Research Article | Vol. 29, No. 13 / 21 June 2021 / Optics Express 19278

Optics EXPRESS

Non-fusion time-resolved depth image reconstruction using a highly efficient neural network architecture

ZHENYA ZANG,[1,2] DONG XIAO,[1,2] AND DAVID DAY-UEI LI[1,2,*]

[1]Strathclyde Institute of Pharmacy and Biomedical Sciences, Faculty of Science, University of Strathclyde, Glasgow, G4 0RE, UK
[2]Department of Biomedical Engineering, Faculty of Engineering, University of Strathclyde, Glasgow, G4 0NW, UK
*david.li@strath.ac.uk

Abstract: Single-photon avalanche diodes (SPAD) are powerful sensors for 3D light detection and ranging (LiDAR) in low light scenarios due to their single-photon sensitivity. However, accurately retrieving ranging information from noisy time-of-arrival (ToA) point clouds remains a challenge. This paper proposes a photon-efficient, non-fusion neural network architecture that can directly reconstruct high-fidelity depth images from ToA data without relying on other guiding images. Besides, the neural network architecture was compressed via a low-bit quantization scheme so that it is suitable to be implemented on embedded hardware platforms. The proposed quantized neural network architecture achieves superior reconstruction accuracy and fewer parameters than previously reported networks.

1. Introduction

Depth imaging has been an essential tool in various applications, such as autonomous vehicles [1], vision-guided robotic systems [2], and augmented reality applications [3]. Many strategies have been proposed to obtain depth information from captured intensity images. For example, Zhan *et al.* used machine-learning and monocular depth perception [4] principles to retrieve depth information from RGB images. However, the reconstruction fidelity deteriorates because of the scale ambiguity. The stereovision technique mimicking human vision systems [5] is also popular. It uses the triangulation principle to understand spatial information. Conventionally, stereo-based cameras are not photon-sensitive; few reflected photons are detected. Besides, the sensing accuracy of the stereovision approach deteriorates when it works in dark conditions or performs long-distance measurements, whereas LiDAR can overcome the limitations. LiDAR has become popular in ranging applications. Unlike Radar systems [6] that use radio waves to measure the time-of-flight (ToF) between transmitted and reflected signals, LiDAR systems adopt pulsed light with a much shorter wavelength to detect an object's range. Therefore, LiDAR can obtain more accurate spatial information than Radar for seeing objects at a longer distance [7].

Single-photon avalanche diodes (SPADs) [8] are effective LiDAR sensors due to their single-photon sensitivity and excellent temporal resolutions. Richardson *et al*. developed and applied low-noise SPAD sensors [9] to time-resolved imaging [10, 11]. Recent advances in silicon manufacturing have introduced more compelling high fill-factor devices [12]. Figure 1 shows a conventional SPAD-based LIDAR system, including a pulsed laser, a SPAD sensor, and a time-correlated single-photon counting (TCSPC) module. The TCSPC module includes a picosecond time-to-digital converter (TDC) to measure and time-stamp reflected photons from the target. The system also contains a histogramming module to establish a time of arrival

#425917 https://doi.org/10.1364/OE.425917
Journal © 2021 Received 23 Mar 2021; revised 21 May 2021; accepted 21 May 2021; published 7 Jun 2021

Optics EXPRESS

(ToA) profile for estimating the average depth (or the distance). For SPAD-based 3D ranging, convex-optimization [13] and Bayesian inference [14] approaches have been widely adopted to tackle the recovery problem in low photon-flux conditions. Shin *et al*. [15] modelled photon registration behaviors as the rate function [16] of a Poisson process and used the constrained maximum likelihood estimation (MLE) to reconstruct depth images. Later, to improve the reconstruction accuracy, Rapp and Goyal [17] proposed a pixel-wise spatial unmixing strategy to split the signal cluster and background noise. Tachella *et al*. used the Bayesian theory to identify surfaces from pixel-wise histograms [18], using the advanced Markov chain Monte Carlo (MCMC) sampling method to address the maximum-a-posterior (MAP) problem. Although their algorithm yielded outstanding 3D reconstructions with a fast processing speed, some hyperparameters should be adjusted to maintain accuracy and computing efficiency. Quentin *et al*. [19] have used the expectation-maximization (EM) method to estimate multi-spectral and depth profiles. It shows excellent performances in estimating mixed probabilistic models with latent variables.

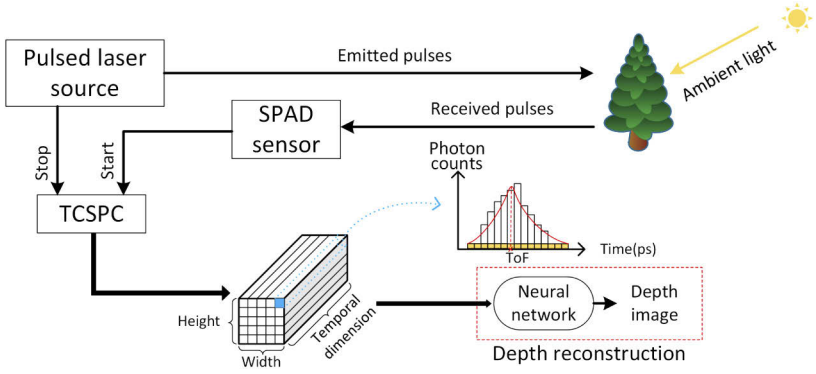

Fig. 1. Time intervals between emitted and detected photons are measured and digitalized by the TCSPC system. Detected photons are accumulated to generate a 3D tensor. Each data cube (blue) is a histogram, and neural network postprocessing is adopted to retrieve the time-bin index representing the average distance.

Deep neural networks (DNNs) feature hierarchical structure, showing powerful learning ability from complex data. Using DNN with sensor-fusion strategies has been a new trend for extracting ToA depth maps. Lindell *et al*. [20] first introduced a U-net neural network [21] merging 2D intensity images and 3D SPAD tensor data to reconstruct depth images. To make the hardware platform more compact, a dedicated neural network was proposed [22] for a SPAD array that can simultaneously generate intensity images and ToA data. Another sensor fusion architecture [23] uses monocular depth perception principles to retrieve a coarse-grain depth map in advance. The corresponding ToA data is fused with the up-projected monocular depth map. This fusion strategy achieved more accurate results than the SPAD-intensity version. A non-local neural network model [24] was introduced to explore the long-range correlation along spatial and temporal dimensions. However, the networks mentioned above have large model sizes and redundant parameters. Therefore, they show a long training time and slow inference speed, not suitable for real-time scenarios. Also, it is nonviable to implement them on embedded hardware such as field-programmable gate arrays (FPGA) or application-specific integrated circuits (ASIC).

In this work, we proposed a photon-efficient 3D convolutional neural network (CNN) architecture without using sensor-fusion strategies. Unlike the U-net based models that only fuse the data with the same dimension [20, 22, 23], our proposed architecture fuses multi-dimensional spatial and temporal features to extract more information. Spatial correlations of depth images can be effectively explored from SPAD data only using short- and long-range connections and multiple

Optics EXPRESS

up-sampling processes. We also compressed the architecture using the low-bit parametric quantization strategy. It shows a good compression ratio over existing sensor-fusion neural network models while maintaining high reconstruction quality. Results show that our architecture is feasible to enhance the accuracy without auxiliary RGB and monocular depth images. Compared with previously reported algorithms, our architecture yields superior reconstruction results in low photon-flux conditions. We evaluated our compressed model over various signal-background ratio (SBR) levels in terms of extensive evaluation metrics [25] for synthetic and captured data [20] with a high spatial resolution. The result shows that this work is suitable for the single-photon LiDAR applications in low-light scenarios.

2. Problem definition

For raster-scanning and wide-field ranging systems, a laser source generates periodic pulses $s(t)$ to illuminate the target scene. The reflected photon flux can be detected by individual pixels with the quantum efficiency $\eta \in [0, 1)$. We assume that there are only a few detected photons (less than 1 photon per pixel). Therefore, pile-up effects and carriers' crosstalk [26] are negligible. The spatial resolution of the SPAD array is $(m_1, m_2) \in \{1, 2, \ldots, M\}^2$. According to previously published reports [16, 18–20], the number of recorded photons in the time interval $n \in \{1, 2, \ldots, N\}$ of each pixel can be formulated as:

$$r_{m_1,m_2}[n] = \int_{n\Delta t}^{(n+1)\Delta t} \eta(g * s)(t - \frac{2d_{m_1,m_2}}{c})\, dt + b_\lambda, \tag{1}$$

where g denotes the instrument response function (IRF), Δt is the time bin-width of the TDC, $\mathbf{d} \in \mathbb{R}_+^{M\times M}$ is the scene's depth profile, c is the speed of light, and b is the ambient light with a wavelength λ. The photon arrival behavior can be formulated as an inhomogeneous Poisson process [27]. The dark count [26] triggered by thermally-generated carriers in the SPAD sensor is also considered the time-varying factor in the Poisson process's rate function. Consequently, the histogram in one pixel with I illumination periods can be formulated by a Poisson process $(\cdot)$ with a time-varying arrival function:

$$h[n] \sim \mathcal{P}(I(\gamma r[n] + b_d)), \tag{2}$$

where the constant γ represents the attenuation factor caused by photon scattering on the surface and b_d is the dark count rate. Suppose $f(\cdot)$ is the neural network's feedforward function, and the input is a noise-corrupted tensor composed of 1D pixel-wise histograms with a 2D spatial resolution M^2. Therefore, pixel-wise denoised ToA data can be modelled as

$$\hat{h}^{(m_1,m_2)} = f(h^{(m_1,m_2)}; \theta), \tag{3}$$

where θ is the parameter set to be learned, and we can use MLE to calculate it as

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(\hat{h}^{(m_1,m_2)} | h^{(m_1,m_2)}; \theta). \tag{4}$$

Equation (4) can be solved by the neural network's learning phase to be detailed in the next section.

3. Non-fusion ToA denoising model

This section introduces the proposed network architecture, loss function, and low-bit quantization scheme. To conduct a fair comparison with peers' work, we focus on depth images retrieved in low photon-flux conditions with a low SBR.

Optics EXPRESS

3.1. Neural network architecture

As shown in Fig. 2, the U-net++ [28] was used as the network's backbone, modified to a 3D version to denoise the ToA tensor. The network consists of two parts: a main feature extraction module and a refinement module. In Fig. 2, $x^{i,j}$ represents the nodes where $i \in \{0, 1, 2, \ldots, l\}$ is the row number along the down-sampling path (blue arrows), $j \in \{0, 1, 2, \ldots, l\}$ is the number of nodes in skipped connections in each row, and l is the level or depth of the down-sampling. The refinement and a differentiable argmax function after the last up-sampled node are included. The procedure for calculating a feature map can be formulated by

$$x^{i,j} = \begin{cases} Conv(\mathcal{M}(x^{i-1,j})), & i>0, j=0 \\ Conv([[x^{i,k}]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1})]), & 0<j<l \\ \mathcal{S}(Conv(x^{i,j})), & i=0,\ j=l \end{cases}, \tag{5}$$

where $Conv\ (\cdot)$ denotes the convolution operation, $\mathcal{M}\ (\cdot)$ the max-pooling to perform down-sampling, $\mathcal{U}\ (\cdot)$ the transposed convolution to perform up-sampling, $[\cdot]$ the concatenating layer's operation $\mathcal{S}\ (\cdot)$ is the soft argmax function to find the bin index.

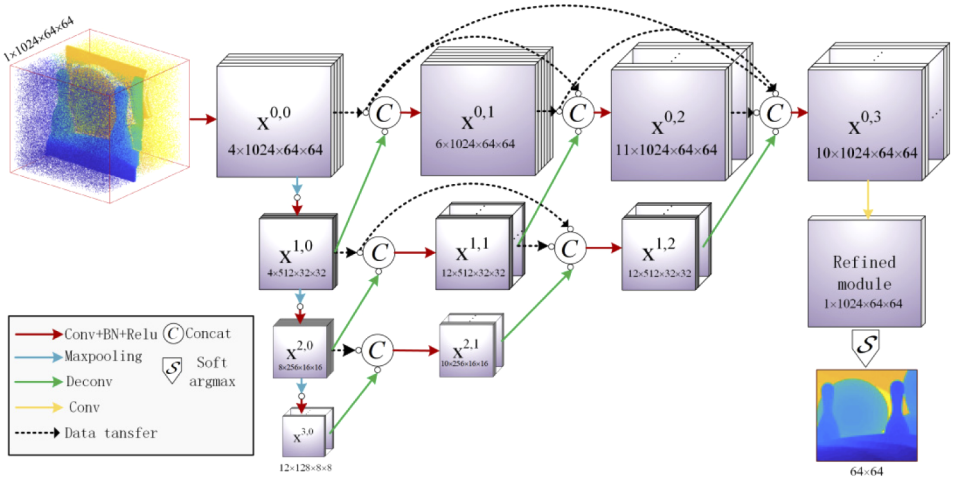

Fig. 2. The proposed model's architecture. The input is a ToA tensor corrupted by background noise. By adopting dense connections in each row, short- and long-range information can be fully explored. A more robust depth map can, therefore, be obtained.

By adding dense connections in each horizontal node, our architecture can compensate for information lost due to the down-sampling. Besides, long-range and short-range concatenations (black dot lines in Fig. 2) between horizontal convolutional layers can fuse different non-local spatial and temporal information of ToA measurements. Within a pixel, a soft-argmax function [20] is applied to find the index corresponding to the ToF (or the distance) (see Fig. 2). So the ToA tensor's noise can be censored during the learning phase, and the network generates a squeezed 2D depth map from the denoised ToA tensor. This architecture does not need any guiding images (monocular and intensity) due to the above features, thereby saving processing time and parameters than fusion approaches. The max-pooling as down-sampling operations along the down-sampling path is applied to reduce network parameters and computing time.

To make our network portable to embedded hardware, a model compression strategy was applied to simplify it. The bottleneck of embedding neural networks in reconfigurable hardware is due to limited on-chip memory for storing pre-trained parameters and the large memory

Optics EXPRESS

bandwidth for data transfer. Therefore, we generalize a 2D low-bit parametric quantization scheme [29] for 3D data quantization to compress the model. Multiplication operations of floating-point numbers can be converted to bitwise operations of fixed-point numbers (in binary). Briefly, suppose x and y are two fixed-point integers coded with M-bit and K-bit binary digits, respectively. The conversions are subjects to $x = \sum_{m=0}^{M} c_m(x)2^m$ and $y = \sum_{k=0}^{K} c_k(x)2^k$, where $(c_m(x))_{m=0}^{M-1}$ and $(c_k(x))_{k=0}^{K-1}$ are corresponding binary digits. The dot product of x and y can be therefore indicated by

$$x \cdot y = \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} 2^{m+k} bitcount[and(c_m(x), c_k(y))], \tag{6}$$

where all operations can be performed via economic bitwise operations in FPGAs. More details are provided in [29]. As for our neural network, weights (W) and activation functions' outputs (A) are quantized with different bit-widths, denoted as $WXAY$, where X and Y are bit-widths. The quantization process in each convolutional layer is depicted in Fig. 3. The weights and activations can be quantized with an arbitrary low bit-width during the forward propagation. The weights are firstly normalized via a $tanh(\cdot)$ function, and a quantization function $Q_k(\cdot)$ converts the normalized floating-point weights to a fixed-point format, defined in Eq. (7).

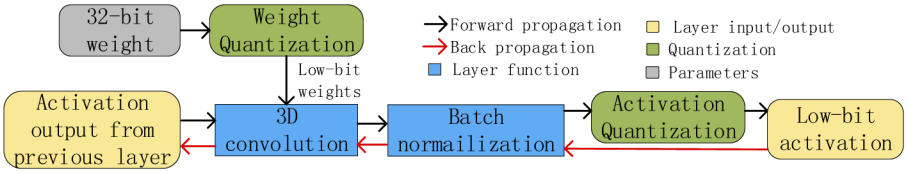

Fig. 3. Dataflow and quantization process in each convolutional layer. In the forward propagation phase, the weight and activation are quantized with a low-bit accordingly.

To avoid information loss, we keep weight parameters in the first and last layers in the floating-point format and compress internal layers' parameters. k is the bit-width of parameters' fixed-point representations, w_i and w_o are the floating-point weight and the quantized version, respectively, and a_i and a_o are floating-point activation functions' output and the corresponding quantized version. w_o and a_o are

$$\begin{cases} w_o = 2Q_k\left(\frac{tanh(w_i)}{2max(|tanh(w_i)|)} + \frac{1}{2}\right) - 1 \\ a_0 = 2Q_k(a_i) \end{cases} \tag{7}$$

where $Q_k(r)$ is defined as

$$Q_k(r) = \frac{1}{2^k - 1} round((2^k - 1)r). \tag{8}$$

The combination of Eq. (7) and (8) is a 'straight-through estimator' [29] that is also used in the back-propagation. And floating-point parameters are quantized to the k-bit fixed-point format in the forward propagation. The quantization procedure was emulated in Pytorch, and it is a prototype for future hardware implementations. The performance of the proposed quantization strategy is evaluated in Fig. 4 in the next section.

3.2. Training detail

Since solving the MLE problem in Eq. (4) is mathematically equivalent to minimizing the Kullback-Leibler (KL) divergence [30]. The KL divergence is adopted to minimize the KL

Optics EXPRESS

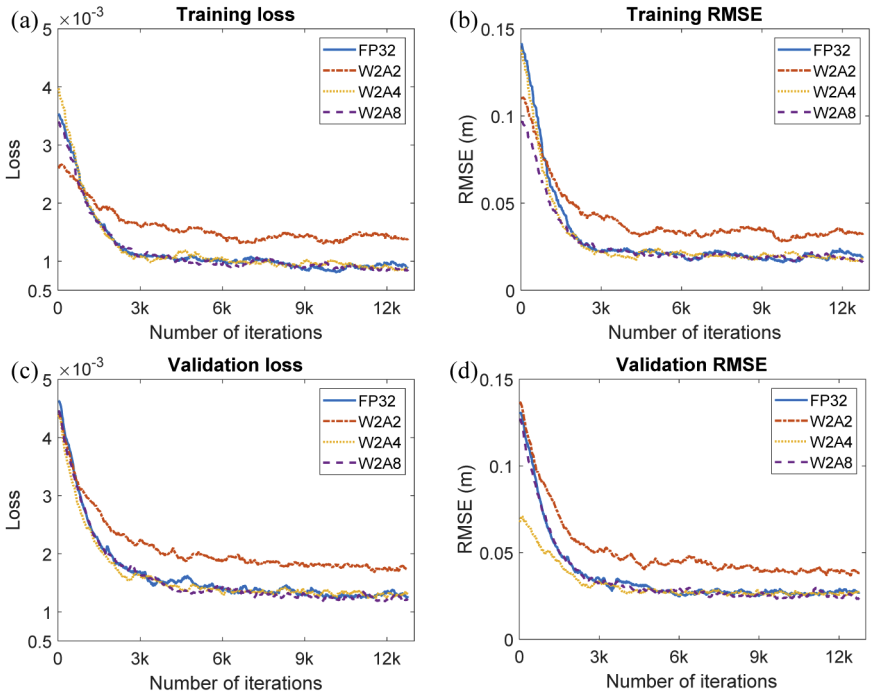

Fig. 4. (a) and (b). Training and validation loss; (c) and (d). RMSE plots with different quantization cases. Each plot contains the original floating-point 32-bit (FP32) format and three quantization cases.

distance between the ground-truth (GT) histograms and the network's output histograms over the spatial dimension:

$$\begin{aligned}\mathcal{L}(H,\hat{H},\theta) &= \frac{1}{M^2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\sum_{n\in N} D_{KL}(h^{(m_1,m_2)}[n] \,||\, \hat{h}_{\theta}^{(m_1,m_2)}[n]) \\ &= \frac{1}{M^2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\sum_{n\in N} h^{(m_1,m_2)}[n]\log\frac{h^{(m_1,m_2)}[n]}{\hat{h}^{(m_1,m_2)}[n]},\end{aligned} \tag{9}$$

where H is the histogram tensor composed of the pixel-wise GT histograms (denoted as h) and H is the network output containing the predicted pixel-wise histograms (denoted as h). To enhance the efficiency of 2D spatial denoising, the total variation (TV) regularization [31] can minimize the spatial variation. The whole loss function becomes

$$\mathcal{L}(H,\hat{H}) = \mathcal{L}_{KL}(H,\hat{H}) + \beta TV(\mathcal{S}(\hat{H})), \tag{10}$$

where β is a tunable hyperparameter before training. We use the stochastic gradient descent (SGD) to learn the parameters during the back-propagation, expressed as

$$J(\theta) = \nabla_{\theta}\mathcal{L}(H,\hat{H},\theta). \tag{11}$$

As for preparing training data, the NYUv2 [32] and Middlebury [33] datasets were utilized as the training and testing datasets, respectively. The training dataset contains nine types of indoor scenes with 13k synthetic ToA tensors, and the validation dataset contains 1.3k. The network's input is the data cube with the size of 512×512×1024, and the SGD with the ADAM

Optics EXPRESS

algorithm [34] in the Pytorch library was employed to execute the back-propagation. Loss curves of training and validation shown in Fig. 4 were generated and fetched from the Tensorboard toolkit. To achieve better visualization, we apply a smooth factor of 0.8 to alleviate fluctuations. For each convolutional module except the last, batch normalization and ReLU operations are added after the convolution operations to alleviate the vanishing gradient. We used a mini-batch to optimize the gradient descent and save memory with a batch size of 5. The hyperparameter β of the TV loss function is 10^{-5}. 4 epochs (each contains 3200 iterations) are configured to train the network, guaranteeing a converging loss. The training dataset is randomly shuffled before each training epoch to generalize our model. Our architecture's training time is 17 hours, 7 hours shorter than the existing sensor-fusion networks. For the inference phase, we aim to retrieve depth images with a high spatial resolution of 688×552. Due to limited GPU memory, the whole tensor (with the size 688×552×1024) was divided into small ones for the network's input with the size 64×64×1024. Since the loss curve of validation fluctuates significantly, it is difficult to use the early-stop method to cease the training to prevent over-fitting. Our approach is to smooth the loss curve and obtain an approximate range that contains the smallest loss. Then we apply the synthetic testing dataset on these saved models, and we pick the one generating the minimum loss or RMSE. The selected model is employed to deduce the depth images from captured SPAD data. Finally, the reconstructed result with a high spatial resolution can be generated by seaming the individual low-resolution ones consecutively. We used SPAD and intensity data captured by the LinoSPAD system [35] as real-word test datasets. The tensor's size is 256×256×1536, and the average bin-width of embedded TDCs is 26 ps. Moreover, for a fair comparison with the monocular-SPAD architecture [23], DenseDepth [36] was used to reproduce monocular depth images to be fused with ToA data in the SPADnet [23] model.

4. Evaluation

4.1. Loss evaluation

The training and validation loss of the floating-point and the quantified model are shown in Fig. 4. We implemented three quantization cases and compared them with the floating-point version. Since the training and inference are susceptible to the activation's bit-width, we tried different bit-widths to quantize the activations' output and selected the best case. As depicted in Fig. 4, only Case *W2A2* shows degraded performances. Therefore, considering the accuracy and consumption of computing resources, *W2A4* works the best. Table 1 compares the model sizes of existing sensor-fusion models and the proposed architecture, where 'Lindell w/ intensity' and 'Lindell w/o intensity' indicate the training with and without intensity images. The compression rate is determined by comparing the model size (obtained from Pytorch) with existing models. Although our network was trained by less powerful GPUs (NVIDIA RTX 5000, whereas NVIDIA Titan V and 1080Ti were used in [20], [23], [24]), it can still achieve the shortest training time. Moreover, after the quantization process, the network obtains a remarkable compression ratio compared with the original floating-point model and existing network architectures. It is suitable for hardware-embedded solutions, as it requires much less on-memory to pre-load the model. The performance of the compressed model will be detailed in the following subsection.

4.2. Synthetic data

We first evaluated the reconstruction quality for simulated data using five different metrics to assess the depth reconstruction. They are the accuracy of a given threshold *thr*, RMSE, RMSE (log), the absolute relative difference (*Abs rel*), and the squared relative difference (*Sq rel*), defined to be

$$RMSE(\mathbf{d},\hat{\mathbf{d}}) = \sqrt{\frac{1}{M^2}\sum_{m_1}^{M}\sum_{m_2}^{M}(d_{m_1,m_2}-\hat{d}_{m_1,m_2})^2}, \quad RMSE_{\log}(\mathbf{d},\hat{\mathbf{d}}) = \sqrt{\frac{1}{M^2}\sum_{m_1}^{M}\sum_{m_2}^{M}(\log_{10}(d_{m_1,m_2})-\log_{10}(\hat{d}_{m_1,m_2}))^2}$$

Optics EXPRESS

Table 1. Compression rates of our floating-point model and existing networks in terms of the parameter size, training time and compression rate.

	Parameter size	Training time	Compression rate
Lindell w/ intensity [20]	3.95M	24h	21.99×
Lindell w/o intensity [20]	3.93M	24h	21.83×
Sun *et al.* [23]	3.95M	24h	21.99×
Non-local [24]	1.01M	36h	5.61×
Proposed (floating-point 32-bit)	2.19M	17h	12.17×
Proposed (*W2A4*)	0.18M	16h	-

$$Abs\ rel(\mathbf{d},\hat{\mathbf{d}}) = \frac{1}{M^2}\sum_{m_1}^{M}\sum_{m_2}^{M}\frac{|d_{m_1,m_2}-\hat{d}_{m_1,m_2}|}{d_{m_1,m_2}}, \quad Sq\ rel(\mathbf{d},\hat{\mathbf{d}}) = \frac{1}{M^2}\sum_{m_1}^{M}\sum_{m_2}^{M}\frac{(d_{m_1,m_2}-\hat{d}_{m_1,m_2})^2}{d_{m_1,m_2}},$$

and accuracy with threshold *thr*:

$$\text{percentage of } \hat{\mathbf{d}} \text{ s.t. } \frac{1}{M^2}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}\max\left(\frac{d_{m_1,m_2}}{\hat{d}_{m_1,m_2}},\frac{\hat{d}_{m_1,m_2}}{d_{m_1,m_2}}\right) = \delta{<}thr,$$

where $\mathbf{d}$ and $\hat{\mathbf{d}}$ denote the ground truth and predicted depth images.

Seven indoor target scenes of the Middlebury dataset were assessed. Table 2 shows the averaged values across seven scenes with three SBRs. Compared with existing algorithms, our compressed architecture obtains comparable or improved accuracy to Sun *et al.* [23] using RGB-SPAD fusion strategy and Peng *et al.* [24] using long- and short-range residual connections architecture. Despite the similar performances, we should notice that Sun *et al.* adopted a log-scale binning method merging fewer time-bins for the front indexes and more time-bins for the ending indexes; thereby, the accuracy deteriorates when sensing long-distance objects. We use a simulated indoor scene (a lecture theatre) to prove this point (shown in Fig. 5) that the binning method cannot reconstruct relatively long objects highlighted in red boxes. Table 3 shows our results are better than [23] . As for Peng *et al.*'s architecture, long-range correlations of feature maps were effectively explored, and dilated convolutions were employed to enlarge the reception field. Both methods can enhance the accuracy and generate fewer parameters. However, Peng *et al.*'s method has many vectorized operations during the training and inference phases, resulting in a longer training time (36 hours) than ours (17 hours with a less powerful GPU). Moreover, owing to the complicated residual connections of [24], it is difficult to reconfigure the structure for different tasks with a proper trade-off. Instead, our architecture has good scalability where the number of down- and up-sampling can be configured without redesigning the connections. Lastly, our network maintains high reconstruction accuracy despite a much smaller model size. To better indicate the statistical significance, we further calculated the *p*-value of our compressed model and the models in [23] and [24] in test datasets in terms of *Abs rel*. We obtained a list of *Abs rel* (in total 21 elements, representing 7 scenes in 3 SBR levels) from each pair of GT and reconstructed depth images from each algorithm. The *p*-value of *Abs rel* is 2.94×10^{-5} between the model in [23] and ours (and 0.0162 between the model in [24] and ours). Both are smaller than 0.05, meaning *Abs rel* of our model is statically significant versus [23] and [24].

Since the compressed version (*W2A4*) can obtain almost the same performance as the floating-point version, we employed it to conduct further visual comparisons hereafter. As shown in Fig. 6, we chose one exemplar image named *Art* as an example. Rapp and Goyal's algorithm achieves better recovering results than the LM filter and Shin *et al.*'s approach because they used pixel-wise signal refining and spatial regularization to smooth objects' boundaries. However,

Optics EXPRESS

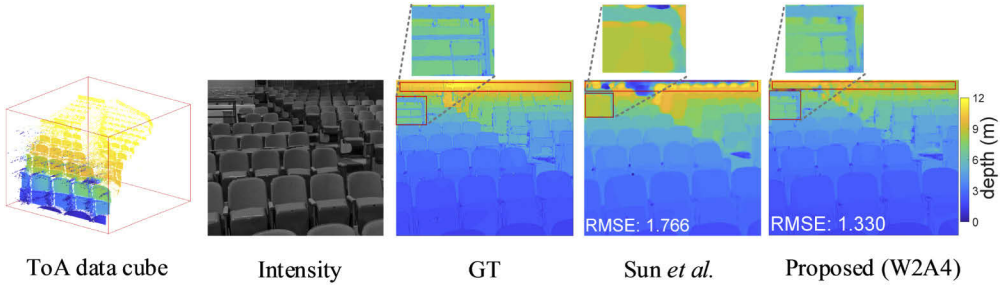

Fig. 5. Depth images reconstructed from the ToA tensor. The fidelity differences between monocular-SPAD fusion and the proposed model are marked in red boxes. The proposed model can reveal more details for long-distance objects and obtain a lower overall RMSE when SBR equals 0.04.

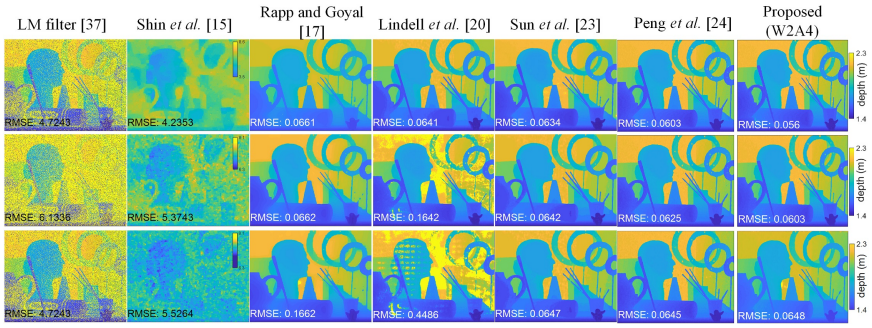

Fig. 6. Reconstructed images using different algorithms. The SBRs from the first row to the last row are 0.2 (2 target photons and 10 background photons), 0.04 (2 target photons and 50 background photons), and 0.02 (2 target photons and 100 background photons). Our model can obtain lower RMSE than other algorithms when SBR equals 0.2 and 0.04 and achieves a comparable RMSE with Sun *et al*., even without fusion strategies..

their method is not robust in distinguishing object boundaries, whereas the neural networks in the last five columns (see Fig. 6) can identify the boundaries through learning numerous training scenes. The proposed compressed model can achieve the smallest RMSE among existing methods.

4.3. Captured data

We also tested the proposed compressed model (*W2A4*) and existing networks on five real-world image datasets. These scenes were captured in a low light condition (less than 1 photon per pixel) with low SBRs. In the first row of Fig. 7(c), the reconstructed depth results from the log-matched (LM) filter [37] and Shin *et al*. [15] show the shadow that did not receive active illumination. In contrast, the other four algorithms can in-paint the shadow and generate relatively detailed depth maps. Our model can retrieve a more precise structure for the lamp scene. Although Rapp and Goyal adopted the super-pixel method to obtain robust depth estimations, some essential signals are lost, like the lamp's circle part.

Similarly, in the second and the third rows, a rolling ball and an elephant toy are presented, as highlighted in red boxes. The human's thumb and the ivory with distinct boundaries can be identified in our model's depth images. However, these two small objects are over-smoothed by the monocular-SPAD fusion owing to the monocular perception. In the fourth row, the depth images of a bouncing ball from stairs are compared. The intense sunlight illumination at the top

Optics EXPRESS

Table 2. Quantitative analysis of the proposed and existing algorithms over seven indoor scenes. Three tables are for three different SBR levels - 0.2, 0.04, and 0.02. And the underlined numbers mean that they are comparable to the best existing results.

	Accuracy (Higher is better)			Error (Lower is better)			
	δ<1.25	δ<1.25^2	δ<1.25^3	RMSE(log)	RMSE (m)	Abs rel	Sq rel
Signal photons: 2; Background Photons: 10; SBR: **0.2**							
LM filter	0.5081	0.6115	0.6415	1.183	4.690	1.403	11.606
Shin *et al.*	0	0	0.0052	1.180	4.306	2.259	9.566
Rapp and Goyal	0.8691	0.9982	0.9998	0.102	0.063	0.0778	0.0310
Lindell *et al.*	0.9962	0.9982	0.9999	0.031	0.066	0.011	0.020
Peng *et al.*	0.9966	0.9981	0.9999	0.029	0.062	0.007	0.0015
Sun *et al.*	0.9966	**0.9987**	0.9999	0.030	0.064	0.0087	0.0019
Proposed (floating-point)	**0.9968**	0.9983	0.9999	**0.027**	**0.059**	0.0069	**0.0013**
Proposed (*W2A4*)	**0.9967**	0.9980	0.9999	**0.028**	**0.061**	**0.0056**	**0.0013**
Signal photons: 2; Background Photons: 50; SBR: **0.04**							
LM filter	0.3492	0.4317	0.4737	1.184	5.774	2.070	17.615
Shin *et al.*	0	0	0	1.362	5.401	2.916	15.671
Rapp and Goyal	0.8614	0.9976	0.9995	0.106	0.236	0.0780	0.0334
Lindell *et al.*	0.9827	0.9951	0.9999	0.064	0.149	0.026	0.0120
Peng *et al.*	0.9948	0.9971	0.9998	0.034	0.073	0.0082	0.0026
Sun *et al.*	0.9961	0.9980	0.9999	0.030	0.064	0.0087	0.0019
Proposed (floating-point)	0.9961	0.9980	0.9999	**0.029**	**0.063**	**0.0067**	**0.0017**
Proposed (*W2A4*)	**0.9962**	0.9980	0.9999	0.030	0.064	**0.0060**	**0.0016**
Signal photons: 2; Background Photons: 100; SBR: **0.02**							
LM filter	0.2691	0.3439	0.3924	1.270	6.210	2.283	20.382
Shin *et al.*	0	0	0	1.383	5.620	2.990	16.568
Rapp and Goyal	0.8610	0.9974	0.9994	0.111	0.301	0.0790	0.0414
Lindell *et al.*	0.9357	0.9729	0.9902	0.129	0.321	0.0580	0.0600
Peng *et al.*	0.9952	0.9978	0.9999	0.033	0.069	0.0081	**0.0021**
Sun *et al.*	0.9961	0.9981	0.9999	0.031	0.065	0.0087	0.0020
Proposed (floating-point)	**0.9963**	0.9980	0.9999	**0.030**	**0.064**	**0.0060**	0.0017
Proposed (*W2A4*)	**0.9963**	0.9981	0.9999	**0.030**	0.065	**0.0060**	0.0017

Table 3. Extensive evaluations between Sun et al.'s [23] and the proposed models with SBR = 0.04. The proposed model achieves more robust results across five evaluations metrics.

	δ<1.25	δ<1.25^2	δ<1.25^3	RMSE(log)	RMSE (m)	Abs rel	Sq rel
Signal photons: 2; Background Photons: 50; SBR: **0.04**							
Sun *et al.* [23]	**0.9087**	0.9503	0.9625	0.292	1.766	0.071	0.269
Proposed	0.8545	**0.9618**	**0.9927**	**0.249**	**1.330**	**0.065**	**0.246**

part of the scene leads to tremendous background noise. Rapp and Goyal's spatially averaging strategy achieves relatively robust results. And for the last row, Peng *et al.*'s strategy can in-paint the spots corrupted by the ambient light; this might be due to the dense residual connections that can extract more features.

Optics EXPRESS

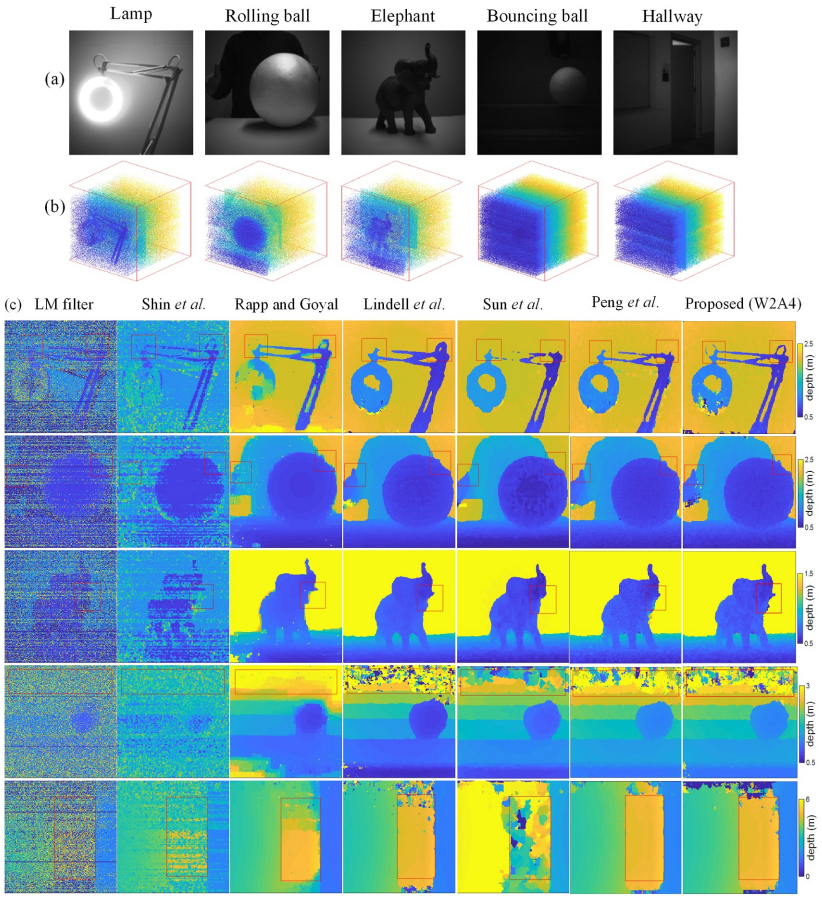

Fig. 7. (a). Intensity images of the five scenes used for fusing with Lindell *et al.*'s and Sun *et al.*'s models. (b). Detected ToA data cubes of five scenes. The bouncing ball and a hallway (the 4th and 5th rows) scenes contain intense ambient light, leading to blurred impact. (c). Reconstructed depth maps. Lindell *et al.* used the intensity-SPAD fusion, and Sun *et al.* used the monocular-SPAD fusion. The processing time is proportional to the network's depth, and the RMSE decreases for a deeper network.

5. Discussion

We evaluated the proposed model with three depth levels, meaning the input tensor is down-sampled three times. The fewer levels we use, the shorter processing time it costs (with fewer parameters). Figure 8 shows the network's depth versus the processing speed and RMSE. L1 is a naive U-net architecture with one down-sampling module without long-range connections. As our architecture becomes deeper with both long- and short-range connections, the reconstruction error decreases significantly. The RMSEs for L4 to L2 are acceptable when the model's depth reduces since they are better than existing studies. However, the processing time increases as more layers are added. Notably, due to the structural regularity of our model, it is easy to configure the numbers of down- and up-sampling without modifying the connections in the architecture. The log-scale binning method introduced by Sun *et al.* can significantly reduce the GPU memory consumption and reduce the processing time. The binning method, however, would lead to critical degradation of performance. The configurability of our network can also be deemed the model pruning since fewer nodes are involved in computations when reducing

Optics EXPRESS

the number of down-sampling. Therefore, our network is more robust to achieve a reasonable trade-off between processing time and accuracy. The processing time is measured by inferencing one depth map from a ToA tensor with the size 552×668×1024 under SBR = 0.04. RMSE is an average value of seven scenes from the Middlebury dataset that are the same data used in Table 2.

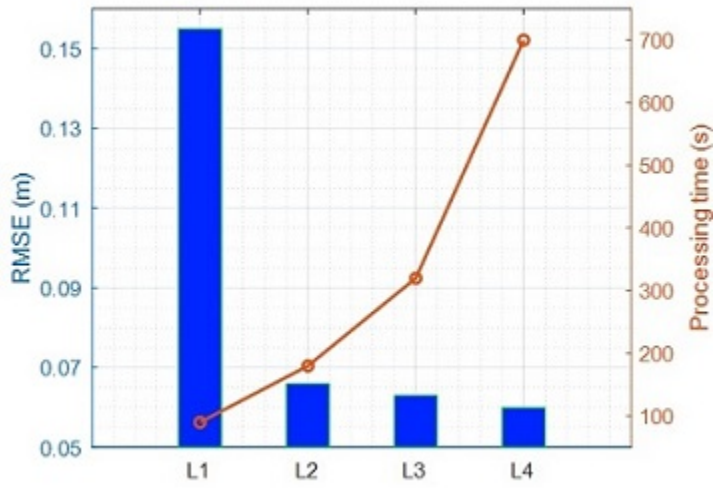

Fig. 8. The processing time is proportional to the network's depth, and the RMSE decreases for a deeper network.

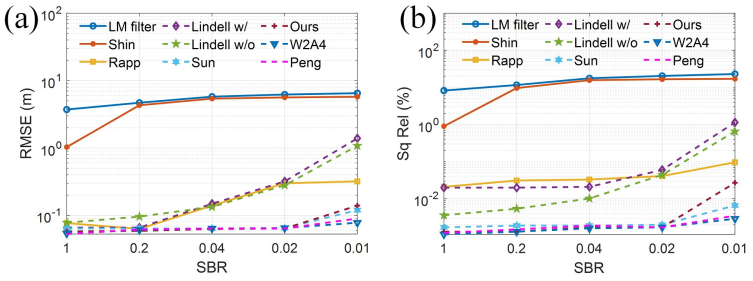

Fig. 9. (a) RMSE and (b) squared relative difference (Sq rel) plots in terms of SBR. The dashed lines indicate neural network-based algorithms, and the rest represent optimization-based algorithms. 'Sun' and 'Lindell w/' shown in the legend use the monocular-SPAD and intensity-SPAD fusion strategy, respectively.

It should be noted that there is no acceleration for our quantized architecture because the GPU cannot handle fixed-point operations due to the unified intrinsic hardware and instruction set architecture. However, FPGA can manipulate quantized numbers and accelerate the analysis; the hardware implementation is regarded as future work. Additionally, our architecture contains dense connections and multiple 3D convolution modules that consume the most processing time. ToA tensors are sparse with many zero elements involved in computing. Thus, more aggressive compression approaches like model pruning, sparse convolution strategies can help further accelerate the forward-propagation process.

The proposed network under different SBR conditions was further investigated in Fig. 9. The algorithms using neural networks are more sensitive to background noise, especially when the

Optics EXPRESS

SBR decreases from 0.02 to 0.01. We can increase the training dataset to cover a broader range of SBRs to alleviate this problem, although it costs more processing time. Nevertheless, the other three methods (LM filter, Shin *et al.*, and Rapp and Goyal) are more robust across all the SBRs. The LM filter and Shin *et al.*'s approach have maximum errors across all SBR levels. Rapp and Goyal's and Lindell *et al.*'s fusion models achieve relatively lower errors, and the former is more robust across all SBR levels. Sun *et al.*'s and the proposed network obtain the lowest errors and are robust for relatively low SBRs except the level 0.01. Therefore, it is feasible for the proposed neural network to achieve low prediction errors at standard SBR levels, and remain robust without using complementary training images from another camera. And optimization-based algorithms can be more stable than neural network-based methods over various SBR levels.

Despite relative better reconstruction results, our method still has limitations. For the scenes corrupted with enormous background noise, the time-bin index retrieved by the soft-argmax might be not the correct ToF profile. Whereas Rapp and Goyal [17] separated the signal and noise, a pixel-wise noise censoring algorithm and a windowing approach were introduced to remove noise and obtain the ToF information accurately. The reconstructed bouncing ball in Fig. 7 shows that Rapp and Goyal's algorithm can produce relatively high fidelity in a noisy scenario.

6. Conclusion

We developed a non-fusion network that does not need an additional camera for depth estimation. It is much more practical for real-world applications. The proposed model was quantized to achieve a smaller model size without degrading performances. Due to the model's flexible structure, the trade-off between accuracy and the processing speed can be balanced by manipulating the network's depth. It achieves excellent reconstruction performances in low light conditions and poor SBR conditions, and it can be extended to biological applications where samples are generally dim.

Funding. Medical Research Scotland (1179-2017); Scottish Funding Council (Datalab); Engineering and Physical Sciences Research Council (EP/L01596X/1).

Acknowledgments. The Quadro P5000 GPU used in this research was supported by the NVIDIA Corporation.

Disclosures. The authors declare no conflict of interest.

Data availability. The raw data supporting this article's conclusions will be made available by the authors without undue reservation.

References

1. M. Beer, O. M. Schrey, J. F. Haase, J. Ruskowski, W. Brockherde, B. J. Hosticka, and R. Kokozinski, "SPAD-based flash LiDAR sensor with high ambient light rejection for automotive applications," Proc. SPIE **10540**, 85 (2018).
2. N. Lazaros and A. Gasteratos, "Stereo Vision Depth Estimation Methods for Robotic Applications." Depth Map and 3D Imaging Applications: Algorithms and Technologies," IGI Global, 397–417, (2012).
3. A. N. Angelopoulos, H. Ameri, D. Mitra, and M. Humayun, "Enhanced Depth Navigation Through Augmented Reality Depth Mapping in Patients with Low Vision," Sci. Rep. **9**(1), 11230 (2019).
4. H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018).
5. Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. V. D. Maaten, M. Campbell, and K. Q. Weinberger, "Anytime Stereo Image Depth Estimation on Mobile Devices," *2019 International Conference on Robotics and Automation (ICRA)*, (2019).
6. E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-Depth Sensors for Fall Detection: A Review," IEEE Sens. J. **17**(12), 3585–3604 (2017).
7. J. Rapp, J. Tachella, Y. Altmann, S. McLaughlin, and V. K. Goyal, "Advances in Single-Photon Lidar for Autonomous Vehicles: Working Principles, Challenges, and Recent Advances," IEEE Signal Process. Mag. **37**(4), 62–71 (2020).
8. C. Niclass, A. Rochas, P. -A. Besse, and E. Charbon, "Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes," IEEE J. Solid-State Circuits **40**(9), 1847–1854 (2005).

Optics EXPRESS

9. J. A. Richardson, L. A. Grant, and R. K. Henderson, "Low Dark Count Single-Photon Avalanche Diode Structure Compatible With Standard Nanometer Scale CMOS Technology," IEEE Photonics Technol. Lett. **21**(14), 1020–1022 (2009).
10. C. Veerappan, J. A. Richardson, R. Walker, D-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 312–314, (2011).
11. D. D. Li, D. J. Arlt, D. Tyndall, R. Walker, J. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm," J Biomed Opt., **16**(9), 096012 (2011).
12. R. K. Henderson, N. Johnston, F. M. D. Rocca, H. Chen, D. D. Li, G. Hungerford, R. Hirsch, D. McLoskey, P. Yip, and D. J. S. Birch, "A 192×128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," IEEE J. Solid-State Circuits, **54**(7), 1907–1916, (2019).
13. Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction Algorithms—Theory and Practice," IEEE Trans. on Image Process. **21**(3), 1084–1096 (2012).
14. K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. (The MIT Press, 2012).
15. D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-Efficient Computational 3-D and Reflectivity Imaging With Single-Photon Detectors," IEEE Trans. Comput. Imaging **1**(2), 112–125 (2015).
16. J. Kingman, *Poisson processes*, (The Clarendon Press Oxford University Press 1993).
17. J. Rapp and V. K. Goyal, "A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging," IEEE Trans. Comput. Imaging **3**(3), 445–459 (2017).
18. J. Tachella, Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, S. McLaughlin, and J.-Y. Tourneret, "Bayesian 3D Reconstruction of Complex Scenes from Single-Photon Lidar Data," SIAM J. Imaging Sci. **12**(1), 521–550 (2019).
19. Q. Legros, S. Meignen, S. McLaughlin, and Y. Altmann, "Expectation-Maximization Based Approach to 3D Reconstruction From Single-Waveform Multispectral Lidar Data," IEEE Trans. Comput. Imaging **6**, 1033–1043 (2020).
20. D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-Photon 3D Imaging with Deep Sensor Fusion," ACM Trans. Graph. **37**(4), 1–12 (2018).
21. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,"*International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
22. A. Ruget, S. McLaughlin, R. K. Henderson, I. Gyongy, A. Halimi, and J. Leach, "Robust super-resolution depth imaging via a multi-feature fusion deep network," arXiv preprint arXiv: 2011, 11444 (2020).
23. Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein, "SPADnet: deep RGB-SPAD sensor fusion assisted by monocular depth estimation," Opt. Express **28**(10), 14948–14962 (2020).
24. J. Peng, Z. Xiong, X. Huang, Z-P. Li, D. Liu, and F. Xu, "Photon-Efficient 3D Imaging with A Non-local Neural Network," in *Computer Vision - ECCV 2020*, 12351 A. Vedaldi, H. Bischof, T. Brox, and JM. Frahm, eds. (Springer BerlinHeidelberg, Berlin, Heidelberg, 2020) pp. 225–241.
25. Y. Cao, Z. Wu, and C. Shen, "Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks," IEEE Trans. Circuits Syst. Video Technol. **28**(11), 3174–3182 (2018).
26. J. Arlt, D. Tyndall, B. R. Rae, D. D.-U. Li, J. A. Richardson, and R. K. Henderson, "A study of pile-up in integrated time-correlated single photon counting systems," Rev. Sci. Instrum. **84**(10), 103105 (2013).
27. D. L. Snyder, *Random point processes* (Wiley, 1975).
28. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2020).
29. S. Zhou, Y. Wu, Z. Ni, X Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training Low Bitwidth Convolutional Neural Net-works with Low Bitwidth Gradients," arXiv preprint arXiv:1606.06160, (2018).
30. I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press 2016).
31. J. Liu, Y. Sun, X. Xu, and U. S. Kamilov, "Image Restoration Using Total Variation Regularized Deep Image Prior," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2019).
32. N. Silberman, D. Hoiem, P. Kohil, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Computer Vision - ECCV 2012, 7576A. Fitzgibbon*, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, (Springer BerlinHeidelberg, Berlin, Heidelberg, 2012) pp. 746–760.
33. D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," in *The IEEE Conferenceon Computer Vision and Pattern Recognition (CVPR)*, (2019).
34. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv: 1412. 6980 (2014).
35. S. Burri, C. Bruschini, E. Bruschini, and Charbon, "LinoSPAD: A Compact Linear SPAD Camera System with 64 FPGA-Based TDC Modules for Versatile 50 ps Resolution Time-Resolved Imaging," Instruments **1**(6), (2017).
36. I. Alhashim and P. Wonka, "High Quality Monocular Depth Estimation via Transfer Learning," arXiv preprint arXiv: 1812. 11941v2, (2019).
37. I. Bar-David, "Communication under the Poisson regime," IEEE Trans. Inform. Theory **15**(1), 31–37 (1969).