Photon-Efficient 3D Imaging with A Non-Local Neural Network

Jiayong Peng, Zhiwei Xiong(✉), Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu

University of Science and Technology of China
jiayong@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
{hx9711,lizhp}@mail.ustc.edu.cn, {dongeliu,feihuxu}@ustc.edu.cn

Abstract. Photon-efficient imaging has enabled a number of applications relying on single-photon sensors that can capture a 3D image with as few as one photon per pixel. In practice, however, measurements of low photon counts are often mixed with heavy background noise, which poses a great challenge for existing computational reconstruction algorithms. In this paper, we first analyze the long-range correlations in both spatial and temporal dimensions of the measurements. Then we propose a non-local neural network for depth reconstruction by exploiting the long-range correlations. The proposed network achieves decent reconstruction fidelity even under photon counts (and signal-to-background ratio, SBR) as low as 1 photon/pixel (and 0.01 SBR), which significantly surpasses the state-of-the-art. Moreover, our non-local network trained on simulated data can be well generalized to different real-world imaging systems, which could extend the application scope of photon-efficient imaging in challenging scenarios with a strict limit on optical flux. Code is available at https://github.com/Jiayong0-0/PENonLocal.

Keywords: Photon-efficient imaging, long-range correlation, non-local network, depth reconstruction.

1 Introduction

Active 3D imaging systems have broad applications including biology, robotics, vehicle navigation and remote sensing. Typically, a large number of photons per pixel (ppp), e.g., 10^3 ppp in a 1 megapixel image, is required to suppress the background noise inherent in the optical detection process [18]. Important progress has been made for image sensors, where single-photon detectors [17] and arrays [38, 46] can provide extraordinary optical sensitivity and timing resolution. Together with the advanced computational algorithms, new photon-counting light detection and ranging (LiDAR) systems, which detect only a single photon per pixel on average [21], have demonstrated dramatic improvements in photon efficiency [3]. However, in certain scenarios, such as remote sensing of a dynamic scene at a long standoff distance [31, 24], non-line-of-sight imaging [30, 39, 28], as well as microscope imaging of delicate biological samples [22, 41], limitations

on the optical flux and integration time preclude the collection of the effective signal photons. Consequently, the raw measurements with extremely low photon counts and low signal-to-background ratio (SBR) pose great challenges on the reconstruction algorithms.

Recently, a number of algorithms have been proposed for 3D imaging with a small number of photons [21, 3, 43, 2, 44, 35, 26]. One of the earliest attempts is first-photon imaging [21], in which 3D structures and reflectivity can be recovered from the first detected photon at each pixel. Afterwards, there emerge other approaches dealing with the measurements captured with array detectors [44, 36, 7]. By exploiting scene structures, recent algorithms [43, 2, 44, 35] build probabilistic models for individual photon detections and use photon-by-photon processing to remove the detections that are likely to be background noise. These algorithms are more effective in low-light scenarios where conventional histogram techniques [6, 1] perform poorly. Still, their performance degrades significantly with the decrease of photon counts and SBR [35].

As in various computer vision tasks [23, 14, 13, 34], deep learning has boosted computational imaging [5, 10, 49, 48, 42, 33], encouraging remarkable progress in this field. Lindell et al. [26] first introduce deep learning to single photon 3D imaging under a sensor fusion configuration, which utilizes an additional high-resolution intensity measurement. However, the intensity image of the target scene is usually not available in practice, which restricts the application scope of this method. On the other hand, the deep neural network lacks specific designs to cope with the large-volume yet sparse photon-efficient measurements, making it less competitive to the state-of-the-art non-learning-based method [35].

In this paper, we first analyze that the photon-efficient measurements contain long-range correlations in both spatial and temporal dimensions. Then we propose an end-to-end deep learning method for depth reconstruction from the measurements with utilization of the long-range correlations. Since the measured raw photon counts are contaminated with background noise, we build our network from a denoising backbone [8, 9]. Most importantly, we integrate the non-local operator to exploit the correlations within the measurements. To make it sufficiently effective for large-volume 3D measurements, we deploy a subsequent downsampling operation along the temporal dimension in the feature space, which promotes the reconstruction performance by further enlarging the receptive field for the backbone network. As a general assumption in previous literature [26, 35, 44, 43], we focus on low photon flux regimes, which copes with our long-distance imaging system where the returning photons are weak. Comprehensive simulations demonstrate the significantly improved accuracy of our non-local network over state-of-the-art methods, and this advantage is even larger under extremely low photon counts (e.g., 1 ppp) and low SBR (e.g., 0.01). In addition, the network trained on simulated data achieves superior performance for outdoor scenes (over ranges up to 21 km, with about 1 ppp and 0.1 SBR) captured by our long-distance imaging system. This advantage is demonstrated again on real-world measurements from another indoor imaging system.

The main contributions of this work can be summarized into three aspects:

1) An end-to-end network for depth reconstruction from photon-efficient measurements, especially those with extremely low photon counts and low SBR;

2) Analysis of long-range correlations in the measurements and exploitation of the correlations with our specifically designed non-local neural network;

3) Superior reconstruction performance on both simulated and real-world measurements and improved generalization capability to unseen noise levels as well as across different imaging systems.

2 Related Work

Single-photon sensors. Photon-efficient imaging has attracted increasing attention recently. To name a few, O'Toole et al. [29] design an imaging system which builds on single-photon avalanche diode (SPAD) sensors to capture multi-path responses with active illuminations. Instead of capturing the distance, Ingle et al. [19] propose the passive free-running SPAD imaging, which uses SPADs to acquire 2D intensity images without any active light source. The captured intensity images are with unprecedented dynamic range under ambient lighting. Gupta et al. [16] study the correlations between photon flux and the distortion of captured temporal waveform. They then derive a closed form expression for the optimal flux of a SPAD-based LiDAR system, and propose a simple adaptive approach to achieve the optimal flux. Furthermore, Gupta et al. [15] propose an asynchronous single-photon 3D imaging system to mitigate the distortions caused by the ambient light.

Computational reconstruction algorithms. Depth reconstruction from photon-efficient measurements is an active research topic. As an embodiment of maximum likelihood estimation, conventional log-matched filter [4] can be effective for high-light scenarios with a large number of data samples. To tackle with decreased photon counts and SBR, Shin et al. [43] develop a robust method for estimating depth and reflectivity using fixed dwell time per pixel. They [44] also develop an array-specific algorithm to recover depth and reflectivity by exploiting both the transverse smoothness and longitudinal sparsity of the natural scenes. Rapp et al. [35] introduce a novel method that emphasizes the unmixing of contributions from signal and noise sources, which achieves promising results. With exploitation of high-resolution intensity images, Lindell et al. [26] propose a deep learning-based method for photon-efficient 3D imaging under a sensor fusion configuration. These advanced algorithms can promote the reconstruction performance in low-light scenarios. Still, their performance degrades significantly under extremely low photon counts and low SBR due to a lack of specifically designed mechanism, which hinders the application of photon-efficient imaging in challenging scenarios.

Image denoising and non-local mechanism. As a representative image denoising method, BM3D [11] searches similar patches in a global manner to exploit the non-local correlations in the whole image. Our work is inspired by this simple yet effective idea, together with the observation that the raw photon-efficient measurements have long-range correlations across both spatial

and temporal dimensions. To accomplish the depth reconstruction task with an advanced architecture, we build our network on the basis of the deep boosting denoising model [8, 9] as well as the non-local operator [47, 50]. The latter has been demonstrated effective in various tasks, such as super-resolution [12] and sequence learning [27]. However, different from ordinary images and videos, the photon-efficient measurements are in large 3D volume, sparse in temporal dimension and contaminated with heavy noise. To the best of our knowledge, this is the first time that the non-local mechanism is adopted to deal with such high dimensional and sparse measurements. Our specifically designed non-local neural network excavates the long-range correlations in both spatial and temporal dimensions and significantly improves the depth reconstruction performance especially in the challenging low photon counts and low SBR scenarios.

3 The Proposed Method

3.1 Forward Model

We depict an image formation model for SPAD-based pulsed LiDAR imaging system, which is then used to generate simulated data for training our network. Such a system generally contains a pulsed laser source and a SPAD detector, as shown in Fig. 1 (a). The pulsed laser source transmits periodic short light pulses $s(t)$ with repetition period T_r to illuminate the scene in a raster-scanned manner. To avoid distance aliasing, we assume $T_r > 2z_{max}/c$, where z_{max} is the maximum scene depth and c is the speed of light. The SPAD detector observes the reflected light pulses by detecting at most one photon per pulse repetition period, and builds a temporal histogram with recorded photons.

Note that the system operates in low photon flux regimes, which is a general assumption in previous literature [26, 35, 44, 43]. This means that the returning photons are very weak (far less than 1 photon) within each repetition period T_r and the pile-up effect [16, 32] can be negligible. Thus for each illumination position (i, j), the photon flux arrived at the detector at time interval n can be described as

$$r_{i,j}[n] = \int_{n\Delta t}^{(n+1)\Delta t} \Phi_{i,j} \cdot s(t - \frac{2z_{i,j}}{c}) dt + b_\gamma, \tag{1}$$

where $r_{i,j}[n]$ denotes the photon flux arrived at the detector at time interval n, Δt is the bins of duration, and $\Phi_{i,j}$ encapsulates the distance fall-off, scene reflectance and BRDF. $s(t)$ denotes the transmitted light pulses and $z_{i,j}$ is the scene depth of illumination position (i, j). b_γ denotes the photon flux caused by the ambient light with optical frequency γ.

For a SPAD detector, the arrived photon flux is attenuated by the detector's quantum efficiency $\eta \in [0, 1)$, which describes the probability that an incident photon can be detected by the device [37]. Besides, the detector has a non-zero dark count b_d (numbers of false detections) as well. Therefore, the number of photons measured by the SPAD detector in response to N illumination periods

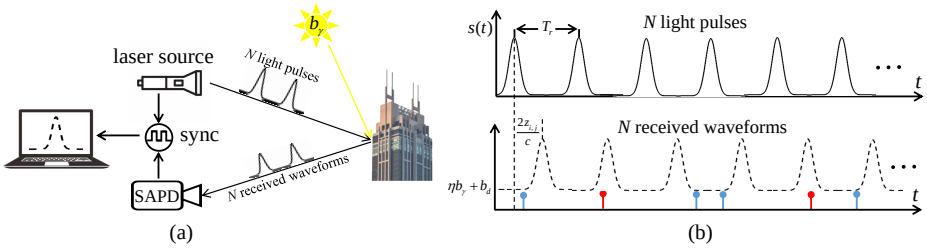

Fig. 1. (a) SPAD-based pulsed LiDAR imaging system, which contains a pulsed laser source and a SPAD detector. (b) In low photon flux regimes, the photon detections within N illumination periods can be described as the sum of signal photon detections (red) and background photon detections (blue). The signal photon detections are from the light pulses with the same distribution and they are correlated with each other, while the background photon detections are randomly distributed and not correlated

of light pulses can be represented by a temporal histogram

$$\mathbf{h}_{i,j}[n] \sim \mathbf{P}\{N[\eta r_{i,j}[n] + b_d]\}, \tag{2}$$

where $\mathbf{h}_{i,j}[n]$ represents the temporal histogram at time interval n for position (i, j) within N illumination periods of light pulses. The measurements are modeled as a Poisson process $\mathbf{P}$ with a time-varying arrival function.

3.2 Long-range Correlations

As shown in Fig. 1 (b), in temporal dimension, the photon detections within N illumination periods can be described as the sum of signal photon detections (red) and background photon detections (blue) under low photon flux regimes. The signal photon detections are from the light pulses with the same distributions and thus they are correlated with each other. However, the background photon detections are randomly distributed in time, which have no correlations with each other or the signal photon detections. Since the signal photons will reach the SPAD at any timestamps, the correlations should be considered across the whole temporal dimension. In spatial dimension, for most natural scenes, the neighborhoods that have similar geometry have correlations with each other. Since these neighborhoods may appear at any spatial positions, the correlations should be considered across the whole spatial dimension. Thus, the photon-efficient measurements have correlations in both spatial and temporal dimensions.

3.3 Network Architecture

Aiming at high fidelity depth reconstruction from photon-efficient measurements especially those with extremely low photon counts and low SBR, we propose an end-to-end deep neural network with dedicated components to exploit the long-range correlations within the measurements. The flowchart of our proposed

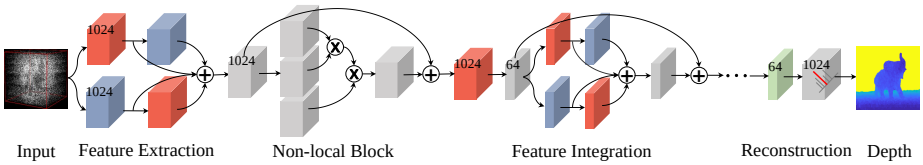

Fig. 2. The flowchart of our proposed network for depth reconstruction from the input raw photon-efficient measurements. The cuboids denote features which are in 3D volume (spatial and temporal). The temporal dimension of features is 1024 originally, and 64 after downsampling. Note that we only show one channel of features for simplification. The red, blue, and green colors denote the 3D convolution, dilated convolution and deconvolution with kernel size of 3×3×3, respectively. The gray color denotes the 3D convolution with kernel size of 1×1×1. "+" and "×" with circular blocks denote the concatenation and matrix multiplication, respectively. Each layer (except for the last one) adopts ReLU as the activation function, which is omitted here for simplification. For more details about our network, please refer to the supplementary material

network is shown in Fig. 2, the backbone of which is an advanced denoising model called dense dilated fusion network [8, 9].

Given a photon-efficient measurement for depth reconstruction, the first step is to extract features with a *feature extraction* block. After that, a *non-local* block is adopted to capture long-range spatial-temporal correlations within the measurement. Then, a *feature integration* block, that contains a downsampling operator and several 3D dilated dense fusion sub-blocks, is performed to downsample features in temporal dimensions and integrate them across channels. The last *reconstruction* block first estimates the denoised histogram $\widehat{\mathbf{h}}$, then generates the 2D depth map by reporting the bin index of the maximum value of $\widehat{\mathbf{h}}$.

In order to make our network training more efficient, we adopt two loss terms to constrain the network. One is the Kullback-Leibler (KL) divergence at each spatial position (i, j) between the denoised histogram $\widehat{\mathbf{h}}$ and the normalized groundtruth histogram $\mathbf{h}$, which can be written as

$$D_{KL}(\mathbf{h}_{i,j}, \widehat{\mathbf{h}}_{i,j}) = \sum_n \mathbf{h}_{i,j}[n] \log \frac{\mathbf{h}_{i,j}[n]}{\widehat{\mathbf{h}}_{i,j}[n]}. \tag{3}$$

The other is a total variation (TV) term for regularization on the output 2D depth map, which is to improve the robustness of the network. We apply a differentiable argmax operator S to $\widehat{\mathbf{h}}$ to find the bin index of the maximum value through a simple weighted sum calculation for each spatial location (i, j)

$$S(\widehat{\mathbf{h}}_{i,j}) = \sum_n n \cdot \widehat{\mathbf{h}}_{i,j}[n]. \tag{4}$$

Thus the final loss function to train our depth reconstruction network is

$$L(\mathbf{h}, \widehat{\mathbf{h}}) = \sum_{i,j} D_{KL}(\mathbf{h}_{i,j}, \widehat{\mathbf{h}}_{i,j}) + \lambda TV(S(\widehat{\mathbf{h}})), \tag{5}$$

where λ is a hyper-parameter giving the ratio of the two loss terms.

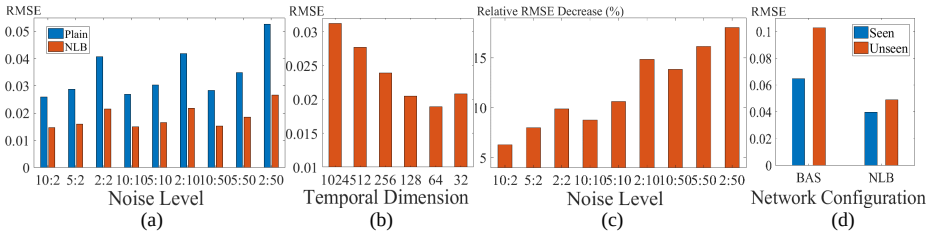

Fig. 3. (a) Performance comparisons of different network configurations against noise levels. (b) Reconstruction performance against temporal dimension of features after downsampling. (c) Improvements of NLB over BAS against noise levels. (d) Quantitative comparisons of NLB and BAS in generalization to unseen noise levels

3.4 Non-local Block

Since the photon-efficient measurements contain long-range correlations in both spatial and temporal dimensions, we propose our non-local block to exploit the correlations. Due to the high dimensionality and sparsity of the measurements, integrating the non-local block and making it sufficiently effective is non-trivial. Thus, we deploy a downsampling operation along the temporal dimension in the feature space after the non-local block to make the training efficient for such high dimensional and sparse 3D measurements.

Global information exploration. The global information across both spatial and temporal dimensions explored by our non-local block remarkably improves the reconstruction performance, especially in low photon counts and low SBR scenarios. To verify this, we conduct an ablation between two networks: "Plain" denotes the backbone network [8, 9], "NLB" denotes our non-local neural network exploiting both spatial and temporal correlations. Here we report some intermediate results. As shown in Fig. 3 (a), our non-local neural network achieves performance improvements on each noise level, and the improvements are more prominent in low photon counts and low SBR (e.g., 49% on 2:50) compared with that in higher ones (e.g., 42% on 10:2), which demonstrates the effectiveness of our non-local network. Note that the less the photon counts are, the more difficult the reconstruction is. For example, it is much more challenging to reconstruct on 2:2 than 10:10, although they are with the same SBR.

Downsampling scale investigation. We also investigate the correlations between the downsampling scale and the reconstruction performance. Different downsampling scales result in different temporal dimensions of the features. As shown in Fig. 3 (b), the reconstruction performance improves with the decrease of the temporal dimension of features (originally 1024) until it hits a certain value. After this turning point (64 channels here), the performance begins to degrade due to the substantial information loss caused by such heavy downsampling. In the following simulation and experiments, we uniformly set the downsampling scale to 16 (corresponding to 64 channels) in our network, which guarantees superior performance as well as training efficiency.

A closer look at the low end. As demonstrated above, the long-range spatial-temporal dependencies within the raw measurements can be effectively captured by our non-local block. Here we take a closer look at the performance improvement, where the network containing the downsampling operator but not the non-local block is adopted as the baseline (denoted as "BAS"). In this way, we can see the role of the non-local block in our network more clearly. Specifically, we compare the improvements of our non-local neural network (with both non-local block and downsampling, denoted as "NLB") over the above baseline on various noise levels. As shown in Fig. 3 (c), the non-local block itself achieves a notable improvement on each noise level, and the improvements are more prominent in low photon counts and low SBR (e.g., 18% on 2:50) compared with those in high ones (e.g., 6% on 10:2), which demonstrates the effectiveness of the non-local block itself.

Generalization capability improvement. Due to the exploration of the global information across both spatial and temporal dimensions, the non-local block also helps to improve the generalization capability of the network to unseen noise levels. To verify this, we first train two networks, i.e., BAS and NLB, on a dataset with a large range of noise levels (9 typical noise levels plus 3 extremely low SBR cases, see Sec. 4.3 for more details), and the obtained models are denoted as "Seen". The two networks are then retrained on a dataset with a small range of noise levels (the 9 typical noise levels in the previous large range), and the obtained models are denoted as "Unseen". We compare the performance of the above models on the test data in the 3 extremely low SBR cases. As shown in Fig. 3 (d), the performance of the BAS network degrades 60% when a trained model generalizes to unseen noise levels, while the NLB network only drops 24% in the same situation. It thus demonstrates the effectiveness of the non-local block, once again.

In summary, our non-local block effectively captures the long-range spatial-temporal dependencies within 3D photon-efficient measurements, which is beneficial for depth reconstruction especially in low photon counts and low SBR scenarios. Moreover, it helps improve the generalization capability of the network to unseen noise levels.

4 Experiments

4.1 Data Simulation and Evaluation Metric

We simulate SPAD measurements for a variety of scenes and illumination conditions using RGB-D images from the NYU v2 dataset [45] captured with the Microsoft Kinect sensor, by sampling the corresponding inhomogeneous Poisson process in Eq. 2. To vary the signal and background noise levels across the dataset, we simulate an average of 2, 5, and 10 signal photons detected per pixel, with 2, 10, and 50 background photons at each signal level. Each measurement has 1024 bins for every histogram on a certain spatial position with a bin size of 80 ps, and a detected illumination pulse with a full width at half maximum (FWHM) of 400 ps. A total of 13800 and 2800 measurements are generated from

the NYU v2 dataset for training and validation, respectively. The test data is simulated on a set of 8 scenes from the Middlebury dataset [40] under different noise levels. The evaluation metric is the generally used root mean square error (RMSE) between the recovered depth map and the ground truth, which is averaged over 8 test scenes under a number of selected noise levels.

4.2 Implementation Details

We implement our method using PyTorch, and make comparisons with conventional log-matched filter (LM Filter) [4] and several state-of-the-art approaches including Shin et al. [44], Rapp et al. [35], and Lindell et al. [26] (Lindell_I and Lindell denote the networks trained with and without intensity maps, respectively) on both simulated and real-world measurements. The implementation of these methods follows their publicly available codes. For our network, we adopt eight 3D dilated dense fusion sub-blocks in the feature integration block, and the hyper-parameter λ in the loss function is set to 10^{-5}. We train the networks in [26] with the same training data as ours, during which we extract patches of size $32\times32\times1024$, with a batch size of 4. We initialize the network randomly and use the ADAM [20] solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-4} with a learning rate decay of 0.9 after each epoch. The training is conducted on NVIDIA 1080Ti GPU, which takes about 35 hours for the network to converge. Limited by the large GPU memory required for 3D convolution especially in [26], we test on the measurements with a relatively low spatial resolution of 72×88 and a uniform temporal resolution of 1024 (unless noted otherwise), yet higher spatial resolution input can be processed in a patch-by-patch manner.

4.3 Simulation Results

Quantitative evaluation. We first evaluate our method on the 8 test scenes under 9 typical noise levels generally reported in literature (10:2, 5:2, 2:2, 10:10, 5:10, 2:10, 10:50, 5:50, 2:50) with comparison of LM Filter [4], Shin et al. [44], Rapp et al. [35], and Lindell et al. [26]. The quantitative results are listed in the upper part of Table 1. As can be seen, our method achieves the best performance in terms of all noise levels and significantly surpasses previous approaches. Compared with the two most recent methods Rapp and Lindell, our method improves the reconstruction performance by a large margin (over 50%), which indicates the effectiveness of exploiting long-range correlations within the measurements using the non-local block.

A closer look at the low end. To further evaluate the performance of our method under extremely low photon counts and low SBR, we conduct simulations with noise levels of 3:100, 2:100, and 1:100, which are seldom investigated before since they are too challenging[1]. The results are listed in the lower part of

[1] Note that we enlarge the illumination periods N to ensure that the returning photons in each period are weak enough so that our image formation model in Eq. 2 can still be valid without suffering from the pile-up effect in simulating these noise levels.

Table 1. Quantitative comparisons of several depth reconstruction methods under different noise levels (signal photon: noise photon). All results are reported as an average root mean square error (RMSE) over the test set containing 8 scenes with spatial resolution of 72×88. Lindell_I and Lindell denote that the networks in [26] are trained with and without intensity maps, respectively

Noise Levels	LM Filter	Shin	Rapp	Lindell	Lindell_I	Ours
10:2	0.8023	0.0274	0.0226	0.0296	0.0278	**0.0147**
5:2	1.8994	0.0380	0.0268	0.0346	0.0334	**0.0160**
2:2	3.7632	0.0677	0.0376	0.0470	0.0454	**0.0216**
10:10	1.2328	0.0385	0.0232	0.0303	0.0286	**0.0150**
5:10	2.5967	0.0532	0.0282	0.0354	0.0342	**0.0165**
2:10	4.8231	0.0892	0.0570	0.0485	0.0479	**0.0218**
10:50	1.7839	0.0764	0.0267	0.0317	0.0295	**0.0153**
5:50	3.4985	0.1060	0.0359	0.0380	0.0345	**0.0185**
2:50	5.7514	0.1514	0.0890	0.0748	0.0681	**0.0266**
Ave.	2.9057	0.0720	0.0385	0.0411	0.0388	**0.0184**
3:100	5.4568	1.2593	0.0614	0.0655	0.0487	**0.0250**
2:100	6.2437	1.3648	0.1163	0.2435	0.1311	**0.0328**
1:100	6.9180	2.1753	0.6605	1.3650	1.2702	**0.0893**
Ave.	6.2062	1.5998	0.2794	0.5580	0.4833	**0.0490**

Table 1. Quantitatively, our method achieves 82% and 89% improvements over the second and third best, respectively. Specifically, the performance of Rapp and Lindell decrease dramatically from 2:100 to 1:100, while our method behaves much better with an elegant degradation. Note that, the three noise levels are not involved in the training data. It demonstrates the generalization capability of our network to unseen test data, which is essential to guarantee that the network trained from simulated data is applicable to real-world scenarios.

Qualitative evaluation. We provide qualitative comparisons of depth reconstruction for different methods on various noise levels, with two exemplar scenes shown in Fig. 4 and Fig. 5. As can be seen, existing state-of-the-art methods recover accurate depth maps under high SBR (e.g., 10:2), but they all fail to recover decent depth under SBR as low as 1:100 (0.01). Specifically, LM Filter generates noisy results, Shin loses informative depth information, Rapp, Lindell, and Lindell_I fail to recover structures of the scene. In contrast, our method still reconstructs decent depth even in this challenging case. The distinct improvement over previous approaches under extremely low photon counts and low SBR demonstrates the superiority of our method.

Running Time. We further compare the running time of our method with existing approaches. LM Filter, Shin, and Rapp are tested on Intel Core i7-6700k @4GHz CPU, while Lindell, Lindell_I, and our method are tested on NVIDIA 1080Ti GPU. As shown in Table 2, our method is much faster than others, achieving nearly 6 and 35 times acceleration compared with Lindell and Rapp.

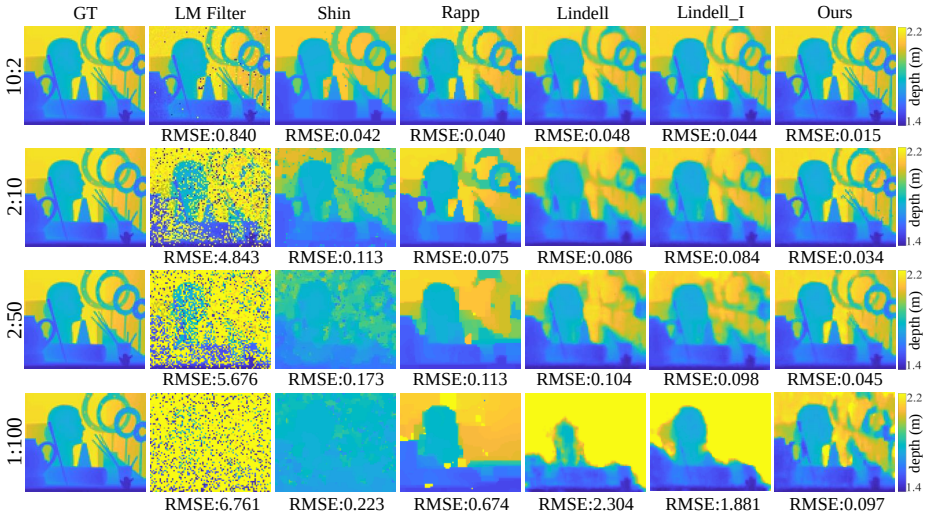

Fig. 4. Comparisons of several depth reconstruction methods for the *Art* scene on different noise levels: 10:2, 2:10, 2:50, and 1:100. Depth maps are with a spatial resolution of 72×88. GT denotes the groundtruth depth map provided in the dataset

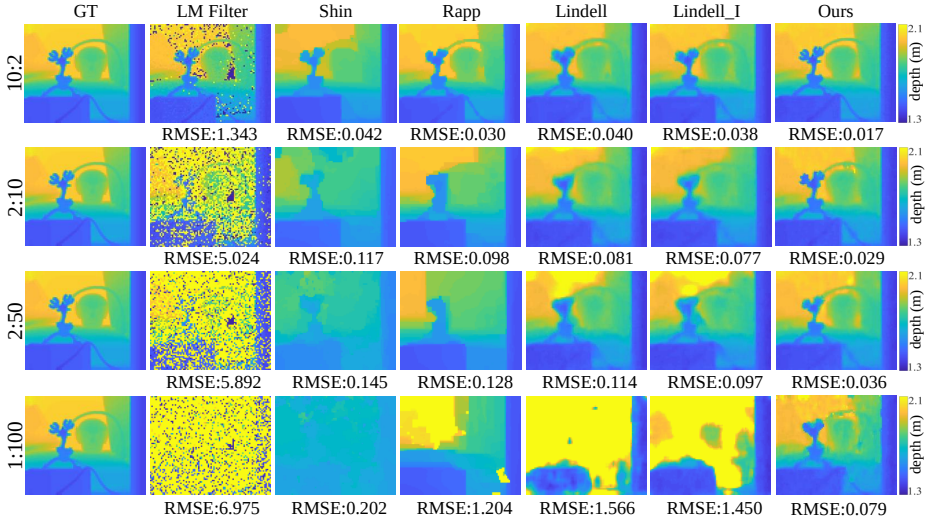

Fig. 5. Comparisons of several depth reconstruction methods for the *Reindeer* scene on different noise levels: 10:2, 2:10, 2:50, and 1:100. Depth maps are with a spatial resolution of 72×88. GT denotes the groundtruth depth map provided in the dataset

4.4 Real-world Results

Besides the simulated test set, we also conduct outdoor experiments to verify the performance of our method in real-world scenarios. Our long-distance coax-

Table 2. The running time of different methods averaged on 8 test scenes with a resolution of 72×88×1024

Methods	LM Filter	Shin	Rapp	Lindell	Lindell_I	Ours
Time(s)	8.47	9.90	19.19	2.91	2.95	0.55

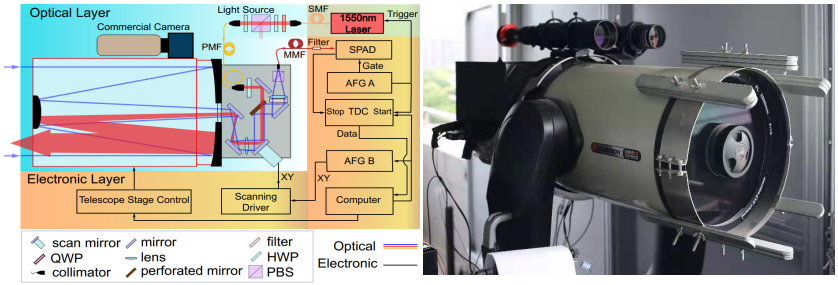

Fig. 6. (a) Imaging optics. (b) Photo of imaging setup

ial single-photon imaging setup is shown in Fig. 6. The laser transmitted from the collimator, passing through the perforated mirror and the scanning mirror, comes out from the telescope. The photons reflected by the target are collected with the same telescope, then delivered through the scanning mirror, perforated mirror and polarization beam splitter, detected by the single photon detector at last. This system was initially proposed in [25] yet with a traditional reconstruction algorithm. For more details about the imaging system, please refer to the supplementary material.

We capture three different scenes over 1 km, 4 km and 21 km away, respectively, and make comparisons among different depth reconstruction methods. Here the networks of Lindell and ours are both trained on the aforementioned simulated dataset in Sec. 4.1, which are adopted to process the real-world measurements directly. The qualitative comparisons are shown in Fig. 7. As can be seen, both LM Filter and Shin fail to reconstruct the main structures of the scenes, resulting heavy noise or missing components. For Rapp and Lindell, they fail to reconstruct the fine structures in the scenes. For example, they can hardly reconstruct the windows in the second and third scenes. In contrast, our network recovers both main and fine structures in the scenes even under heavy noise. For a further comparison, we also provide quantitative results in terms of RMSE computed with the groundtruth depth maps which are captured by our system with a long acquisition time. The quantitative results clearly demonstrate the superior performance of our method over previous approaches.

In addition to our long-distance coaxial single-photon imaging system, we also test on real-world measurements captured by another indoor single-photon imaging prototype [26], which consists of synchronization electronics, off-the-

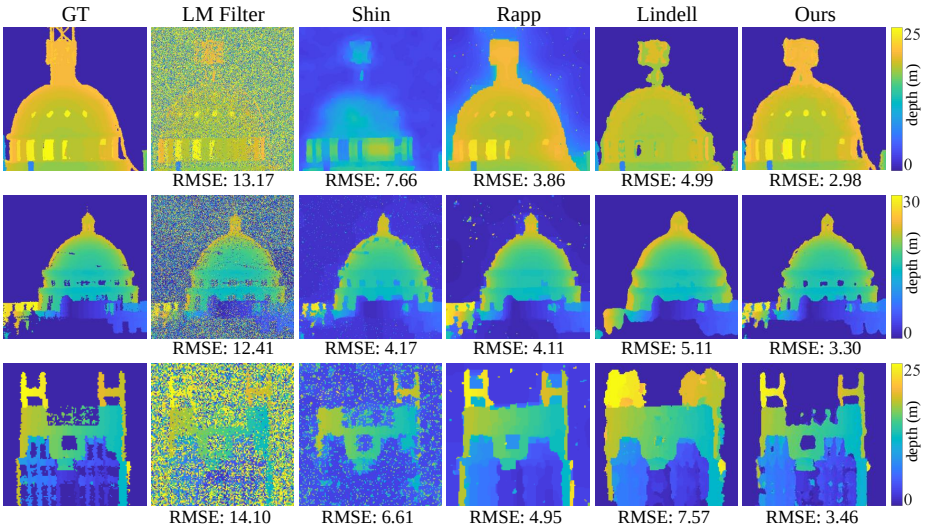

Fig. 7. The reconstruction results for three outdoor scenes. First row: a *Hotel*, that locates 1 km away with a spatial resolution of 320×320, SBR=0.1, and about 1 ppp. Second row: a *Castle*, that locates 4 km away with a spatial resolution of 256×256, SBR=0.9, and about 1 ppp. Third row: a tall building named *K11*, that locates 21 km away with a spatial resolution of 128×128, SBR=0.1, and about 1 ppp. GT denotes the groundtruth depth maps captured by our system with a long acquisition time. Limited by the GPU memory, the input measurements are cropped into 64x64 patches in the spatial dimensions, and the reconstruction results are stitched together to obtain final depth maps with the same spatial resolution as the inputs

shelf optical and optomechanical components, a standard vision camera, a picosecond laser, and a linear array of 256 SPADs. The qualitative comparisons are shown in Fig. 8, which demonstrate the superiority of the proposed method over previous approaches again. Specifically, as marked by the red boxes, one can easily observe grid-like errors in the first scene, a bump-like error below the elephant's head in the second scene, and missing structures in the third scene for different methods in comparison, yet our method gets rid of these errors. In the last scene, our method makes the lamp circle larger which is an error due to the extremely high brightness of the lamp (see the intensity image). However, other methods also encounter errors in this region which are even more severe. It is worth mentioning that the networks used in the above two single-photon imaging systems are trained on the same simulated dataset, which demonstrates the generalization capability of our method across different imaging systems.

5 Conclusion

We analyze the long-range correlations across spatial and temporal dimensions within the photon-efficient measurements and propose an end-to-end deep neu-

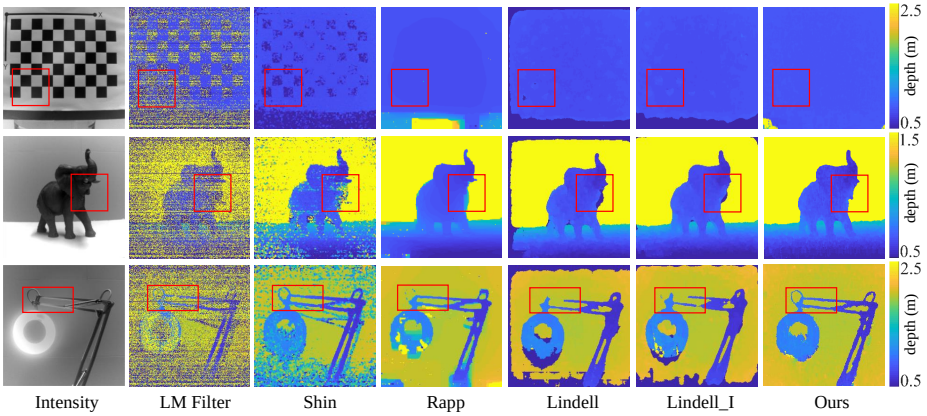

Fig. 8. The reconstruction results for real-world scenes captured by an indoor single-photon imaging prototype [26]. The input measurements with a resolution of 256×256×1536 are cropped into 64×64 patches in the spatial dimensions, and the reconstruction results are stitched together to obtain final depth maps with a spatial resolution of 256×256

ral network for depth reconstruction from the measurements by exploiting the correlations with the non-local block and the downsampling operator. Comprehensive simulations demonstrate the significantly improved accuracy of the proposed method over existing state-of-the-art approaches, and this advantage is even larger under extremely low photon counts (e.g., 1 ppp) and low SBR (e.g., 0.01). In addition to the superior performance on simulated data, our method also generalizes well in real-world experiments ranging up to 21 km, with about 1 ppp and 0.1 SBR. We believe that the proposed method could extend the application scope of photon-efficient imaging especially in challenging scenarios, e.g., long-range imaging at a few hundreds of kilometers, non-line-of-sight imaging, and biological imaging with a strict limit on optical flux.

Acknowledgements

We acknowledge funding from National Key R&D Program of China under Grants 2017YFA0700800 and 2018YFB0504300, National Natural Science Foundation of China under Grants 61671419 and 61771443, the Shanghai Municipal Science and Technology Major Project (2019SHZDZX01), the Shanghai Science and Technology Development Funds (18JC1414700), and the Fundamental Research Funds for the Central Universities (WK2340000083).

References

1. Abreu, E., Lightstone, M., Mitra, S.K., Arakawa, K.: A new efficient approach for the removal of impulse noise from highly corrupted images. IEEE Transactions on Image Processing **5**(6), 1012–1025 (1996)
2. Altmann, Y., Ren, X., Mccarthy, A., Buller, G., Mclaughlin, S.: Lidar waveform based analysis of depth images constructed using sparse single-photon data. IEEE Transactions on Computational Imaging **25**(5), 1935–1946 (2016)
3. Altmann, Y., McLaughlin, S., Padgett, M.J., Goyal, V.K., Hero, A.O., Faccio, D.: Quantum-inspired computational imaging. Science **361**(6403), 2298 (2018)
4. Bar-David, I.: Communication under the poisson regime. IEEE Transactions on Information Theory **15**(1), 31–37 (1969)
5. Barbastathis, G., Ozcan, A., Situ, G.: On the use of deep learning for computational imaging. Optica **6**(8), 921–943 (2019)
6. Buller, G.S., Wallace, A.M., Mccarthy, A., Lamb, R.A.: Ranging and three-dimensional imaging using time-correlated single-photon counting. IEEE Journal of Selected Topics in Quantum Electronics **13**(4), 1006–1015 (2007)
7. Chan, S., Halimi, A., Zhu, F., Gyongy, I., Henderson, R.K., Bowman, R., McLaughlin, S., Buller, G.S., Leach, J.: Long-range depth imaging using a single-photon detector array and non-local data fusion. Scientific Reports **9**(1), 8075 (2019)
8. Chen, C., Xiong, Z., Tian, X., Wu, F.: Deep boosting for image denoising. In: European Conference on Computer Vision. pp. 3–18 (2018)
9. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Real-world image denoising with deep boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
10. Cheng, Z., Xiong, Z., Liu, D.: Light field super-resolution by jointly exploiting internal and external similarities. IEEE Transactions on Circuits and Systems for Video Technology (2019)
11. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing **16**(8), 2080–2095 (2007)
12. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11065–11074 (2019)
13. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2015)
14. Girshick, R.: Fast r-cnn. In: IEEE International Conference on Computer Vision. pp. 1440–1448 (2015)
15. Gupta, A., Ingle, A., Gupta, M.: Asynchronous single-photon 3d imaging. In: IEEE International Conference on Computer Vision. pp. 7909–7918 (2019)
16. Gupta, A., Ingle, A., Velten, A., Gupta, M.: Photon-flooded single-photon 3d cameras. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6770–6779 (2019)
17. Hadfield, R.H.: Single-photon detectors for optical quantum information applications. Nature Photonics **3**(12), 696 (2009)
18. Holst, G.C.: Ccd arrays, cameras, and displays. SPIE Optical Engineering (1998)
19. Ingle, A., Velten, A., Gupta, M.: High flux passive imaging with single-photon sensors. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6760–6769 (2019)

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kirmani, A., Venkatraman, D., Shin, D., Colaço, A., Wong, F.N., Shapiro, J.H., Goyal, V.K.: First-photon imaging. Science **343**(6166), 58–61 (2014)
22. Köllner, M., Wolfrum, J.: How many photons are necessary for fluorescence-lifetime measurements? Chemical Physics Letters **200**(1-2), 199–204 (1992)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems. pp. 1097–1105 (2012)
24. Li, Z.P., Huang, X., Cao, Y., Wang, B., Li, Y.H., Jin, W., Yu, C., Zhang, J., Zhang, Q., Peng, C.Z., et al.: Single-photon computational 3d imaging at 45 km. arXiv preprint arXiv:1904.10341 (2019)
25. Li, Z.P., Huang, X., Cao, Y., Wang, B., Li, Y.H., Zhang, J., Zhang, Q., Peng, C.Z., Xu, F., Pan, J.W.: All-time single-photon 3d imaging over 21 km. In: Conference on Lasers and Electro-Optics. p. SM1N.1 (2019)
26. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3d imaging with deep sensor fusion. ACM Transactions on Graphics **37**(4), 113 (2018)
27. Liu, P., Chang, S., Huang, X., Tang, J., Cheung, J.C.K.: Contextualized non-local neural networks for sequence learning. In: Association for the Advancement of Artificial Intelligence. pp. 6762–6769 (2019)
28. Liu, X., Guillén, I., La Manna, M., Nam, J.H., Reza, S.A., Le, T.H., Jarabo, A., Gutierrez, D., Velten, A.: Non-line-of-sight imaging using phasor-field virtual wave optics. Nature **572**(7771), 620–623 (2019)
29. O'Toole, M., Heide, F., Lindell, D.B., Zang, K., Diamond, S., Wetzstein, G.: Reconstructing transient images from single-photon sensors. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1539–1547 (2017)
30. OToole, M., Lindell, D.B., Wetzstein, G.: Confocal non-line-of-sight imaging based on the light-cone transform. Nature **555**(7696), 338 (2018)
31. Pawlikowska, A.M., Halimi, A., Lamb, R.A., Buller, G.S.: Single-photon three-dimensional imaging at up to 10 kilometers range. Optics Express **25**(10), 11919–11931 (2017)
32. Pediredla, A.K., Sankaranarayanan, A.C., Buttafava, M., Tosi, A., Veeraraghavan, A.: Signal processing based pile-up compensation for gated single-photon avalanche diodes. arXiv preprint arXiv:1806.07437 (2018)
33. Peng, J., Xiong, Z., Liu, D., Chen, X.: Unsupervised depth estimation from light field using a convolutional neural network. In: International Conference on 3D Vision. pp. 295–303 (2018)
34. Peng, J., Xiong, Z., Wang, Y., Zhang, Y., Liu, D.: Zero-shot depth estimation from light field using a convolutional neural network. IEEE Transactions on Computational Imaging **6**, 682–696 (2020)
35. Rapp, J., Goyal, V.K.: A few photons among many: Unmixing signal and noise for photon-efficient active imaging. IEEE Transactions on Computational Imaging **3**(3), 445–459 (2017)
36. Ren, X., Connolly, P.W., Halimi, A., Altmann, Y., McLaughlin, S., Gyongy, I., Henderson, R.K., Buller, G.S.: High-resolution depth profiling using a range-gated cmos spad quanta image sensor. Optics Express **26**(5), 5541–5557 (2018)
37. Renker, D.: Geiger-mode avalanche photodiodes, history, properties and problems. Nuclear Instruments and Methods in Physics Research **567**(1), 48–56 (2006)
38. Richardson, J.A., Grant, L.A., Henderson, R.K.: Low dark count single-photon avalanche diode structure compatible with standard nanometer scale cmos technology. IEEE Photonics Technology Letters **21**(14), 1020–1022 (2009)

39. Saunders, C., Murray-Bruce, J., Goyal, V.K.: Computational periscopy with an ordinary digital camera. Nature **565**(7740), 472 (2019)
40. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)
41. Schwartz, D.E., Charbon, E., Shepard, K.L.: A single-photon avalanche diode array for fluorescence lifetime imaging microscopy. IEEE Journal of Solid-State Circuits **43**(11), 2546–2557 (2008)
42. Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F.: Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)
43. Shin, D., Kirmani, A., Goyal, V.K., Shapiro, J.H.: Photon-efficient computational 3-d and reflectivity imaging with single-photon detectors. IEEE Transactions on Computational Imaging **1**(2), 112–125 (2015)
44. 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**, 12046 (2016)
45. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision. pp. 746–760 (2012)
46. Villa, F., Lussana, R., Bronzi, D., Tisa, S., Tosi, A., Zappa, F., Mora, A.D., Contini, D., Durini, D., Weyers, S.: Cmos imager with 1024 spads and tdcs for single-photon timing and 3d time-of-flight. IEEE Journal of Selected Topics in Quantum Electronics **20**(6), 364–373 (2014)
47. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
48. Xiong, Z., Shi, Z., Li, H., Wang, L., Liu, D., Wu, F.: Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2017)
49. Yao, M., Xiong, Z., Wang, L., Liu, D., Chen, X.: Spectral-depth imaging with deep learning based reconstruction. Optics Express **27**(26), 38312–38325 (2019)
50. Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F.: Compact generalized non-local network. In: International Conference on Neural Information Processing Systems. pp. 6510–6519 (2018)