photonics

MDPI

Article

TSDSR: Temporal–Spatial Domain Denoise Super-Resolution Photon-Efficient 3D Reconstruction by Deep Learning

Ziyi Tong [1], Xinding Jiang [1], Jiemin Hu [1], Lu Xu [1], Long Wu [1], Xu Yang [1,2,*] and Bo Zou [3,*]

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; ptrayi@163.com (Z.T.); echoity@163.com (X.J.); hujiemin@zstu.edu.cn (J.H.); xlhit@126.com (L.X.); wulong@zstu.edu.cn (L.W.)
[2] Key Laboratory of Optical Field Manipulation of Zhejiang Province, Zhejiang Sci-Tech University, Hangzhou 310018, China
[3] Institute of Land Aviation, Beijing 101121, China
* Correspondence: yangxu@zstu.edu.cn (X.Y.); zoubodr@sina.com (B.Z.)

Abstract: The combination of a single-photon avalanche diode detector with a high-sensitivity and photon-efficient reconstruction algorithm can realize the reconstruction of target range image from weak light signal conditions. The limited spatial resolution of the detector and the substantial background noise remain significant challenges in the actual detection process, hindering the accuracy of 3D reconstruction techniques. To address this challenge, this paper proposes a denoising super-resolution reconstruction network based on generative adversarial network (GAN) design. Soft thresholding is incorporated into the deep architecture as a nonlinear transformation layer to effectively filter out noise. Moreover, the Unet-based discriminator is introduced to complete the high-precision detail reconstruction. The experimental results show that the proposed network can achieve high-quality super-resolution range imaging. This approach has the potential to enhance the accuracy and quality of long-range imaging in weak light signal conditions, with broad applications in fields such as robotics, autonomous vehicles, and biomedical imaging.

Keywords: LiDAR; GaAs; photon-efficient imaging; 3D reconstruction; deep learning

check for updates

Citation: Tong, Z.; Jiang, X.; Hu, J.; Xu, L.; Wu, L.; Yang, X.; Zou, B. TSDSR: Temporal–Spatial Domain Denoise Super-Resolution Photon-Efficient 3D Reconstruction by Deep Learning. *Photonics* **2023**, *10*, 744. https://doi.org/10.3390/photonics10070744

Received: 25 April 2023
Revised: 20 June 2023
Accepted: 27 June 2023
Published: 28 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Active optical imaging has lower environmental requirements and is more versatile than passive imaging detection. However, optical detection typically requires a large number of photons to reduce background noise. This can be challenging in remote sensing [1], non-visual imaging [2], and other applications [3], where the amount of signal photons is limited by light flux and integration time. Traditional systems based on photomultiplier tubes struggle to accurately reconstruct scene information under these conditions. single-photon avalanche diode (SPAD) is an avalanche photodiode operating in a Geiger mode, which achieves detection sensitivity at the single-photon level by utilizing the avalanche breakdown phenomenon. SPAD-based laser radar systems are highly sensitive and can detect very weak echo signals [4]. The SPAD detector records the photon arrival time, and a 3D reconstruction algorithm calculates the target range information based on the photon count histogram.

During the measurement process, dark counts can introduce additional noise and uncertainties, impacting the accuracy of imaging. Moreover, sparse echo signal scenes are often accompanied by strong background noise, such as sunlight, fog, rain, and snow. To improve detection accuracy in various scenarios, it is necessary to reduce the impact of background noise on the system using a reconstruction algorithm. Several methods have been proposed to reconstruct high-noise SPAD detection data [5–10]. Moreover, in recent years, deep learning-based methods have advanced photon-efficient reconstruction imaging even further [11–15].

For large-scale SPAD arrays, the packaging technology poses difficulties, leading to a mutual restriction between temporal and spatial resolution. To achieve better range accuracy and reduce noise from crosstalk between pixels, the spatial resolution of the SPAD array is often sacrificed, resulting in a lack of detail in reconstructed images. However, there is currently a lack of methods that utilize deep learning to improve the resolution of photon-efficient range imaging. Single-image super-resolution (SISR) is a widely used computer vision technique in astronomical and biomedical imaging. Neural networks, particularly GANs, offer a wide range of uses in SISR and have expanded quickly [16–20]. The introduction of SISR technology into photon-efficient imaging can further enhance reconstruction image resolution [21–26].

In this paper, a photon-efficient super-resolution convolutional neural network is proposed. The network uses the Non-Local Sparse Attention (NLSA) module [15] to extract features and addresses the high noise typically present in SPAD detector data by directly solving 3D data and introducing adaptive soft thresholding for denoising [27]. However, soft thresholding may filter out weak echo signals, so a more reasonable loss function is needed to balance noise removal and signal preservation. To achieve this, the proposed method uses a Unet-based discriminator to output both local and global discriminant information [28]. The discriminator guides the network to generate more details using pixel-level discrimination information, and the total hybrid loss function includes Mean Absolute Error (MAE), Structural Similarity (SSIM), and adversarial loss from the Unet-based discriminators. Simulation experiments conducted in different noise environments and scenarios demonstrate the proposed method's excellent reconstruction results through qualitative and quantitative analysis and comparison with previously reported algorithms. Ablation experiments explore the influence of each module on network performance. The proposed network's 3D super-resolution reconstruction performance is further validated through experiments on a pulsed lidar system based on an SPAD array in the real world.

2. Related Work

2.1. Reconstruction Photon-Efficient Imaging

Recently, there has been a significant amount of work focused on improving the quality of reconstructed images by suppressing background noise. This has been achieved through various techniques such as leveraging the distribution characteristics of signal photons or integrating multiple sensors. Shin et al. [5,6] suppressed Poisson noise on a pixel-by-pixel basis by establishing a SPAD trigger probability model. Rapp et al. [7] improved the great likelihood estimation method by incorporating an adaptive super-resolution algorithm that leverages the signal photon distribution law. In challenging conditions, Halimi et al. [8] utilized a layered Bayesian algorithm combined with multi-spectral single-photon lidar to estimate range image. Chen et al. [9] proposed a deep domain adversarial adaptation technique that addresses the domain shift issue in photon-efficient imaging by making advantage of a potent network structure.

With the development of deep learning technique, neural network has become a new direction to solve the problem of efficient photon imaging with its powerful feature extraction ability [11]. Peng et al. [12] proposed an end-to-end reconfigurable network with a denoising module as the main architecture. Non-local attention module was introduced to extract the temporal-spatial domain long-range correlation feature of SPAD detection data. Zang et al. [13] proposed a lightweight neural network that can be implemented on an embedded hardware platform. Unet was used as the main body of the network to integrate multidimensional space-time spatial characteristics. Zhao et al. [14] constrained the network with gradient regularization function and introduced the ADAG module into the Unet network model to recover more accurate edge details. Yang et al. [15] designed a convolutional encoder to directly reconstruct the incoming 3D data into range and reflectivity images. NLSA module was used to extract non-local features with low computational cost.

2.2. Single-Image Super-Resolution

Super-Resolution Convolutional Neural Network (SRCNN) [16] first applies deep convolutional neural networks to SISR. Over the years, researchers have made significant advancements in optimizing the network architecture by incorporating various techniques such as stacking convolutional layers, adding jump connections and integrating attention modules [17–20]. For instance, Zhang et al. [17] incorporated the channel attention module into extremely deep residual network, which improved reconstruction accuracy significantly. Wang et al. [20] explored the use of sparsity in SISR by incorporating a sparse mask that identifies important and unimportant regions in images, leading to further enhancements in the reconstruction process.

GANs have been used in SISR to improve the perceived quality of the output images. Ledig et al. [21] proposed the SRGAN network, which uses adversarial losses and content losses to strengthen the generator. Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) [22] refine the adversarial training process by introducing perceived loss and a Residual-in-Residual Dense Block (RRDB). Ma et al. [23] introduced a gradient guidance via an additional branch in the network to alleviate the problem of structural distortion and inconsistency. Li et al. [24] addressed potentially undesired image artifacts using a regional awareness adversarial learning strategy. Liang et al. [25] proposed an LDL framework to regularize adversarial training by clearly distinguishing between visual artifacts and realistic details. These advancements have significantly improved the performance of GAN-based SISR methods, making GANs a popular choice in the field.

3. Method

3.1. Forward Model

Figure 1 illustrates a typical pulsed lidar system based on photon-efficient imaging technology. The system comprises a pulsed laser that emits a pulsed laser beam $s(t)$ at a fixed period T. A SPAD is used to collect the arrival time of the echo signal. In extremely weak echo scenes, the amount of light that can enter the detector during each repetition period is very small, typically less than one photon per pixel. Under these conditions, the impact of pulse pile-up on the detection result is negligible. For each pixel q of the SPAD array during a time period Δt, the expected number Φ_q of photons that can be detected is

$$\Phi_q = \int_0^{\Delta t} [\sigma \alpha_q s(t - \frac{2z_q}{c}) + B] dt, \tag{1}$$

where σ contains the quantum efficiency of SPAD and distance attenuation, c is the speed of light, α_q represents the reflectivity of the target area corresponding to the pixel cell, z_q is the distance between the target area and the detector, and the response caused by environmental noise and dark count is denoted as B.

To avoid producing distance aliasing during detection, it is essential to ensure that $T > 2z_{\max}/c$, where $z_{\max}$ represents the farthest distance of the target. During each illumination pulse repetition cycle, detections by the SPAD are treated as separate events. The probability density function of the avalanche multiplication event of pixel q in a full detection process can be written as follows:

$$f_q(t, z_q) = \frac{\sigma \alpha_q s(t - 2z_q/c) + B}{\sigma \alpha_q \int_{T_0}^{T_1} s(t - 2z_q/c) dt + B(T_1 - T_0)}, \tag{2}$$

where T_0 and T_1 represent the start time and end time of detection.

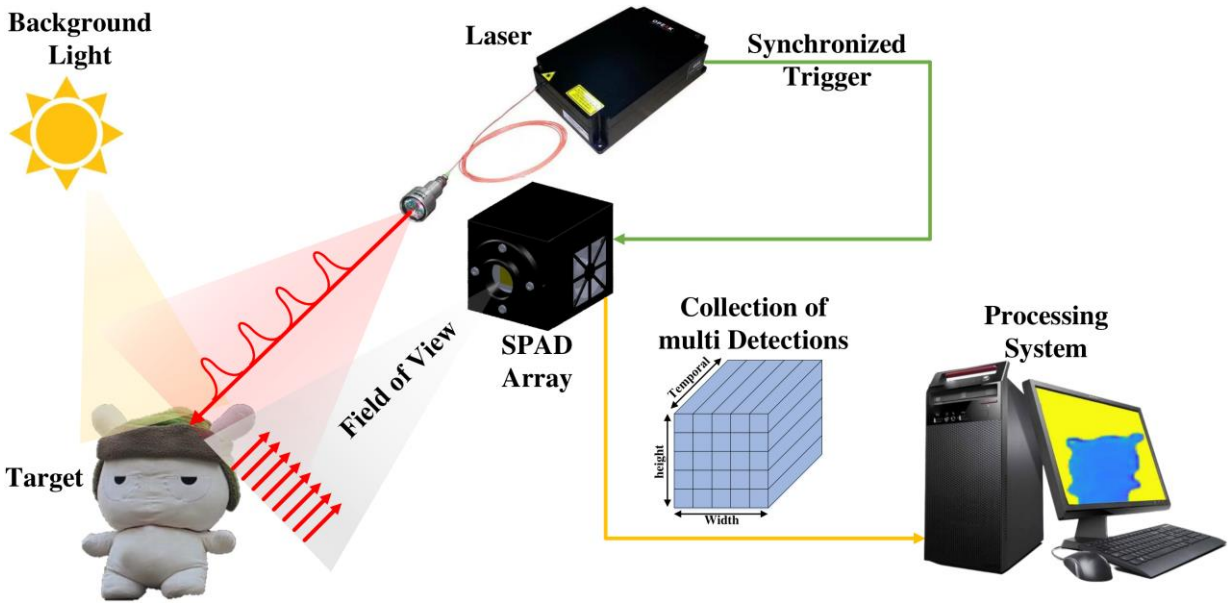

Figure 1. Pulsed LiDAR imaging system based on SPAD.

The SPAD detector generates point cloud data **M** in the low light flux scenario by accumulating many detection results. **M** contains the response of the detector to the echo signal photon and the background noise photon, which can be reconstructed to obtain the range image. By employing the maximum likelihood estimation (MLE) method [29], the range information on q can be expressed as follows:

$$z_q^{ML} = \underset{z_q \in [0,cT)}{\operatorname{argmax}} \sum_{i=1}^{k_q} \log[s(t^{(i)} - 2z_q/c)], \tag{3}$$

where k_q is the total number of photons that the detector responds to on q. The proposed method is detailed in the next section.

3.2. Network Architecture

This research proposes the Temporal–Spatial domain Denoise Super-Resolution (TS-DSR) neural network architecture based on GAN which comprises a generator and a discriminator in order to produce greater quality super-resolution range reconstruction. To reconstruct an image using a neural network, the estimation process can be expressed as follows:

$$\left\{ \mathbf{Y}_{range}^{HR} \right\} = \underset{\theta=\theta^*}{G}(\mathbf{M}), \tag{4}$$

where $G(.)$ represents the generator neural network, θ is the parameters of the neural network, θ^* is the parameters after training, and $\mathbf{Y}_{range}^{HR}$ is the High Resolution (HR) range image of the target scene.

3.2.1. Generator

The structure of the Residual Channel-wise Soft Thresholding (RCST) unit is illustrated in Figure 2. The main body is composed of residual units, and adaptive soft thresholding is added. Equation (2) indicates that in the presence of an echo signal, photon detection events have a higher probability density, and there is a greater chance of obtaining a high value in the statistical detection results. The reactions caused by background noise and dark counts are asymmetric and randomly dispersed. In photon reconstruction imaging tasks with low Signal-to-Background Ratio (SBR), it is essential to eliminate the impact of background noise and dark count.

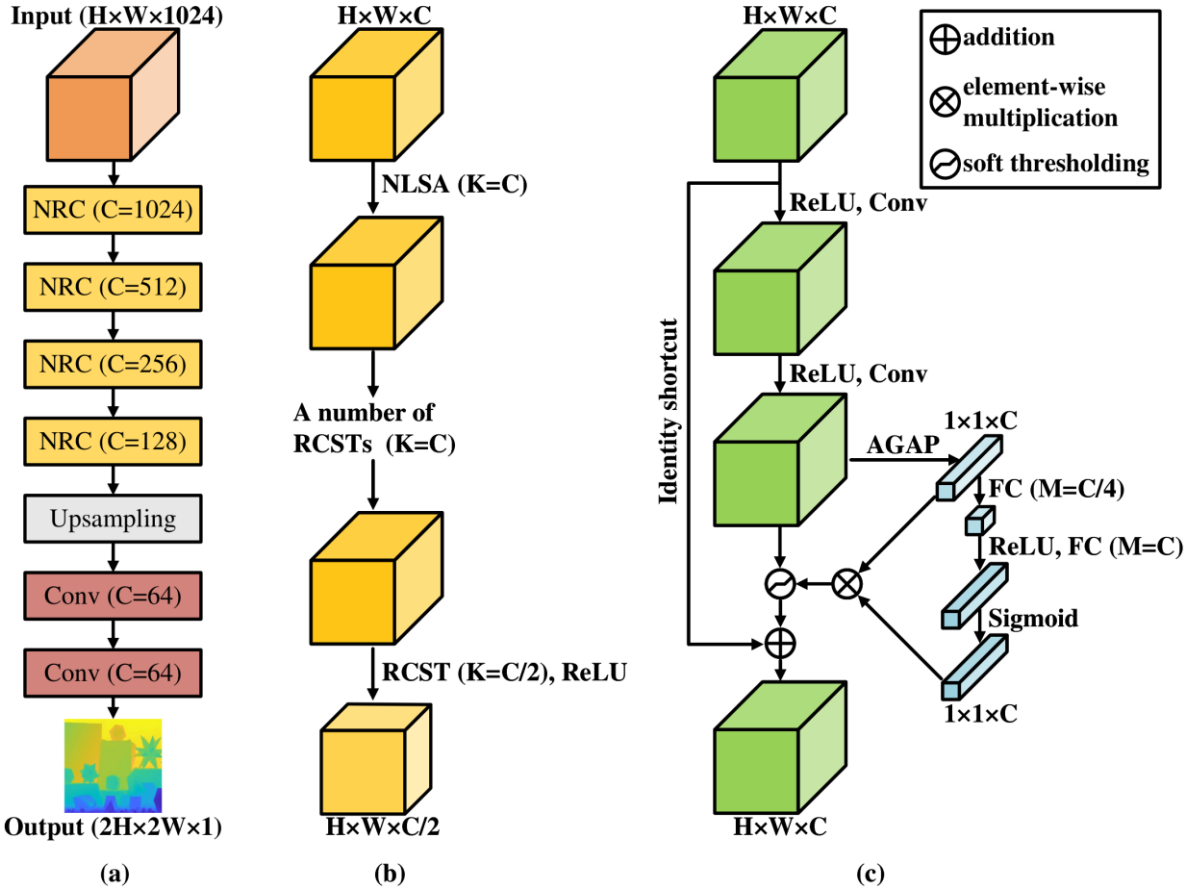

Figure 2. (**a**) Generator general architecture. (**b**) Non-local sparse attention Residual Channel-wise soft thresholding (NRC) module. (**c**) RCST unit.

To achieve denoising, the noise information is transformed into a feature close to zero through convolution, which is then further transformed into zero through soft thresholding. The soft thresholding function can be expressed as follows:

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \le x \le \tau. \\ x + \tau & x < \tau \end{cases} \tag{5}$$

The selection of threshold τ is critical to ensuring optimum denoising performance in RCST. RCST combines soft thresholding with residual units. The feature information is transformed into a C-dimensional vector by absolute value global mean pooling after Rectified Linear Unit (ReLU) activation and convolution processing (AGAP). The vector value is then scaled to the range (0,1) with a sigmoid after two layers of fully connected (FC) processing. The output is multiplied element-wise with the vector output from AGAP processing to obtain a vector of size $1 \times 1 \times C$, where each value represents a separate threshold for the corresponding channel. These thresholds are all positive and can be learned continuously during back propagation. RCST units avoid discarding too much useful information in a single filtering step by employing shortcut connections. By reusing RCST units in generators, noise in features is gradually reduced.

The NRC module integrates multiple functions, including denoising, feature extraction and data compression. With NRC as the main component of the network, the spatial resolution of the data is expanded by the upsampling module when the data are compressed to 64 channels. Finally, the feature data are multiplied by two 3×3 filters to produce an HR range image of the target scene.

3.2.2. Discriminator

In the GAN, the discriminator obtains discriminant information according to the input data distribution and uses the difference between the real image and the generated image discriminant information as the loss guidance generator. Improving the accuracy of the information expressed by the discriminator can prompt the generator to produce more realistic output. However, as a classification network, most discriminators cannot extract different features from the input data simultaneously. When trying to learn global semantics and details at the same time, these discriminators often lose their ability to express themselves effectively. Therefore, global and local feature extraction can be regarded as different tasks. Some studies use multiple independent discriminators to perceive the global and local information of images separately. The network proposed in this paper chooses U-net as the main body of the discriminator and obtains the global and local feature information from it. As shown in Figure 3, the D^U structure is divided into two parts: the encoder performs the classification task to extract the global image features, while the decoder performs the semantic segmentation task to output the discrimination information pixel by pixel.

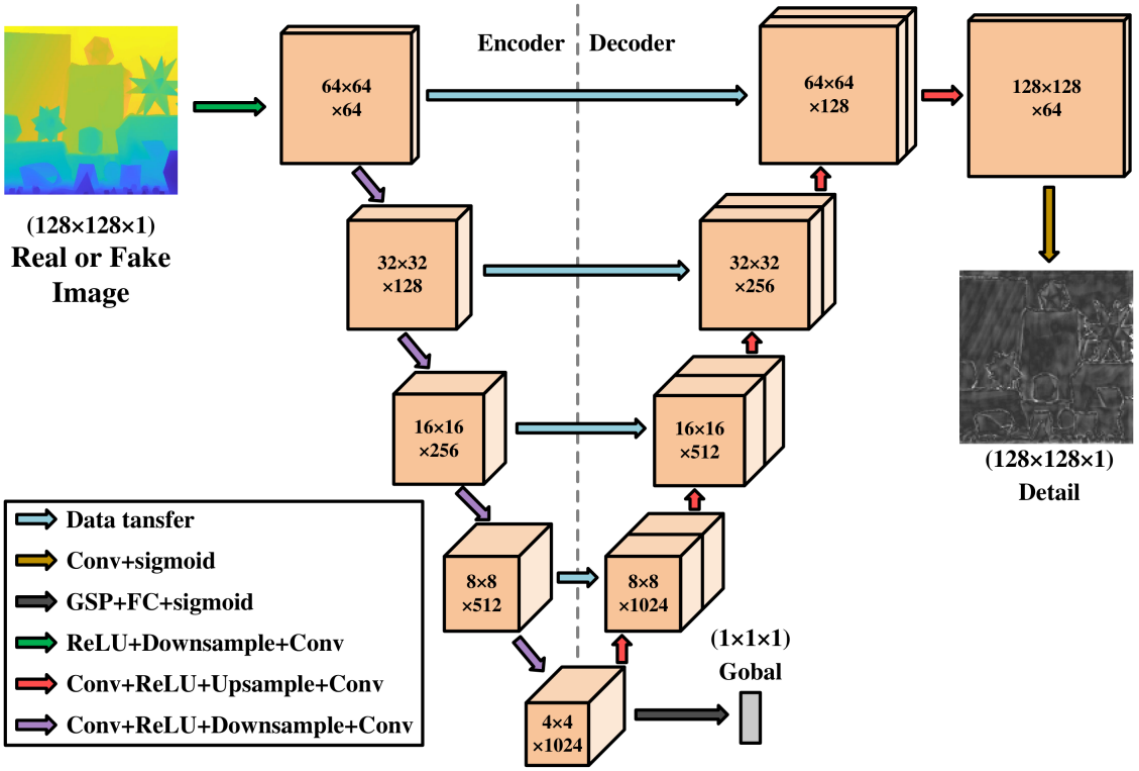

Figure 3. U-Net-based discriminator. Brighter colors in the decoder output correspond to lower confidence (discriminator considers fake).

The encoder consists of multiple convolutions, activation functions, and downsampling operations to map a 128 × 128 × 1 input to a 4 × 4 × 1024 size feature. Global discriminant information is obtained through global summation pooling (GSP) and FC layers. The decoder uses a similar structure to gradually up-sample the encoder output features to the size of the discriminator input to complete the semantic segmentation task. During this process, the features from each layer of the encoder are merged with the decoder output features of the same resolution. By fusing shallow and deep information in this way, details can be better recovered. The global discriminant information is based on the similarity of deep image features. However, in super-resolution tasks, more attention needs to be paid to the image details. D^U fuses encoder and decoder information to better guide the generator in narrowing the detail gap between real and fake samples.

3.2.3. Loss Function

The super-resolution mission is essentially an ill-posed problem. The use of mixed loss function can avoid imposing rigid constraints on Low Resolution (LR) to HR and reduce the difficulty of network training. MAE and adversarial loss are used as hybrid loss functions to guide generators in training. The total loss of the generator can be expressed as

$$L_G = MAE - \lambda_1 \mathbb{E}_{\mathbf{Y}}\left[\lambda_2 \log D^U_{en}(G(\mathbf{M})) + \lambda_3 \log D^U_{de}(G(\mathbf{M}))\right], \tag{6}$$

where $D^U_{en}(\cdot)$ represents the encoder output of the discriminator, $D^U_{de}(\cdot)$ represents the discriminant matrix output by the decoder of the discriminator, n is the total number of pixels, and λ is the hyperparameter. In the process of network training in this paper, λ_1 = 0.005, λ_2 = 0.3 and λ_3 = 0.7. This hyperparameter setting encourages the generator to pay more attention to local details. The loss of the discriminator consists of $L_{D^U_{en}}$ calculated from the output of the encoder and $L_{D^U_{de}}$ calculated from the output of the decoder:

$$L_{D^U_{en}} = -\mathbb{E}_{\mathbf{R}}\left[\log D^U_{en}(\mathbf{R})\right] - \mathbb{E}_{\mathbf{Y}}\left[1 - \log D^U_{en}(G(\mathbf{M}))\right], \tag{7}$$

$$L_{D^U_{de}} = -\mathbb{E}_{\mathbf{R}}\left[\log D^U_{de}(\mathbf{R})\right] - \mathbb{E}_{\mathbf{Y}}\left[1 - \log D^U_{de}(G(\mathbf{M}))\right], \tag{8}$$

where $\mathbf{R}$ is HR ground truth. The total discriminator loss L_D is

$$L_D = L_{D^U_{en}} + L_{D^U_{de}}. \tag{9}$$

4. Experiments

4.1. Dataset and Training Detail

Firstly, the histogram of photon detection probability is established by combining ground truth images of range and reflectivity in the NYU v2 [30] dataset according to Equation (2) Then, the histogram is sampled using a non-homogeneous Poisson process to obtain SPAD measurements. The spatial resolution of simulation data used in training is 64 × 64, with time divided into 1024 intervals, each interval being 80 ps apart. The measurements included 9 different noise scenes, with an average of 2, 10 and 50 background photons per pixel and an average of 2, 5 and 10 signal photons per pixel. A total of 38,624 and 2372 measurements were generated for training and validation, respectively. The deep learning neural network code was implemented using PyTorch. ADAM optimizer was used for iteration in the training process, with a learning rate of 2×10^{-4} and each epoch decaying to 0.95 of the previous one. Network training requires approximately 20 epochs, with a batch size of 4 and approximately 193 k iterations.

4.2. Numerical Simulation

A total of 72 sets of 128 × 128 × 1024 testing data were generated in the simulation experiments, which included 8 scenarios and 9 different noise levels (SBR = 10:2; 10:10; 10:50; 5:2; 5:10; 5:50; 2:2; 2:10; 2:50) provided by the Middlebury [31] dataset. The proposed method was compared with Peng's [12] method, Zhao's [14] method, and NLSA-Encoder [15]. In the simulation experiment, the proposed method achieved end-to-end reconstruction and super-resolution. In contrast, the compared methods first reconstructed the test data to obtain a 128 × 128 LR range image and then performed super-resolution using other methods.

The result of range image reconstruction in low SBR (2:50) Art scene is shown in Figure 4. The first row shows the super-resolution using the Bicubic method, while the second row shows the super-resolution using the ESRGAN neural network method trained on the range reconstruction results. In the numerical simulation experiments, the super-resolution results obtained using bicubic interpolation appear relatively blurry from a human visual perception standpoint.

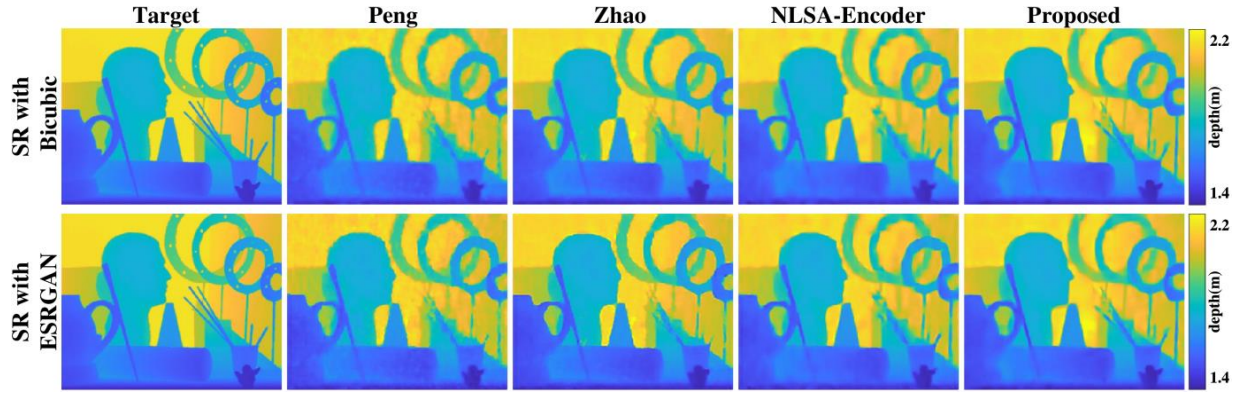

Figure 4. Comparison of different methods for HR range image reconstruction in Art scene. The image resolution is 256 × 256 and the noise environment is 2:50. The "Target" represents the ground truth range image provided in the dataset.

In order to better compare the various reconstruction methods, this paper shows more results with ESRGAN in Figure 5. The Peak Signal-to-Noise Ratio (PSNR) was calculated using the ground truth range image and the HR range reconstruction result. Peng's method outperformed in lower SBR environments due to its use of total variation (TV) as part of the loss. However, TV constraints obscure local details, resulting in less detail in the panes of the refactoring results in the Laundry scene compared to the other methods. Zhao's method uses gradient regularization to penalize the neural network to obtain more accurate edges. In the high-noise environment, more singular points appear in Zhao reconstruction results. This situation shows that this method enhances the expression of some noise as well as the boundary information. The NLSA-Encoder method yields more balanced reconstruction results but lacks some details, such as the panes in the Laundry scene, the small square in the middle of the Moebius scene, and the lower-right corner of the Dolls scene. The proposed method provides better denoising performance and produces more detailed reconstruction results.

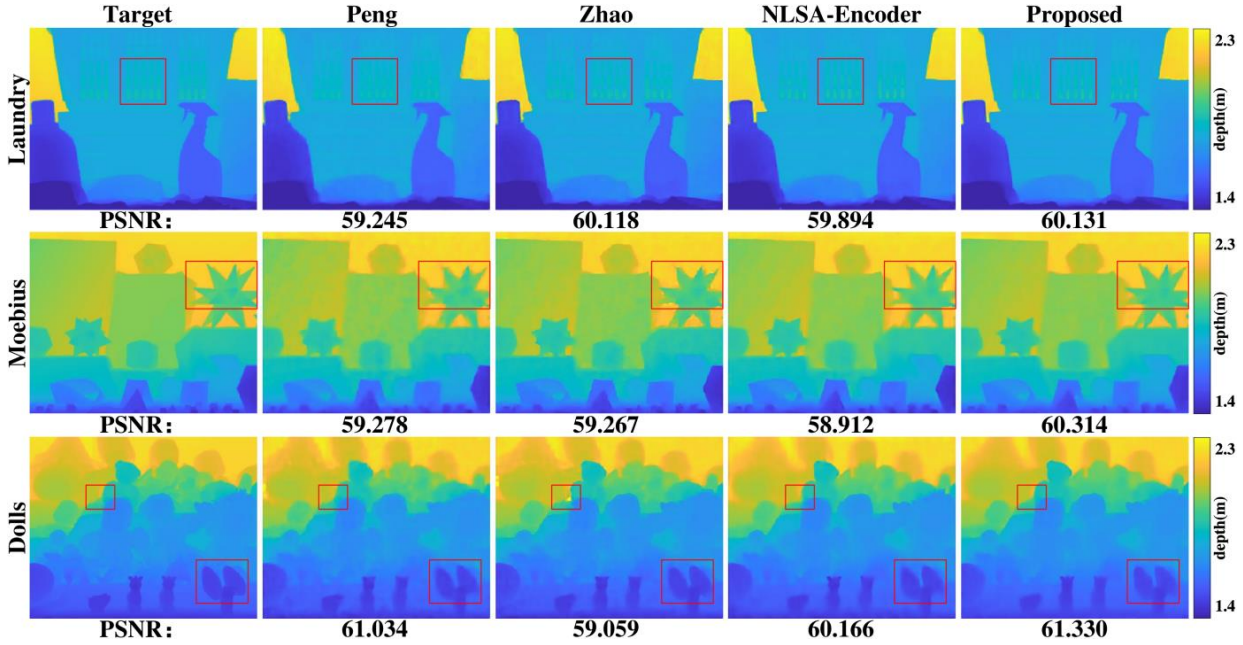

Figure 5. Three scenes comparing different methods of HR range reconstruction image. The image resolution is 256 × 256. The noise environment is 5:10 (Laundry), 2:10 (Moebius) and 5:50 (Dolls). The "Target" represents the ground truth range image provided in the dataset. The content enclosed in the square highlights significant differences observed among the methods.

To compare the reconstruction performance of various methods, four quantitative indicators were selected. The following is the formula for each quantitative index.

Root Mean Square Error (*RMSE*) is represented by the following formula:

$$RMSE(\mathbf{R},\mathbf{Y}) = \sqrt{\frac{1}{n}\sum_1^n (\mathbf{R}_q - \mathbf{Y}_q)^2}, \tag{10}$$

accuracy is represented as follows:

$$Accuracy(\mathbf{C}) = \frac{1}{n}\sum_1^n \mathbf{C}_q, \tag{11}$$

PSNR is formulated as

$$PSNR(\mathbf{R},\mathbf{Y}) = 10\log_{10}\left[z_{\max}^2 / \frac{1}{n}\sum_1^n (\mathbf{R}_q - \mathbf{Y}_q)^2\right], \tag{12}$$

and the Universal Image Quality Index (*UIQI*) [32] can be expressed as follows:

$$UIQI(\mathbf{R},\mathbf{Y}) = \frac{4\sigma_{\mathbf{RY}}\mu_{\mathbf{R}}\mu_{\mathbf{Y}}}{(\sigma_{\mathbf{R}}^2 + \sigma_{\mathbf{Y}}^2)(\mu_{\mathbf{R}}^2 + \mu_{\mathbf{Y}}^2)}, \tag{13}$$

where σ is variance, μ is mean value, n is total number of pixels. The value at any position of the confidence matrix $\mathbf{C}$ determined by the threshold δ can be represented as follows:

$$\mathbf{C}_q(\mathbf{R},\mathbf{Y}) = \begin{cases} 0 & \max(\mathbf{R}_q/\mathbf{Y}_q, \mathbf{Y}_q/\mathbf{R}_q) > \delta \\ 1 & \max(\mathbf{R}_q/\mathbf{Y}_q, \mathbf{Y}_q/\mathbf{R}_q) < \delta \end{cases}, \tag{14}$$

Table 1 lists the average values of the indicators for each method in each noise environment. In terms of quantitative indexes, Peng's method has poor overall performance. Zhao's method performs well in environments with high signal photon numbers but is susceptible to noise. NLSA-Encoder is more prominent in low-SBR environments. The average performance of the proposed network is the best. This method performs best or second-best in all kinds of noise environments. The results show that TSDSR has the ability to accurately super-resolution reconstruct the range image.

Table 1. Under 9 different SBRs conditions, the average performance of various reconstruction methods was quantitatively evaluated across 8 different test scenes. A lower value of RMSE (Root Mean Squared Error) indicates a closer resemblance of the reconstructed image to the ground truth image. Higher values of accuracy, PSNR, and UIQI indicate a closer resemblance of the reconstructed image to the ground truth image. Optimal performance is represented by bold characters. The abbreviation "NSE" refers to the NLSA-Encoder method.

SBR	RMSE				Accuracy (δ = 1.03)			
	Peng	Zhao	NSE	Proposed	Peng	Zhao	NSE	Proposed
10:2	0.0205	**0.0174**	0.0211	0.0189	0.9763	0.9786	0.9798	**0.9788**
10:10	0.0204	**0.0177**	0.0211	0.0191	0.9759	0.9785	0.9796	**0.9789**
10:50	0.0207	**0.0198**	0.0213	0.0194	0.9759	0.9779	0.9791	**0.9784**
5:2	0.0218	**0.0189**	0.0214	0.0201	0.9743	0.9770	0.9781	**0.9783**
5:10	0.0220	0.0212	0.0217	**0.0205**	0.9740	0.9765	0.9770	**0.9783**
5:50	0.0233	0.0278	0.0229	**0.0217**	0.9723	0.9698	0.9747	**0.9771**
2:2	0.0245	0.0232	0.0234	**0.0224**	0.9694	0.9720	0.9739	**0.9756**
2:10	0.0262	0.0355	0.0248	**0.0241**	0.9660	0.9672	0.9700	**0.9728**
2:50	0.0318	0.0364	0.0308	**0.0304**	0.9534	**0.9624**	0.9538	0.9592
AVG	0.0235	0.0242	0.0232	**0.0218**	0.9708	0.9733	0.9740	**0.9753**

Table 1. *Cont.*

SBR	PSNR				UIQI			
	Peng	**Zhao**	**NSE**	**Proposed**	**Peng**	**Zhao**	**NSE**	**Proposed**
10:2	60.729	**62.188**	61.009	61.192	0.9791	**0.9825**	0.9811	0.9823
10:10	60.691	**62.119**	60.837	61.194	0.9788	**0.9835**	0.9813	0.9831
10:50	60.316	58.082	60.644	**60.757**	0.9740	0.9803	0.9804	**0.9809**
5:2	59.874	60.159	60.265	**60.903**	0.9758	0.9799	0.9792	**0.9813**
5:10	59.780	57.824	59.973	**60.661**	0.9752	0.9781	0.9790	**0.9805**
5:50	59.004	57.576	59.336	**59.926**	0.9717	0.9705	0.9760	**0.9786**
2:2	58.219	58.753	58.933	**59.616**	0.9692	0.9727	0.9752	**0.9761**
2:10	57.465	55.249	58.062	**58.388**	0.9649	0.9563	0.9696	**0.9717**
2:50	55.093	54.574	**55.415**	55.264	0.9492	0.9508	0.9523	**0.9524**
AVG	59.019	58.503	59.386	**59.767**	0.9709	0.9727	0.9749	**0.9763**

4.3. Real-World Experiments

To validate the performance of the proposed method, a SPAD-based pulsed laser system was constructed as shown in Figure 1 to collect real-world experimental data. The fiber pulse laser emitted pulses with a pulse width of 1 ns, a peak power of 500 mW (near-field) or 1.5 mJ (far-field), and a wavelength of 1064 nm at a repetition rate of 20 KHz. The laser beam was diffused by an external lens, resulting in a beam divergence angle of 25 mrad (near-field) or 15 mrad (far-field). The laser transmitter triggered the SPAD synchronously with the pulse to detect the flight time of the echo photon. The SPAD model used was the GD5551 InGaAs, with a trigger exposure time of 4096 ns, time resolution of 1 ns, and spatial resolution of 64 $\times$ 64.

The target scene photos and reconstruction results of the real experiment data are illustrated in Figure 6. In the reconstruction process, MLE utilized 20,000 frames of data to reconstruct 64 $\times$ 64 LR range images as a reference, while the other methods used only 400 frames to reconstruct HR range images. Peng's method failed to filter out the noise, whereas Zhao's method restored more edge information, but the left-hand reconstruction of the middle doll was less effective. Although the NLSA-Encoder did not have much noise, it struggled to accurately capture details in certain weak echo regions, such as the edges of the dolls and flagpoles. The NLSA-Encoder-reconstructed rabbit doll on the left is smaller than the result of MLE method.

TSDSR showed the highest quality for the reconstructed data in the table part of the image. The shape of the right side of the table and the signboard on the rooftop were well restored. These details were difficult to observe in the LR reconstructed images of the MLE method, indicating the importance of super-resolution. The qualitative comparison of the reconstruction results of real-world experiments proved that the proposed method outperformed other methods in super-resolution range reconstruction.

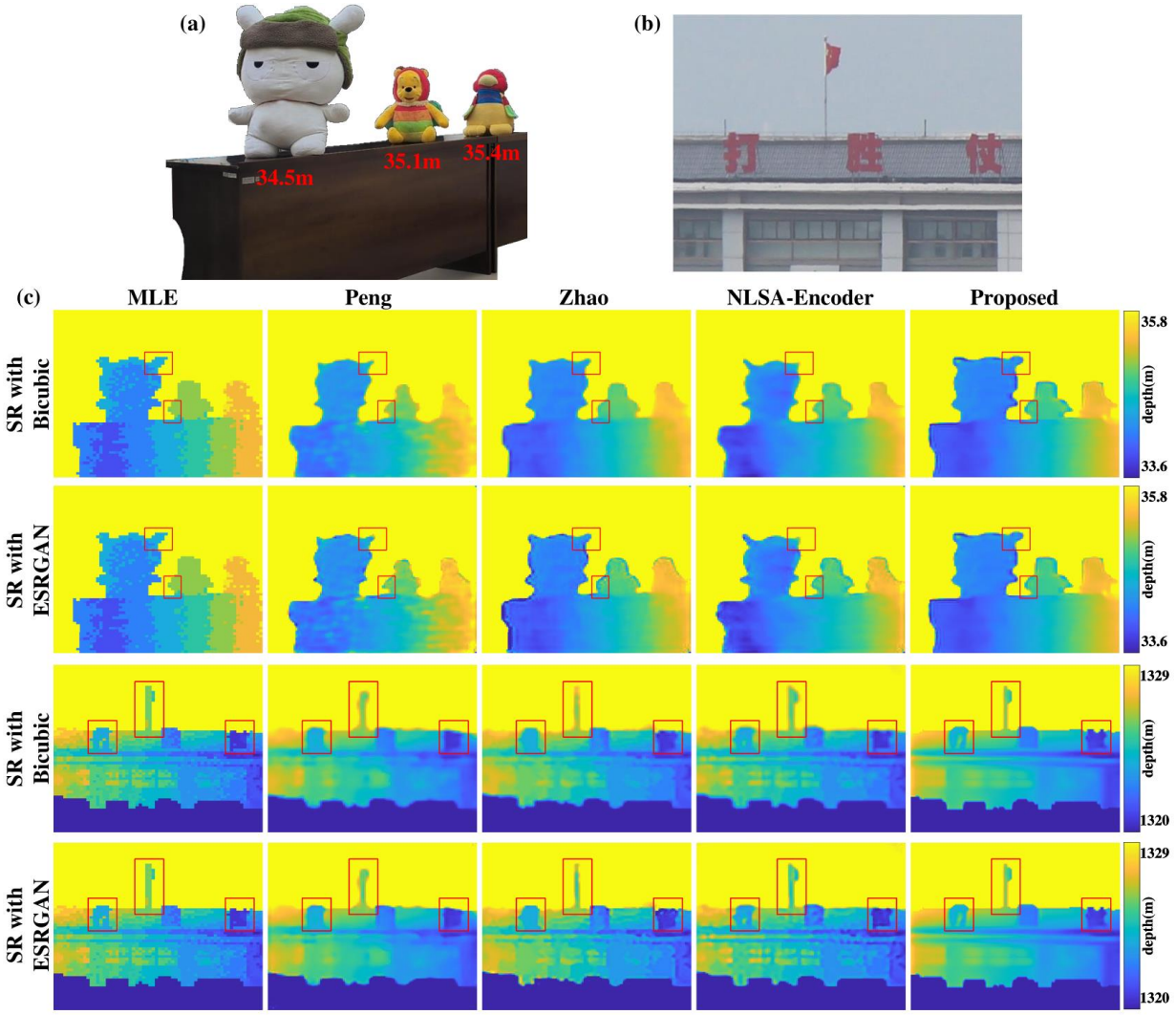

Figure 6. Real-world experiment target scenes and reconstruction results. The resolution of the range image by MLE method is 64×64. The resolution of the other images is 128×128. The last column of images was reconstructed by TSDSR. The content enclosed in the square highlights significant differences observed among the methods. (**a**) Near-field experimental target scene. (**b**) Far-field experimental target scene. (**c**) Reconstruction results of near-field experimental detection data by various methods.

5. Discussion

Photon-efficient imaging is limited by significant noise and photon detection hardware. In this paper, an end-to-end network is proposed for super-resolution range reconstruction from photon-efficient measurements. To thoroughly examine the contribution of each module to the network's performance, three network architectures were compared: one without adaptive soft thresholding (A w/o AST), one without D^U (B w/o D^U), and one containing all components (C Proposed).

The ablation experiment used the same test data as the simulation experiment. Figure 7 shows the reconstruction results under low SBR (2:50) Art and Laundry scenes. The comparison highlights that the network's reconstruction results without using adaptive soft thresholding introduce a lot of noise. While the adaptive soft thresholding module's denoising capabilities are evident, it also leads to the problem of detail loss. Among the panes in the Laundry scene, the A network, which does not use adaptive soft thresholding, displays the most detailed results. During adaptive soft thresholding, some useful information may be accidentally filtered out. Choosing an appropriate threshold can reduce the likelihood

of signal photon filtering, resulting in a higher-quality reconstructed image. In both test scenarios, the proposed network demonstrated outstanding performance in terms of image reconstruction quality and PSNR. This is because the D^U penalty helps the network achieve a more reasonable threshold, which better balances the trade-off between detail recovery and denoising performance.

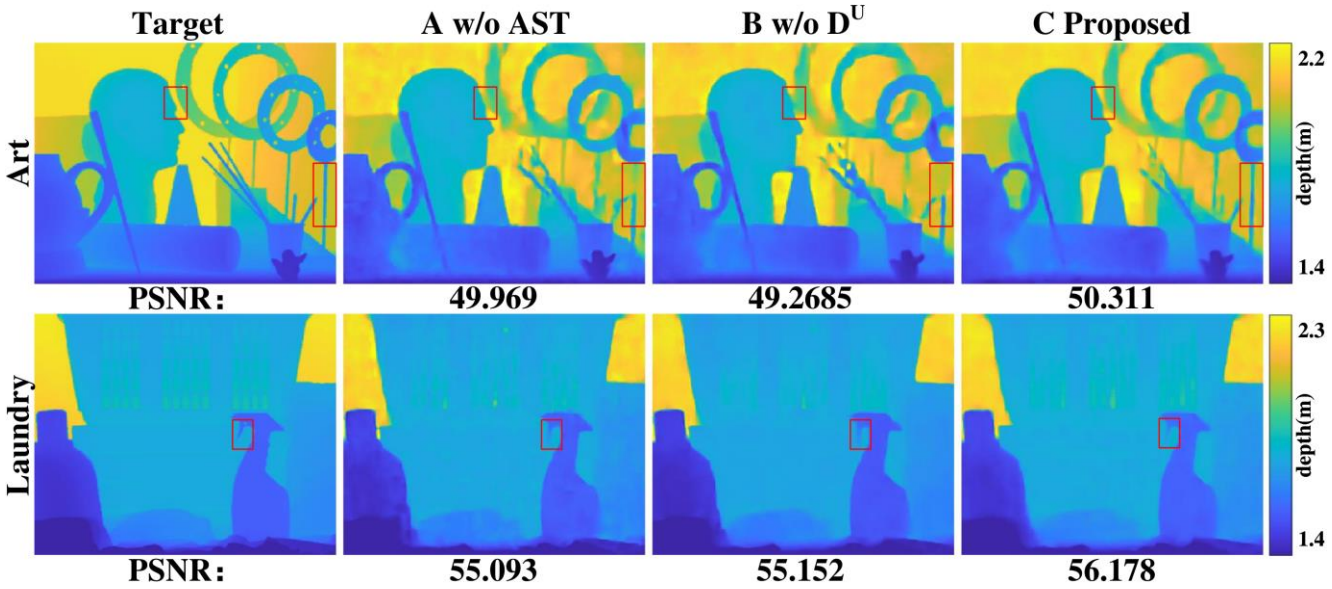

Figure 7. Art and Laundry scenes comparing different methods of HR range reconstruction image. The image resolution is 256 × 256 and the noise environment is 2:50. The "Target" represents the ground truth range image provided in the dataset. The content enclosed in the square highlights significant differences observed among the methods.

Compared to existing convolutional neural network approaches, the employed technique involves learnable soft thresholding to effectively filter out the substantial amount of background noise present in the original data captured by the SPAD detector. Furthermore, a Unet-based discriminator is introduced to guide the network in generating high-fidelity, high-resolution range images. This discriminator serves as both a classifier and a segmenter, enhancing the preservation of the signal while minimizing potential information loss due to soft thresholding. In numerical simulation experiments, the network effectively preserves fine details. However, in real-world experiments, there is an occurrence of unexpected filtering of some echo signals. For instance, in the window part of the far-field experiment (Figure 6), the resulting image appears excessively blurred, potentially due to the negative impact of the introduced soft thresholding. Future work can focus on further optimizing the thresholding process or selecting more suitable loss functions to enhance the network's reconstruction performance.

6. Conclusions

To address the problem of poor single photon imaging quality caused by the limited spatial resolution of the SPAD array, a GAN-based neural network called TSDSR is utilized for super-resolution reconstruction. This method uses a residual structure and an NLSA module for feature extraction. The discriminator employs the Unet architecture, with the encoder performing classification by image and the decoder by pixel. This architectural improvement results in a stronger discriminator, which is encouraged to maintain a more powerful data representation, making it more difficult for the generator to deceive the discriminator and thereby enhancing the quality of generated samples. Soft threshold processing is added to the residual module to lessen the negative impact of background noise on reconstruction quality. The network's high reconstruction performance is vali-

dated through numerical simulations of various scenarios, and ablation experiments are conducted to examine the influence of each module on network performance. Furthermore, the proposed method has been evaluated in real-world tests. Compared to other existing deep learning methods, the proposed network can reconstruct HR range images of the target with higher quality.

Author Contributions: Conceptualization, Z.T. and X.Y.; methodology, Z.T.; software, Z.T.; validation, Z.T., X.J. and X.Y.; formal analysis, J.H.; investigation, L.X.; resources, X.Y.; data curation, L.W.; writing—original draft preparation, Z.T.; writing—review and editing, X.Y. and X.J.; visualization, X.J.; supervision, B.Z.; project administration, X.Y.; funding acquisition B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Z.-P.; Ye, J.-T.; Huang, X.; Jiang, P.-Y.; Cao, Y.; Hong, Y.; Yu, C.; Zhang, J.; Zhang, Q.; Peng, C.-Z.; et al. Single-Photon Imaging over 200 km. *Optica* **2021**, *8*, 344–349. [CrossRef]
2. Yang, J.; Li, J.; He, S.; Wang, L.V. Angular-Spectrum Modeling of Focusing Light inside Scattering Media by Optical Phase Conjugation. *Optica* **2019**, *6*, 250–256. [CrossRef] [PubMed]
3. Maccarone, A.; McCarthy, A.; Ren, X.; Warburton, R.E.; Wallace, A.M.; Moffat, J.; Petillot, Y.; Buller, G.S. Underwater Depth Imaging Using Time-Correlated Single-Photon Counting. *Opt. Express* **2015**, *23*, 33911–33926. [CrossRef] [PubMed]
4. 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 Technol. Lett.* **2009**, *21*, 1020–1022. [CrossRef]
5. Shin, D.; Kirmani, A.; Goyal, V.K.; Shapiro, J.H. Photon-Efficient Computational 3-D and Reflectivity Imaging With Single-Photon Detectors. *IEEE Trans. Comput. Imaging* **2015**, *1*, 112–125. [CrossRef]
6. Shin, D.; Xu, F.; Venkatraman, D.; Lussana, R.; Villa, F.; Zappa, F.; Goyal, V.K.; Wong, F.N.C.; Shapiro, J.H. Photon-Efficient Imaging with a Single-Photon Camera. *Nat. Commun.* **2016**, *7*, 12046. [CrossRef] [PubMed]
7. Rapp, J.; Goyal, V.K. A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging. *IEEE Trans. Comput. Imaging* **2017**, *3*, 445–459. [CrossRef]
8. Halimi, A.; Maccarone, A.; Lamb, R.A.; Buller, G.S.; McLaughlin, S. Robust and Guided Bayesian Reconstruction of Single-Photon 3D Lidar Data: Application to Multispectral and Underwater Imaging. *IEEE Trans. Comput. Imaging* **2021**, *7*, 961–974. [CrossRef]
9. Chen, S.; Halimi, A.; Ren, X.; McCarthy, A.; Su, X.; McLaughlin, S.; Buller, G.S. Learning Non-Local Spatial Correlations To Restore Sparse 3D Single-Photon Data. *IEEE Trans. Image Process.* **2020**, *29*, 3119–3131. [CrossRef] [PubMed]
10. Chen, Y.; Yao, G.; Liu, Y.; Su, H.; Hu, X.; Pan, Y. Deep Domain Adversarial Adaptation for Photon-Efficient Imaging. *Phys. Rev. Appl.* **2022**, *18*, 54048. [CrossRef]
11. Lindell, D.B.; O'Toole, M.; Wetzstein, G. Single-Photon 3D Imaging with Deep Sensor Fusion. *ACM Trans. Graph.* **2018**, *37*, 111–113. [CrossRef]
12. Peng, J.; Xiong, Z.; Huang, X.; Li, Z.-P.; Liu, D.; Xu, F. Photon-Efficient 3d Imaging with a Non-Local Neural Network. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Volume 16, pp. 225–241.
13. Zang, Z.; Xiao, D.; Li, D.D.-U. Non-Fusion Time-Resolved Depth Image Reconstruction Using a Highly Efficient Neural Network Architecture. *Opt. Express* **2021**, *29*, 19278–19291. [CrossRef] [PubMed]
14. Zhao, X.; Jiang, X.; Han, A.; Mao, T.; He, W.; Chen, Q. Photon-Efficient 3D Reconstruction Employing a Edge Enhancement Method. *Opt. Express* **2022**, *30*, 1555–1569. [CrossRef] [PubMed]
15. Yang, X.; Tong, Z.; Jiang, P.; Xu, L.; Wu, L.; Hu, J.; Yang, C.; Zhang, W.; Zhang, Y.; Zhang, J. Deep-Learning Based Photon-Efficient 3D and Reflectivity Imaging with a 64 64 Single-Photon Avalanche Detector Array. *Opt. Express* **2022**, *30*, 32948–32964. [CrossRef] [PubMed]
16. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Volume 13, pp. 184–199.
17. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.

18. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
19. Hussein, S.A.; Tirer, T.; Giryes, R. Correction Filter for Single Image Super-Resolution: Robustifying off-the-Shelf Deep Super-Resolvers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1428–1437.
20. Wang, L.; Dong, X.; Wang, Y.; Ying, X.; Lin, Z.; An, W.; Guo, Y. Exploring Sparsity in Image Super-Resolution for Efficient Inference. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4917–4926.
21. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
22. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European conference on computer vision (ECCV) workshops, Munich, Germany, 8–14 September 2018.
23. Ma, C.; Rao, Y.; Cheng, Y.; Chen, C.; Lu, J.; Zhou, J. Structure-Preserving Super Resolution with Gradient Guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7769–7778.
24. Li, W.; Zhou, K.; Qi, L.; Lu, L.; Lu, J. Best-Buddy Gans for Highly Detailed Image Super-Resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1412–1420.
25. Liang, J.; Zeng, H.; Zhang, L. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5657–5666.
26. Mei, Y.; Fan, Y.; Zhou, Y. Image Super-Resolution with Non-Local Sparse Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3517–3526.
27. Zhao, M.; Zhong, S.; Fu, X.; Tang, B.; Pecht, M. Deep Residual Shrinkage Networks for Fault Diagnosis. *IEEE Trans. Ind. Informatics* **2019**, *16*, 4681–4690. [CrossRef]
28. Schonfeld, E.; Schiele, B.; Khoreva, A. A U-Net Based Discriminator for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8207–8216.
29. Houwink, Q.; Kalisvaart, D.; Hung, S.-T.; Cnossen, J.; Fan, D.; Mos, P.; Ülkü, A.C.; Bruschini, C.; Charbon, E.; Smith, C.S. Theoretical Minimum Uncertainty of Single-Molecule Localizations Using a Single-Photon Avalanche Diode Array. *Opt. Express* **2021**, *29*, 39920–39929. [CrossRef] [PubMed]
30. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from Rgbd Images. *ECCV* **2012**, *7576*, 746–760.
31. Scharstein, D.; Pal, C. Learning Conditional Random Fields for Stereo. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
32. Wang, Z.; Bovik, A.C. A Universal Image Quality Index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.