FoveaSPAD: Exploiting Depth Priors for Adaptive and Efficient Single-Photon 3D Imaging

Justin Folden, Atul Ingle, and Sanjeev J. Koppal

***Abstract*—Fast, efficient, and accurate depth-sensing is important for safety-critical applications such as autonomous vehicles. Direct time-of-flight LiDAR has the potential to fulfill these demands, thanks to its ability to provide high-precision depth measurements at long standoff distances. While conventional LiDAR relies on avalanche photodiodes (APDs), single-photon avalanche diodes (SPADs) are an emerging image-sensing technology that offer many advantages such as extreme sensitivity and time resolution. In this paper, we remove the key challenges to widespread adoption of SPAD-based LiDARs: their susceptibility to ambient light and the large amount of raw photon data that must be processed to obtain in-pixel depth estimates. We propose new algorithms and sensing policies that improve signal-to-noise ratio (SNR) and increase computing and memory efficiency for SPAD-based LiDARs. During capture, we use external signals to *foveate*, i.e., guide how the SPAD system estimates scene depths. This foveated approach allows our method to "zoom into" the signal of interest, reducing the amount of raw photon data that needs to be stored and transferred from the SPAD sensor, while also improving resilience to ambient light. We show results both in simulation and also with real hardware emulation, with specific implementations achieving a 1548-fold reduction in memory usage, and our algorithms can be applied to newly available and future SPAD arrays.**

***Index Terms*—Foveation, single-photon avalanche diode (SPAD), time-of-flight, computational imaging.**

I. Introduction

BIOLOGICAL vision systems have the remarkable ability to *foveate* — i.e. redistribute cognitive resources towards "salient" features or objects in a scene, depending on context. Unfortunately, most conventional cameras and computer vision systems today capture scene information in a non-adaptive fashion, spending power and bandwidth on sensing scene components that may not help the overall imaging task. In fact, the current framework for deep learning-based systems assumes uniform sampling of the scene and overcomes these limitations through data-driven pipelines that focus on interesting regions of the scene [1], [2] in the input RGB images.

While this inefficient but popular framework for conventional RGB sensors may be difficult to change, our proposed method, called FoveaSPAD, can impact the next wave of single-photon avalanche diode (SPAD) sensor technology. SPADs can capture scene information at the granularity of individual photons, at timescales as small as 10's of picoseconds. Recent advances in CMOS-compatible SPAD pixel designs has enabled real-time in-pixel processing of these photon timestamp streams. Thus, SPADs are a natural candidate for designing efficient depth cameras — individual pixels can be reprogrammed on-the-fly to adaptively accept or reject a spatio-temporal subset of the photon stream.

Our FoveaSPAD algorithms enable capturing scene information at higher granularity in regions that are most relevant to a downstream vision task. *In this sense, we generalize the term "foveation" in the context of adaptive SPAD spatio-temporal sampling to allow both depth and memory efficiencies.* For robots, remote sensor nodes, and other resource-constrained systems, foveation for SPAD sensors can allow accurate depth sensing under constraints on power and bandwidth (see Fig. 1).

The raw data captured by an array of SPAD pixels can be thought of as a spatio-temporal photon stream. Each photon detection is represented as an (x, y, t) coordinate, where the $x - y$ coordinates denote the pixel location and the t coordinate denotes the photon detection timestamp. Each SPAD pixel captures the round trip time of a laser pulse to and from a given scene point, constructing a photon timing *histogram* which records the number of photons captured at various time delays with respect to the time the laser pulse was transmitted. Each pixel must construct one such histogram, typically with 1000's of bins, which causes a severe data bottleneck for today's SPAD cameras. To illustrate the severity of the bottleneck, consider a 1-megapixel SPAD array with a 1000-bin histogram per pixel, storing 1 B per bin. At 30 frames per second, this setup generates a staggering 30GB of data every second

Our algorithms foveate across the spatio-temporal histogram space to efficiently recover the peak, providing the time-delay t for depth computation. We adaptively capture subranges to locate laser photons, rejecting ambient photons. Note that our proposed algorithms are not exhaustive; rather, we aim to define a class of algorithms that rely on a depth prior. In this work, we propose three methods for acquiring priors, though many other

Received 13 June 2024; revised 30 September 2024; accepted 5 November 2024. Date of current version 6 December 2024. The work of Justin Folden and Sanjeev J. Koppal was supported in part by National Science Foundation under Grant 1942444 and Grant 2330416, and in part by the Office of Naval Research under Grant N000142312429 and Grant N000142312363. The work of Atul Ingle was supported by National Science Foundation Award under Grant 2138471. The associate editor coordinating the review of this article and approving it for publication was Prof. David Lindell. *(Corresponding author: Justin Folden.)*

Justin Folden and Sanjeev J. Koppal are with the Department of Electrical and Computer Engineering, University of Florida, Gainsville, FL 32611 USA (e-mail: jfolden@ufl.edu).

Atul Ingle is with the Department of Computer Science, Portland State University, Portland, OR 97201 USA.

This article has supplementary downloadable material available at https://doi.org/10.1109/TCI.2024.3503360, provided by the authors.

Digital Object Identifier 10.1109/TCI.2024.3503360

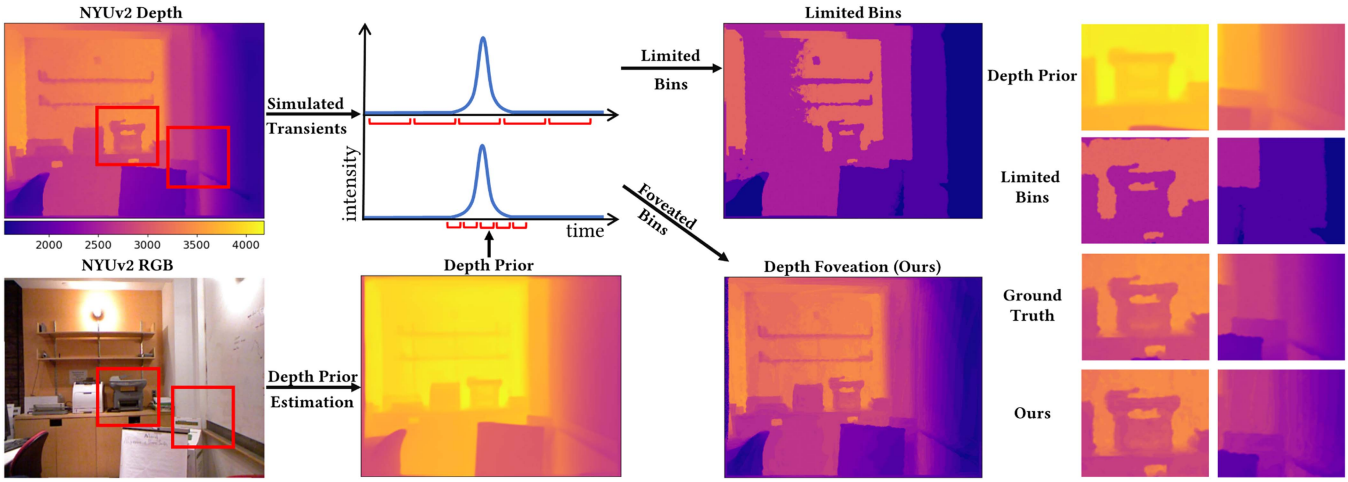

Fig. 1. Depth Prior Driven SPAD Depth Foveation: SPAD sensors suffer from a data bottleneck, since thousands of histogram bins are used to generate depth as shown in the top left. If fewer bins are used, this reduces depth resolution, as shown in the limited bins depth result. Our idea is to use additional information, such as a color image (Section IV, VII) or optical flow (Section VI), to foveate the SPAD bins. Therefore, for the same memory cost we can place the bins near where the histogram peak should be, results in accurate depth, as shown in the depth foveation result. The insets show that our method achieves the accuracy and resolution of ground truth, with fewer bins. They also show that the depth prior, in this case monocular estimation, by itself cannot provide the correct depth, and foveation is required.

methods exist, such as depth from stereo, depth from defocus [3], or non-vision-based methods such as sonar. Each method has trade-offs, and it is up to the user to determine which method best suits their use case. Our contributions in this work are as follows:

- We present a theoretical model for expected gains (in terms of increased signal-to-noise-ratio and depth resolution) from foveation with SPADs.
- We explore the question of how to foveate in space and time at a single time instant by leveraging monocular depth estimates, which can come either from the SPAD-generated image or a cheap, external color camera. We propose different flavors of practical FoveaSPAD designs that optimize for memory/bandwidth and depth resolution.
- For images of moving scenes, we demonstrate how to use optical flow cues to direct SPAD foveation.
- We show results both in simulation and using recently available real SPAD datasets.

A. Hardware Emulation

Time-correlated single photon counting is the technique that enables SPAD cameras to build histograms and control binning on-sensor. Our work is limited to simulation experiments and hardware emulation of existing SPAD LiDAR data. Hardware emulation refers to leveraging real-world data captured using a single SPAD or line arrays [4], [5], which we then use to emulate the performance of larger SPAD arrays. While SPAD sensor arrays with native support for foveation are not yet available commercially, we believe our proposed techniques could be implemented at the pixel or camera level. This is supported by recent proofs-of-concept in kilopixel-resolution reconfigurable SPAD arrays with in-pixel timestamping, gating, and histogramming capabilities [6], [7]. We anticipate that this work will inspire future hardware designs, leading to more efficient and versatile SPAD sensor arrays.

B. Scope: Simulation and Emulations

In this work, we anticipate future hardware advancements that will enhance SPAD-based depth sensing. Our simulations and emulations are intended to project the performance of emerging SPAD sensor technologies, focusing on adaptive and efficient bin sampling to mitigate memory bottlenecks with minimal loss of accuracy, which is particularly advantageous for flash-based SPAD LiDARs systems. Potential future implementations could feature a shared "macropixel" architecture and a dynamic gating system, allowing pixel groups to adjust to appropriate gating signals in real time. We explore these ideas further in Section VIII and present a speculative "macropixel" array design in Fig. 10, which includes a variable-resolution TDC'a key component for one of the proposed methods. These simulations play a critical role in validating our algorithms and highlighting their potential impact on future sensor designs, even in the absence of current hardware.

II. Related Work

Our research takes inspiration from biology, since many animals have a region of high spatial acuity, i.e. the fovea, which they scan over the scene. In this sense, we are allied with foveated imaging research in computer vision and computational photography, and we now outline these related efforts:

Efficiency in Single-photon 3D Cameras: The data bottleneck issue in SPADs due to high-resolution sampling in histograms is well-known. Research that attempts to mitigate this issue include novel statistical representations [8] as well as compressive histograms [9], [10], [11] that use a small number of bins at maximum resolution to recover the entire scene. In contrast,

our approach works and scales easily with a large number of SPAD pixels. Other efforts include partial histogram methods such as using sliding windows for sub-range gating has been investigated [12], [13], [14], [15] which have linear efficiency and two-stage coarse-to-fine resolution scaling [16] which provide logarithmic efficiency. Our method uses context from cues such as optical flow to provide $\approx O(1)$ near-constant time efficiency. Finally, other work has used external sensors for guided upsampling or upscaling, [4], [17], but these are post-capture processes. In contrast, we perform foveation during capture and this gives us SNR and compute efficiencies that we have theoretically analyzed. A complementary approach to foveation is to use adaptive "equi-depth" histogramming approach for the signal peak [18]. Our approach is also complementary to adaptive gating approaches for SPAD LiDARs [19], with adaptive gating and exposure techniques working with or without a prior.

Foveated Depth Sensors: Our work is related to post-capture methods for upsampling and superresolution shown on data from many modes, such as depth images, color photographs etc. [20], [21], [22], [23], [24] and many of these have blended deep learning algorithms into the process of deciding where to sample [24], [25], [26], [27], [28], [29]. In fact, some of these algorithms are mature enough that commercial depth and LIDAR sensors allow post-capture foveation of the 3D point cloud through, for example, LIDAR-RGB fusion. In contrast, FoveaSPAD adapts during capture, and the efficiencies can impact small autonomous systems with power constraints. Directionally controlled LIDAR systems foveate spatially [30], [31], [32], [33]. These results complement our work on temporal foveation of SPAD sensors, including spatio-temporal foveation results (Section V).

Foveation in Display Graphics: Foveation is an important research topic in computer graphics, where data displayed to a viewer on AR/VR glasses, for example, is rendered in a way that reduces bandwidth [34]. Most of the work in this area does not focus on data capture but only on data visualization post-capture [35], [36]. Foveated light-field optics have been proposed [37] and these can be integrated with algorithms that foveate which portions of the scene to render at high resolution to reduce rendering resource consumption. Algorithms include perceptually guided foveation [38], [39] and hardware-optimized rendering [40]. Unlike our depth sensor, these use passive displays and cameras to optimize bandwidth, storage, and compute.

SPAD Histogram Techniques: Various techniques have been recently proposed to reduce the memory and bandwidth required to capture high-resolution photon timing histograms. Compressive histogramming techniques rely on a lower-dimensional linear projection of the high resolution histogram [5], [11] and estimating scene distances directly from the compressed representation. Algorithms that rely on "sketching" [41] attempt to directly estimate a parametric form of the true underlying waveform. These compressive acquisition approaches can be combined with foveation techniques developed here to further reduce the bandwidth required to store histograms. Differential capture methods [42], [43] can provide large reduction in bandwidth, but unlike foveation-based techniques, differential capture methods require additional post-processing to recover absolute scene depths. Recently, photon processing techniques that bypass the need for constructing a histogram have also been proposed, but they only work in the case of a single strong peak [44]. Sun et al.'s optical coding and super-resolution techniques leverage a phase plate and deep learning to achieve super-resolved images with minimal photon counts, further optimizing SPAD-based imaging [45]. Such optical techniques can work synergistically with our foveated capture approach, collectively reducing data transfer and computational demands.

III. Imaging Model and the Foveation Advantage

In this section, we present the imaging model and the concept of foveation, specifically focusing on how foveation can enhance the efficiency and effectiveness of (SPAD) LiDAR systems. We will delve into the specifics of how the imaging model is constructed, including assumptions about the behavior of laser pulses and photon detection, and how these factors influence the design and performance of SPAD sensors. Furthermore, the impact of ambient light on signal-to-noise and signal-to-background ratios will be examined, demonstrating how foveation can mitigate these effects. The theoretical foundations laid out in this section will serve as the basis for the foveation techniques proposed in the subsequent sections, where we will develop and analyze algorithms to optimize the selection of foveated bins in SPAD imaging.

A. Foveation and Scene Priors

We propose two methods of foveation, specifically memory foveation and depth foveation, are designed to optimize the efficiency of SPAD LiDAR systems by leveraging a priori knowledge about the scene's depth. Both methods require adaptive per-pixel gating, for which the hardware has yet to be developed.

Memory foveation focuses on reducing the amount of data that needs to be stored and processed by concentrating on a subset of histogram bins where the depth information is most likely to reside. Depth foveation, on the other hand, aims to improve depth resolution by reallocating histogram bins into a smaller, more focused region around the expected depth. The strategies proposed are fundamentally dependent on the accuracy and reliability of the scene depth prior, which guide the allocation of sensor resources.

Depth priors may be derived from any variety of means, including coarse initial scans, external sensors, or deep learning models. In this paper, we explore a few options, namely monocular estimation in Section IV, optical flow warping in Section VI, and coarse initial scans in Section VII. The quality of the prior directly impacts the success of foveation, with inaccurate priors potentially misallocating memory resources into incorrect regions. This dependence implies a trade-space between depth prior accuracy, and the amount of resources foveation stands to reduce. Exploring this trade-space is out of the scope of this paper, rather, we focus on using priors that are prone to error or are otherwise lower quality.

In the following subsections, we will define the image formation model, detailing the assumptions and mechanics of photon

detection. We will then explore the effects of ambient light on SPAD histogram formation and discuss how the proposed foveation techniques provide an advantage.

B. Image Formation Model

We assume that each pixel in the SPAD sensor array is co-located with a pulsed laser illumination source with a Gaussian pulse shape. Assuming no multi-path or sub-surface scattering effects, the photon flux incident on each pixel consists of a superposition of laser photons (that arrive in a short time window corresponding to the round-trip time-of-flight to and from the scene point) and background photons due to ambient light (that arrive uniformly randomly distributed throughout the capture duration). The laser repetition period (T) determines the maximum depth range of the SPAD LiDAR. We assume that this period is discretized into N bins (N is often on the order of 1000's of bins in conventional SPAD cameras). The number of photons captured by the SPAD pixel in the n^{th} bin ($1 \leq n \leq N$) is Poisson distributed with a mean of $\Phi_{\text{sig}}\mathbf{1}(n = i) + \Phi_{\text{bkg}}$ where i is the bin location corresponding to the true scene depth. Various sources of noise such as dark counts and afterpulsing are assumed to be absorbed in the Φ_{bkg} term. A complete histogram captured by this SPAD pixel over C laser cycles is given by a Poisson random vector with mean $C\Phi_{\text{sig}}\mathbf{1}(n = i) + C\Phi_{\text{bkg}}$ for $1 \leq i \leq N$.

The simplified imaging model assumes all laser photons arrive in a single bin i. In practice, the laser pulse spans several bins "smearing" the signal photons over more than one bin. The laser peak is often modeled as a Gaussian shaped pulse; we use a 1 ns full width at half maximum (FWHM) in our simulation results. Since the peak can span more than one histogram bin location, the defined Gaussian pulse may be used to estimate depth through match filtering. It is also possible to obtain a pseudo-intensity image by aggregating photon counts across histograms for each pixel which can be used in lieu of a co-located RGB or monochrome camera image for monocular depth cues.

C. Effects of Ambient Light

The integration time taken for all experiments is consistent. In this scenario, we show how foveation saves memory or improves depth resolution, and how the signal-to-noise ratio changes depending on ambient light, bin width, and the number of laser cycles or exposure time. Consider a SPAD pixel imaging a scene point illuminated by a pulsed laser. Initially, let us assume there are no multi-bounce effects and no ambient light, although we address these issues later on.

Photon detections from the SPAD pixel generate a histogram of arrival times. A conventional approach would use all N bins across the full histogram, whereas we propose methods to foveate attention onto a subset $M \leq N$ of these bins, where M is a window or gate with a user defined width (number of bins). Therefore, it is not surprising that, in the SNR analysis of our system, the ratio $\frac{M}{N}$ appears since this represents the advantage due to foveation.

In the analysis below, we will not make any assumption as to how the foveated bins M were obtained and instead just characterize the advantage of these, given that the desired histogram peak is captured by these bins. The analysis is not specific to any one method of acquiring a depth prior. In Sections IV, V, VI, and VII we propose algorithms to drive the selection of the foveated bins M and in Section VIII we provide a worst case analysis for whether the foveated M bins capture the histogram peak or not.

1) Low Ambient Light (No Pileup): Now consider the conventional imaging case, where the SPAD sensor detects time-of-arrival of photons and accumulates into a photon timing histogram to find the time that corresponds to the true depth of the scene point.

We assume that the histogram has a full scale range of T seconds which is related to the maximum unambiguous depth range Z as $T = \frac{2Z}{c}$ where c is the speed of light. Consider N histogram bins that are uniformly distributed across the full scale range T. The width of each bin is $\frac{T}{N}$. Since narrower bins produce fewer photons, the SNR for each bin is proportional to the width of that time bin:

$$\text{SNR} \propto C\sqrt{\frac{T}{N}}, \tag{1}$$

where C denotes the number of laser cycles (i.e., the total exposure time) that was used to capture the histogram.

We now consider two types of foveation. In *memory foveation*, only a limited number of bytes in memory can be dedicated to the task of finding the histogram peak, and therefore placing these at the peak is most efficient. In *depth foveation*, memory allocation remains fixed but is concentrated in the foveated region, bringing the bins closer together near the histogram peak, thereby improving depth resolution.

Memory foveation: In memory foveation, we identify M bins $M \ll N$ where the true depth exists. The width of the bins remains the same $\frac{T}{N}$, and therefore the SNR is also identical to the conventional case:

$$\text{SNR} \propto \sqrt{\frac{M\frac{T}{N}}{M}} \propto \sqrt{\frac{T}{N}} \tag{2}$$

Depth foveation: In depth foveation, we concentrate the N bins that would have been distributed over the entire depth range, into a small region. The region is the same region used in memory foveation, and is given by multiplying the number of memory foveation bins M with the original bin width to give $M\frac{T}{N}$. This region is divided into N bins, and therefore the new bin width is $\frac{MT}{N^2}$. As before, the SNR is proportional to the bin width, and therefore much lower,

$$\text{SNR} \propto C\sqrt{\frac{MT}{N^2}} = \sqrt{\frac{M}{N}\frac{T}{N}} \tag{3}$$

Therefore, we have improved depth resolution but at the cost of SNR. To increase the SNR of the foveated depth we can increase C, the number of cycles the laser pulses through to create the histogram. The new cycle number must be equal to or greater

Algorithm 1: Memory and Depth Foveation.

Require: Total histogram bins N, Temporal Volume T, Number of foveated bins M, Total histogram bins for depth foveation N'

1: **Calculate bin widths**
$\Delta t = \frac{T}{N}$, $\Delta t_{depth} = \frac{T}{N'}$
2: **Acquire a depth prior:**
Monocular Section IV, Optical-Flow Section VI, Low-Resolution Super-Pixel Sampling Section VII
3: **for** $(x, y) \in S$ **do**
4: Utilize the depth prior to find $\hat{d}(x, y)$
5: Center foveation window M around $\hat{d}(x, y)$
Memory Foveation:
6: Capture histogram in the foveated window with bin width Δt and M number of bins
Depth Foveation:
7: Capture histogram in the foveated window with bin width Δt_{depth} and N' number of bins
8: **end for**
9: **return** Histogram image H
10: Decode depth image D. $H \rightarrow D$
Optional Spatio-Temporal steps:
11: **Quantization Based Sampling** Section V
12: Quantize depth prior into discrete buckets B
13: Select several pixels in each bucket at random. $S \rightarrow \hat{S}$
14: Complete steps 3–10 with $\hat{S}$
15: Quantize sparse depth map. $D(B) = \min(D(\hat{S}) \in B)$
16: **SuperPixel Based Sampling** Section VII
17: Acquire a pseudo-intensity map through photon counting
18: Apply the superpixel algorithm to segment the pseud-intensity map
19: Sample the centroid of each superpixel segment at full histogram resolution. $\hat{d}_{\text{SP}}$
20: Complete steps 3-10 with S and $\hat{d}_{\text{SP}}$

than $\frac{C_{\text{new}}}{C} \geq \frac{N^2}{M^2}$, then,

$$\text{SNR}_{\text{new}} \propto C_{\text{new}} \sqrt{\frac{MT}{N^2}} = C \sqrt{\frac{T}{N}}. \tag{4}$$

In summary, memory foveation reduces memory usage with no change in SNR. Depth foveation increases depth resolution but with reduced SNR that can be compensated by more laser photons (i.e. longer exposure).

Below, in Algorithm 1, we define the general algorithm for memory and depth foveation. Note that the algorithms are independent of depth prior, and the spatio-temporal step, which we show in Section V, is optional.

2) Strong Ambient Light (Pileup): With strong ambient light, we now focus on the signal-to-background ratio (SBR), defined in [46] for SPADs as the ratio of the total number of signal photons to the total number of background photons received over each laser cycle. W.l.o.g, here we note that the SBR is proportional to the probability of receiving signal photons divided by the probability of receiving background photons.

With ambient light, photons from both the laser source and the ambient illumination may be measured by the SPAD. Each time a photon is detected, the SPAD sensor resets creating a pause. It is this pause that creates a binomial model for image capture in SPADs [46], [47].

Therefore, the SBR analysis cannot simply compare the photon bin widths as in the prior section for the full resolution (N bins) and the foveated resolution (M bins). Instead, SBR calculations must include the *probability* of photons from the source vs. the background.

Conventional scenario: Let us first consider the SBR in the conventional case, with no foveation. From [47], using the Poisson model for photon distribution, we can write the probability of a photon from the laser incident on the bin corresponding to the correct depth as $p_{\text{laser}} = (1 - e^{-\Phi_{\text{sig}}})$. Correct depth detection will happen even if an ambient photon is detected at the correct depth, so the probability of correct depth detection is $p_{\text{correct}} = (1 - e^{-(\Phi_{\text{sig}} + \Phi_{\text{bkg}})})$.

Let i be the location of the bin corresponding to the correct depth of the scene point. This photon is only detected at i if, in addition, no photon from the laser is detected at any prior bin. Since the laser photons only show up at bin i, constrained by depth, the probability of the photon showing up at any other bin is zero. However, in this conventional scenario, photons from ambient light could show up at any prior bin to i, pausing detection at bin i. Therefore, the probability that the photon from the laser is detected at the correct depth is $p_{\text{sig}} = (1 - e^{-(\Phi_{\text{sig}} + \Phi_{\text{bkg}})})\, e^{-\Sigma_1^{i-1} \Phi_{\text{bkg}}}$.

The situation is different for ambient photons, which can arrive at any time instant before photons from the i^{th} bin arrive. We can write the probability that an ambient photon is detected at location q as $p^q_{\text{bkg}} = (1 - e^{-\Phi_{\text{bkg}}})\, e^{-\Sigma_1^{q-1} \Phi_{\text{bkg}}}$. We can therefore write the SBR proportionality for the conventional imaging case as:

$$\text{SBR} \propto \frac{p_{\text{sig}}}{p_{\text{bkg}}} \propto \frac{(1 - e^{-(\Phi_{\text{sig}} + \Phi_{\text{bkg}})})\, e^{-\Sigma_1^{i-1} \Phi_{\text{bkg}}}}{\Sigma_{q=1}^{i} p^q_{\text{bkg}}}. \tag{5}$$

FoveaSPAD with Ambient Light: We now consider both memory foveation and depth foveation where the foveated bins N are given to us. In both these scenarios, we model the arrival of photons from both ambient and laser sources.

Memory foveation: Consider the foveated bins N, which we assume contain the bin with the histogram peak. Suppose the closest index for these bins is j. Then, the SBR increases, since the histogram sensitivity is unaffected by photons that impact the sensor before bin j.

$$\text{SBR} \propto \frac{(1 - e^{-(\Phi_{\text{sig}} + \Phi_{\text{bkg}})})\, e^{-\Sigma_j^{i-1} \Phi_{\text{bkg}}}}{\Sigma_{q=j}^{i} p^q_{\text{bkg}}}. \tag{6}$$

In the extreme case, where we have perfect foveation, and $i = j$, then the terms for ambient light before bin i become 1,

$$\text{SBR} \propto (1 - e^{-(\Phi_{\text{sig}} + \Phi_{\text{bkg}})}). \tag{7}$$

i.e. in other words, the effect of foveation is to remove the dependence on prior photon arrival for detection, since these no longer delay the measurement of photons at the ith bin. This "perfect foveation" SBR term is dependent on the ratio of the strength of the laser and ambient signal directly and is not constrained by the binomial nature of SPAD photon capture.

Depth foveation: Since we concentrate all N bins into the foveation window, we are again susceptible to the binomial nature of SPAD photon capture. In addition, the bins are smaller to fit within the window, and as described in the non-ambient light section, the bin width is reduced as $\frac{M}{N}$.

We can write the probability that an ambient photon is detected at location q as $p^q_{\text{bkg}} = \left(1 - e^{-\frac{M}{N}\Phi_{\text{bkg}}}\right) e^{-\Sigma_1^{q-1}\frac{M}{N}\Phi_{\text{bkg}}}$. The SBR proportionality also shows the effect of reduced signal strength as:

$$\text{SBR} \propto \frac{p_{\text{sig}}}{p_{\text{bkg}}} \propto \frac{\left(1 - e^{-\left(\frac{M}{N}(\Phi_{\text{sig}}+\Phi_{\text{bkg}})\right)}\right) e^{-\Sigma_1^{i-1}\frac{M}{N}\Phi_{\text{bkg}}}}{\Sigma_{q=1}^{i} p^q_{\text{bkg}}}. \quad (8)$$

In summary, memory foveation increases SBR. While depth foveation has the same SBR as conventional capture, it improves depth resolution. It is this theory that motivates the remaining simulation results in the paper, where we explore different ways of creating depth and memory foveation for SPAD sensors.

TABLE I
MATHEMATICAL SYMBOLS USED IN THIS PAPER TO STUDY THE FOVEATED SPAD IMAGING MODEL

Symbol	Meaning
N	Number of bins across full histogram
M	Number of bins across foveated histogram
i	Bin location of corresponding to true scene depth
Z	Working volume of the sensor
T	Temporal volume calculated from Z and speed of light
SNR	Signal-to-noise ratio
SBR	Signal-to-background ratio
C	Number of cycles to create histogram
Φ_{sig}	Mean number of signal photons received per bin
Φ_{bkg}	Mean number of background photons received per bin
p_{gt}	Probability that a detected photon originated from the laser
$p_{\text{multipath}}$	Probability that a detected photon experienced multipath bounces
p_{floor}	Probability of a low noise floor
S	Number of pixels in the camera

IV. SPAD FOVEATION FROM MONOCULAR DEPTHS

With the imaging model defined, we proceed to our first experiment, demonstrating how our memory and depth foveation techniques can effectively work with a monocular depth prior.

Monocular depth estimation is inherently brittle due to biases in training datasets, whereas SPADs provide high-accuracy sensor measurements. In this section, we leverage the less accurate monocular depth to reduce the number of SPAD bins needed for capturing data, thereby saving memory and improving depth resolution.

Simulation Details: We conducted our simulations using the SPAD simulation framework provided in Gutierrez-Barragan et al. [5], [48], utilizing the code available on GitHub. While the simulations are initialized with RGBD datasets, all "ground truth" depth images presented in this paper result from SPAD simulation on full high-resolution histograms.

Monocular depth estimation algorithms use visual cues from 2D images to infer depth information and are trained on annotated datasets such as NYU Depth v2 [49] and KITTI [50]. We employed ZoeDepth [51], a monocular depth estimator chosen for its performance and ability to produce metric depth estimates. The monocular depth is used to guide a foveation window consisting of M bins in the histogram. The window size is a hyper-parameter, with larger sizes offering better accuracy at the cost of reduced efficiency.

For effective use of the monocular estimate as a prior, it must provide metric depth, and to enhance foveation performance, it needs to be scaled to match the scene. ZoeDepth fulfills the metric depth requirement, and we ensure compatibility with the dataset through appropriate scaling and bounding.

We chose a polynomial fit for scaling, observing that a majority of points in a randomly selected subset of the monocular output for the NYUv2 dataset exhibited a linear relationship. This scaling can be performed either locally, fitting the data to a specific scene, or generally across the dataset. In both cases, a small set of pixels is sampled at full histogram resolution, and the relationship between the monocular estimate and the SPAD estimate at these pixels is modeled. The fit is then applied to the entire monocular estimate, with bounds enforced for the minimum and maximum values across the dataset, which are 0m and 10m for NYUv2.

We now describe our results shown in Fig. 2 and evaluated in Table II which are calibrated locally. The first two columns in the figure show the ground truth from the NYUv2 dataset. The depth is not simply the depth from the NYUv2 dataset, but the output of full-resolution SPAD simulation followed by the detection of the histogram peak. The third column shows the **scaled** monocular output.

Memory Foveation: The fifth column in Fig. 2 shows our memory foveation results. Here, most bins are not used, saving memory for the same SNR. The foveated window is given at the right of the figure as a fraction of the original number of bins N, with N set to 1000 bins for all experiments. The results are visually indistinguishable from ground truth, in some cases with a $\frac{1}{16}$ save in memory. In Table II we show the change in accuracy with these memory savings. Unsurprisingly, there is an inverse relationship between memory usage and depth error.

Depth Foveation: In Fig. 2 the foveated window around the estimated monocular depth is packed with a limited number of bins. With no foveation, as in the fourth column, a limited number of bins N' are distributed over the entire SPAD volume. The depth foveation in the last column shows what happens when these limited number of bins are packed into the foveated window. Note that the depth resolution has increased from the limited bins case because the samples are placed within a foveated window where we expect to find the histogram peak. In Table II, entries with the same memory usage demonstrate the effects of depth foveation, where higher depth resolution

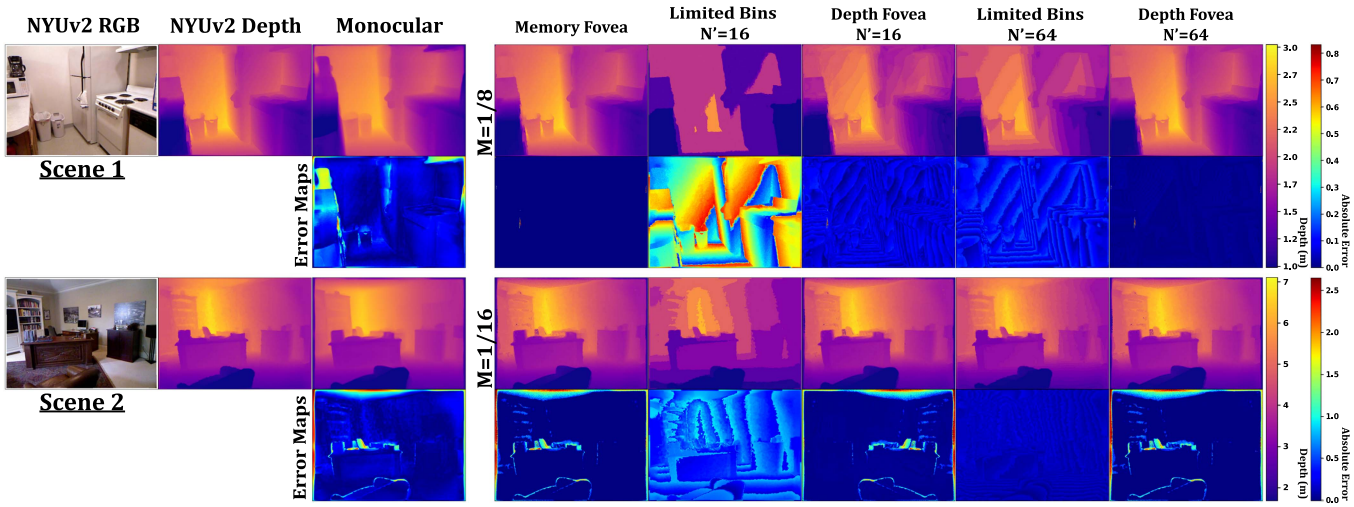

Fig. 2. Qualitative Comparison on NYUv2 Our memory and depth foveation techniques produce quality depth reconstructions with a fraction of the memory usage. Each row consists of the NYUv2 ground truth images, the monocular depth output from ZoeDepth, a simulated SPAD output with N′ bins, and our foveation techniques. The rows show different combinations of M and N′, where M is the number of bins in the foveated histograms, and N′ is the limited number of bins used for depth foveation. Monocular estimation is just one method of obtaining a depth prior in a class of methods, in Section VI and Section VII we show two more methods.

TABLE II
MEMORY AND DEPTH FOVEATION EVALUATION - LOCAL SCALE

M (Fraction)	RMSE↓ (m)	$\log_{10}$↓ (m)	REL↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)	N′ (Num. Bins)	RMSE↓ (m)	Lim. Bins↓ RMSE (m)	$\log_{10}$↓ (m)	REL↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)
1/16	0.211	0.0106	0.0211	97.07	99.13	99.55	16	0.235	0.504	0.0173	0.0360	96.55	98.96	99.48
							32	0.211	0.250	0.0119	0.0241	97.1	99.14	99.55
							64	0.211	0.121	0.012	0.0242	96.44	99.01	99.54
1/8	0.151	0.005	0.0109	98.36	99.42	99.79	16	0.201	0.509	0.018	0.0418	97.87	99.26	99.71
							32	0.184	0.250	0.011	0.0254	98.1	99.38	99.77
							64	0.152	0.121	0.0064	0.0141	98.36	99.45	99.81
1/4	0.117	0.0032	0.00686	99.24	99.57	99.79	16	0.221	0.501	0.0326	0.0714	98.77	99.6	99.82
							32	0.166	0.2497	0.015	0.0355	99.15	99.59	99.82
							64	0.145	0.123	0.0087	0.0195	99.01	99.52	99.78

This table shows a quantitative comparison of RMSE and depth inlier metrics for different depth and memory foveation strategies for the NYUv2 dataset and a monocular estimation prior. For each memory foveation fraction, we vary the number of histogram bins in the foveated sub-window to achieve depth foveation. metrics used from left to right: root-mean-squared errror, absolute log_{10} error, absolute relative error, $\delta < 1.25$, $\delta < 1.25^2$,$\delta < 1.25^3$.

consistently produces better results. These depth foveation outcomes are directly dependent on the memory foveation results, as both algorithms place fovea windows based on the same depth prior, with the depth foveation experiments having a lower depth resolution. Meaning, the memory foveation results establish a lower bound for the depth foveation error. Additionally, the limited bins case, which is not confined to a foveated window and thus reliant on a depth prior, shows that the error continues to decrease as depth resolution increases.

V. SPATIO-TEMPORAL SPAD FOVEATION

The previous section seeks to reduce the SPAD histogram bottleneck by reducing the number of bins to examine per-pixel with a monocular estimate prior. This section aims to improve these savings by incorporating spatial foveation. By exploiting depth coherencies and applying foveated windows to a small selection of pixels we show an order of magnitude increased bandwidth savings.

Foveated LiDAR systems [31], [32], [33] can place samples onto depth edges and recover the rest of the scene, post-capture, through algorithmic estimation such as deep guided upsampling or gradient-based reconstruction. Similarly, here, we place samples *across* depth edges and, rather than use an algorithm, we use the SPAD measurement to provide correct depths in redundant areas.

Quantized Sampling: Our approach to spatial sampling begins by quantizing the prior through thresholding, resulting in digitized regions that we refer to as 'buckets.' We make the *assumption* that the values within each quantized bucket are redundant. From each bucket, we randomly select pixels and use the SPAD to measure these points in the scene, applying memory foveation in the process. These measurements provide a sparse depth map, which we subsequently sort and quantize based on the buckets defined by the depth prior.

In Fig. 3 we show examples of our approach, where the first two columns show the scene and ground truth depths. The third column is a quantized version of the monocular depth estimation,

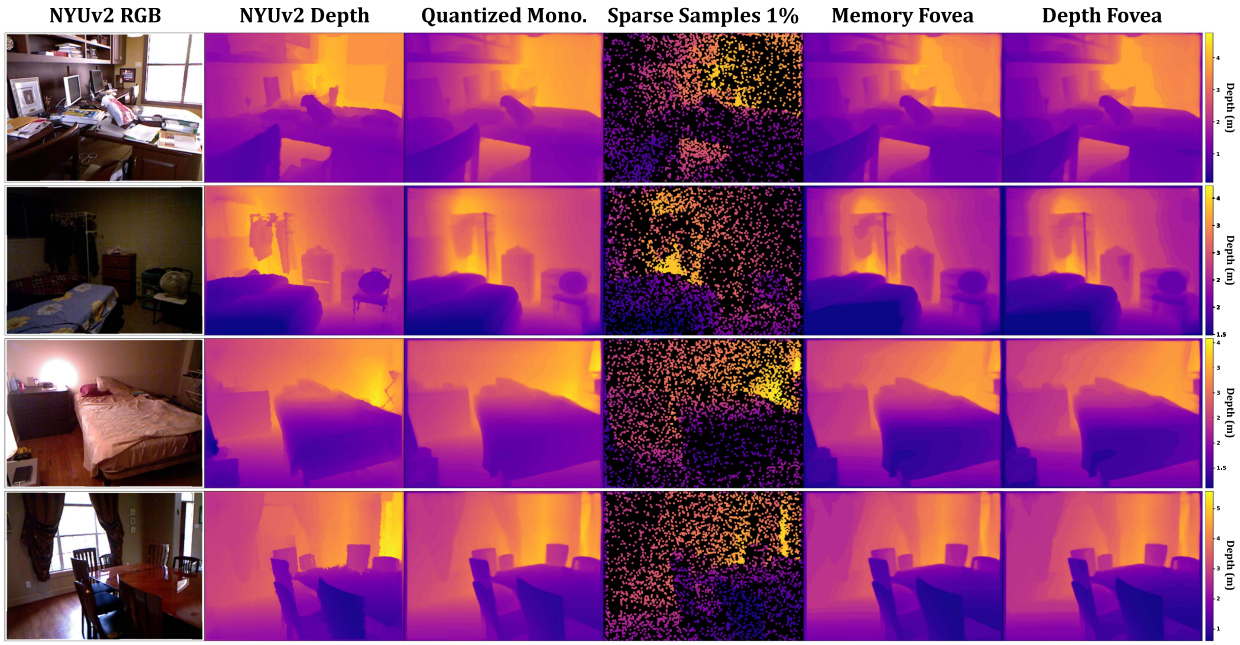

Fig. 3. Spatio-temporal foveation The first two columns display the scene's color and ground truth depth. Using the quantized monocular depth in the third column, we select certain pixels in the fourth column. Processing only histograms at these locations with foveated windows generates results in the last column, indicating a 1548-fold reduction in memory usage. This is calculated by measuring memory allocation for full-res and spatio-temporal histograms. The results shown are with **M=1/16N** and $\mathbf{N'} = \mathbf{16}$.

TABLE III
SPATIO-TEMPORAL FOVEATION EVALUATION - LOCAL SCALE

Sparsity = 0.26 %																
M (Fraction)	**RMSE**↓ (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)	**N′** (Num. Bins)	**RMSE**↓ (m)	**Lim. Bins**↓ **RMSE** (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)		
1/16	0.39	0.06	0.124	84.901	97.054	99.429	16	0.649	0.509	0.102	0.15	83.788	95.189	96.514		
							32	0.687	0.251	0.103	0.151	81.556	95.272	96.972		
1/8	0.392	0.068	0.137	80.154	94.812	99.046	16	0.738	0.502	0.129	0.19	71.23	91.362	95.817		
							32	1.055	0.269	0.17	0.202	69.595	89.694	92.852		
1/4	0.355	0.054	0.10	88.244	98.114	99.186	16	0.756	0.497	0.131	0.199	67.472	92.431	96.184		
							32	0.837	0.25	0.137	0.202	68.609	86.771	93.232		
Sparsity = 0.52 %																
M (Fraction)	**RMSE**↓ (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)	**N′** (Num. Bins)	**RMSE**↓ (m)	**Lim. Bins**↓ **RMSE** (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)		
1/16	0.414	0.07	0.12	87.672	97.008	98.139	16	0.582	0.505	0.092	0.134	86.543	96.111	97.068		
							32	0.484	0.25	0.07	0.119	87.664	96.492	98.158		
1/8	0.387	0.051	0.108	87.292	99.162	99.919	16	0.518	0.519	0.071	0.136	84.049	98.177	99.255		
							32	0.587	0.248	0.074	0.142	78.714	94.945	97.969		
1/4	0.38	0.049	0.0996	90.254	96.965	98.365	16	0.734	0.518	0.121	0.184	74.702	93.679	96.225		
							32	0.553	0.256	0.068	0.127	85.968	96.942	98.09		
Sparsity = 1.04 %																
M (Fraction)	**RMSE**↓ (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)	**N′** (Num. Bins)	**RMSE**↓ (m)	**Lim. Bins**↓ **RMSE** (m)	**log10**↓ (m)	**REL**↓	δ_1↑ (%)	δ_2↑ (%)	δ_3↑ (%)		
1/16	0.288	0.039	0.0855	94.214	99.582	99.935	16	0.364	0.508	0.048	0.0959	93.693	99.248	99.646		
							32	0.412	0.254	0.051	0.0933	93.048	98.179	99.145		
1/8	0.313	0.04	0.0881	91.782	99.443	99.841	16	0.386	0.495	0.056	0.111	90.719	99.057	99.474		
							32	0.432	0.257	0.053	0.106	89.662	98.472	99.276		
1/4	0.274	0.035	0.0786	94.264	99.045	99.875	16	0.471	0.503	0.072	0.148	82.311	97.104	98.821		
							32	0.399	0.25	0.063	0.111	91.482	97.966	98.643		

Here we look at a quantitative comparison between the size of the foveation window (memory usage), the number of bins in depth foveation, and the number of total samples per the spatio-temporal algorithm.

where the number of quantized buckets is 64. For each of these buckets, we picked 50 points at random and recovered the SPAD depths of these points. Note that these transients were also foveated in time, using the method described in the previous section. The fourth column in Fig. 3 depicts exactly those points in the SPAD camera that were sampled, with the number of bins sampled at $\frac{1}{16}$ of the original histogram. This is a factor of 1548 memory savings, compared to the ground truth measurement, with depth results in the last column. These efficiencies are evaluated in Table III.

VI. OPTICAL FLOW DRIVEN SPAD FOVEATION

In previous sections, we focused on static scenes. However, one of the key advantages of using SPAD arrays is their fast capture speed, making them ideal for dynamic environments, such as when mounted on a vehicle. In this section, we demonstrate how our techniques can be applied to moving scenes by utilizing optical flow to guide the foveation process.

Consider a SPAD sensor on a moving platform, say an autonomous vehicle, where high-frame rate and efficient depth

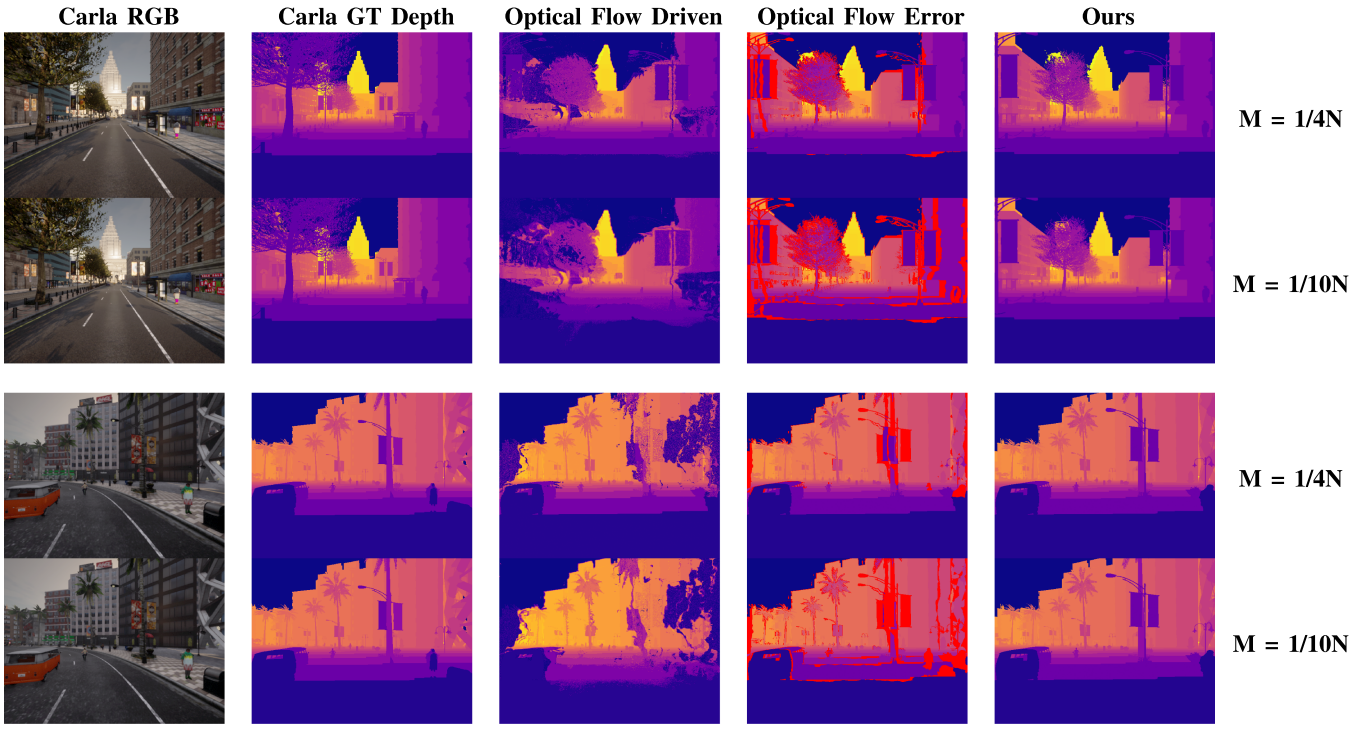

Fig. 4. Optical Flow Driven Foveation Here we see our optical flow driven SPAD foveation using the Carla simulator whose color and ground-truth depth are shown in the first two columns. Directly using optical flow, as shown in the third column, creates errors that propagate over time. We correct for the optical flow error by detecting those pixels whose foveated windows are close to the noise floor. The last column shows the final optical flow driven foveated depth at different window sizes. Please see the supplementary for video results.

capture are important [52], [53]. The foveation algorithm described in the previous section analyses pixels in each frame, reducing the bins in the histogram that need to be processed. Here we consider an approach to reduce the computation even further, using temporal information by transferring foveation information from previous frames to subsequent frames.

Consider a sequence of frames containing both depth and reflectance information from a scene. Assume that the depth in the first frame is reconstructed at high quality, such as from full-resolution SPAD histograms. Now, for a subsequent frame, we can calculate optical flow between the frames (color or grayscale), producing a vector (u, v) for each pixel at a given time t. These vectors satisfy the brightness consistency principle, meaning that $I(x + u \cdot \delta t, y + v \cdot \delta t, t + \delta t) = I(x, y, t)$. holds true. We use the depth information from the previous frame to guide the positioning of the foveating window in the current frame, by warping the previous frame based on the vector (u, v). Although the object may move and the histogram peak will shift from frame to frame, it will remain within a nearby range, allowing a window of pixels to recover the histogram peak in the current frame.

However, optical flow is never perfect, often having errors at the edges of a frame. Further, these propagate incorrect depths through time, since our optical flow method only considers the depths in the previous frame. To remove this error, we compare the distribution of the photons under a foveated region to that from a noise floor. If they match, we ignore the erroneous optical flow, and recompute depth from the full histogram. In practice, this is done by thresholding the values in the foveated window.

In Fig. 4, we show some optical flow results. Please see the supplementary video for all of our video results. These were created on the CARLA simulator [54] and the results show two street scenes with ground truth depths. We found the native optical flow in CARLA to be noisy, and so we used OpenCV's in-built optical flow estimator. The third and fourth columns show first the incorrect results from optical flow, and our method to detect these regions, shown in red. The optical flow driven depth foveation results are shown in the last column. Calculating errors using a running average across all video frames reveals compounding errors over time. In the first scene, at $\frac{1}{10}N$, RMSE and SSIM are 101.9 m and 0.530, and at $\frac{1}{4}N$, 38.6 m and 0.884. In the second scene, RMSE and SSIM are 0.164 m and 0.87 for both $\frac{1}{10}N$ and $\frac{1}{4}N$.

Time Delay: In dynamic scenes, unlike static ones, time delay becomes a potential issue. The process of calculating optical flow may introduce a delay between capturing the prior frame and the current one, which could lead to increased errors. Additionally, the proposed error correction method in this section relies on resampling pixels, which could introduce time-delay artifacts if the frame rate is not sufficiently high. However, with a fast enough capture speed, as provided by SPAD arrays, these delays become negligible, maintaining the accuracy of the output depth map.

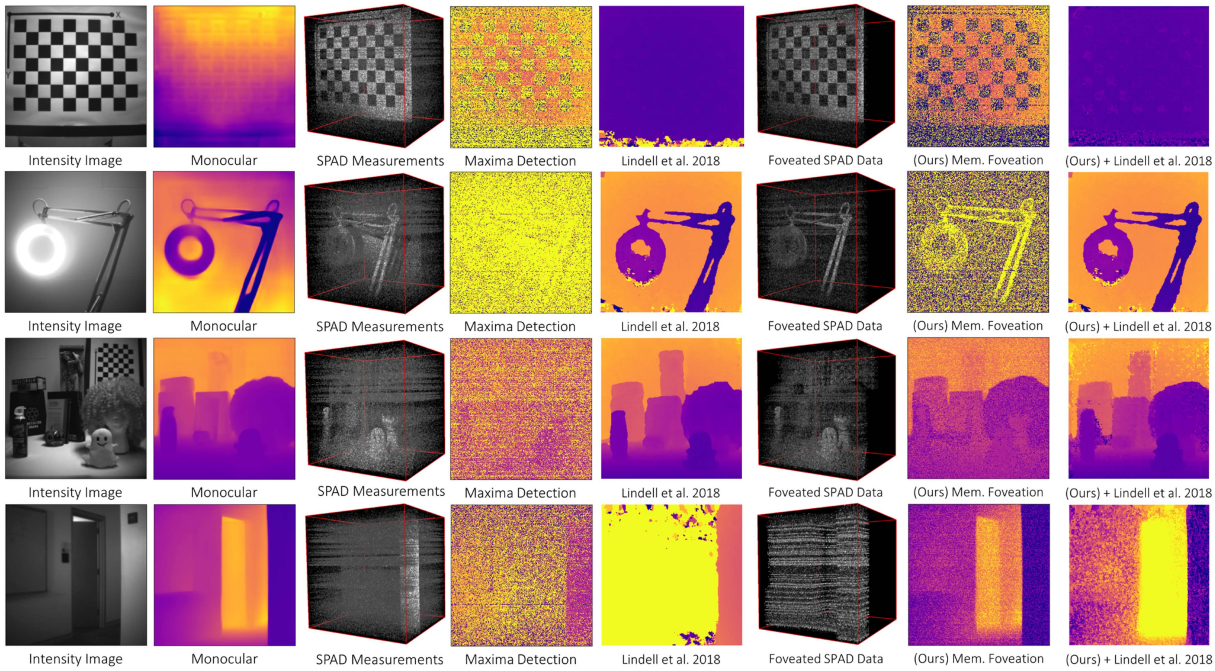

Fig. 5. Hardware emulation results for scenes from Lindell et al. [4]. (Column 1) The Lindell dataset consists of monochrome images captured by a camera co-aligned with the SPAD sensor that captures photon data cubes. (Column 2) We obtain monocular depth maps from these monochrome images. (Column 3) Raw photon data cube without foveation shows a "cloud" of background photon detections. (Column 4) Maxima detection on low SBR photon clouds leads to unusable depth maps. (Column 5) The CNN-based algorithm of Lindell et al. improves depth map reconstruction. (Column 6) Our approach relies on memory foveation in a 1/4th size sub-window around an estimate of the true depth obtained from monocular depth maps. Observe that the photon data cubes are less noisy. (Column 7) Even a simple max-estimator provides better depth map estimates after foveation. (Column 8) Providing foveated clouds to the CNN denoiser of Lindell et al. further improves reconstructions.

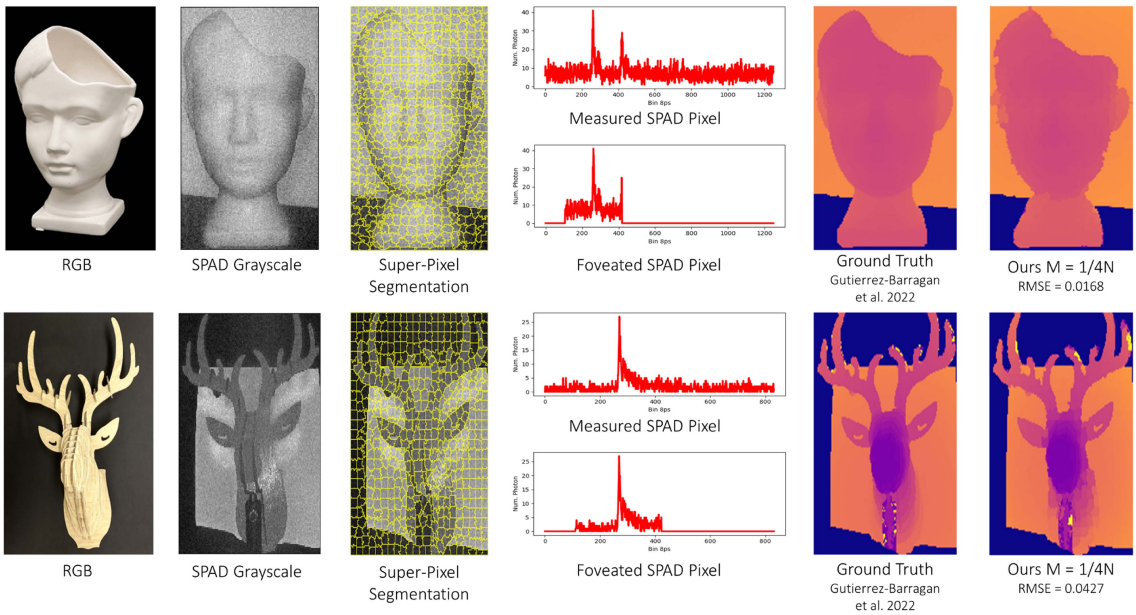

Fig. 6. Hardware emulation results for scenes without co-aligned monochrome camera [5]. (Column 1) RGB images of the "face-vase" and "reindeer" scenes shown for visualization. (Column 2) A pseudo-intensity image is estimated by accumulating photon counts for each pixel. (Column 3) Pseudo intensity maps are converted into superpixel representations, and a single pixel in each superpixel is used for measuring complete histograms. (Column 4) The peak location of the chosen pixel is used to apply foveation windows of 1/4th the total temporal extent for the remaining pixels in each superpixel. (Column 5) Ground truth depth maps obtained using matched filtering. (Column 6) Our result requires $64\times$ less memory per pixel for $> 99\%$ of the pixels in these scenes.

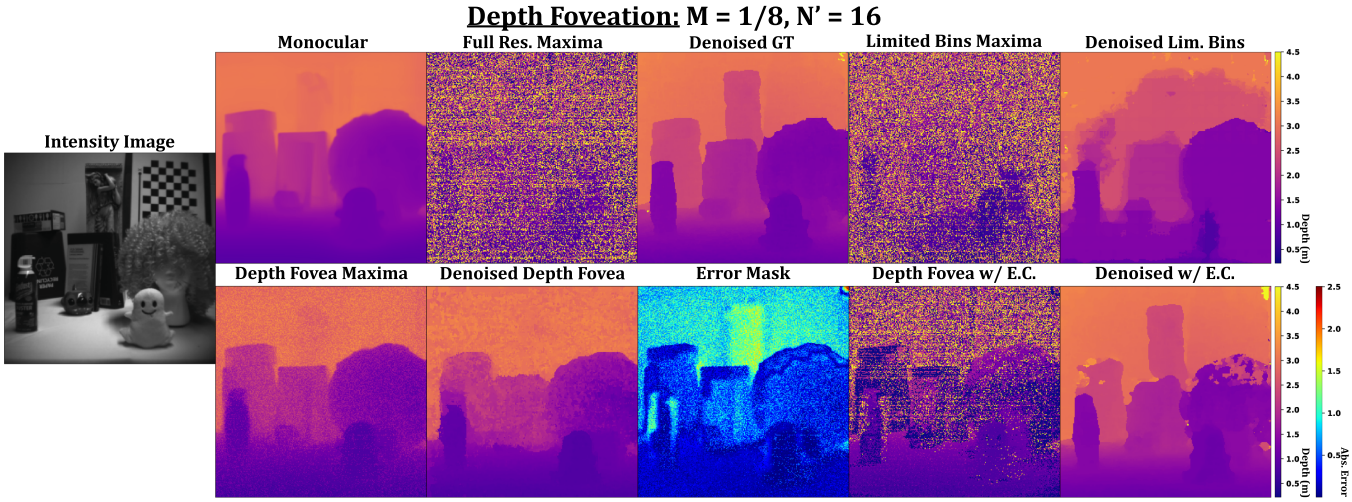

Fig. 7. Additional Results: Depth Fovea. This figure demonstrates the application of the depth foveation technique described in Section IV to the Lindel dataset, along with the error correction technique presented in the supplementary material. A window size of M = 1/8 and a bin count of N' = 16 were used. The results were subsequently processed using the sensor fusion denoising network [4].

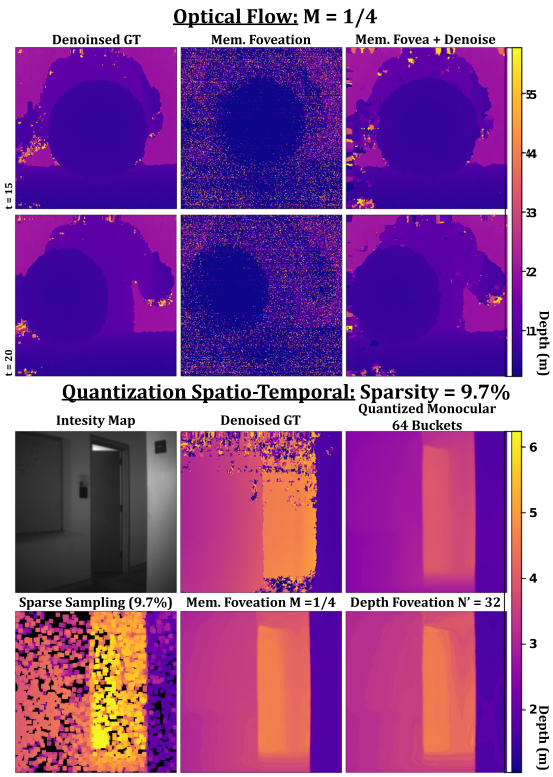

Fig. 8. Additional Results: Optical Flow and Quantization Spatio-Temporal. This figure illustrates the application of the techniques described in Section VI and Section V to the Lindel dataset. The top portion showcases our optical flow algorithm on the "roll" scene. The first column displays the denoised ground truth, followed by the optical-flow-driven memory foveation result using maxima detection, and finally the denoised memory foveation result. The bottom portion of the figure presents our quantization spatio-temporal foveation technique, utilizing 9.7% sampling to mitigate the high levels of noise and the abundance of pixels with no photon counts in the scene.

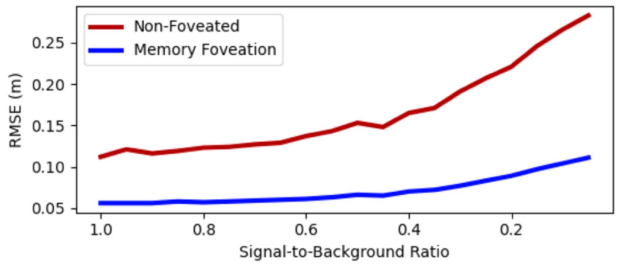

Fig. 9. Effect of increasing background illumination. The conventional (non-foveated) depth map quality degrades more rapidly as background illumination increases. Using memory foveation allows reliable depth map recovery for the "deer" scene for a wider range of SBR levels.

VII. Hardware Emulation Results

In this section, we present hardware emulation results for depth and memory foveation using SPAD data captured using real hardware. The goal of hardware emulation study is to de-risk future in-pixel implementations of foveation algorithms. We use datasets by Lindell et al. [4] and Gutierrez-Barragan et al. [5] from prior sources [46], [47].

A. Using Monocular for Memory Foveation

We'll start by showcasing how our memory foveation technique works on the dataset by Lindell et al. [4] by using monocular as a prior. The Lindell dataset consists of scenes under different ambient illumination conditions captured using a linear SPAD pixel array [55] co-aligned with a monochrome camera that captures intensity images.

We use these intensity images to obtain a monocular depth prior. Because the performance of monocular estimation networks is dependent on the dataset, we perform a calibration step by using the "elephant" scene in the dataset to define a global scaling function. We place foveation windows of 1/4th the total temporal extent of the full histograms centered around these scaled monocular depth estimates for each pixel.

Memory foveation improves the overall SBR, in a scene-adaptive manner, by focusing on regions of the spatio-temporal photon cube where signal photons arrive. Comparing columns 3

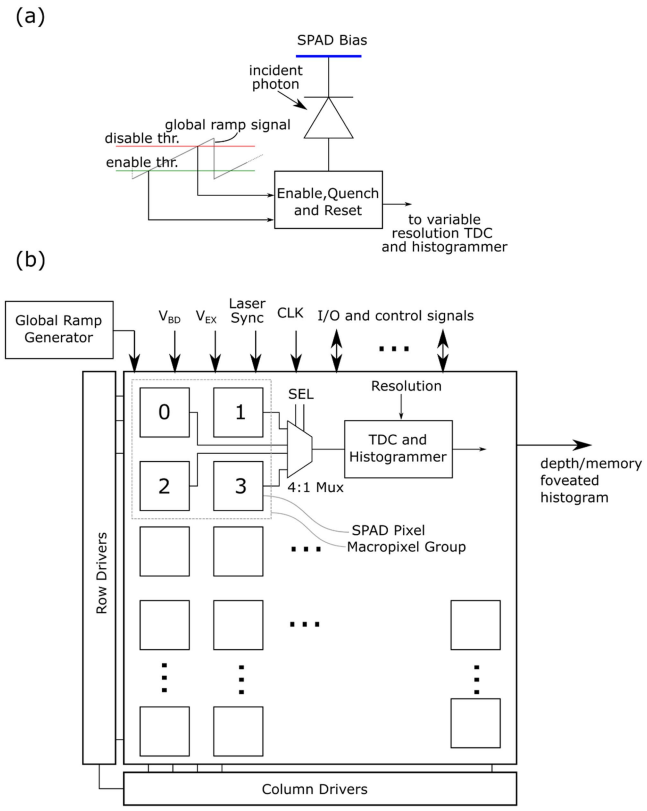

Fig. 10. Future pixel and array designs for foveated single-photon 3D imaging. (a) A speculative pixel design where individual SPADs are gated on or off based on thresholds set with respect to a linear ramp signal. Pixels only need to store the thresholds; the ramp signal is generated externally. (b) A possible array of SPAD pixels with per-pixel gating. Observe that the ramp signal is generated globally, simplifying pixel design. Variable-resolution TDCs and histogrammers are shared by small pixel neighborhoods (e.g., 2×2 multiplexed "macropixels") to improve fill factor.

and 6 in Fig. 5, foveated SPAD measurement cubes show fewer background photon detections, with clear 3D object structure in the photon cubes. Depth estimates are improved even with a simple maxima-detection approach — observe that the lamp is barely visible in the non-foveated maxima-detection-based depth map in column 5, but is visible after memory foveation in column 7. Running memory foveated measurements through the denoising algorithm of Lindell et al. further improves the depth map, as seen in the last column of Fig. 5.

B. A Different Approach to Spatio-Temporal Foveation

To illustrate the flexibility of our foveation techniques and their independence from external sensors as a prior, we propose an alternative spatio-temporal method, which we apply to two scenes from the Gutierrez-Barragan et al. dataset [5], for which there is no co-located camera. The dataset is captured using a single-pixel point scanned SPAD detector co-aligned with a pulsed laser. Fig 6 shows the results of the alternate approach for the single object "face-vase" and "reindeer" scenes, with the RGB images shown in column 1 for visualization purposes.

SuperPixels: Because there is no intensity map captured in the dataset, we instead obtain a pseudo-intensity map by summing the raw photon data cubes along the temporal axis for each pixel. In a real hardware implementation, this process would be achieved by utilizing a counter in each SPAD pixel, a feature commonly available in existing commercial SPAD arrays. We then run a superpixel algorithm [56] on the pseudo-intensity maps to obtain coarse segmentations of the scene, as shown in column 3. For each superpixel segment, we capture a complete (non-foveated) histogram of the centroid pixel. By identifying the true peak location in this histogram, we can then foveate within a 1/4th sub-window centered around this peak for all remaining pixels in the superpixel segment, reducing the overall bandwidth requirement per pixel by a factor of 64.

In the "face-vase" scene, with a spatial resolution of 174×154 pixels, the segmentation reduces the data to 473 superpixels. Similarly, the "deer" scene, originally at 204×116 pixels, is reduced to 515 superpixels. This reduction translates to a 3/4 reduction in memory requirement for approximately 99.98% pixels in both scenes. Examples of foveated histograms in column 4 show that the laser impulse response function has a non-ideal shape which departs significantly from the commonly assumed Gaussian shape used in simulation studies. (The second peak is likely due to optical inter-reflections in the hardware setup). Yet, our method is able to produce reliable depth maps (columns 5 and 6).

We also examine the impact of reconstruction error under increasing background noise for the "deer" scene. As shown in Fig. 9, foveation allows for the accurate selection of the correct depth peak, even in the presence of strong background illumination, thereby expanding the operable SBR range in practice.

VIII. Limitations and Discussion

Worst Case Stochastic Limits: We explored the limitations of our approach by analyzing the worst-case scenario where depth is incorrectly detected due to various errors, such as monocular depth calibration issues, ambient light interference, and global effects like multipath inter-reflections. We characterized these errors using a probabilistic framework. Specifically, we defined the probability p_{gt} as the chance that a detected photon originates from the laser i.e. single-bounce photons, $p_{\text{multipath}}$ as the probability of multipath photon detection, and p_{floor} as the probability of spurious peaks due to sensor noise. The overall probability of accurate depth detection is given by

$$p_{\text{gt}}(1 - p_{\text{gt}}p_{\text{multipath}})^{M-1}p_{\text{floor}}, \tag{9}$$

where M is the number of foveated bins. We further derived the probability p_{worst} for the worst-case scenario, where none of the S pixels detect the correct depth, expressed as

$$p_{\text{worst}} = (1 - p_{\text{gt}}(1 - p_{\text{gt}}p_{\text{multipath}})^{M-1}p_{\text{floor}})^{S}. \tag{10}$$

Through optimization, we identified two conditions that lead to this worst-case scenario, linked to specific relationships between p_{gt}, $p_{\text{multipath}}$, and M.

- The first condition occurs when $p_{\text{gt}} = \frac{1}{p_{\text{multipath}}}$. This situation arises when the probability is 1 for every bin to contain both direct photons from the laser and photons that have

undergone multipath effects, indicating a degenerate scene, such as one made entirely of mirror-like surfaces.

- The second condition occurs when $p_{\text{gt}} = \frac{1}{M \cdot p_{\text{multipath}}}$. This scenario implies that the number of foveated bins M and the probability of multipath effects $p_{\text{multipath}}$ must satisfy this relationship, under the constraint that $0 \leq p_{\text{gt}} \leq 1$. This suggests that it is possible to avoid the worst-case scenario by adjusting the number of bins M for scenes with specific global illumination characteristics.

In order to illustrate the findings of this analysis, consider a toy example with a number of bins $M = 1000$ and pronounced multipath effects, such as $p_{\text{multipath}} = 0.1$. In the worst case, the probability of depth recovery would be significantly hindered $p_{\text{gt}} = 0.01$, but can be improved by changing the number of bins M at the cost of depth resolution. The detailed derivations of these results are provided in the supplementary material.

Quality of depth priors: Our algorithms can enable memory-efficient SPAD sensing while maintaining depth accuracy. However, our method strongly relies on the accuracy of the depth prior. If the prior is incorrect, our algorithms may produce errors, highlighting the importance of robust error correction mechanisms. We can correct for such errors by trading off efficiency. For instance, in Section VI, we discuss a correction method for low-quality depth priors, where compounded errors arise from optical flow warping over time. Additionally, in the supplementary material, we present an error mask correction technique. This method enables adjustments, like enlarging the foveation windowâeven to the extreme of covering the entire transient spanâfacilitating more robust error management.

Hardware complexity: A key limitation of our approach is the lack of available hardware that fully supports our algorithms, necessitating more complex pixel architectures and driving up costs. Each SPAD pixel in the 2D array requires a programmable gate, along with a variable TDC and histogrammer, which increases the complexity and expense of the hardware. This presents a significant challenge to the widespread adoption and practical implementation of our method. In Fig. 10, we propose a potential array design with per-pixel gating capability, where a global ramp generator provides individualized on/off thresholds for each pixel. To enhance the fill factor, the TDC and histogrammer are shared among groups of neighboring pixels, forming "macropixels".

We believe the next generation of programmable and software-defined SPAD cameras [57], [58] will be key enablers for in-pixel and on-chip implementation of memory- and energy-efficient foveated sensing schemes. As SPAD cameras become low-cost and widely available [59], the integration of in-pixel foveated sensing algorithm proposed here will reduce memory consumption while maintaining depth accuracy, or alternatively, provide more accurate depth estimates without increasing memory usage.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[3] Y. Xiong and S. Shafer, "Depth from focusing and defocusing," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 1993, pp. 68–73.

[4] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3D imaging with deep sensor fusion," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, 2018.

[5] F. Gutierrez-Barragan, A. Ingle, T. Seets, M. Gupta, and A. Velten, "Compressive single-photon 3D cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17854–17864.

[6] S. W. Hutchings et al., "A reconfigurable 3-D-stacked SPAD imager with in-pixel histogramming for flash LiDAR or high-speed time-of-flight imaging," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 2947–2956, Nov. 2019.

[7] C. Zhang, S. Lindner, I. M. Antolović, J. M. Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252×144 SPAD flash LiDAR with 1728 dual-clock 48.8-ps TDCs, and pixel-wise integrated histogramming," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, Nov. 2019.

[8] F. Heide, S. Diamond, D. B. Lindell, and G. Wetzstein, "Sub-picosecond photon-efficient 3D imaging using single-photon sensors," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 17726.

[9] I. Gyongy et al., "High-speed 3d sensing via hybrid-mode imaging and guided upsampling," *Optica*, vol. 7, no. 10, pp. 1253–1260, 2020.

[10] G. Gariepy, J. Leach, R. Warburton, S. Chan, R. Henderson, and D. Faccio, "Picosecond time-resolved imaging using SPAD cameras," *Proc. SPIE*, vol. 9992, pp. 130–137, 2016.

[11] F. Gutierrez-Barragan et al., "Learned compressive representations for single-photon 3D imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 10756–10766.

[12] X. Ren et al., "High-resolution depth profiling using a range-gated CMOS SPAD quanta image sensor," *Opt. Exp.*, vol. 26, no. 5, pp. 5541–5557, 2018.

[13] N. A. Dutton et al., "A SPAD-based QVGA image sensor for single-photon counting and quanta imaging," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 189–196, Jan. 2016.

[14] A. T. Erdogan et al., "A CMOS spad line sensor with per-pixel histogramming TDC for time-resolved multispectral imaging," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1705–1719, Jun. 2019.

[15] N. A. Dutton, I. Gyongy, L. Parmesan, and R. K. Henderson, "Single photon counting performance and noise analysis of cmos spad-based image sensors," *Sensors*, vol. 16, no. 7, 2016, Art. no. 1122.

[16] C. Zhang et al., "A 240× 160 3D-stacked SPAD dToF image sensor with rolling shutter and in-pixel histogram for mobile devices," *IEEE Open J. Solid-State Circuits Soc.*, vol. 2, pp. 3–11, 2021.

[17] F. Taneski, I. Gyongy, T. A. Abbas, and R. K. Henderson, "Guided direct time-of-flight LiDAR using stereo cameras for enhanced laser power efficiency," *Sensors*, vol. 23, no. 21, 2023, Art. no. 8943.

[18] A. Ingle and D. Maier, "Count-free single-photon 3D imaging with race logic," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 07, 2023, doi: 10.1109/TPAMI.2023.3302822.

[19] R. Po, A. Pediredla, and I. Gkioulekas, "Adaptive gating for single-photon 3D imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16354–16363.

[20] R. Battrawy, R. Schuster, O. Wasenmüller, Q. Rao, and D. Stricker, "LiDAR-flow: Dense scene flow estimation from sparse LiDAR and stereo images," in *Proc. 2019 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 7762–7769.

[21] Z. Chen, V. Badrinarayanan, G. Drozdov, and A. Rabinovich, "Estimating depth from RGB and sparse sensing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 167–182.

[22] F. Mal and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proc. 2018 IEEE Int. Conf. Robot. Automat.*, 2018, pp. 1–8.

[23] J. Lu and D. Forsyth, "Sparse depth super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2245–2253.

[24] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. 2017 Int. Conf. 3D Vis.*, 2017, pp. 11–20.

[25] G. Riegler, M. Rüther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 268–284.

[26] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 268–284.

[27] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy LiDAR completion with RGB guidance and uncertainty," in *Proc. 16th Int. Conf. Mach. Vis. Appl. (MVA)*, 2019, doi: 10.23919/MVA.2019.8757939.
[28] T. Gruber, F. Julca-Aguilar, M. Bijelic, W. Ritter, K. Dietmayer, and F. Heide, "Gated2depth: Real-time dense LiDAR from gated images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.
[29] B. Tilmon and S. J. Koppal, "SaccadeCam: Adaptive visual attention for monocular depth sensing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6009–6018.
[30] T. Yamamoto, Y. Kawanishi, I. Ide, H. Murase, F. Shinmura, and D. Deguchi, "Efficient pedestrian scanning by active scan LiDAR," in *Proc. Adv. Image Technol.*, 2018, pp. 1–4.
[31] Z. Tasneem, D. Wang, H. Xie, and K. Sanjeev, "Directionally controlled time-of-flight ranging for mobile sensing platforms," in *Proc. Robot.: Sci. Syst.*, Pittsburgh, PA, USA, Jun. 2018, doi: 10.15607/RSS.2018.XIV.011.
[32] A. Bergman, D. Lindell, and G. Wetzstein, "Deep adaptive LiDAR: End-to-end optimization of sampling and depth completion at low sampling rates," in *Proc. IEEE Int. Conf. Comput. Photography*, 2020, pp. 1–11.
[33] F. Pittaluga, Z. Tasneem, J. Folden, B. Tilmon, A. Chakrabarti, and S. J. Koppal, "Towards a mems-based adaptive LiDAR," in *Proc 2020 Int. Conf. 3D Vis.*, 2020, pp. 1216–1226.
[34] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, "Foveated 3D graphics," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.
[35] R. Albert, A. Patney, D. Luebke, and J. Kim, "Latency requirements for foveated rendering in virtual reality," *ACM Trans. Appl. Percep.*, vol. 14, no. 4, pp. 1–13, 2017.
[36] O. T. Tursun et al., "Luminance-contrast-aware foveated rendering," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–14, 2019.
[37] F.-C. Huang, K. Chen, and G. Wetzstein, "The light field stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues," *ACM Trans. Graph*, vol. 34, no. 4, Jul. 2015, Art. no. 60, doi: 10.1145/2766922.
[38] Q. Sun, F.-C. Huang, J. Kim, L.-Y. Wei, D. Luebke, and A. Kaufman, "Perceptually-guided foveation for light field displays," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–13, 2017.
[39] A. Patney et al., "Perceptually-based foveated virtual reality," in *Proc. ACM SIGGRAPH 2016 Emerg. Technol.*, 2016, pp. 1–2.
[40] X. Meng, R. Du, J. F. JaJa, and A. Varshney, "3D-kernel foveated rendering for light fields," *IEEE Trans. Visualization Comput. Graph.*, vol. 27, no. 8, pp. 3350–3360, Aug. 2021.
[41] M. Sheehan, J. Tachella, and M. Davies, "A sketching framework for reduced data transfer in photon counting LiDAR," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 989–1004, 2021, doi: 10.1109/TCI.2021.3113495.
[42] T. Zhang et al., "First arrival differential LiDAR," in *Proc. 2022 IEEE Int. Conf. Comput. Photogr.*, 2022, pp. 1–12.
[43] M. White, S. Ghajari, T. Zhang, A. Dave, A. Veeraraghavan, and A. Molnar, "A differential SPAD array architecture in 0.18 um CMOS for HDR imaging," in *Proc. 2022 IEEE Int. Symp. Circuits Syst.*, 2022, pp. 292–296.
[44] A. Tontini, S. Mazzucchi, R. Passerone, N. Broseghini, and L. Gasparini, "Histogram-less LiDAR through SPAD response linearization," *IEEE Sensors J.*, vol. 24, no. 4, pp. 4656–4669, Feb. 2024.
[45] Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. Peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution spad camera," *ACM Trans. Graph.*, vol. 39, no. 2, Mar. 2020, doi: 10.1145/3372261.
[46] A. Gupta, A. Ingle, and M. Gupta, "Asynchronous single-photon 3D imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7909–7918.
[47] A. Gupta, A. Ingle, A. Velten, and M. Gupta, "Photon-flooded single-photon 3D cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6770–6779.
[48] F. Gutierrez-Barragan, H. Chen, M. Gupta, A. Velten, and J. Gu, "iToF2dToF: A robust and flexible representation for data-driven time-of-flight imaging," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 1205–1214, 2021.
[49] P. K. N. Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.
[50] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[51] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," 2023, *arXiv:2302.12288*.
[52] J. Lee, A. Ingle, J. V. Chacko, K. W. Eliceiri, and M. Gupta, "Caspi: Collaborative photon processing for active single-photon imaging," *Nat. Commun.*, vol. 14, no. 1, 2023, Art. no. 3158.
[53] M. Beer et al., "Spad-based flash LiDAR sensor with high ambient light rejection for automotive applications," *Proc. SPIE*, vol. 10540, pp. 320–327, 2018.
[54] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. robot Learn.*, 2017, pp. 1–16.
[55] S. Burri, C. Bruschini, and E. Charbon, "LinoSPAD: A compact linear SPAD camera system with 64 FPGA-based TDC modules for versatile 50ps resolution time-resolved imaging," *Instruments*, vol. 1, no. 1, 2017, Art. no. 6.
[56] R. Achanta and S. Susstrunk, "Superpixels and polygons using simple non-iterative clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4895–4904.
[57] A. Ardelean, "Computational imaging SPAD cameras," Ph.D. dissertation, Swiss Federal Inst. Technol. Lausanne, Lausanne, Switzerland, 2023.
[58] V. Sundar, A. Ardelean, T. Swedish, C. Bruschini, E. Charbon, and M. Gupta, "Sodacam: Software-defined cameras via single-photon imaging," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8165–8176.
[59] C. Callenberg, Z. Shi, F. Heide, and M. B. Hullin, "Low-cost SPAD sensing for non-line-of-sight tracking, material classification and depth imaging," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–12, 2021.