Single-Photon 3D Imaging with Equi-Depth Photon Histograms

Kaustubh Sadekar, David Maier, and Atul Ingle

Portland State University, Portland OR 97201, USA
{ksadekar, maier, ingle2}@pdx.edu

Abstract. Single-photon cameras present a promising avenue for high-resolution 3D imaging. They have ultra-high sensitivity—down to individual photons—and can record photon arrival times with extremely high (sub-nanosecond) resolution. Single-photon 3D cameras estimate the round-trip time of a laser pulse by forming equi-width (EW) histograms of detected photon timestamps. Acquiring and transferring such EW histograms requires high bandwidth and in-pixel memory, making SPCs less attractive in resource-constrained settings such as mobile devices and AR/VR headsets. In this work we propose a 3D sensing technique based on equi-depth (ED) histograms. ED histograms compress timestamp data more efficiently than EW histograms, reducing the bandwidth requirement. Moreover, to reduce the in-pixel memory requirement, we propose a lightweight algorithm to estimate ED histograms in an online fashion without explicitly storing the photon timestamps. This algorithm is amenable to future in-pixel implementations. We propose algorithms that process ED histograms to perform 3D computer-vision tasks of estimating scene distance maps and performing visual odometry under challenging conditions such as high ambient light. Our work paves the way towards lower bandwidth and reduced in-pixel memory requirements for SPCs, making them attractive for resource-constrained 3D vision applications.

1 Introduction

The demand for high-resolution, low-cost 3D sensing is growing for a wide range of applications, from autonomous robots to augmented reality, machine vision, surveillance, and industrial inspection. Thanks to their extreme sensitivity, high spatial resolution (exceeding 100's of kilo-pixels [1, 16]) and increasing commercial availability, single-photon cameras (SPCs) based on single-photon avalanche diode (SPAD) technology are increasingly popular for dense 3D-sensing LiDARs and on mobile devices [24, 28]. SPAD-based SPCs combine the extreme sensitivity of SPADs and the time-of-flight principle to capture scene distance maps with sub-centimeter resolution.

Each SPC pixel measures the round-trip time taken for a laser pulse to travel from the camera to the scene and back. A conventional SPC pixel must repeatedly sample hundreds-to-thousands of photon timestamps to reconstruct the

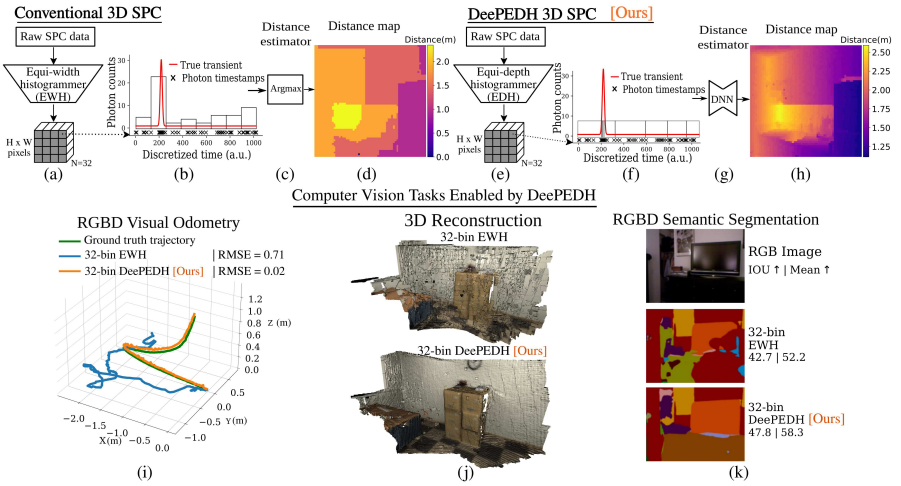


Fig. 1: Our proposed DeePEDH pipeline enables SPCs to be used in resource-constrained applications. (a–c) Conventional 3D sensing pipeline constructs equi-width (EW) histograms that require high in-pixel memory and cause a data bottleneck. (d) Existing methods resort to low resolution EW histograms at the cost of poor distance resolution. (e–g) DeePEDH uses more efficient on-sensor compression scheme through equi-depth (ED) histograms combined with a deep neural network distance map estimator. (h) DeePEDH provides accurate high resolution distance maps with $10-100\times$ lower bandwidth. (i–k) Various downstream vision tasks can benefit from high quality distance maps generated using DeePEDH.

shape of the true laser peak (which we call the *transient distribution*) to estimate the round-trip time-of-flight. The raw data captured by an SPC can be thought of as a stream of photon timestamps at each pixel location, generating a spatio-temporal "photon data cube" [18] shown in Fig. 1(a). Each SPAD pixel can detect millions of photon timestamps per second. Due to hardware constraints, it is impossible to practically process this raw photon data in the sensor or transfer it to any off-sensor compute module.

Conventional SPC pixels summarize the timestamp data by creating equi-width (EW) histograms that keep track of the number of photons received over equally spaced time bins, spanning the entire distance range. The peak of this EW histogram serves as an estimate of the time-of-flight and hence the scene distance. In practice, these EW histograms require ~ 1000 bins per pixel. For recent SPAD-based SPCs that have approached megapixel resolutions, the amount of EW histogram data adds up to several gigabytes per second when operating at video rates. For SPCs to become mainstream, especially for resource-constrained applications, such as low-power mobile robots and smartphone cameras, it is desirable to reduce the volume of the photon data cube without compromising the spatial or temporal resolution of the final distance maps.

In this work, we propose a hardware-compatible method that can compress the timestamp data to 32 or even fewer bins per pixel as opposed to the 1000's of bins needed in conventional EW histogram-based processing techniques. Our method can still retain useful information for reconstructing high-fidelity distance estimates. Our work is inspired by recent work on constructing equi-depth (ED) histograms instead of the conventional EW histograms [15]. We estimate the bin-boundary locations of these ED histograms in a count-free fashion, without explicitly storing the photon timestamps. Through publicly available datasets, we demonstrate that these ED bin-boundary representations can support high-quality distance maps for a variety of 3D scenes and under challenging illumination conditions. Our method achieves significant in-pixel memory reduction and lower bandwidth requirements relative to conventional EW histogram-based SPCs. Our contributions in this paper are threefold: (i) We propose a hardware-friendly, in-pixel algorithm that estimates ED histogram bin boundaries for arbitrary q-quantiles without having to store the entire history of photon timestamps. (ii) We design a learning-based distance-estimation technique that leverages spatio-temporal correlations in the ED boundary locations to generate high-quality distance maps. (iii) We present three case studies—RGBD visual odometry, 3D scene reconstruction, and RGBD semantic segmentation—to demonstrate the value of using ED bin boundaries as a resource-efficient scene representation for various downstream computer vision tasks.

2 Related Work

3D imaging with SPCs: Most of the current methods using SPCs for 3D imaging explicitly construct an EW histogram based on photon-arrival timestamps [13], and use the constructed EW histogram for estimating scene distance [7,8,29, 34]. Such EWH-based SPCs face a trade-off between the accuracy of distance estimates per pixel and the number of SPC pixels, due to practical bandwidth limits. To improve spatial resolution without compromising on accuracy, researchers have proposed more efficient measurement and compression techniques inspired by concepts such as Fourier-domain compression [12], compressed-sensing with random projections [4], estimating low-dimensional parametric models [40], or sketching [33] to track different statistics of the SPC data. In contrast, our approach captures ED histogram bin boundaries using a lightweight online algorithm that does not store the full history of photon timestamps, making it amenable to future in-pixel implementations.

Efficient SPC hardware designs: Improvements in SPC hardware designs have also been proposed considering the practical limitations. While some propose to measure the EW histogram in a coarse-to-fine scheme by iteratively zooming into the region-of-interest [17, 21, 27, 41], others have proposed designs that can share the resources on the sensor [14]. Although not discussed in this work, our ED histogram-based method can further benefit from recent hardware-friendly designs that use adaptive zooming to reject background light. While most of the above-mentioned SPC designs require time-to-digital convert-

ers (TDCs) to measure the photon-arrival timestamps, new TDC-less designs have been proposed that process a stream of photons using in-pixel neural networks [19, 23]. Tontini et al. [39] proposed a novel 3D-imaging method using SPCs that does not require constructing a histogram of timestamps. However, it fails in the presence of strong multipath due to the single-peak assumption.

Memory-efficient photon data processing: The idea of compressing photon timestamps on the fly using different coding techniques has been shown in simulations to provide $> 10\times$ reduction in bandwidth [12, 38]. However, to compress the photon timestamps on the fly, additional information such as a coding matrix [12] needs to be stored on the imaging sensor such that it can be simultaneously accessed by all the pixels, which is challenging to implement using existing pixel designs. In contrast, our technique constructs ED histograms of photon timestamps in an online fashion without explicitly storing them. We use a similar binner element (circuit) as Ingle and Maier [15] to iteratively update the ED histogram, but our optimized update strategy for the binner results in faster convergence, lower uncertainties, and improved distance estimates. Additionally, we also train a deep neural network (DNN) to improve the distance estimates further for challenging scenarios such as high ambient light which benefits downstream computer vision tasks.

3 Image Formation Model and Equi-Depth Histogram

We summarize the image-formation model for individual pixels of a SPAD-based 3D camera that we use for generating simulated SPAD measurements from existing RGBD datasets. We assume that each SPAD pixel is optically co-aligned and operated in synchronization with a pulsed laser source that emits a periodic train of Gaussian-shaped light pulses. In addition to these *signal* photons reflected from each scene point, the SPAD pixel also records *background* photons which consist of spurious events due to ambient light (e.g., sunlight) and other sources of electronic noise (e.g., dark counts).

In a low-photon-flux scenario and absence of multipath reflections, the average number of photons measured by each pixel is given by a continuous inhomogeneous Poisson process time-varying intensity $\Phi(t)$ and a period equal to the laser repetition period. (See supplement.) We discretize the time axis into B locations. In the k^{th} time interval ($1 \leq k \leq B$), $\Phi[k] = \Phi_{\text{sig}}[k] + \Phi_{\text{bkg}}$, where $\Phi_{\text{sig}}[k]$ represents the average number of signal photons for time bin k and Φ_{bkg} represents the average number of background photons per time bin [12, 20, 25, 37, 44]. We call $\Phi[k]$ the *transient distribution*. Assuming the peak of the transient distribution is located at a time delay t with respect to the start of the laser cycle, the scene distance is given by $z = ct/2$ where c is the speed of light. In an ideal scenario of no background light and a narrow laser peak that occupies a single discrete time interval, the time of arrival of a single laser photon would be sufficient to estimate scene distance. In practice, due the Gaussian peak shape and the presence of sensor noise and ambient light, an SPC pixel must sample the true transient over many laser cycles to estimate distance.

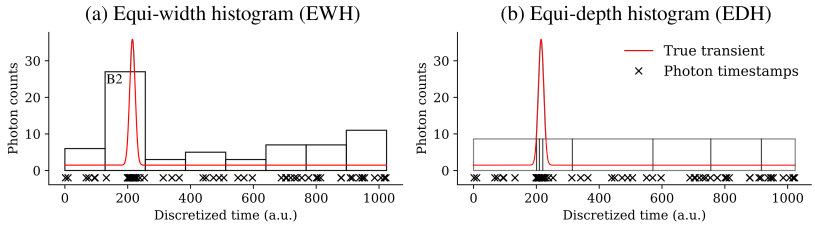


Fig. 2: Equi-width vs. equi-depth histograms for peaky data. (a) Most bins of an 8-bin EW histogram are wasted on background photons; a single bin B2 gives a coarse estimate of the true peak. (b) An ED histogram captures this "peaky" transient distribution better with a cluster of narrower bins around the true peak.

Conventional equi-width (EW) histograms: Most SPAD-based SPC pixels today construct EW histograms of photon counts; the location of the peak of this histogram is used to estimate the scene distance. The number of photons detected in each EW-histogram time bin is stochastic and is governed by a Poisson process. The number of photons in the k^{th} histogram bin over N laser cycles is given by $\widehat{\Phi}[k] \sim \text{Poisson}(N\Phi[n])$. The volume of this histogram data for a 1 megapixel SPC operating at 30 fps exceeds several gigabytes per second—much higher than can be reasonably handled by today's data-transfer buses. Today's SPCs resort to using fewer pixels and fewer numbers of EW histogram bins to reduce this bandwidth requirement, which lead to poor distance-map estimates that have low spatial resolution and suffer from quantization artifacts.

Bandwidth-efficient equi-depth (ED) histograms: Instead of storing photon counts as an EW histogram with equal sized bins, an ED histogram uses variable-width bins such that each bin contains (approximately) equal photon counts. The term "depth" in equi-depth refers to the photon counts in each histogram bin and should not be confused with scene distance. Fig. 2 shows an example (synthetic) transient distribution consisting of a narrow Gaussian peak. An 8-bin EW histogram is resource inefficient as seven out of these eight bins are spent on capturing photons from background light. Moreover, the laser photons all arrive in a single bin $B2$, resulting in an extremely coarse estimate of the scene distance. In contrast, three of the eight bins in the 8-bin ED histogram cluster around the true peak location.

Recently ED histograms were proposed as a more resource-efficient alternative to EW histograms for capturing "peaky" transient distributions such as that of a laser pulse [15]. ED-histogram bin boundaries can be estimated using an equi-depth histogrammer (EDH) built using a recursive tree of median-finding *binner circuits*. A 4-stage tree-based EDH, for example, tracks 15 ED boundaries that constitute 16 quantiles. A binner circuit maintains an estimate of the median in the form of a *control value (CV)* and iteratively updates it at every laser cycle. The general update step for the binner in n^{th} laser cycle is given by $C_{n+1} = C_n + S_n$ where C_n and S_n denote the CV and the step size

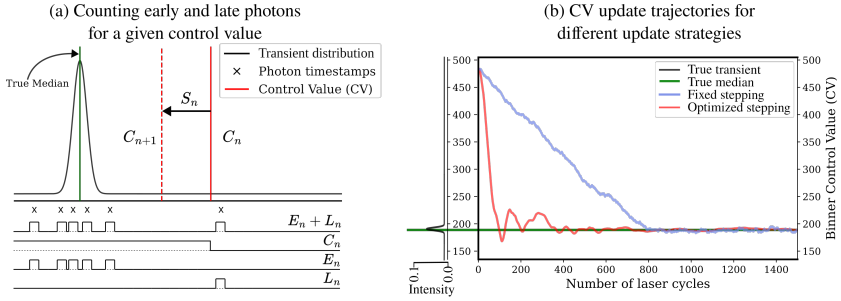


Fig. 3: A binner element used for tracking ED histogram boundaries. (a) A median-finding binner splits incident photons into early (E_n) and late photons (L_n) compared to the current CV (C_n). As the CV is greater than the true median, $E_n/(E_n+L_n) > 1/2$, hence the step size S_n is negative. (b) Our optimized stepping strategy uses a sequence of step sizes to achieve faster convergence and lower variance compared to earlier fixed-stepping binner design [15].

for the n^{th} laser cycle. The step size depends on the relative numbers of *early* (E_n) and *late* (L_n) photons, i.e., photons that arrive earlier or later than the current CV. The step size is negative when more photons are seen earlier than later, as seen in the example in Fig. 3(a). Ingle and Maier [15] proposed different stepping strategies and the possibility of binner tracking arbitrary q-quantiles. However, several questions remained unexamined in their work and their experimental results were limited to a combination of median-tracking binners that used fixed-stepping. We build upon the ideas presented in [15] and propose an optimized binner design called the *proportional binner* that tracks arbitrary q-quantiles of the transient distribution with optimized stepping strategies that speed up convergence and reduce variance.

Proportional binner: The stepping strategy of the proportional binner relies on the following key observation: For the CV to track the j^{th} quantile (out of q), the CV must be updated such that a fraction j/q of photons arrive earlier than the CV and the remaining $1 - j/q$ arrive later than the CV. If the binner had access to the entire history of photon arrival times (say, in the form of a fine-grain EW histogram), the j^{th} quantile can be estimated simply by locating the value where the cumulative distribution function crosses j/q. We call a bank of such (hypothetical) binners the *oracle ED histogrammer (OEDH)* and note that in practice, the binner circuit operates in a resource-constrained hardware setting which precludes the possibility of storing the full histogram.

Our proposed stepping strategy for the proportional binner involves iteratively updating its CV at every laser cycle such that the fraction of photons arriving before the CV approaches j/q. At the n^{th} laser cycle the control value is updated by a step size proportional to $\Delta_n = {}^{j}/_{q} - {}^{E_n}/_{(E_n+L_n)}$ where E_n and L_n denote the number of photons arriving earlier and later than the current CV. Although this method does not guarantee deterministic convergence to the j^{th} quantile, the stepping strategy is self-correcting — the larger the magnitude of

Δ_n, the larger is the step in the direction that minimizes the discrepency from the j^{th} quantile. This straightforward stepping strategy of updating CV by Δ_n suffers from two drawbacks in practice due the presence of strong Poisson noise: (i) The CV progresses towards the true quantile location, but continues to wander around that true quantile location. (ii) Large jumps in the "wrong" direction due to background photons cause a large variance in the estimated quantile location. We propose an optimized proportional-stepping strategy that mitigates these drawbacks. First, we apply a *temporal-decay* parameter that gradually reduces the maximum step size and allows the final CV to settle at or close to the true quantile value. Second, we apply *exponential smoothing* to reduce the variance in the step sizes used for adjusting CV. The optimized step size is computed using the following equations:

$$\begin{aligned} S_n &= \beta_2 S_{n-1} + (1-\beta_2)\gamma^n \widetilde{\Delta}_n \\ \text{where } \widetilde{\Delta}_n &= \beta_1 \widetilde{\Delta}_{n-1} + (1-\beta_1)\Delta_n. \end{aligned} \tag{1}$$

Here $0 < \gamma < 1$ is the temporal decay parameter, $\widetilde{\Delta}_n$ is an exponentially smoothed version of the step size Δ_n, and β_1 and β_2 are additional tuning parameters that control the level of exponential smoothing applied. We determine values for these tuning parameters empirically by running extensive simulations of the proportional binner over a wide range of transient distributions. We find that $\beta_1 = 0.95$, $\beta_2 = 0.8$, and $\gamma = 0.99902$ work well for a wide range of transient distributions over various combinations of signal and background strengths. (See supplement for details.) We simulate median-tracking binners with fixed and optimized stepping strategies for a wide range of scene distances and illumination conditions. Fig. 3(b) shows CV trajectories for a single run. Experimental results demonstrate that our optimized stepping binner provides faster convergence over a wide range of illumination conditions than the fixed-stepping binner used by Ingle and Maier [15]. (See supplement.)

A single quantile is often not sufficient to reliably track the peak of a transient distribution. We need a bank of binners to track several different quantiles of the underlying distribution. Ingle and Maier [15] use a *hierarchical equi-depth histogrammer (HEDH)*, which consists of a recursive tree of median-tracking binners with fixed-stepping to track different quantiles. However, in HEDH the binners must run sequentially for each level in the tree; convergence of lower-level binners depends on higher-level binners.

We propose to use a combination of our optimized proportional-stepping binners and configure each one to track a specific quantile of the underlying distribution. We call this design the *proportional equi-depth histogrammer (PEDH)*. The ability of our optimized proportional-stepping binner to track arbitrary quantiles enables the PEDH to run all the binners in parallel and use the complete exposure time. Our optimized parameter combination allows the PEDH to track different quantiles with faster convergence and lower variance as compared to HEDH. The binner-trajectory plots of 8-bin HEDH and 8-bin PEDH (Fig. 3(b)) illustrate improved accuracy and increased robustness to changing illumination

conditions when using PEDH over HEDH. (See supplement for detailed comparison results.)

4 Distance Estimation from Equi-Depth Histograms

This section considers the low-level computer-vision task of reconstructing scene-distance maps in a resource-constrained setting. We first begin with some intuition for how ED-histogram bin boundaries (obtained from a PEDH) can be used to estimate scene distances on a per-pixel basis. Next, we propose a data-driven approach, called *DeePEDH*, which exploits correlations in ED histogram bins over pixel neighborhoods to further improve the distance-map estimates.

4.1 Single-Pixel Distance Estimators

The ED histograms bin boundaries inherently capture the density of the underlying transient distribution $\Phi[n]$. Clusters of narrowly spaced ED histogram bin boundaries correspond to a higher concentration of photon arrivals (such as around the true laser peak location), whereas sparser, widely spaced ED-histogram bin boundaries correspond to a lower concentration of photon arrivals, such as those due to ambient illumination. Since the ED histogram bin widths are inversely proportional to the local values of the underlying transient distribution, we use the reciprocal of these bin widths as an estimate (up to an unknown scaling factor) of the true underlying transient distribution. We call these the *local photon density* estimates. We propose two different distance estimators based on piecewise-constant and piecewise-linear interpolation-based local-photon-density estimates.

For an ED histogram that tracks arbitrary q-quantiles, with bin boundary locations given by $\{t_i\}_{i=0}^{q}$, we define the piecewise constant local photon density estimator as $\rho_0(t) = {}^{1}/_{(t_j - t_{j-1})}$ for $t_{j-1} \leq t < t_j$, $1 \leq j \leq q$ and $0 \leq t \leq B$. The extreme ED bin boundaries are, by definition, located at $t_0 = 0$ and $t_q = B$. Assuming that the transient distribution contains a single sharp peak, we use the midpoint of the narrowest piece of this piecewise constant function as an estimate of the round-trip time-of-flight t_0. Ties are broken arbitrarily. Let $j^* = \arg\max_{1 \leq j \leq q} \frac{1}{t_j - t_{j-1}}$. We define the estimate $\widehat{t_0} = {}^{(t_{j^*} - t_{j^*-1})}/_{2}$.

The narrowest-bin-midpoint estimator is computationally lightweight and has the advantage of being amenable to in-pixel implementation. However, it suffers from a bias away from the true peak location if, for instance, the narrowest and the second narrowest ED bins split the peak, or several narrow bins of equal width cluster around the peak (as seen in the example in Fig. 4(a)). Therefore we propose a linearly interpolated local photon density estimator $\rho_1(t)$ which is obtained by taking the sequence of non-uniformly spaced pairs of points $\left\{\left({}^{(t_j - t_{j-1})}/_{2}, {}^{1}/_{(t_j - t_{j-1})}\right\}_{j=1}^{q}\right.$ interpolated on a grid of 1024 uniformly spaced discrete time locations between 0 and B. (See supplement.) We define an alternative time-of-flight estimate as $\widehat{t_1} = \frac{1}{2} \arg\max_{t \in [0,B]} \rho_1(t)$. Fig. 4(a) shows an example

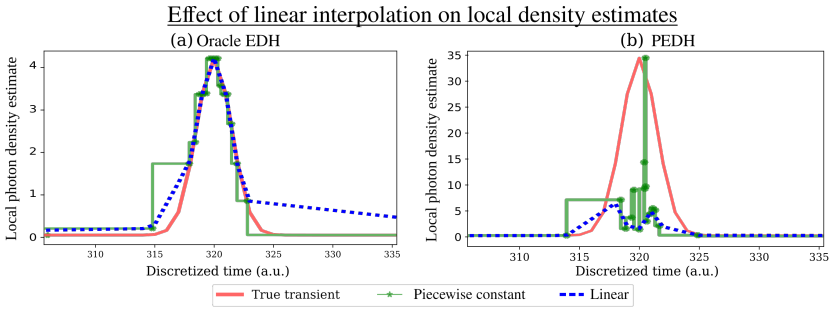


Fig. 4: Examples of interpolated local density estimates obtained using an oracle EDH and a PEDH. (a) The narrowest bin and linear interpolation-based estimators perform equally well for this example transient when using the oracle EDH. (b) Linear interpolation-based estimate is often worse than the narrowest-bin estimator due to noise in PEDH bin boundary locations.

of the interpolated photon density estimate $\rho_1(t)$ where the peak lines up with the true peak location, resulting in better distance estimates.

Simulations: To evaluate the performance of the narrowest-bin-midpoint $\widehat{t_0}$ estimator and linear interpolation-based $\widehat{t_1}$ estimator, we simulate photon timestamps over $L = 5000$ laser cycles for different scenes with varying signal and background levels. We use simulated timestamps to estimate ED-histogram bin boundaries using a simulated single-pixel PEDH. Additionally, we also simulate a single-pixel oracle ED histogrammer (OEDH), which serves as a lower bound for the distance-estimation error achievable with the PEDH. We find that the performance of the $\widehat{t_0}$ estimator is almost identical when using OEDH and PEDH over a wide range of signal and background conditions, which serves as a validation for the optimized PEDH stepping strategies developed in Sec. 3. We observe that the $\widehat{t_1}$ estimator always performs better than the $\widehat{t_0}$ estimator with OEDH. Whereas with the PEDH, $\widehat{t_1}$ is slightly worse than $\widehat{t_0}$. The plots for ρ_0 and ρ_1 in Fig. 4(b) provide some intuition. The converged boundaries for the PEDH are quite close to the OEDH but are not equal due to the inherent randomness of the incident photons and the count-free design of PEDH. This error in the PEDH boundaries, although small, results in strong noise near the peak of ρ_0, which adversely affects the interpolation results, and the peak of ρ_1 shifts away from the true peak resulting in sub-optimal distance estimates, increasing the mean absolute error by $\approx$0.1 cm. See supplement for quantitative accuracy metrics.

4.2 DeePEDH Distance Estimator

The per-pixel distance estimators described in the previous section do not exploit correlations across pixel neighborhoods. Most real scenes have well defined 3D structures, where pixels that are close are likely to have similar distance values. Moreover, deriving a rule-based method to denoise the PEDH estimates can be difficult, as there is a complex correlation between the scene properties and the

Table 1: Distance estimation results comparing the average performance of EWH-based (conventional) SPCs, CSPH [12], HEDH [15] (narrowest bin), narrowest-bin PEDH distance estimator, and DeePEDH for 10 NYUv2 test set images and 8 different (signal, background) photon-level pairs.

	EWH (1024-bin)	EWH (32-bin)	CSPH [12] (32-bin)	HEDH [15] (32-bin)	PEDH (32-bin)	DeePEDH [Ours] (32-bin)
RMSE (cm) ↓	6.89	29.80	15.01	219.91	18.05	10.92
MAE (cm) ↓	1.06	24.58	2.03	105.9	2.40	1.84
2% Inliers ↑	99.89	10.69	99.45	81.06	97.87	98.51
10% Inliers ↑	99.90	55.86	99.62	86.27	99.71	99.81

sensor properties that determines the variance in the PEDH values. Hence, we develop a data-driven approach and train a DNN that exploits the spatial (within neighboring pixels) and temporal (within PEDH boundaries of individual pixels) correlations to generate high quality distance maps.

Instead of an end-to-end model (ED histogram boundaries to distance map prediction), we prefer to take a modular approach. We first compute the per-pixel photon density estimates ρ_1 from the measured PEDH boundaries. Next, we train a DNN to denoise ρ_1 and predict the scene distance maps. We call this deep learning-based PEDH-to-distance-map method the *DeePEDH distance estimator*. Our approach is inspired by the "PENonLocal Deep Boosting" DNN proposed by Peng et al. [25] with one key difference — instead of training the model on noisy 1024-bin EW histograms, our model is trained on 1024-dimensional linearly interpolated photon density estimates ρ_1 obtained from PEDH output. The loss function used for training the network is a combination of Kullback-Leibler (KL) divergence between the histograms and total variation (TV) distance between the distance maps, identical to the loss function used by Peng et al. [26].

5 Experiments

PEDH dataset generation. We use the image-formation model presented in Sec. 3 to generate a simulated photon timestamp dataset from existing NYUv2 and Middlebury RGBD datasets [32, 35]. We assume a Gaussian laser pulse of 100 ns repetition period (corresponding to a maximum distance range of 15 m) and full width at half maximum (FWHM) of 0.32 ns. We simulate the PEDH output with $q = 32$ ED bins on a per-pixel basis. We run each proportional binner for $L = 5000$ laser cycles with the following optimized stepping parameters: $\gamma = 0.99902$, $m_1 = 0.95$, and $m_2 = 0.8$. (See supplement for details on parameter selection.) We also generate 1024-bin EW histograms for baseline comparisons.

Training and validation. To simulate training and validation datasets, we generate PEDH output for a wide range of scenes and illumination conditions from the NYUv2 dataset [35]. The training and validation datasets are generated using the RGBD images from the NYUv2 dataset. The training dataset consists

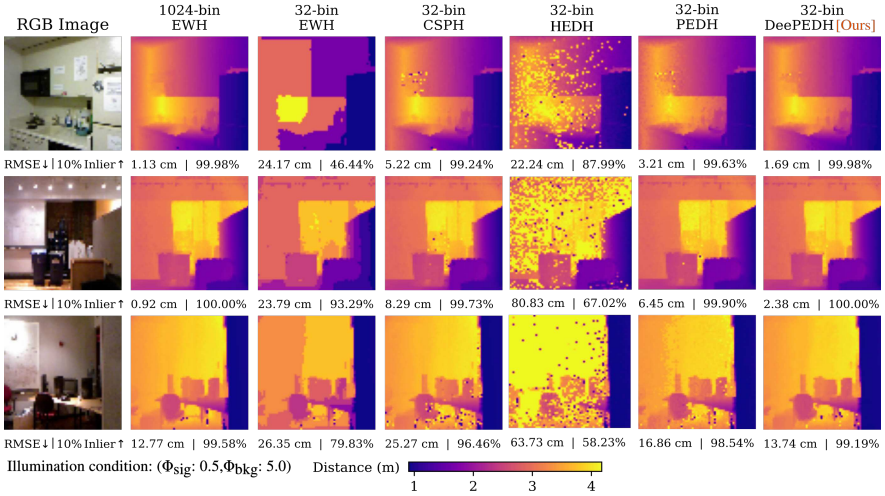


Fig. 5: Comparison of distance map reconstructions on NYUv2 test images. The CSPH [12] algorithm, HEDH [15] (narrowest bin), and the PEDH (narrowest bin) methods suffer from noisy distance maps in darker regions and farther distances. The 32-bin EWH suffers from quantization artifacts. The proposed DeePEDH method provides high spatial-and-distance resolution even in regions where other methods fail.

of 2000 RGBD images from the NYUv2 training set of the dataset and the validation dataset consists of 200 RGBD images from the NYUv2 test set. To make our DeePEDH model robust to a wide range of signal and background levels, we simulate the PEDH output by randomly choosing a pair of average signal and background photon-levels per cycle from the following pairs: $(\Phi_{\text{sig}}, \Phi_{\text{bkg}}) \in \{(1.0, 1.0), (1.0, 2.0), (1.0, 5.0), (0.5, 0.5), (0.5, 1.0), (0.5, 2.5)\}$. Since we do not make any architectural changes to the "PENonLocal Deep Boosting" model [25], the chosen validation dataset is only used to determine training convergence.

Performance evaluation. We evaluate the performance of the narrowest-bin distance estimator $\widehat{t}_0$ presented in Sec. 4.1 and the DeePEDH distance estimator and compare the results with four baseline methods — peak of the full resolution (conventional) 1024-bin EWH, peak of a coarsely binned (conventional) 32-bin EW histogram, distance estimated from the compressive single-photon histogram (CSPH) method [12], and 32-bin HEDH [15] (narrowest bin). We evaluate the performance on two different test sets. The first test set consists of 10 images from the Middlebury dataset [32] and the second test set consists of 10 images from the test set of the NYUv2 dataset [35]. For each test image, we test the distance estimators for the following eight (signal, background) photon-level pairs: $(\Phi_{\text{sig}}, \Phi_{\text{bkg}}) \in \{(1.0, 1.0), (1.0, 2.0), (1.0, 5.0), (1.0, 10.0), (0.5, 0.5), (0.5, 1.0), (0.5, 2.5), (0.5, 5.0)\}$. The performance is compared based on four different evaluation metrics: RMSE, MAE, 2% inlier metric, and 10% inlier metric. Fig. 5 shows qualitative and quantitative results for three images from NYUv2 test set [35]. Observe the distance images from HEDH [15] are grainy and contain

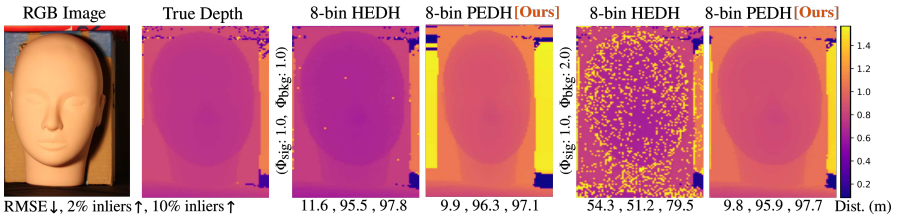


Fig. 6: PEDH improves distance estimates over HEDH. Experimental results with real-world data show improved distance estimates (RMSE cm, 2% inliers, 20% inliers) with our proposed PEDH compared to the HEDH of Ingle & Maier [15].

large errors for darker objects at farther distances. The distance images from the narrowest-bin PEDH estimator are less grainy, demonstrating the robustness of PEDH over HEDH to challenging illumination conditions. DeePEDH reconstructs high quality distance maps even for darker regions in the scenes at farther distances and does not suffer from quantization artifacts. Table 1 shows quantitative results for 10 images of NYUv2 test set averaged over the 8 different (signal, background) pairs. These results highlight that the DeePEDH can exploit spatio-temporal correlations present in PEDH boundary estimates, to generalize well on such challenging scenarios and produce high quality distance maps. Additional results for both NYUv2 and Middlebury datasets are shown in the supplement.

Hardware emulation results. We conducted hardware emulation experiments on a publicly available hardware dataset by Gupta et al. [7] containing real-world noise sources (dark counts, afterpulsing, dead-time). This data was captured with a single-pixel SPAD LiDAR setup. We emulated (8-bin) HEDH and PEDH for 5000 laser cycles for 5 illumination conditions. Fig. 6 shows results for one of the scenes. (See supplement for additional comparisons.)

6 Equi-Depth Histograms for Computer Vision Tasks

Do PEDH-based SPCs work well for downstream computer vision tasks in resource-constrained applications? We show results for three computer vision tasks: RGBD visual odometry (VO), 3D reconstruction using truncated signed distance function (TSDF) fusion, and RGBD semantic segmentation. These tasks are common components of resource-constrained applications such as autonomous robots, augmented reality applications for mobile devices, and virtual reality headsets. For all three applications, we simulate distance estimates using 32-bin EWH (conventional SPC) and 32-bin PEDH, so that both use the same amount of in-pixel memory. The PEDH output is processed using DeePEDH for distance estimation which is then fed to existing processing pipelines.

Application 1: RGBD-visual odometry. Visual odometry [31,36] is the task of estimating camera pose and motion from image sequences captured by the camera. In the case of RGBD cameras, the sequences consist of RGB images

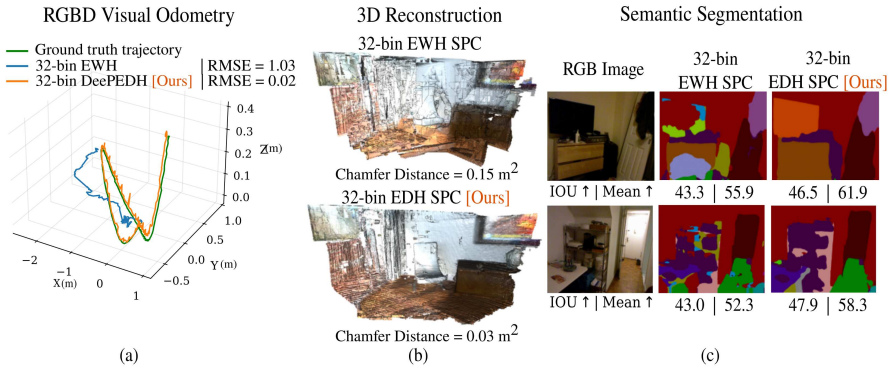


Fig. 7: DeePEDH SPCs enable better camera tracking, 3D reconstruction, and improve performance of deep RGBD semantic segmentation models. (a) The camera motion trajectory reconstructed from DeePEDH distance maps (green) closely tracks the ground truth (green) and provides $> 10\times$ lower RMSE than 32-bin EWH (blue). (b) 3D surface reconstruction obtained using a 32-bin EWH suffers from severe quantization artifacts. The proposed 32-bin DeePEDH method generates high quality 3D reconstructions both qualitatively and in terms of quantitative metrics. (c) semantic segmentation results for a pre-trained CEN network [42] using distance estimates from 32-bin EWH and 32-bin DeePEDH. As the distance maps from DeePEDH are significantly closer to the ground truth, the segmentation results are better than using distance images from 32-bin EWH.

and corresponding distance maps. The key component of a VO pipeline is the feature-matching step where the movement of interesting visual features is estimated between consecutive frames and used to compute the camera trajectory. The quality of distance estimates in each frame directly affects the feature-matching step. For resource-constrained applications there are strict bandwidth and memory limitations that affect the quality of distance measurements and thus the performance of the VO pipeline. We simulate distance maps from 32-bin EWH and 32-bin DeePEDH and pass the RGBD sequences from the Redwood dataset [3] to the existing RGBD VO pipeline using the Open3D library [46]. Fig. 7(a) shows camera trajectories obtained from the VO pipeline using distance maps from both methods. The 32-bin EWH suffers from large tracking errors due to inaccurate distance estimates that accumulate over the motion trajectory. In contrast, our method closely tracks the ground truth.

Application 2: Dense 3D reconstruction. Dense 3D reconstruction is an important computer vision task for applications such as virtual reality, digital twinning of 3D assets, photo-realistic telepresence, and autonomous mobile robots. 3D reconstruction pipelines have two major components: (i) estimating the camera pose per frame and, (ii) fusing the RGBD information to reconstruct a 3D scene that is consistent with a maximum number of RGBD frames and the corresponding camera poses. A common method to fuse the RGBD information is the TSDF fusion method [5]. We use a similar experimental setup as the VO

application and use the distance maps from EWH and DeePEDH as input to the TSDF pipeline of Open3D [46]. We pass the ground truth camera poses to the TSDF pipeline instead of the VO estimates. For quantitative evaluation, the *chamfer distance* is computed between the point cloud reconstructed from each method and the point cloud reconstructed using the ground truth distance maps for TSDF fusion. Fig. 7(b) shows qualitative and quantitative results for TSDF fusion. Observe that the EWH results suffer from strong quantization artifacts whereas the DeePEDH-based 3D reconstruction preserves finer details.

Application 3: RGBD deep semantic segmentation. In semantic segmentation, each pixel of the image is assigned a class label. Multiple deep-learning methods have been proposed in the past that perform semantic segmentation, some only use RGB image [2, 22, 30, 43, 45] whereas some use the scene distance map in addition to the RGB image [6, 9, 10, 42]. We use the Channel-Exchanging Network (CEN) proposed by Wang et.al. [42] to compare the effect of using distance maps from the 32-bin EWH SPC and the 32-bin PEDH SPC, on the task of deep RGBD semantic segmentation. We use the NYUv2 dataset as our test set. We generate CEN output using the ground-truth distance map, and distance maps obtained from the 32-bin EWH SPC, and DeePEDH. The CEN output using a ground-truth distance map is used as the ground-truth for the segmentation result to compare the output for the other two cases. Fig. 7(c) shows the output for all three cases for different scenes. We observe that segmentation results are more accurate for DeePEDH as compared to the conventional 32-bin EWH SPC.

7 Discussion

In this work, we proposed a resource-efficient method for producing a compact representation for single-photon camera outputs and studied the effect of using these representations for 3D perception in power-and data-constrained applications. We proposed different distance estimators to predict scene distances from the PEDH output and validate them in both simulation and hardware emulation. Although we did not design a hardware binner circuit, the key operations involved in computing the step size at each laser cycle involve only $\sim$3 multiply-add operations that can be performed cheaply in hardware. The initial step-size computation requires a division operation, which could be performed through a look-up table shared across groups of pixels. Our experimental results highlight that using our proposed PEDH-based SPCs can achieve superior performance over conventional EWH-based SPCs for popular computer vision tasks like distance estimation, visual odometry, dense 3D reconstruction, and RGBD semantic segmentation. Our DeePEDH results highlight the power of using deep learning to exploit spatio-temporal correlations captured in PEDH output to improve 3D perception in resource-constrained settings.

Acknowledgements

This work was supported in part by NSF ECCS-2138471.

References

1. Canon Inc.: Canon Launches MS-500 - The World's First Ultra-High-Sensitivity Interchangeable-Lens SPAD Sensor Camera. https://www.usa.canon.com/newsroom/2023/20230801-ms500, Canon Press Release 8/1/2023. Accessed 2/25/2024. 1
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2018). https://doi.org/10.1109/TPAMI.2017.2699184 14
3. Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 13
4. Colaco, A., Kirmani, A., Howland, G.A., Howell, J.C., Goyal, V.K.: Compressive depth map acquisition using a single photon-counting detector: Parametric signal processing meets sparsity. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. pp. 96–102. IEEE Computer Society (2012). https://doi.org/10.1109/CVPR.2012.6247663 3
5. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. p. 303–312. SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA (1996). https://doi.org/10.1145/237170.237269 13
6. Deng, Z., Todorovic, S., Latecki, L.J.: Semantic Segmentation of RGBD Images with Mutex Constraints. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). p. 1733–1741. ICCV '15, IEEE Computer Society, USA (2015) 14
7. Gupta, A., Ingle, A., Gupta, M.: Asynchronous Single-Photon 3D Imaging. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 7908–7917 (2019) 3, 12, 9, 10
8. Gupta, A., Ingle, A., Velten, A., Gupta, M.: Photon-Flooded Single-Photon 3D Cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
9. Gupta, S., Arbeláez, P., Malik, J.: Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 564–571 (2013). https://doi.org/10.1109/CVPR.2013.79 14
10. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 345–360. Springer International Publishing, Cham (2014) 14
11. Gutierrez-Barragan, F., Chen, H., Gupta, M., Velten, A., Gu, J.: itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. IEEE Transactions on Computational Imaging **7**, 1205–1214 (2021). https://doi.org/10.1109/TCI.2021.3126533 9, 10

12. Gutierrez-Barragan, F., Ingle, A., Seets, T., Gupta, M., Velten, A.: Compressive Single-Photon 3D Cameras. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17833–17843 (2022). https://doi.org/10.1109/CVPR52688.2022.01733 3, 4, 10, 11, 1
13. Gyongy, I., Dutton, N.A.W., Henderson, R.K.: Direct time-of-flight single-photon imaging. IEEE Transactions on Electron Devices **69**(6), 2794–2805 (2022). https://doi.org/10.1109/TED.2021.3131430 3
14. Hutchings, S.W., Johnston, N., Gyongy, I., Al Abbas, T., Dutton, N.A.W., Tyler, M., Chan, S., Leach, J., Henderson, R.K.: A Reconfigurable 3-D-Stacked SPAD Imager With In-Pixel Histogramming for Flash LIDAR or High-Speed Time-of-Flight Imaging. IEEE Journal of Solid-State Circuits **54**(11), 2947–2956 (2019). https://doi.org/10.1109/JSSC.2019.2939083 3
15. Ingle, A., Maier, D.: Count-Free Single-Photon 3D Imaging with Race Logic. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–12 (2023). https://doi.org/10.1109/TPAMI.2023.3302822 3, 4, 5, 6, 7, 10, 11, 12, 9
16. J Yoshida: Breaking Down iPad Pro 11's LiDAR Scanner. https://www.eetimes.com/breaking-down-ipad-pro-11s-lidar-scanner/, EE Times 6/5/2020. Accessed 5/6/2021. 1
17. Kim, B., Park, S., Han, S.H., Kim, S.J.: CMOS SPAD-Based LiDAR Sensors with Zoom Histogramming TDC Architectures. ITE technical report **46**(41), 77–80 (2022) 3
18. Lee, J., Ingle, A., Chacko, J.V., Eliceiri, K.W., Gupta, M.: CASPI: collaborative photon processing for active single-photon imaging. Nature Communications **14**(1), 3158 (2023) 2
19. Lin, Y., Charbon, E.: Spiking Neural Networks for Active Time-Resolved SPAD Imaging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8147–8156 (January 2024) 4
20. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-Photon 3D Imaging with Deep Sensor Fusion. ACM Trans. Graph. **37**(4) (Jul 2018). https://doi.org/10.1145/3197517.3201316 4, 1
21. Lindner, S., Zhang, C., Antolovic, I.M., Wolf, M., Charbon, E.: A 252 × 144 SPAD Pixel Flash Lidar with 1728 Dual-Clock 48.8 PS TDCs, Integrated Histogramming and 14.9-to-1 Compression in 180NM CMOS Technology. In: 2018 IEEE Symposium on VLSI Circuits. pp. 69–70 (2018) 3
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015). https://doi.org/10.1109/CVPR.2015.7298965 14
23. MacLean, J., Stewart, B., Gyongy, I.: TDC-less Direct Time-of-Flight Imaging Using Spiking Neural Networks (2024), arXiV preprint 2401.10793 (2024) 4
24. Ouster: Fully autonomous turbine inspection with Clobotics and Ouster,. https://ouster.com/blog/ (2022), [Online; accessed 2-June-2022] 1
25. Peng, J., Xiong, Z., Huang, X., Li, Z.P., Liu, D., Xu, F.: Photon-Efficient 3D Imaging with A Non-Local Neural Network. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI. pp. 225–241. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58539-6_14 4, 10, 11, 1
26. Peng, J., Xiong, Z., Tan, H., Huang, X., Li, Z.P., Xu, F.: Boosting photon-efficient image reconstruction with a unified deep neural network. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4180–4197 (2022) 10

27. Po, R., Pediredla, A., Gkioulekas, I.: Adaptive Gating for Single-Photon 3D Imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16354–16363 (June 2022) 3
28. Rangwala, S.: The iPhone 12 - LiDAR At Your Fingertips,. https://www.forbes.com/sites/sabbirrangwala/2020/11/12/ (2022), [Online; accessed 2-July-2022] 1
29. Rapp, J., Ma, Y., Dawson, R.M.A., Goyal, V.K.: High-flux single-photon lidar. Optica **8**(1), 30–39 (Jan 2021). https://doi.org/10.1364/OPTICA.403190 3
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015) 14
31. Scaramuzza, D., Fraundorfer, F.: Visual Odometry [Tutorial]. IEEE Robotics and Automation Magazine **18**(4), 80–92 (2011). https://doi.org/10.1109/MRA.2011.943233 12
32. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007). https://doi.org/10.1109/CVPR.2007.383191 10, 11
33. Sheehan, M.P., Tachella, J., Davies, M.E.: A sketching framework for reduced data transfer in photon counting lidar. IEEE Transactions on Computational Imaging **7**, 989–1004 (2021). https://doi.org/10.1109/TCI.2021.3113495 3
34. Shin, D., Xu, F., Venkatraman, D., Lussana, R., Villa, F., Zappa, F., Goyal, V.K., Wong, F.N., Shapiro, J.H.: Photon-efficient imaging with a single-photon camera. Nature communications **7**(1), 1–8 (2016). https://doi.org/10.1038/ncomms12046 3
35. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 746–760. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 10, 11
36. Steinbrücker, Frank, Sturm, Jürgen, Cremers, D.: Real-time visual odometry from dense RGB-D images. IEEE Robotics and Automation Magazine pp. 719–722 (11 2011). https://doi.org/10.1109/ICCVW.2011.6130321 12
37. Sun, Z., Lindell, D.B., Solgaard, O., Wetzstein, G.: SPADnet: Deep RGB-SPAD sensor fusion assisted by monocular depth estimation. Opt. Express **28**(10), 14948–14962 (May 2020). https://doi.org/10.1364/OE.392386 4, 1
38. Tachella, J., Sheehan, M.P., Davies, M.E.: Sketched RT3D: How to Reconstruct Billions of Photons Per Second. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1566–1570 (2022) 4
39. Tontini, A., Mazzucchi, S., Passerone, R., Broseghini, N., Gasparini, L.: Histogram-Less LiDAR Through SPAD Response Linearization. IEEE Sensors Journal **24**(4), 4656–4669 (2024). https://doi.org/10.1109/JSEN.2023.3342609 4
40. Valentin, P., William, G., David, C., Gilles, S.: A 2-Stage EM Algorithm for Online Peak Detection, an Application to TCSPC Data. IEEE Transactions on Circuits and Systems II: Express Briefs **69**(9), 3625–3629 (2022). https://doi.org/10.1109/TCSII.2022.3181687 3
41. Vornicu, I., Darie, A., Carmona-Galan, R., Rodriguez-Vazquez, A.: ToF Estimation Based on Compressed Real-Time Histogram Builder for SPAD Image Sensors. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–4 (2019) 3

42. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 4835–4845. Curran Associates, Inc. (2020) 13, 14, 15
43. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Neural Information Processing Systems (NeurIPS) (2021) 14
44. Zang, Z., Xiao, D., Li, D.D.U.: Non-fusion time-resolved depth image reconstruction using a highly efficient neural network architecture. Opt. Express **29**(13), 19278–19291 (Jun 2021). https://doi.org/10.1364/OE.425917 4, 1
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.660 14
46. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018) 13, 14