of a domain shift, which likely explains why NeTF++ outperforms our method on the **CMU** dataset (Table 4.3 and Figure 4.7). In the meantime, NeTF++ and our method compare favorably against NeTF, confirming the strength of our proposed transient rendering recipe.

Hardware Prototype

Our hardware prototype, depicted in Figure 4.8, consists of an ultrafast pulsed laser (OneFive Katana HP, 700mW average power, 532nm, 35ps pulse width, 5MHz repetition rate) and two 1D SPAD arrays each with 14 available pixels (75ps FWHM) (Renna et al., 2020). The laser raster-scans

a 190×24 grid of locations that cover a 1.9m $\times$ 1.9m square on the wall. The SPAD arrays approximately co-localize and focus on the same 1cm $\times$ 9cm patch on the wall near the center of the scanned area. Photon counts from overlapping pixels are summed together, yielding 14 histograms per scanned location. The histograms have 768 bins with a temporal bin size of 32ps. We apply the pixel remapping algorithm from Nam et al. (2021) alongside nearest-neighbor interpolation to convert the raw transients into a measurement of size $128 \times 128 \times 768$ (height $\times$ width $\times$ time) for subsequent reconstruction. Further details about our imaging hardware are discussed in Nam et al. (2021).

Sensor Modeling

For the real-world experiments, it is challenging to obtain the pulse shape, power and FoV of the *virtual* laser, as it not only depends on the characteristics of the actual light source, but also the distance to the relay wall, the incident angle of light, and the wall's geometry and reflectance. Likewise, calibrating the *virtual* detector is also extremely difficult, if not infeasible.

Fortunately, the high-quality laser of our system has a small FWHM ($\sim$ 25ps) with respect to the bin size of transients (32ps), and the wall is chosen to be flat and diffuse. Hence, we simply assume that the laser and detector both have a hemispherical FoV, and the laser impulse g is a time Dirac delta function. We infer ϕ^{scale} from data through trial and error.

Further, NLOS imaging operates in the *low-light regime*; a laboratory-grade detector with minimal internal noise captures indirectly scattered light with extremely weak incident flux, and data collection is performed in a dark environment for improved signal-to-noise ratio. It is well-known that pile-up and time jitter become insignificant in this regime. We thus assume $\phi^{bkgd} \approx 0$, and directly scale the rendered flux by the inferred ϕ^{scale} and the number of cycles C to obtain an estimate of the transient $\mathbf{h}$.

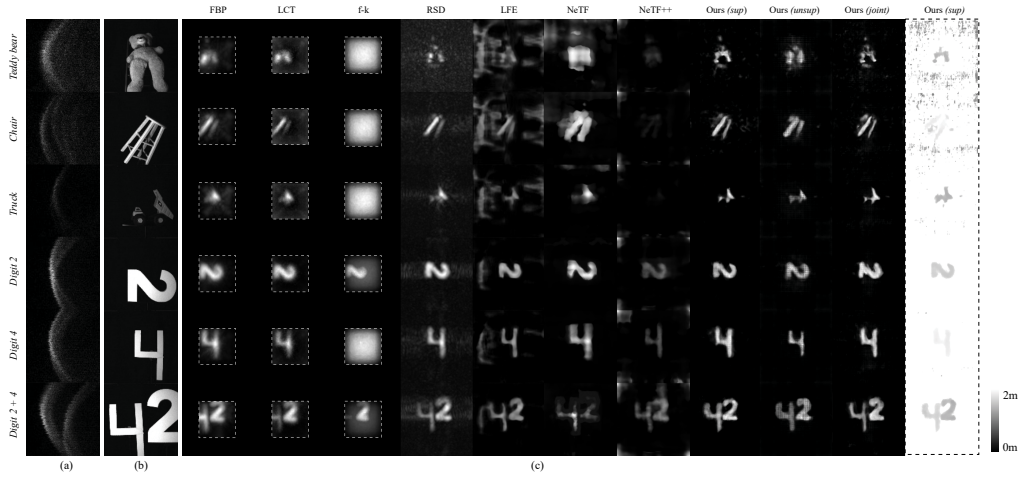


Figure 4.9: **Reconstruction results on real-world captures.** (a) An *x*-*t* slice of the measurement volume. Note the rough frontier of returning photons due to pixel remapping. (b) A reference image of the hidden scene (not used for inference). (c) Intensity and depth reconstruction. Our method is robust to approximations in the lighting model and produces strong reconstructions on real-world captures. The rightmost column presents depth estimation of our supervised model.

Real-World Experiments

Experiment Setup

We collect NLOS measurements for a few real-world scenes using our imaging hardware. These datasets include digits "2" and "4" for evaluating the *in-class* generalizability of models on real-world data, and "chair","truck" and "Teddy bear" that represent more challenging scenes not present in training. The objects are placed approximately 1m away from the relay wall in various poses. We capture a reference intensity image for each scene for the qualitative assessment of reconstruction quality.

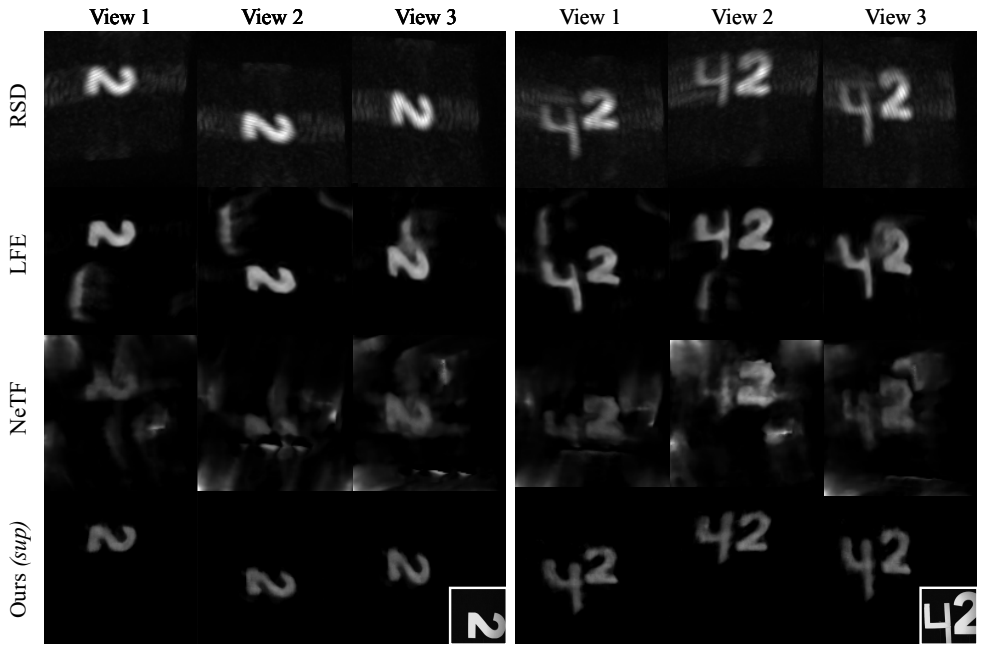


Figure 4.10: **Novel view synthesis results.** Our method learns accurate 3D scene geometry and can render intensity images beyond the frontal view. A reference view of the hidden scene is displayed in the inset.

Reconstruction Results

Figure 4.9 presents qualitative results of our method in comparison to the baselines. Our method reliably reconstructs simple objects and yields strong results on complex scenes. By contrast, LFE cannot accommodate the large domain gap between simulated data and real captures, thus producing distorted and noisy reconstructions that are substantially worse than RSD. NeTF and NeTF++ are both sensitive to the approximations in their lighting models. As a result, their reconstructions are prone to artifacts especially on complex scenes. Finally, FBP, LCT and f-k yield the worst results, confirming that these methods are not well-suited for non-confocal NLOS reconstruction.

	RSD	LFE	Ours (*sup*)	Ours (*unsup*)	Ours (*joint*)
Accuracy	82.2%	81.0%	**85.4%**	**85.8%**	**85.2%**

Table 4.4: **Object recognition results using learned features.** Our model learns strong feature representations with more discriminative power.

Novel View Synthesis

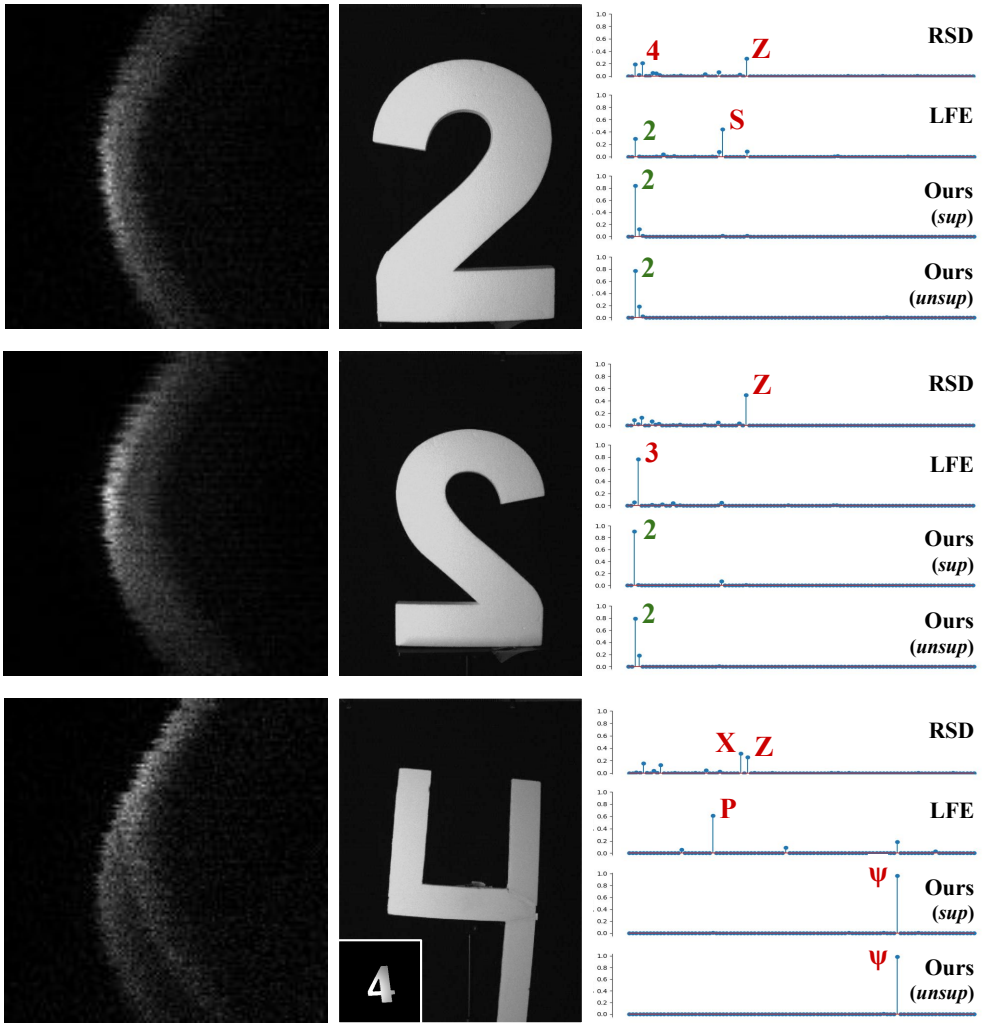
NLOS reconstruction algorithms are often evaluated on the *frontal* view of the scene observed from behind the wall. We hypothesize that accurate 3D reconstruction of a hidden scene would facilitate view synthesis beyond the frontal view. We thus task RSD, LFE, NeTF and our method for novel view synthesis to assess the quality of 3D reconstruction.

We render three random views given real-world datasets of two digits ("2" and "4"), and compare the results of different methods in Figure 4.10. Our method produces crispy, artifact-free renderings compared to the noisy and blurry outputs of the baselines.

Object Recognition

We further compare our method to RSD and LFE by adapting the learned features for object recognition. This surrogate task helps reveal these models' capacity to encode scene priors, which may be useful for downstream recognition tasks.

Recall that both our method and LFE encode and propagate a transient measurement to a spatial feature cube. For both models trained on the **alphanumerics** dataset, we project the feature cube to a 2D feature map by taking the per-channel maximum along the depth axis. Similarly, we project the volumetric reconstruction of RSD and interpret the 2D projection as a feature map. We then train a ResNet-18 (He et al., 2016) to take these 2D feature maps for 100-way alphanumeric classification. All classifiers are trained for 50 epochs with a mini-batch size of 32 using stochastic



(a) Measurements (x-t) (b) Hidden Scene (c) Class Prediction

Figure 4.11: **Object recognition results.** (a) An x-t slice of the input measurement volume. (b) A reference image of the hidden scene (not used for inference). (c) Predicted class probabilities. Taller lines indicate higher probability values. Green for correct predictions and red for incorrect predictions.

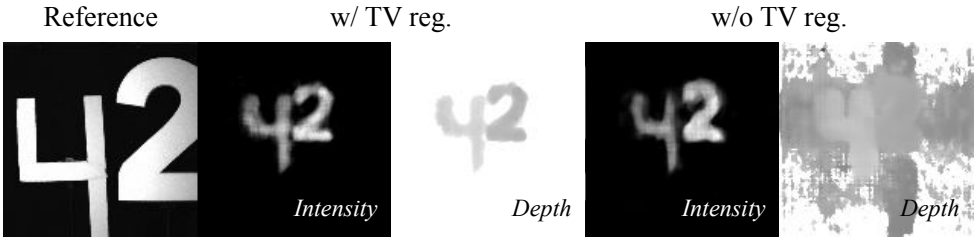


Figure 4.12: **Ablation on total variation prior.** Our total variation regularizer eliminates floaters in empty space.

gradient descent (SGD) with a learning rate of 0.1, a momentum of 0.9 and a weight decay of 0.0005. We report classification accuracy on the "Unseen Poses" test split (Table 4.4), along with the softmax confidence scores for the real-world measurements of digits "2" and "4" (Figure 4.11).

The classifiers taking features from our models achieve 85.4% (supervised) and 85.8% (unsupervised) test accuracy, outperforming those taking RSD and LFE features by more than 3% and 4%, respectively. Moreover, our classifier predicts the correct labels with high confidence for the real-world measurements of "2", whereas the baselines confuse "2" with the morphologically similar letters "S" and "Z" and digit "3". All classifiers fail on the measurement of "4" since it looks quite different from the simulated digit "4" in the training set, yet the classifiers taking our features yield the most plausible prediction. These results suggest that our encoder learns rich scene priors with more discriminative power.

Ablation Study

Total Variation Prior

Our training objectives include a total variation term to encourage sparsity in scene opacity (Equation 4.7). A model trained without the regularizer

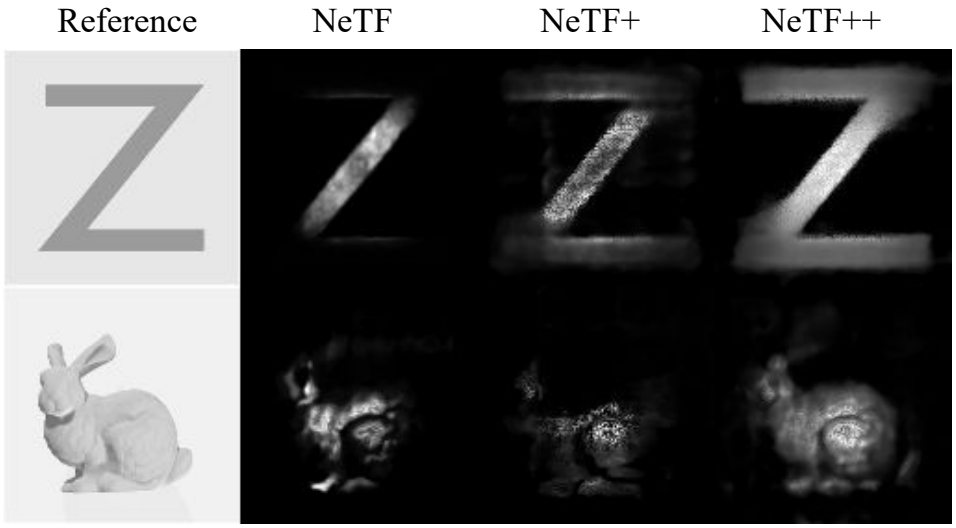


Figure 4.13: **Ablation on transient rendering recipe.** Our NeTF++ employs a principled sampling strategy for estimating transmittance, whereas NeTF and NeTF+ omit transmittance and yield worse reconstruction.

renders plausible intensity images yet introduces spurious density (*i.e.*, floaters) in empty space (Figure 4.12).

Transient Rendering Recipe

NeTF models reflectance at every scene point. Unfortunately, the transmittance term in their rendering equation is omitted for tractable numerical integration. Our renderer instead models outgoing radiance without factoring it into illumination and reflectance. Importantly, this enables transmittance estimation with a tractable sampling strategy (Figure 4.4). To understand the impact of this design choice, we drop the transmittance term in NeRF++ and dub this variant NeRF+. One may further recover NeTF from NeTF+ by learning reflectance as opposed to radiance.

We compare NeTF, NeTF+ and NeTF++ on the two simulated scenes from the **Z-NLOS** dataset (Galindo et al., 2019). The results in Figure 4.13

show that proper evaluation of transmittance is key to high-quality reconstruction, while the choice of modeling radiance as opposed to reflectance has minor impact on reconstruction quality.

4.5 Conclusion and Discussion

We presented a novel learning-based method for non-confocal NLOS reconstruction. Our key innovation is a method that embeds strong domain knowledge in the deep model in the form of an inverse propagation module and a volume renderer, both physics based, to navigate the learning of a conditional neural scene representation. Moreover, our model can be flexibly trained using diverse supervision signals including multi-view target images, and more importantly transient measurements themselves. We demonstrated superior reconstruction quality of our model in comparison to state-of-the-art physics and learning-based methods. In particular, our method, despite being trained on synthetic data, generalizes well on real measurements. We anticipate that our method alongside the fast imaging system will lay the foundation for exciting applications of NLOS imaging that require high-speed imaging. We hope our method will provide a solid step towards the challenging problem of NLOS reconstruction, and shed light on a broader spectrum of inverse problems in imaging sciences.

Incorporating Shape Prior

Our model learns rich scene priors from data to address the fundamental missing cone problem of NLOS imaging. Nevertheless, the generalization of our model depends on the training data distribution, and the recovery of complex shapes (*e.g.*, human) remains challenging due to a lack of constraint on the learned shapes. One potential solution is to combine our analysis-by-synthesis framework with parametric shape models (*e.g.*, SMPL (Loper et al., 2015)). These models have orders-of-magnitude

lower degrees of freedom compared to non-parametric NeRF-like scene representations, and are thus well-suited for reconstructing certain scene categories. Similar ideas have been explored for human reconstruction from posed RGB images (*e.g.*, HumanNeRF (Weng et al., 2022)). These methods may inform the development of NLOS reconstruction algorithms with strong shape priors.

Imaging Motion

Moving beyond static scenes, imaging moving objects and human actions is more indicative of the use case of a high-speed imaging system. Our current approach operates on one frame at a time, and thus does not leverage the strong prior of motion continuity. Intuitively, sharing information among adjacent frames may compensate for the short exposure time of individual frames and provide additional cues to the reconstruction algorithm. We envision that both optimization-based and learning-based methods could leverage this smoothness constraint for improved reconstruction under low per-frame signal-to-noise ratio. A promising direction is to adapt NeRF-like approach for video modeling (*e.g.*, HyperNeRF (Park et al., 2021)) and neural architectures for temporal reasoning (*e.g.*, Transformers (Vaswani et al., 2017)) for NLOS reconstruction of moving scenes.

NLOS Imaging in the Wild

While our method represents a solid step towards practical NLOS imaging, all of our experiments are run in a highly controlled laboratory environment with an expensive room-sized hardware prototype. Future endeavor may focus on building miniature imaging systems with low-cost sensors (*e.g.*, those from Chapter 3) that supports experimentation under more realistic imaging conditions (*e.g.*, non-planar relay surface, sparse scanning pattern, and strong ambient light).

5 CONCLUSION AND FUTURE WORK

In this dissertation, I presented an unconventional imaging paradigm for single-photon 3D vision (Chapter 2). Central to this paradigm is to capture a spatially distributed set of transient histograms using SPAD sensors with diffuse lighting and a wide-FoV detector. The reconstruction of 3D scenes can be then achieved by analyzing these recorded transient histograms. I described the transient formation model, and developed a general algorithmic framework for the reconstruction of complex 3D scenes. The reconstruction approach follows the analysis-by-synthesis principle and combines expressive neural scene representations with differentiable transient volume rendering.

The effectiveness of this reconstruction approach is demonstrated through extensive qualitative and quantitative results for both simulations and real-world imaging systems. Specifically, my dissertation studied two key applications, one for direct line-of-sight reconstruction with low-cost proximity sensors (Chapter 3), the other for high-speed non-line-of-sight reconstruction (Chapter 4).

Overall, the algorithmic development and empirical results of my work validates the hypothesis raised in my dissertation statement. Specifically, my dissertation work had successfully demonstrated that transient histograms from a distributed set of single-photon cameras under diffuse lighting can be used to accurately reconstruct complex 3D scene geometry for line-of-sight and non-line-of-sight imaging. I hope my work can shed light on the design and implementation of practical 3D vision systems with single-photon cameras.

Future Directions

I envision that my work can be extended in the following directions:

Illumination and Lens Engineering

Our imaging paradigm assumes a diffuse light source and a wide-FoV SPAD detector. An active line of research in computational imaging studies the design of illumination patterns (*e.g.*, in structured-light imaging) and detector point spread functions (PSF) (*e.g.*, in lensless imaging), with strong evidence showing that better 3D reconstruction can be achieved with principled or learned lighting and PSF patterns. Our sensor can be thought of as combining the simplest illumination (*i.e.*, diffuse) and PSF (*i.e.*, lensless) patterns. Future work may investigate, *e.g.*, the joint optimization of scene geometry, lighting, and detector PSF.

Reflectance Modeling and Recovery

Our methods emphasize the reconstruction of scene geometry, and currently make simplified assumptions about scene reflectance (*e.g.*, Lambertian). In principle, our volume rendering framework can support more complex reflectance models, and thus may allow BRDF estimation using transients. In the meantime, faithful modeling of scene reflectance may facilitate the accurate reconstruction of scene geometry. Hence, one important direction for future research is to understand the feasibility of reflectance modeling and recovery using our imaging approach.

Domain-Specific Shape Reconstruction

Single-photon cameras have found major applications in smartphone photography, wearable sensing and virtual reality, where a central goal is to model and reconstruct human face, hands and full body. Our method in its current form relies on non-parametric neural scene representations, thus may not fit delicate and deformable human body parts at high precision. It is thus desirable to bring parametic mesh models (*e.g.*, SMPL (Loper et al., 2015)) into the optimization to constrain the recovered shapes, yet

this would require changes to the reconstruction procedure to allow differentiation through mesh vertices. Future work may thus explore novel differentiable forward models for human reconstruction from transients.

REFERENCES

Ackermann, Jens, Michael Goesele, et al. 2015. A survey of photometric stereo techniques. *Foundations and Trends® in Computer Graphics and Vision*.

AG, AMS OSRAM. *Tmf882x datasheet*. AMS OSRAM AG.

Arellano, Victor, Diego Gutierrez, and Adrian Jarabo. 2017. Fast back-projection for non-line of sight reconstruction. *Optics Express*.

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

Atzmon, Matan, and Yaron Lipman. 2020. Sal: Sign agnostic learning of shapes from raw data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Becker, Cienna N, and Lucas J Koerner. 2023. Plastic classification using optical parameter features measured with the tmf8801 direct time-of-flight depth sensor. *Sensors*.

Callenberg, Clara, Zheng Shi, Felix Heide, and Matthias B Hullin. 2021. Low-cost spad sensing for non-line-of-sight tracking, material classification and depth imaging. *ACM Transactions on Graphics*.

Chan, Eric R, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chang, Angel X, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song,

Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, Wenzheng, 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*.

Coates, PB. 1968. The correction for photonpile-up'in the measurement of radiative lifetimes. *Journal of Physics E: Scientific Instruments*.

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

Faccio, Daniele, Andreas Velten, and Gordon Wetzstein. 2020. Non-line-of-sight imaging. *Nature Reviews Physics*.

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

Galindo, Miguel, Julio Marco, Matthew O'Toole, Gordon Wetzstein, Diego Gutierrez, and Adrian Jarabo. 2019. A dataset for benchmarking time-resolved non-line-of-sight imaging. *ACM SIGGRAPH 2019 Posters*.

Geng, Jason. 2011. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*.

Grau Chopite, Javier, Matthias B Hullin, Michael Wand, and Julian Iseringhausen. 2020. Deep non-line-of-sight reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gropp, Amos, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit geometric regularization for learning shapes. *International Conference on Machine Learning*.

Grossmann, Paul. 1987. Depth from focus. *Pattern Recognition Letters*.

Gupta, Anant, Atul Ingle, and Mohit Gupta. 2019a. Asynchronous single-photon 3d imaging. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Gupta, Anant, Atul Ingle, Andreas Velten, and Mohit Gupta. 2019b. Photon-flooded single-photon 3d cameras. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gutierrez-Barragan, Felipe, Atul Ingle, Trevor Seets, Mohit Gupta, and Andreas Velten. 2022. Compressive single-photon 3d cameras. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gutierrez-Barragan, Felipe, Fangzhou Mu, Andrei Ardelean, Atul Ingle, Claudio Bruschini, Edoardo Charbon, Yin Li, Mohit Gupta, and Andreas Velten. 2023. Learned compressive representations for single-photon 3d imaging. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

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

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Heide, Felix, Steven Diamond, David B Lindell, and Gordon Wetzstein. 2018. Sub-picosecond photon-efficient 3d imaging using single-photon sensors. *Scientific reports*.

Hernandez, Quercus, Diego Gutierrez, and Adrian Jarabo. 2017. A computational model of a single-photon avalanche diode sensor for transient imaging. *arXiv preprint arXiv:1703.02635*.

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

Iseringhausen, Julian, and Matthias B Hullin. 2020. Non-line-of-sight reconstruction using efficient transient rendering. *ACM Transactions on Graphics*.

Isogawa, Mariko, Ye Yuan, Matthew O'Toole, and Kris M Kitani. 2020. Optical non-line-of-sight physics-based 3d human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jarabo, Adrian, Julio Marco, Adolfo Munoz, Raul Buisan, Wojciech Jarosz, and Diego Gutierrez. 2014. A framework for transient rendering. *ACM Transactions on Graphics*.

Jiang, Deyang, Xiaochun Liu, Jianwen Luo, Zhengpeng Liao, Andreas Velten, and Xin Lou. 2021. Ring and radius sampling based phasor field diffraction algorithm for non-line-of-sight reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jungerman, Sacha, Atul Ingle, Yin Li, and Mohit Gupta. 2022. 3d scene inference from transient histograms. *European Conference on Computer Vision*.

Kingma, Diederik P, and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kirmani, Ahmed, Tyler Hutchison, James Davis, and Ramesh Raskar. 2009. Looking around the corner using transient imaging. *Proceedings of the IEEE International Conference on Computer Vision*.

Kutulakos, Kiriakos N, and Steven M Seitz. 2000. A theory of shape by space carving. *International Journal of Computer Vision*.

La Manna, Marco, Fiona Kine, Eric Breitbach, Jonathan Jackson, Talha Sultan, and Andreas Velten. 2018. Error backprojection algorithms for non-line-of-sight imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lagarto, João L, Federica Villa, Simone Tisa, Franco Zappa, Vladislav Shcheslavskiy, Francesco S Pavone, and Riccardo Cicchi. 2020. Real-time multispectral fluorescence lifetime imaging using single photon avalanche diode arrays. *Scientific Reports*.

Liao, Zhengpeng, Deyang Jiang, Xiaochun Liu, Andreas Velten, Yajun Ha, and Xin Lou. 2021. FPGA accelerator for real-time non-line-of-sight imaging. *IEEE Transactions on Circuits and Systems I: Regular Papers*.

Lindell, David B, Matthew O'Toole, and Gordon Wetzstein. 2018. Single-photon 3d imaging with deep sensor fusion. *ACM Transactions on Graphics*.

Lindell, David B, Gordon Wetzstein, and Matthew O'Toole. 2019. Wave-based non-line-of-sight imaging using fast fk migration. *ACM Transactions on Graphics*.

Liu, Steven, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. 2021. Editing conditional radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, Xiaochun, Sebastian Bauer, and Andreas Velten. 2019a. Analysis of feature visibility in non-line-of-sight measurements. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

———. 2020. Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems. *Nature Communications*.

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

Liu, Xinyang, Yijin Li, Yanbin Teng, Hujun Bao, Guofeng Zhang, Yinda Zhang, and Zhaopeng Cui. 2023. Multi-modal neural radiance field for monocular dense slam with a light-weight tof sensor. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lombardi, Stephen, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics*.

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*.

Lorensen, William E, and Harvey E Cline. 1987. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*.

Macario Barros, Andréa, Maugan Michel, Yoann Moline, Gwenolé Corre, and Frédérick Carrel. 2022. A comprehensive survey of visual slam algorithms. *Robotics*.

Malik, Anagh, Parsa Mirdehghan, Sotiris Nousias, Kyros Kutulakos, and David B Lindell. 2023. Transient neural radiance fields for lidar view synthesis and 3d reconstruction. *Advances in Neural Information Processing Systems*.

Max, Nelson. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*.

Mildenhall, Ben, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*.

Mora-Martín, Germán, Stirling Scholes, Alice Ruget, Robert Henderson, Jonathan Leach, and Istvan Gyongy. 2023. Video super-resolution for single-photon lidar. *Optics Express*.

Mu, Fangzhou, 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*.

Mur-Artal, Raul, Jose Maria Martinez Montiel, and Juan D Tardos. 2015. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*.

Musarra, Gabriella, Ashley Lyons, Enrico Conca, Federica Villa, Franco Zappa, Yoann Altmann, and Daniele Faccio. 2019. 3D RGB non-line-of-sight single-pixel imaging. *Imaging Systems and Applications*.

Nam, Ji Hyun, Eric Brandt, Sebastian Bauer, Xiaochun Liu, Marco Renna, Alberto Tosi, Eftychios Sifakis, and Andreas Velten. 2021. Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. *Nature Communications*.

Oren, Michael, and Shree K Nayar. 1994. Generalization of lambert's reflectance model. *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*.

Ortiz, Joseph, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. 2022. isdf: Real-time neural signed distance fields for robot perception. *Robotics: Science and Systems*.

O'Toole, Matthew, Felix Heide, David B Lindell, Kai Zang, Steven Diamond, and Gordon Wetzstein. 2017. Reconstructing transient images from single-photon sensors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

Park, Keunhong, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. 2021. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Transactions on Graphics*.

Pediredla, Adithya K, Aswin C Sankaranarayanan, Mauro Buttafava, Alberto Tosi, and Ashok Veeraraghavan. 2018. Signal processing based pile-up compensation for gated single-photon avalanche diodes. *arXiv preprint arXiv:1806.07437*.

Peng, Jiayong, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. 2020. Photon-efficient 3d imaging with a non-local neural network. *European Conference on Computer Vision*.

Phong, Bui Tuong. 1975. Illumination for computer generated pictures. *Communications of the ACM*.

Pollák, Martin, Marek Kočiško, Dušan Paulišin, and Petr Baron. 2020. Measurement of unidirectional pose accuracy and repeatability of the collaborative robot ur5. *Advances in Mechanical Engineering*.

Renna, Marco, Ji Hyun Nam, Mauro Buttafava, Federica Villa, Andreas Velten, and Alberto Tosi. 2020. Fast-gated 16 $\times$ 1 SPAD array for non-line-of-sight imaging applications. *Instruments*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*.

Ruget, Alice, 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*.

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

Seitz, Steven M, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shen, Siyuan, Zi Wang, Ping Liu, Zhengqing Pan, Ruiqian Li, Tian Gao, Shiying Li, and Jingyi Yu. 2021. Non-line-of-sight imaging via neural transient fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sifferman, Carter, Yeping Wang, Mohit Gupta, and Michael Gleicher. 2023. Unlocking the performance of proximity sensors by utilizing transient histograms. *IEEE Robotics and Automation Letters*.

Smith, Brandon M, Matthew O'Toole, and Mohit Gupta. 2018. Tracking multiple objects outside the line of sight using speckle imaging. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sommerfeld, Arnold. 1964. *Optics: Lectures on theoretical physics*, vol. 4.

Subbarao, Murali, and Gopal Surya. 1994. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*.

Sucar, Edgar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. 2021. imap: Implicit mapping and positioning in real-time. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Tsai, Chia-Yin, Kiriakos N Kutulakos, Srinivasa G Narasimhan, and Aswin C Sankaranarayanan. 2017. The geometry of first-returning photons for non-line-of-sight imaging. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tsai, Chia-Yin, Aswin C Sankaranarayanan, and Ioannis Gkioulekas. 2019. Beyond volumetric albedo–a surface optimization framework for non-line-of-sight imaging. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ureña, Carlos, Marcos Fajardo, and Alan King. 2013. An area-preserving parametrization for spherical rectangles. *Computer Graphics Forum*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Velten, Andreas, 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*.

Wang, Peng, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*.

Wang, Yiming, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Weng, Chung-Yi, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xin, Shumian, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis Gkioulekas. 2019. A theory of fermat paths for non-line-of-sight shape reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yi, Jingang, Junjie Zhang, Dezhen Song, and Suhada Jayasuriya. 2007. Imu-based localization and slip estimation for skid-steered mobile robots. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Yu, Alex, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zappa, Franco, Simone Tisa, Alberto Tosi, and Sergio Cova. 2007. Principles and features of single-photon avalanche diode arrays. *Sensors and Actuators A: Physical*.

Zhang, Ruo, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. 1999. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, Dayu, and Wenshan Cai. 2022. Fast non-line-of-sight imaging with two-step deep remapping. *ACS Photonics*.