Spatial Hierarchy Aware Residual Pyramid Network for Time-of-Flight Depth Denoising

Guanting Dong, Yueyi Zhang(✉), and Zhiwei Xiong

University of Science and Technology of China, Hefei, China
gtdong@mail.ustc.edu.cn, {zhyuey,zwxiong}@ustc.edu.cn

Abstract. Time-of-Flight (ToF) sensors have been increasingly used on mobile devices for depth sensing. However, the existence of noise, such as Multi-Path Interference (MPI) and shot noise, degrades the ToF imaging quality. Previous CNN-based methods remove ToF depth noise without considering the spatial hierarchical structure of the scene, which leads to failures in obtaining high quality depth images from a complex scene. In this paper, we propose a Spatial Hierarchy Aware Residual Pyramid Network, called SHARP-Net, to remove the depth noise by fully exploiting the geometry information of the scene in different scales. SHARP-Net first introduces a Residual Regression Module, which utilizes the depth images and amplitude images as the input, to calculate the depth residual progressively. Then, a Residual Fusion Module, summing over depth residuals from all scales, is imported to refine the depth residual by fusing multi-scale geometry information. Finally, shot noise is further eliminated by a Kernel Prediction Network. Experimental results demonstrate that our method significantly outperforms state-of-the-art ToF depth denoising methods on both synthetic and realistic datasets. The source code is available at https://github.com/ashesknight/tof-mpi-remove.

Keywords: Time-of-Flight · Multi-Path Interference · Spatial hierarchy · Residual pyramid · Depth denoising

1 Introduction

Depth plays an important role in current research, especially in the field of computer vision. In the past decades, researchers have proposed various methods to obtain depth [22,29,30], among which Time-of-Flight (ToF) technology is becoming increasingly popular for depth sensing. Many successful consumer products, such as Kinect One [21], are equipped with ToF sensors, providing high quality depth image. These devices further promote many applications in

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58586-0_3) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12369, pp. 35–50, 2020.
https://doi.org/10.1007/978-3-030-58586-0_3

computer vision areas, for example scene understanding, action recognition and human-computer interaction. However, ToF depth images suffer from various noises, such as Multi-Path Interference (MPI) and shot noise, which limit the applicability of ToF imaging technologies.

ToF depth images are vulnerable to MPI noise, which originates in the fact that numerous multi-bounce lights are collected by one pixel during the exposure time. The existence of MPI breaks the key assumption that the receiving light is only reflected once in the scene and results in serious ToF depth error. Shot noise, a common and inevitable noise caused by sensor electronics, is another source of ToF depth noise. Figure 1 shows the depth error maps caused by shot noise and MPI noise respectively. It can be seen that both shot noise and MPI noise are widespread in ToF depth images but MPI noise is significantly intense in several regions such as corner and edge areas.

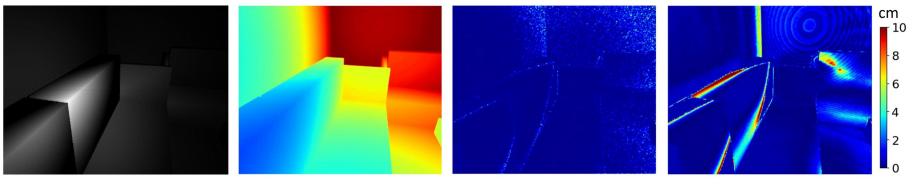

Fig. 1. (a) ToF amplitude image. (b) ToF ground truth depth. (c) Depth error map caused by shot noise. (d) Depth error map caused by MPI. The example comes from a synthetic dataset.

Recently, many Convolutional Neural Networks (CNN) based methods have been proposed for MPI removal in ToF sensors [17,23,26]. The fundamental theory of these CNN based methods is that the MPI noise of a pixel can be estimated as a linear combination of information from surrounding pixels. In the image space, CNN is a proper way to model this linear combination process with spatial convolution and achieves encouraging results. To fit the unknown parameters of convolution kernel, supervised learning is often utilized and the ground-truth labels without MPI of scenes are required. Since it is difficult to get the ground truth depth of realistic scenes, many synthetic ToF datasets are introduced for the training and testing of neural networks. Usually, these datasets consist of ToF depth images as well as corresponding amplitude images. Some datasets even contain the raw measurements of ToF sensors and color images, both of which are usually captured by the calibrated RGBD camera.

The large-scale datasets make it possible to learn the linear combination process of light transport through CNN based methods. However, the existing CNN based methods still have some limitations. Especially, the elimination of MPI noise for a complex scene is not satisfying. Specifically, in a complex scene, many objects with different shapes and sizes are located close to each other. In this case, each pixel of the ToF sensor may collect many light signals which

are from various indirect light paths, which easily leads to intense MPI noise. Eliminating MPI noise in a complex scene still remains a challenging problem and needs more investigation.

A key observation is that in a scene, the objects usually have spatial hierarchical structures. For example, a showcase, a dog toy and the head of the dog toy can formulate a hierarchical relationship. In this case, the depth value of a point located at the surface of any object is usually affected by these three interrelated objects. In a complex scene with large-size shapes and detailed structures, there should be more diverse hierarchical relationships. And previous works have demonstrated that utilizing the hierarchical representations of the scene can lead to improvement in computer vision filed such as scene understanding [24,27], image embedding [4], image denoising [20], object detection [19], depth and 3D shape estimation [6,18]. Aforementioned works inspire us to explicitly utilize the spatial hierarchical relationships to improve the result of the MPI removal for ToF depth.

In this paper, we propose a Spatial Hierarchy Aware Residual Pyramid Network (SHARP-Net) to fully exploit scene structures in multiple scales for ToF depth denoising. The spatial hierarchical structure of the scene, in the forms of a feature pyramid with multiple scales, can provide a proper receptive field and more ample geometric relationships between the objects of the scene for the network, which improves the performance of noise removal.

Within SHARP-Net, a Residual Regression Module is first introduced, which consists of a feature extractor to build a feature pyramid and residual regression blocks to establish a depth residual pyramid in a coarse-to-fine manner. At upper levels of the residual pyramid, the depth residual maps represent MPI noise regressed by utilizing global geometry information. At lower levels, the depth residual maps describe subtle MPI effects by considering local scene structures. The Residual Regression Module pushes every level to utilize the available hierarchical relationships of the current level and deeply extracts the geometric information lying in the corresponding hierarchy of the scene. The geometric information obtained in different scales give excellent hints for estimating the MPI noise. Our proposed Residual Regression Module generates a depth residual map for each level, which is much different from the widely used U-Net structure. After going through Residual Regression Module, a depth residual pyramid is obtained to represent MPI estimation corresponding to the hierarchical structure of the scene. In order to further optimize the performance of SHARP-Net on both large-size shapes and detailed structures, we propose a Residual Fusion Module to explicitly choose predominant components by summing over the depth residuals from all scales. Finally, we employ a Depth Refinement Module, which is based on a Kernel Prediction Network, to remove shot noise and refine depth images.

Combining the Residual Regression Module, Residual Fusion Module and Depth Refinement Module, our SHARP-Net accurately removes noise for ToF depth images, especially MPI noise and shot noise. In short, we make the following contributions:

- We propose a Residual Regression Module to explicitly exploit the spatial hierarchical structure of the scene to accurately remove MPI noise and shot noise in large-size shapes and detailed structures simultaneously.
- We propose a Residual Fusion Module to selectively integrate the geometric information in different scales to further correct MPI noise, and introduce a Depth Refinement Module to effectively eliminate the shot noise.
- The proposed SHARP-Net significantly outperforms the state-of-the-art methods in the quantitative and qualitative comparison for ToF depth denoising on both the synthetic and realistic datasets.

2 Related Work

ToF imaging is affected by noise from different sources, such as shot noise and MPI noise [15,28]. Shot noise is caused by sensor electronics, which appears in all sensors. Shot noise removal for ToF sensors is well investigated. Traditional filtering algorithms, such as bilateral filtering, are able to eliminate shot noise effectively [2,16]. In contrast, MPI removal is a more difficult problem in ToF depth denoising. Many physics-based and learning-based MPI removal methods have been proposed.

For physics-based methods, Fuchs *et al.* conduct a series of studies to estimate MPI noise in the scene, from using single modulation frequency [9] to considering multiple albedos and reflections [10,14]. Feigin *et al.* propose a multi-frequency method to correct MPI through comparing the pixel-level changes of the raw measurements at different frequencies [5]. Gupta *et al.* study the impact of modulation frequencies on MPI and propose a phasor imaging method by emitting two signals with frequencies of great differences [12]. Freedman *et al.* propose a model based on a compressible backscattering representation to tackle the multi-path with more than two paths and achieve real-time processing speed [8].

For learning-based methods, Marco *et al.* exploit the transient imaging technology [13] to simulate the generation of MPI noise in ToF imaging process and produce a large dataset for ToF depth denoising. They also propose a two-stage deep neural network to refine ToF depth images [17]. Su *et al.* propose a deep end-to-end network for ToF depth denoising with raw correlation measurements as the input [26]. Guo *et al.* produce a large-scale ToF dataset FLAT, and introduce a kernel prediction network to remove MPI and shot noise [11]. To overcome the domain shift between the unlabelled realistic scene and the synthetic training dataset, Agresti *et al.* exploit an adversarial learning strategy, based on the generative adversarial network, to perform an unsupervised domain adaptation from the synthetic dataset to realistic scenes [2]. Qiu *et al.* take into account the corresponding RGB images provided by the RGB-D camera and propose a deep end-to-end network for camera alignment and ToF depth refinement [23].

Recently, residual pyramid methods have been adopted for a variety of computer vision tasks. For stereo matching, Song *et al.* build a residual pyramid to solve the degradation of depth images in tough areas, such as non-texture areas, boundary areas and tiny details [25]. For the monocular depth estimation,

Chen *et al.* propose a structure-aware residual pyramid to recover the depth image with high visual quality in a coarse-to-fine manner [6]. For image segmentation, Chen *et al.* propose a residual pyramid network to learn the main and residual segmentation in different scales [7]. For image super-resolution, Zheng *et al.* employ a joint residual pyramid network to effectively enlarge the receptive fields [31]. Our SHARP-Net refers to residual pyramid methods as well and achieve success in ToF depth denoising, which will be explained in detail of the following sections. To the best of our knowledge, SHARP-Net is the first work to apply residual pyramid to ToF depth denoising, which greatly surpasses existing methods by integrating spatial hierarchy.

3 ToF Imaging Model

In this section, we briefly introduce the mathematical models of ToF imaging and MPI.

With a single modulation frequency f_ω and four-step phase-shifted measurements $r_i\,(i = 1, 2, 3, 4)$, the depth d at each pixel is computed as

$$d = \frac{c}{4\pi f_\omega} \arctan\left(\frac{r_4 - r_2}{r_1 - r_3}\right), \tag{1}$$

where c is the speed of light in the vacuum. Under the ideal condition, it is assumed that a single light pulse is reflected only once in the scene and captured by a pixel (x, y) on the sensor. So the raw correlation measurement r_i can be modeled as

$$r_i(x, y) = \int_0^T s(t) b \cos(\omega t - \psi_i) dt, \tag{2}$$

where $s(t)$ is the received signal, $b\cos(\omega t - \psi_i)$ is the referenced periodic signal, ψ_i is the phase offset and T is the exposure temporal range.

In real world, MPI noise always exists. In this case, the received signal is changed to $\hat{s}(t)$, which can be described as

$$\hat{s}(t) = s(t) + \sum_{p \in P} s_p(t), \tag{3}$$

where P is the set of all the light paths p followed by indirectly received signals. Here indirectly received signals $s_p(t)$ represent the captured signals which are reflected multiple bounces after being emitted to the scene. The difference between $s(t)$ and $\hat{s}(t)$ further leads to a deviation to the depth d. In our proposed network, we call this deviation the depth residual. To better regress the depth residual, we bring in a residual pyramid to estimate MPI noise in multiple scales. At different levels of the pyramid, the deviation induced by the set P is regressed and further optimized by our network.

4 Spatial Hierarchy Aware Residual Pyramid Network

Our proposed Spatial Hierarchy Aware Residual Pyramid Network (SHARP-Net) consists of three parts: a Residual Regression Module as the backbone for multi-scale feature extraction, a Residual Fusion Module and a Depth Refinement Module to optimize the performance. The flowchart of SHARP-Net is shown in Fig. 2. The following subsections explain these three parts respectively.

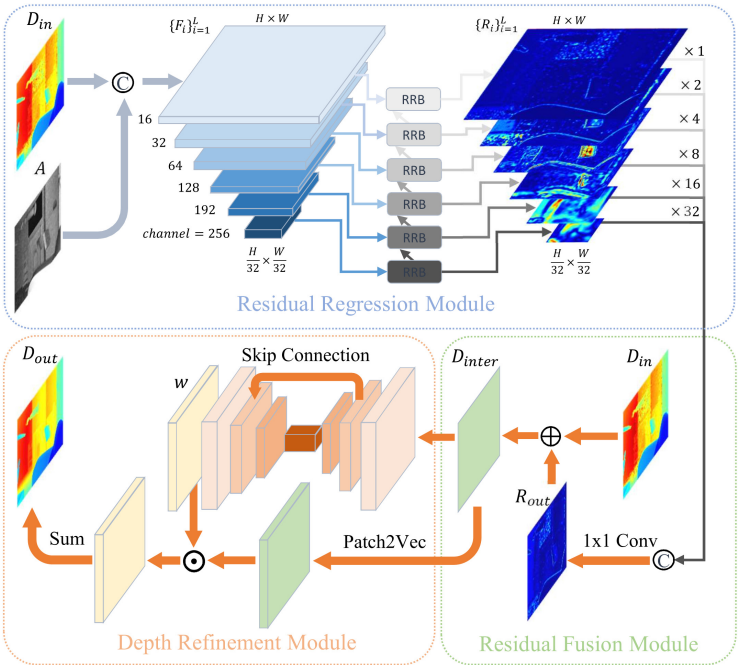

Fig. 2. Flowchart of Spatial Hierarchy Aware Residual Pyramid Network (SHARP-Net). Here $\odot$ means the dot product operation, © is the concatenate operation, and $\bigoplus$ represents the addition operation. The 'Patch2Vec' represents the operation to reshape the neighbourhoods of each pixel to a vector.

4.1 Residual Regression Module

As the backbone of SHARP-Net, Residual Regression Module first introduces a feature encoder to extract a multi-scale feature pyramid $\{F_i\}_{i=1}^{L}$ from the combination of depth image D_{in} and amplitude image A, where F_i indicates the feature map extracted at the i^{th} level, and L is the number of layers in the pyramid. When the size of the input image is $W \times H$, the size of feature maps at the i^{th} level is $\frac{W}{2^{i-1}} \times \frac{H}{2^{i-1}} \times C_i$, where C_i is the number of output channels. In our network, we set $L = 6$ to keep the amount of parameters similar to that of state-of-the-art methods. The corresponding C_i are $16, 32, 64, 128, 192, 256$

respectively. From bottom to top, the feature pyramid gradually encodes the geometric information of the more detailed structure in the scene.

At each level, a Residual Regression Block, as shown in Fig. 3, is proposed to predict the depth residual map. The depth residual map from the lower level R_{i+1} is upsampled by the factor of 2 via bi-cubic interpolation, and then concatenated with the feature map at the current level. The new concatenated volume is the input of five sequential convolutional layers, which output the residual map R_i for the current level. Specifically, for the bottom level, the input of Residual Regression Block is only the feature map with size $\frac{W}{32} \times \frac{H}{32} \times 256$ because there is no depth residual map from the lower level. Different from the previous method [23] that directly regresses a residual map by sequentially up-sampling feature maps, our Residual Regression Module progressively regresses multi-scale residual maps in a coarse-to-fine manner by considering the hierarchical structures of the scene. The residual maps in lower resolutions depict depth noise existing in large-size shapes, while the residual map in higher resolutions focuses on depth noise existing in detailed structures. Finally, we get a residual pyramid $\{R_i\}_{i=1}^{L}$ consisting of the depth residual map at each level.

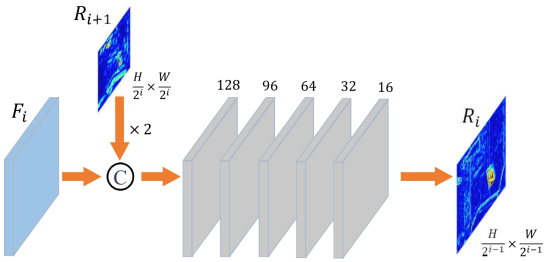

Fig. 3. Flowchart of residual regression block at the i^{th} level

4.2 Residual Fusion Module

The uppermost level of the residual pyramid provides a depth residual map with the original resolution, which can be treated as an estimation of the depth error. However, depth residual map from a single level cannot fully utilize the geometry information of the scene. Although the uppermost level of the residual pyramid contains the information from all the levels below, after the convolutional operation, information from lower resolution levels may get lost. Thus, we propose a Residual Fusion Module to explicitly combine the depth residual maps in all scales. The depth residual map at each level is first upsampled to the original resolution via bi-cubic interpolation. Then all the upsampled depth residual maps are concatenated together. The new residual volume is the input of a 1×1 convolutional layer. After the convolutional operation, we get the final depth residual map R_{out}. The depth residual map is added to the original input depth image, by which the depth image is recovered as D_{inter}. The details of Residual Fusion Module are shown in Fig. 2.

4.3 Depth Refinement Module

After previous two modules, MPI noise is removed to a great extent. In the meantime, shot noise also gets alleviated, but not as much as MPI removal. The existence of shot noise still hinders the application of ToF depth sensing. To address this problem, we propose a Depth Refinement Module, which utilizes Kernel Prediction Network [3] to further remove shot noise.

Depth Refinement Module takes the intermediate depth image D_{inter} as the input, and employs a U-Net model with skip connection to generate a weight matrix. The weight matrix consists of a vectorized filter kernel for each pixel in the depth image. In our experiment, we set the kernel size k as 3 and the size of the weight matrix is $W \times H \times 9$. Next, we generate a patch matrix by vectoring a neighbourhood for each pixel in the depth image. We call the above operation 'Patch2Vec'. When the neighbourhood is a 3×3 area, it is easy to calculate that the size of the patch matrix is also $W \times H \times 9$. Then the weight matrix is multiplied element-wisely with the patch matrix, generating a 3D volume with the same size. By summing over the 3D volume, we finally get the refined depth image D_{out}. Figure 2 shows details of Depth Refinement Module as well.

4.4 Loss Function

To train the parameters in our proposed SHARP-Net, we need to compute the differences between the predicted depth image D_{out} and the corresponding ground truth depth image D_{gt}. The loss function should guide our network to accurately remove depth noise while preserving geometry details. Following [23], our loss function has two components, which are L_1 loss and its gradients on the refined depth image. The formulation of the loss function is depicted as

$$L = \frac{1}{N} \sum \|D_{out} - D_{gt}\|_1 + \lambda \|\nabla D_{out} - \nabla D_{gt}\|_1 , \tag{4}$$

where $\|\cdot\|_1$ represents the L_1 norm, and N is the number of pixels. Here discrete Sobel operator is utilized to compute the gradients. In our experiments, we set $\lambda = 10$.

5 Experiments

5.1 Datasets

Our SHARP-Net is a supervised neural network to remove the noise for ToF depth images. To train all the parameters, we need ToF datasets with ground truth depth. To produce a suitable dataset, the mainstream method is applying the transient rendering technology to simulate the ToF imaging process while introducing MPI and shot noise [13]. Previous CNN based methods on ToF denoising have provided several synthetic datasets with thousands of scenes. In our experiments, we select two large-scale synthetic datasets ToF-FlyingThings3D (TFT3D) [23] and FLAT [11] for training and evaluation.

The TFT3D dataset contains 6250 different scenes such as living room and bathroom. We only utilize the ToF amplitude images and ToF depth images with resolution 640×480 as input for our proposed method. The FLAT dataset provides a total of 1929 scenes, which include the raw measurements and the corresponding ground truth depth. By using the pipeline released by the FLAT dataset, we convert the raw measurements to ToF depth images and ToF amplitude images with resolution 424×512. Furthermore, to evaluate the performance of SHARP-Net on realistic scenes, we also adopt the True Box dataset which is constructed by Agresti *et al.* in [2]. The ground truth depth of the True Box dataset is acquired by an active stereo system jointly calibrated with a ToF sensor. In total, there are 48 different scenes with resolution 239×320 on the dataset.

5.2 Data Pre-processing

We normalize the input depth images according to the range of depth value provided by the dataset, and filter out the pixels whose depth value are not within the range (0, 1]. For the convenience of experiments, we crop the images on the TFT3D dataset and FLAT dataset to size 384×512. For the True Box dataset, we crop the images to size 224×320. In addition, for the FLAT dataset, we exclude scenes without background following the experiment setting in [23]. For all the three datasets, we randomly select 20% scenes as the test set while the rest for training.

5.3 Training Settings

For the TFT3D dataset, the learning rate is set to be 4×10^{-4}, which is reduced 30% after every 2 epochs. We trained SHARP-Net for 40 epochs with a batch size of 2. For the FLAT dataset, we set the learning rate as 1×10^{-4} with conducting the rate decay. We train the SHARP-Net for 100 epochs with a batch size of 8. For the True Box dataset, the training settings are consistent with that of the TFT3D dataset. The network is implemented using TensorFlow framework [1] and trained using Adam optimizer. With four NVIDIA TITAN Xp graphics cards, the training process takes about 20 h for both TFT3D and FLAT datasets, less than half an hour for the True Box dataset.

5.4 Ablation Studies

SHARP-Net is a CNN based method with a 6-level Residual Regression Module as the backbone and two extra fusion and refinement modules. In order to validate the effectiveness of our proposed modules, we design experiments to compare SHARP-Net against its variants.

- WOFusRef: A variant of SHARP-Net without the Depth Refinement Module and the Residual Fusion Module.
- WORefine: A variant of SHARP-Net without the Depth Refinement Module.
- WOFusion: A variant of SHARP-Net without the Residual Fusion Module.

– FourLevel: A variant of SHARP-Net whose backbone has 4 levels.
– FiveLevel: A variant of SHARP-Net whose backbone has 5 levels.

For a fair comparison with FourLevel and FiveLevel, we need to ensure that the amount of parameters of these two variants are nearly the same with SHARP-Net. Therefore, we adjust the number of convolution kernel channels of the variants.

Table 1. Quantitative comparison with the variants of SHARP-Net on the TFT3D dataset.

Model	TFT3D dataset: MAE (cm)/Relative error				
	1st Quan.	2nd Quan.	3rd Quan.	4th Quan.	Overall
WOFusRef	0.12/7.7%	0.49/8.3%	1.08/9.4%	4.90/16.3%	1.69/13.8%
WORefine	0.12/7.7%	0.44/7.5%	0.97/8.4%	4.69/15.6%	1.55/12.7%
WOFusion	0.11/7.1%	0.42/7.2%	0.94/8.2%	5.01/16.7%	1.62/13.2%
FourLevel	0.15/9.6%	0.57/9.7%	1.24/10.7%	5.15/17.2%	1.78/14.5%
FiveLevel	0.12/7.7%	0.46/7.8%	1.00/8.8%	4.55/15.2%	1.53/12.5%
SHARP-Net	**0.09/5.8%**	**0.30/5.1%**	**0.67/5.8%**	**3.40/11.3%**	**1.19/9.7%**

For the quantitative comparison, we use two metrics, Mean Absolute Error (MAE) and relative error, to evaluate the performance. The MAE between the original noisy depth image and the ground truth depth image is depicted as the original MAE. Then, we define the relative error as the ratio of the MAE of each method to the MAE of the corresponding input. The overall and partial MAE/Relative Error at each error level are also calculated. Different denoising methods may have varying performances at different error levels. In our experiment, we adopt an evaluation method that is similar to the method in [23] to comprehensively evaluate our proposed SHARP-Net at different error levels. First, we calculate the per-pixel absolute error value between the input depth image and the ground truth. Then we sort all the per-pixel absolute errors in an ascending order. Next all the pixels in the test set are split into four quantiles (four error level sets). The difference between our evaluation method and the method in [23] is that we sort all the pixels in the test set instead of in a single image. This change makes our evaluation more reasonable because sorting in the whole test set eliminates the depth distinction over images. The pixels in the range of 0%–25% are classified into the 1st error level. In the same way, the pixels in the range of 25%–50% and 50%–75% and 75%–100% are classified into the 2nd, 3rd, and 4th error level. Pixels with depth value beyond the maximum depth for each dataset are considered as outlier here and excluded from any error level sets. Finally, we calculate the partial MAE and overall MAE for different error levels respectively.

For ablation studies, we just utilize the TFT3D dataset to compare our SHARP-Net against its variants. The overall MAE and partial MAE at each

error level are reported in Table 1. From Table 1, it can be observed that SHARP-Net achieves the lowest MAE and relative error at all error levels. In addition, '4th Quan.' contributes the greatest share on the value of overall MAE compared with the remanent three quantiles. Comparing SHARP-Net with FourLevel and FiveLevel variants, we can see that at all error levels, MAE decreases as the total number of pyramid levels increases. This is because the network explicitly divides the scene into a more detailed hierarchical structure if the pyramid has more levels, which results in a more accurate estimation of MPI noise.

Comparing SHARP-Net with WORefine and WOFusion, it can be observed that the employment of Residual Fusion Module and Depth Refinement Module reduce the overall MAE by 26% and 23% respectively, which indicates the necessity of those two modules. Linking the comparison between WORefine and WOFusion at all error levels, it can be seen that either of the two modules facilitates the decline of MAE but the extent is limited. However, considering the difference between WOFusRef and SHARP-Net on the MAE and relative error, we conclude that utilizing these two modules together can greatly improve the performance of the noise removal at all error levels.

Table 2. Quantitative comparison with competitive ToF depth denoising methods on TFT3D, FLAT and True Box datasets.

Model	1st Quan.	2nd Quan.	3rd Quan.	4th Quan.	Overall
	TFT3D dataset: MAE (cm)/Relative error				
DeepToF	0.47/30.1%	1.56/26.6%	3.11/27.0%	9.01/30.0%	3.54/28.9%
ToF-KPN	0.19/12.2%	0.82/13.9%	1.87/16.2%	6.64/21.3%	2.38/19.4%
SHARP-Net	**0.09/5.8%**	**0.30/5.1%**	**0.67/5.8%**	**3.40/11.3%**	**1.19/9.7%**
	FLAT dataset: MAE (cm)/Relative error				
DeepToF	0.09/27.3%	0.44/33.6%	1.13/43.5%	2.74/37.8%	1.10/43.3%
ToF-KPN	0.08/24.2%	0.30/22.9%	0.66/25.4%	2.12/29.3%	0.79/31.1%
SHARP-Net	**0.04/12.1%**	**0.14/10.7%**	**0.32/12.3%**	**1.33/18.4%**	**0.46/18.1%**
	True Box dataset: MAE (cm)/Relative error				
DeepToF	0.31/42.5%	1.06/49.5%	2.15/52.9%	5.75/53.9%	2.32/52.7%
ToF-KPN	0.28/38.4%	0.87/40.6%	1.64/40.4%	4.51/42.3%	1.82/41.4%
SHARP-Net	**0.15/20.5%**	**0.47/21.9%**	**0.91/22.4%**	**3.02/28.3%**	**1.14/25.9%**

5.5 Results on Synthetic Datasets

To evaluate the performance of our proposed SHARP-Net, we compare it with two state-of-the-art ToF depth denoising methods DeepToF [17] and ToF-KPN [23]. The inputs of all selected methods are the concatenation of depth images and corresponding amplitude images. It should be noted that the original DeepToF is smaller than SHARP-Net in term of model size. For a fair comparison, we take the same strategy as [23] to replace the original DeepToF model with the U-Net backbone of ToF-KPN. The quantitative experimental results

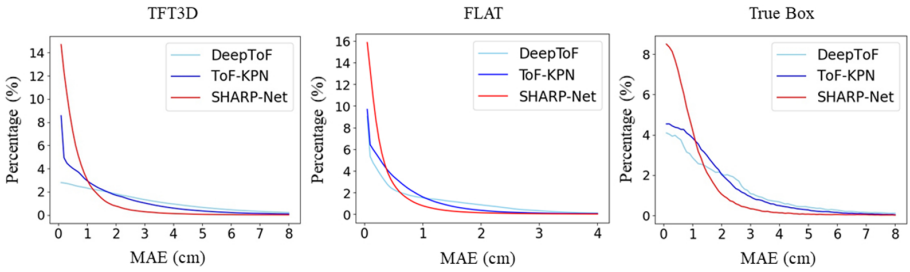

Fig. 4. The per-pixel error distribution curves of different methods on the TFT3D, FLAT and True Box datasets. The distribution curves of these three methods show that our proposed SHARP-Net obtains the optimal error distribution on all the datasets.

on the TFT3D and FLAT datasets are reported in Table 2. It can be seen that SHARP-Net achieves the lowest MAE and relative error at all error levels of the two synthetic datasets. The MAE between the input depth and ground truth depth is 12.24 cm and 2.54 cm for both TFT3D and FLAT datasets. After training on these datasets, SHARP-Net reduces the MAE to 1.19 cm and 0.46 cm in the test sets respectively.

The relative error is also a good indicator to measure the performance for different methods. From Table 2, it can be seen that the DeepToF method gives similar relative errors for all the four error levels, especially on the TFT3D dataset. Compared with two other methods, DeepToF's performance in terms of the relative error indicator is low. For ToF-KPN, the relative error increase as the error level increasing, which means ToF-KPN has better denoising performance for higher error level sets. For SHARP-Net, it can be seen that relative error is much smaller than the other two methods on the TFT3D dataset on the FLAT dataset, SHARP-Net is much better than DeepToF in term of relative error. Compared with ToF-KPN, SHARP-Net performs the same as ToF-KPN at the preceding three error levels, and outperform ToF-KPN at the highest error level.

For an intuitive comparison, in Fig. 4, we illustrate the per-pixel error distribution curves for all the methods on the TFT3D and FLAT datasets. It can be seen that after denoising by our SHARP-Net, the depth errors are mainly concentrated in the lower error region. In Fig. 5, we give several qualitative comparison results for SHARP-Net, ToF-KPN and DeepToF. It can be seen that the depth image corrected by our proposed method is more accurate, preserving more geometry structures in the scene. We observe that ToF-KPN performs better than DeepToF in removing the noise existing in detailed structures. However, the noise removal of ToF-KPN on large-size shapes is not adequate. In contrast, SHARP-Net demonstrates better results for large-size shapes and detailed structures simultaneously. In fact, our SHARP-Net also has some failure cases in depth denoising, for example low reflection areas and extremely complex geometry structures, which are the limitations of our method.

In Fig. 6, we compare the performance of all the methods along a scan line on a depth image selected from the TFT3D dataset. From the ToF amplitude image, we can observe that the scene is located in a living room. Many different objects

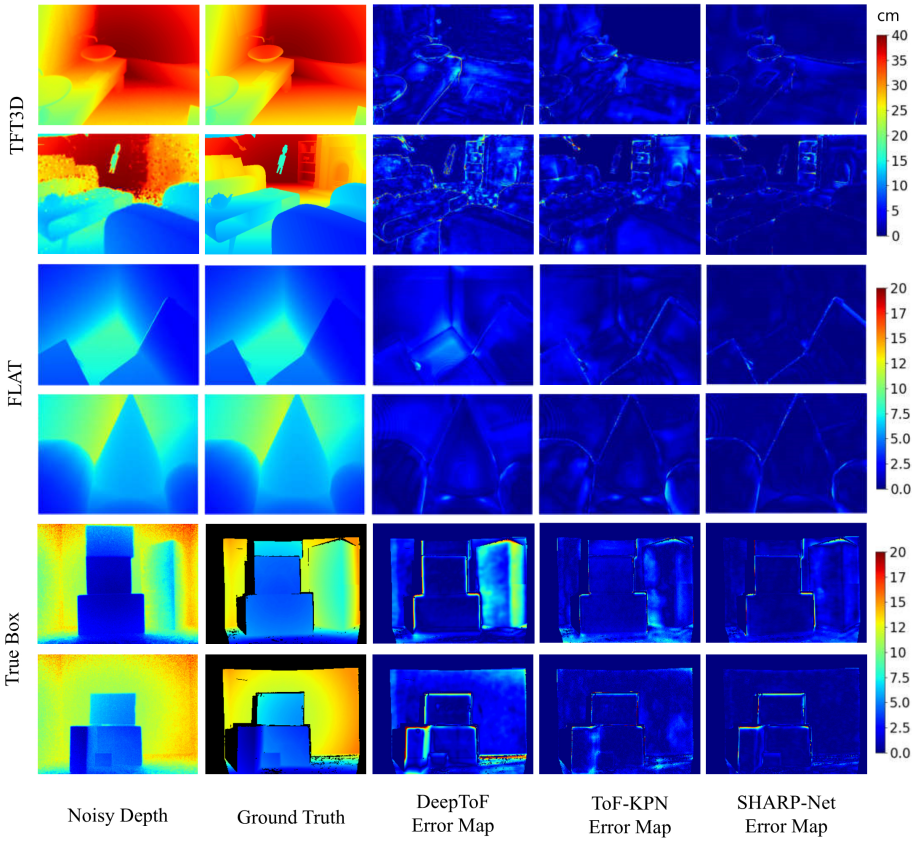

Fig. 5. Qualitative comparison on the TFT3D dataset, the FLAT dataset and the True Box dataset for ToF depth denoising. For each dataset, two scenes are selected for comparison. The colorbars in the right show the color scale for error maps with the unit in cm. (Color figure online)

appear in the living room and demonstrate complex hierarchical structures. The distinct depth variation along this scan line makes it suitable for this comparison. It can be seen that after depth denoising, the depth data corrected by SHARP-Net draw the closest line to the ground truth.

5.6 Results on the Realistic Dataset

Furthermore, we test our proposed SHARP-Net along with previous methods on a realistic dataset. All the tested models are retrained on the True Box training set. The experimental results for all the methods on the True Box dataset are also shown in Table 2. It can be seen that SHARP-Net surpasses other methods at all error levels. We also find that the relative errors at all error levels on the True Box dataset are significantly larger than those on the synthetic datasets. One reason

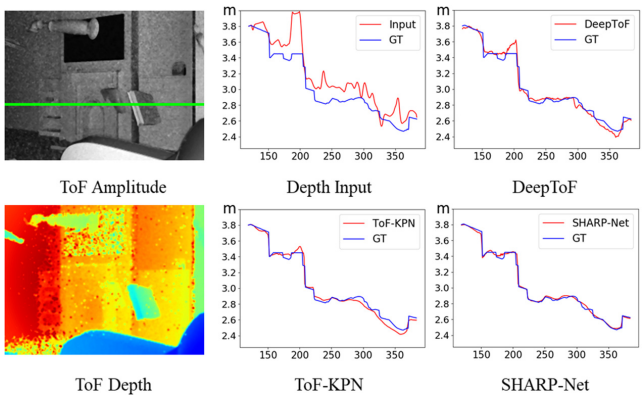

Fig. 6. Quantitative comparison with previous works along a green scan line in a depth image from the TFT3D dataset. 'GT' means the ground truth depth. Our proposed SHARP-Net demonstrates the best performance on depth denoising. (Color figure online)

may be that the noise generation mechanism for realistic ToF depth noise is more complex, which are not accurately modeled on the synthetic datasets. From Fig. 4, it can be observed that the error distribution curve of SHARP-Net on the True Box dataset is similar to those on the two synthetic datasets. Compared with other methods, after denoising by SHARP-Net, the remaining depth error on the dataset is concentrated in the small value area. On the bottom of Fig. 5, we demonstrate the qualitative comparison results on two scenes selected from the True Box test set. We can observe that for this dataset, SHARP-Net presents the best visual effects. Compared with other methods, SHARP-Net performs better in large-size shapes, especially in the background areas.

6 Conclusion

The Multi-Path Interference (MPI) seriously degrades the depth image captured by ToF sensors. In this work, we propose SHARP-Net, a Spatial Hierarchy Aware Residual Pyramid Network for ToF depth denoising. Our SHARP-Net progressively utilizes the spatial hierarchical structure of the scene to regress depth residual maps in different scales, obtaining a residual pyramid. A Residual Fusion Module is introduced to selectively fuse the residual pyramid by summing over the depth residual maps at all levels in the pyramid, and a Kernel Prediction Network based Depth Refinement Module is employed to further eliminate shot noise. Ablation studies validate the effectiveness of these modules. Experimental results demonstrate that our SHARP-Net greatly surpasses the state-of-the-art methods in both quantitative and qualitative comparison on synthetic and realistic datasets.

Acknowledgments. We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, and National Natural Science Foundation of China under Grants 61671419 and 61901435.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th Symposium on Operating Systems Design and Implementation, pp. 265–283 (2016)
2. Agresti, G., Schaefer, H., Sartor, P., Zanuttigh, P.: Unsupervised domain adaptation for ToF data denoising with adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5584–5593 (2019)
3. Bako, S., et al.: Kernel-predicting convolutional networks for denoising Monte Carlo renderings. ACM Trans. Graph. (TOG) **36**(4), 97 (2017)
4. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 638–647. IEEE (2019)
5. Bhandari, A., Feigin, M., Izadi, S., Rhemann, C., Schmidt, M., Raskar, R.: Resolving multipath interference in Kinect: an inverse problem approach. In: 2014 IEEE SENSORS, pp. 614–617. IEEE (2014)
6. Chen, X., Chen, X., Zha, Z.: Structure-aware residual pyramid network for monocular depth estimation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019, pp. 694–700 (2019)
7. Chen, X., Lou, X., Bai, L., Han, J.: Residual pyramid learning for single-shot semantic segmentation. IEEE Trans. Intell. Transp. Syst. **21**, 2990–3000 (2019)
8. Freedman, D., Smolin, Y., Krupka, E., Leichter, I., Schmidt, M.: SRA: fast removal of general multipath for ToF sensors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 234–249. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_16
9. Fuchs, S.: Multipath interference compensation in time-of-flight camera images. In: 2010 20th International Conference on Pattern Recognition, pp. 3583–3586. IEEE (2010)
10. Fuchs, S., Suppa, M., Hellwich, O.: Compensation for multipath in ToF camera measurements supported by photometric calibration and environment integration. In: Chen, M., Leibe, B., Neumann, B. (eds.) ICVS 2013. LNCS, vol. 7963, pp. 31–41. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39402-7_4
11. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3D ToF artifacts through learning and the FLAT dataset. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 381–396. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_23
12. Gupta, M., Nayar, S.K., Hullin, M.B., Martin, J.: Phasor imaging: a generalization of correlation-based time-of-flight imaging. ACM Trans. Graph. (ToG) **34**(5), 156 (2015)
13. Jarabo, A., Marco, J., Muñoz, A., Buisan, R., Jarosz, W., Gutierrez, D.: A framework for transient rendering. ACM Trans. Graph. (ToG) **33**(6), 177 (2014)
14. Jiménez, D., Pizarro, D., Mazo, M., Palazuelos, S.: Modeling and correction of multipath interference in time of flight cameras. Image Vis. Comput. **32**(1), 1–13 (2014)

15. Jung, J., Lee, J.Y., Jeong, Y., Kweon, I.S.: Time-of-flight sensor calibration for a color and depth camera pair. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1501–1513 (2014)
16. Lenzen, F., Schäfer, H., Garbe, C.: Denoising time-of-flight data with adaptive total variation. In: Bebis, G., et al. (eds.) ISVC 2011. LNCS, vol. 6938, pp. 337–346. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24028-7_31
17. Marco, J., et al.: DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Transactions on Graphics (ToG) **36**(6), 219 (2017)
18. Mo, K., et al.: StructureNet: hierarchical graph networks for 3D shape generation. arXiv preprint arXiv:1908.00575 (2019)
19. Nan, Y., Xiao, R., Gao, S., Yan, R.: An event-based hierarchy model for object recognition. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2342–2347. IEEE (2019)
20. Park, B., Yu, S., Jeong, J.: Densely connected hierarchical network for image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
21. Payne, A., et al.: 7.6 a 512 × 424 CMOS 3D time-of-flight image sensor with multi-frequency photo-demodulation up to 130 MHz and 2 gs/s ADC. In: 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 134–135. IEEE (2014)
22. Peng, J., Xiong, Z., Wang, Y., Zhang, Y., Liu, D.: Zero-shot depth estimation from light field using a convolutional neural network. IEEE Trans. Comput. Imaging **6**, 682–696 (2020)
23. Qiu, D., Pang, J., Sun, W., Yang, C.: Deep end-to-end alignment and refinement for time-of-flight RGB-D module. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9994–10003 (2019)
24. Shi, Y., Chang, A.X., Wu, Z., Savva, M., Xu, K.: Hierarchy denoising recursive autoencoders for 3D scene layout prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1771–1780 (2019)
25. Song, X., Zhao, X., Hu, H., Fang, L.: EdgeStereo: a context integrated residual pyramid network for stereo matching. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 20–35. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_2
26. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6383–6392 (2018)
27. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2621–2629 (2019)
28. Zanuttigh, P., Marin, G., Dal Mutto, C., Dominio, F., Minto, L., Cortelazzo, G.M.: Time-of-Flight and Structured Light Depth Cameras: Technology and Applications, pp. 978–983. Springer, Switzerland (2016). ISSBN
29. Zhang, S.: High-speed 3D shape measurement with structured light methods: a review. Opt. Lasers Eng. **106**, 119–131 (2018)
30. Zhang, Y., Xiong, Z., Wu, F.: Fusion of time-of-flight and phase shifting for high-resolution and low-latency depth sensing. In: 2015 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2015)
31. Zheng, Y., Cao, X., Xiao, Y., Zhu, X., Yuan, J.: Joint residual pyramid for joint image super-resolution. J. Vis. Commun. Image Represent. **58**, 53–62 (2019)