Iterative Error Removal for Time-of-Flight Depth Imaging

Zhuolin Zheng[1,2,3], Yinzhang Ding[1,2,3], Xiaotian Tang[1,2,3], Yu Cai[1,2,3], Dongxiao Li[1,2,3(✉)], Ming Zhang[1,2,3], Hongyang Xie[4], and Xuanfu Li[4]

[1] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
{zhengzhuolin,dingyzh,3160101464,yucaimr,lidx,zhangm}@zju.edu.cn
[2] Zhejiang Provincial Key Laboratory of Information Processing Communication and Networking, Hangzhou, China
[3] State Key Laboratory of CAD and CG, Hangzhou, China
[4] Huawei Technologies Co. Ltd., Shenzhen, China
xiehongyang@hisilicon.com, lixuanfu@huawei.com

Abstract. Depth information plays an increasingly important role in computer vision tasks. As one of the most promising depth sensing techniques, Amplitude Modulated Continuous Wave (AMCW)-based indirect Time-of-Flight (ToF) has been widely used in recent years. Unfortunately, the depth acquired by ToF sensors is often corrupted by imaging noise, multi-path interference (MPI), and low intensity. Different methods have been proposed for tackling these issues. Nevertheless, they failed to exploit the characteristics of the ToF depth map to propose a targeted solution, and are unable to achieve various error removal. We present a new iterative method for removing various errors simultaneously through cascaded Convolutional Neural Networks (CNNs). A Synthetic Dataset is created using computer graphics, and a Real-World Dataset is developed via RGBD-based 3D reconstruction, both contain the raw measurement acquired by a certain ToF camera and corresponding dense ground truth depth. Experimental results demonstrate the superior performance of the proposed iterative method in removing various ToF depth errors, compared to state-of-the-art methods, on both the newly developed datasets and existing public datasets.

Keywords: Time-of-flight · Convolutional neural networks · Three-dimensional vision · Depth sensor

1 Introduction

Depth acquisition is not only the key to most 3D vision tasks but also playing an increasingly important role in traditional RGB-based computer vision tasks such as gesture recognition and semantic segmentation in the past few years. Previous representative depth acquisition approaches including structured light and stereo vision have some critical limitations [20], either cannot be used for

I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12892, pp. 92–105, 2021.
https://doi.org/10.1007/978-3-030-86340-1_8

mid-to-long-distance ranging, or require sufficient texture in the scene. Recently, Time-of-Flight (ToF) [11] based depth camera has attracted more and more attention due to its inexpensiveness, lightweight, fair accuracy, and robustness.

The Time-of-Flight technique obtains depth by measuring the time it takes for a wave or pulse to travel from the emitter to object and back to the receiver. For the particular Amplitude Modulated Continuous Waves (AMCW) implementation of the ToF camera, the depth is acquired indirectly by measuring the cross-correlation between emitted and received wave and calculating the phase delay to represent depth. Modern AMCW-ToF camera usually uses multiple modulation frequencies to enlarge the sensing range of the camera while remaining accuracy.

Various errors may occur during this imaging procedure. There are two main types of errors. The first is the common error in digital imaging system, such as Gaussian imaging noise, temperature drift. The second type of error occurs when the real scenario does not follow the assumption of the working principle of the ToF camera, that the light received at each pixel position consists only of the light firstly reflected from that position. When in a complex scene or there are surfaces with low reflectance or mirror reflection, this principle is often violated. Some pixels may receive light reflected multiple times from elsewhere, thus cause the so-called Multipath Interference (MPI). Some pixels may suffer low intensity that the power of light received there is too low to get sufficient Signal-Noise Ratio (SNR) to calculate precise depth, due to low retroreflectivity on the surface. When in a multi-frequency ToF camera, these errors can then be transmitted and amplified by phase-unwrapping and induce a significant error on the final depth map due to incorrect estimation of rounds.

Earlier ToF depth refinement work [7,10] mostly adopts various sparse simplified assumptions about the response characteristics of the scene (e.g. Lambertian [6], Two-bounce light [17] and Two-path-caused MPI [8]), based on which to model the local light path, and then solve it through probability models [8] or optimization methods [5] to find the optimal solution.

In recent years, many learning-based methods have begun to emerge. Despite few other learning-based methods [12], most of them adopt CNN to their methods. Marco et al. [16] designed a network of encoder-decoder structure and used a two-stage training scheme to correct for MPI. Su et al. [22] proposed to directly use the correlation map obtained by the ToF camera as input, and then designed an end-to-end CNN based on U-Net [19] to replace the traditional pipeline, Generative Adversarial Networks (GAN) is also included in consideration of deep image generation. Guo et al. [9] noticed that multiple sampling of ToF cameras could cause artifacts in dynamic scenes, and proposed an encoder-decoder CNN to deal with the artifacts and eliminating MPI meanwhile. Qiu et al. [18] proposed a network that used RGBD as input, utilizing cross-modal dense optical flow for image alignment, and used the aligned depth map to pass a kernel prediction network (KPN) to get a refined depth map. Agresti et al. [1,2] designed a coarse-fine CNN to capture multi-scale information and later developed this approach by performing unsupervised learning on unlabeled data with GAN.

There are also some work focused on ToF imaging problems under special situations via CNN, such as translucent objects [21] and short exposure [3]. Although CNN based methods have achieved good results, these methods did not exploit characteristics of the ToF depth map.

Instead of a usual end-to-end approach, we designed a multi-stage iterative CNN to tackle this problem for three main reasons. First, the end-to-end CNN cannot realize the discontinuous mapping from multi-frequency raw measurement to depth map in the underlying principles. Second, unlike RGB generation, which has a higher tolerance for pixel-level error due to human visual perception, the results completely output by CNN are not perfectly competent for prediction of ToF depth map in a pixel-level millimeter accuracy. And it will be very likely to degrade when odd input is encountered. Third, for different kinds of error, different ideas and principles are needed to remove, it is difficult to predict the nonlinear coupling of different types of error in a single output. Our iterative method can avoid the above issues by continuous residual prediction of the current depth map and retaining the information for the next refinement.

In order to achieve a robust and effective solution for multiple types of ToF depth map error removal. We present a newly designed CNN specifically for this problem. We devise a CNN module and use it to implement this network architecture. Moreover, we also develop two new large ToF datasets, one is a Synthetic dataset made by computer graphics techniques, and the other is a Real-World dataset generated by 3D reconstruction. Both of them consist of plentiful types of data, and all the critical error mentioned above is considered in the proposed datasets.

Our main contributions are summarized as follows:

- We proposed a targeted, iterative-based CNN method that can significantly remove various types of error by utilizing principles of ToF imaging and avoiding the shortcomings of existing methods.
- We created two large ToF datasets via computer graphics and 3D reconstruction, especially the firstly proposed large real-world dataset. Both cover various error types and provide plenty of types of data access, especially the ToF raw measurement and dense ground truth depth.
- Through qualitative and quantitative experiments, we showed that our proposed method is able to remove most of the error on the ToF depth map and surpass the previous method on our datasets and other public avaliable datasets.

Code and datasets are available from the authors upon reasonable request or with permission of Huawei Technologies Co., Ltd.

2 Method

2.1 Formulating for ToF Depth Imaging

The working principle of the ToF camera based on AMCW is briefly introduced as follows. When the camera is imaging, the transmitter on camera will emit

an amplitude modulated near-infrared (NIR) wave. The receiver at each pixel position will receive the echo and calculates the cross-correlation between the received signal and the reference signal from the transmitter. This inner production process will automatically complete the demodulation and filter out the signal on other frequency bands such as ambient light. Since we only focus on the part that actually carries information in the modulated wave, we use the phasor to express it concisely as Eq. 1.

$$C_i = \mathrm{CC}\left(\boldsymbol{S}, \boldsymbol{S}_{ref,i}\right) = \frac{<\boldsymbol{S}, \boldsymbol{S}_{ref,i}>}{\|\boldsymbol{S}_{ref,i}\|} = \frac{\int_0^T \boldsymbol{S}(t) \cdot \boldsymbol{S}_{ref,i}(t)dt}{\sqrt{\int_0^T \boldsymbol{S}_{ref,i}(t)^2 dt}} = I \cdot \cos \Delta\varphi_i \quad (1)$$

where C_i represents for the i-th imaging result of ToF raw measurement. CC denotes for the cross-correlation operation. I represents the intensity of this imaging. $\boldsymbol{S}$ is the phasor of the received signal, $\boldsymbol{S}_{ref,i}$ is the reference signal used for every imaging, $i = 0...\mathrm{n}-1$, n denote times of imaging. This imaging will be performed multiple times by shifting the phase of the reference signal (or received signal) to eliminate ambiguity of cosine. Take four times as an example, $\boldsymbol{S}_{ref,0}$ is the reference signal from the transmitter when received returned signal, the other reference signals will be expressed in this way. j is the imaginary unit.

$$\boldsymbol{S}_{ref,i} = \boldsymbol{S}_{ref,i-1} \cdot e^{-\frac{j\pi}{2}} \quad (2)$$

Four imaging will be performed according to Eq. 1. By basic mathematical transformations we can calculate the radius depth and intensity map, which is the typical output of the ToF camera. * represents for depth or intensity directly output by camera.

$$Depth^* = arctan2\left(C_1 - C_3,\ C_0 - C_2\right) \cdot c/(4\pi f) \quad (3)$$

$$Intensity^* = \sqrt{\left(C_1 - C_3\right)^2 + \left(C_0 - C_2\right)^2}/2 \quad (4)$$

For a multi-frequency ToF camera, the camera needs to first calculate the phase measured at each frequency separately, and then use the half-wavelength at each modulation frequency as the base to establish a linear congruence equation to solve the real depth via a phase unwrapping algorithm [15] based on Chinese Remainder Theorem.

2.2 Input and Output Defining

We first define the input and output of the problem.

For input, most of the previous methods used the direct output of the ToF camera, i.e., the depth map and the intensity map as input. Later other work [9,22] pointed out that using ToF raw correlation information can lead to a more accurate depth prediction, since it avoids losing the luminosity and geometric information contained in it. However, this approach has a major flaw in the bottom layer: If the ToF camera uses multiple modulation frequencies, the

solving of linear congruence equations will be involved. The remainder operation in it make the mapping from correlation maps to the depth map discontinuous and undifferentiable. However, neural networks are not able to fit a discontinuous mapping in principle. Therefore, an end-to-end fashion CNN can only fit some appearance, it is still not able to solve the problem essentially.

For output, previous work [16,22] often adopts the outputs completely generated by CNN as the results, since most visual applications is done by this approach. But there is an important difference between depth map generation and RGB generation, that depth maps are very sensitive to values. A certain number of abnormal pixels may not affect visual perception, but it will surely affect the accuracy of the depth map. Therefore, CNNs based on human vision can generate visually impressive images, but the result is unreliable and unstable in terms of pixel-wise accuracy. On the one hand, it is not capable to achieve high accuracy, on the other hand, it is not robust enough, once odd input is encountered, the output tends to fail.

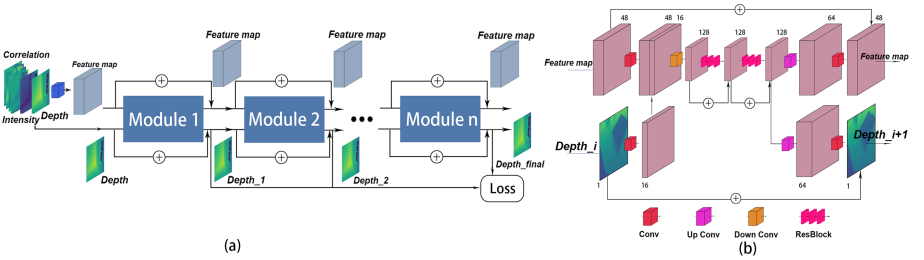

Fig. 1. (a) Overview of our iterative CNN for ToF error removal. Each module takes a depth map and a feature map as input, outputs an residual prediction to refine the depth map, and gives deeper features for the next iteration. The initial feature map is obtained from the ToF raw correlation map, intensity map and depth map. (b) The iterative module structure built for our CNN.

To solve these problems, we keep the depth map and intensity map output by the ToF camera and feed them to CNN together with the raw correlation maps. This approach directly provides the solution of the congruence equation, frees the CNN from the complicated remainder problem and enables the network to focus on the information contained in the raw measurement that contributes to the error removal. Also, we turn the task of the network into predicting the errors in the ToF depth map, making the input depth as anchor. This approach makes the network robust and prevents possible performance degradation of CNN.

2.3 Proposed Iterative CNN

This problem has the following characteristics. First, the input depth map and the output depth map are signals in the same domain, which means that if a naive method is defined to eliminate the error in the depth map as much as

possible, then a better result could be expected by using the output as the input. This inspired us to design an iterative method. Second, the entropy contained in this problem is large, since this is a dense prediction at a relatively high spatial resolution [12]. Besides, with a full range 4 m to 10 m and accuracy in millimeter [11], the SNR required for each pixel in this task reaches 60 dB, which is much higher than that in the image generation task e.g. image super-resolution with state-of-the-art PSNR around 35 dB [23]. Therefore, residual prediction is necessary and enough spatial information should be retained during the whole processing. Third, there are various errors in the ToF depth map, e.g. the imaging noise, MPI and low intensity, the way and level they affect the depth map vary [15,20]. It's difficult for CNN to make a single residual prediction to the highly coupled nonlinear superposition of all different kinds of errors. So based on these features, we proposed our method.

First, we designed a CNN module as shown in Fig. 1(b) The input of this module is feature maps of 48 channels and a depth map of one channel. Depth map passes through a convolution layer and fuse with feature maps in channel dimension, then a down convolution is used to scale down these feature maps, and doubled channel to 128. The feature will pass through two cascaded Res-blocks. Then the resulted feature maps will proceed to two different sets of up convolution and standard convolution, respectively, to obtain feature maps and a residual prediction. Feature maps and residual depth map are of the same size as the input feature maps and depth map, and will be added to them to form the outputs of module.

Proposed CNN is built using this module, as shown in Fig. 1(a). First, the correlation maps at different frequencies as well as depth map and intensity map are pre-fused via atrous and standard convolution to obtain a rough feature map. This feature map will be input in this module together with the ToF depth map. The module architecture will be iterated several times with different parameters while the outputs of previous module serve as the inputs of the next module. The depth map output by each module will be constrained by loss function. And naturally, the depth map from the last module will be the final depth map output.

From overview, the flow of feature maps followed a multi-level residual network fashion; from a detailed perspective of each module, the flow of feature maps and depth maps has followed the spirit of U-Net [19]: Multi-scale information extraction and reusing. The module continuously up and down sampling to extract multi-scale contexts and creating a shortcut from shallow to deep to retain spatial information of the original depth map space.

For loss function, we adopt L1-loss as data term loss for its simplicity.

$$L_{l1} = \sum_{i} ||Depth_{gt} - Depth_{CNNi}||_1 \tag{5}$$

Here, i represents for the depth map output from the i-th module.

And we utilize second order total variation as regularization term to impose an artificial prior constraint, since depth are mostly first-order smooth within occlusion boundary.

$$L_r = ||f(Depth_{CNN} \circledast LoG)||_1 \tag{6}$$

$$f(x) = \begin{cases} x, & if \quad x < threshold \\ 0, & else \end{cases} \tag{7}$$

This term only act on the final output depth. LoG represents for the Laplace of Gaussian filter. $f(x)$ and $threshold$ are there to prevent the impact on object edges. The weight ratio of L1-loss to second order total variation loss is $1 * 10^{-4}$.

We built this iterative network based on the module. The input of each module of the network is a depth map and a feature map containing depth information, the task of each module is to make a residual prediction, add it to input depth of the module to obtain a finer depth map, and meanwhile output a feature map with deeper information for the next module to exploit. Depth map and feature maps are both iteratively improved with the cascading of modules. Our network architecture is also efficient, containing only 2.08 million parameters with two Resblocks in each module and five modules cascading to form the CNN.

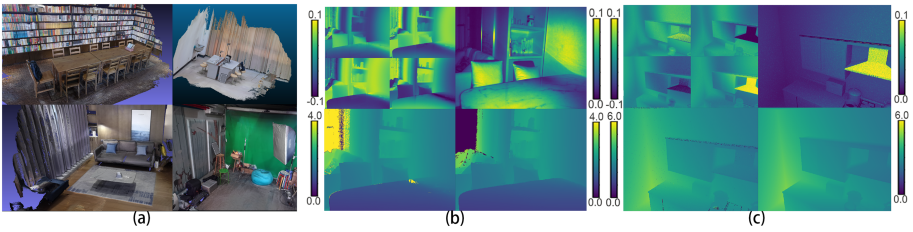

Fig. 2. (a) 3D scene reconstruction result. UpperLeft: Bookstore, UpperRight: Laboratory, LowerLeft: Living-room, LowerRight: PhotoBooth (b) (c) Partial datatype of our Real-World Dataset and Synthetic Dataset respectively. UpperLeft: Four ToF raw correlation maps; UpperRight: ToF Intensity map; LowerLeft: ToF Depth map; LowerRight: Ground Truth Depth map.

3 Datasets

Data is crucial for CNN, however, most of the previous work uses only the synthetic dataset with limited error simulation, and lacks real-world dataset for training. We successfully generated a large Synthetic dataset and a large Real-World dataset by computer graphics and 3D reconstruction techniques, respectively. Both contain dense ground truth depth and ToF raw correlation measurements (Fig. 2).

3.1 Synthetic Dataset

Our Synthetic dataset contains six main indoor scenes: living room, bedroom, bathroom, dining room, kitchen, and staircase. There are 1379 sets of data contained in total, we split them into a training set of size 1260 and a test set of size 119. We simulated the imaging system of the LUCID Helios camera. Except for the RGB images that LUCID cannot capture, all our simulations use the same internal parameters and resolution as the camera, i.e. 480 * 640. Each set of data includes the raw correlation measurements on 100 MHz and 75 MHz, with each frequency containing two imaging results of phase shift $\pi/2$, the ToF depth map, the intensity map, the ground truth depth, and RGB image of resolution 1440 * 1920. In terms of error simulation, raw measurements in the dataset include imaging noise, MPI, low reflectivity, and specular reflections which are typical errors that may occur in the actual application of ToF cameras, the depth map and intensity map generated by correlation maps will also be introduced to these errors.

The scenes of the dataset are mainly taken from the 3D models shared by the Blender community. The simulation of the correlation map and the ground truth depth is attributed to the work of Jarabo et al. [14] on transient rendering. By rendering the light received by the camera in every time interval, then amplitude modulating the rendering result and integrating them to obtain the correlation maps. Support for ray tracing allows us to simulate MPI. Simulation for low reflectivity, and specular reflection are done by adjusting the material and reflectivity of different objects. Imaging noise is treated as additive Gaussian noise. The surface normal map and RGB map are generated thanks to Blender and its Cycles Renderer.

3.2 Real-World Dataset

Our Real-World dataset contains 1060 sets of data in total and is split into a train/test sets of size 964 and 96, which is significantly larger than the Real-World dataset [2,22] used only for testing. We reconstruct several different indoor scenes, namely bathroom, bedroom, bookstore, cafe, dining room, library, laboratory, kitchen, living room, and read room, to ensure the diversity of the dataset and take common errors of ToF cameras into consideration. Each set of data includes the same data in the same resolution as the Synthetic dataset except for the surface normal map.

Modern portable RGB-Depth sensors are used in this 3D reconstruction. We attached an Azure Kinect camera with a Lucid Helios camera, to form a data acquisition system. The transformation of two camera poses can be estimated by SfM (Struct from Motion) algorithm [13]. Azure Kinect provides RGB-Depth streams for reconstruction while Lucid Helios camera is set to capture ToF raw measurement stream in both 75 MHz and 100 MHz. To the best of our knowledge, BundleFusion [4], with reconstruction error in millimeter level, is the start-of-the-art work to reconstruct a scene from RGB-D stream input. We reconstructed

scenes from dense frames, estimated the camera poses of every frame via BundleFusion and projected the reconstructed scene onto the camera pose to get a dense depth map.

4 Experiments

We will show that our method achieves the comprehensive correction of various errors in the ToF depth map, and surpasses the state-of-the-art work in quantitative evaluations, in both our datasets and other publicly available datasets. We use five modules to build the CNN used in this work, because this is the minimum number to achieve convergence of performance on iteration.

4.1 Error Removal

We first trained our CNN on the Synthetic dataset and tested it on its test set in different scenes. And for the Real-World dataset, we adopted the model trained on the Synthetic dataset and performed a fine-tuning. The model trained for 150 epochs on our Synthetic dataset and fine tuning for 30 epochs on our Real-World dataset. We compared our result to the ToF depth map directly output by camera without additional processing, which serves as the baseline.

Table 1. The performance of the depth map output by each module on test sets of Synthetic and RealWorld Dataset, expressed in absolute error and relative error (in the form of **cm/%**).

	Module1	Module2	Module3	Module4	Module5
Synthetic	2.59/1.10	1.74/0.78	1.58/0.72	1.49/0.68	1.46/0.67
RealWorld	2.51/1.56	2.10/1.30	1.94/1.22	1.84/1.15	1.79/1.12

Figure 3(a) qualitatively demonstrates the effectiveness of our method. There are many typical errors in the original ToF depth map. The first is the common imaging noise that exists in both cases. Secondly, there is a bathtub with high mirror reflectivity and concave shape in case 1, which forms a very typical local MPI; the corner in case 2 has the same but slighter effect. Thirdly, the mirror in case 1 has a very high specular reflectance and a very low retroreflectivity, resulting in most of the echo received in this direction is the signal which from other surfaces, this phenomenon also occurs in other surfaces that have a large angle with imaging plane. The last is low intensity, the black velvet curtain in case 2 with a very low reflectivity absorbed most of the light, lead to missing information in this part of the ToF depth map. All the errors mentioned above are strongly removed by our network architecture, and we could obtain a corrected depth map very close to the ground truth.

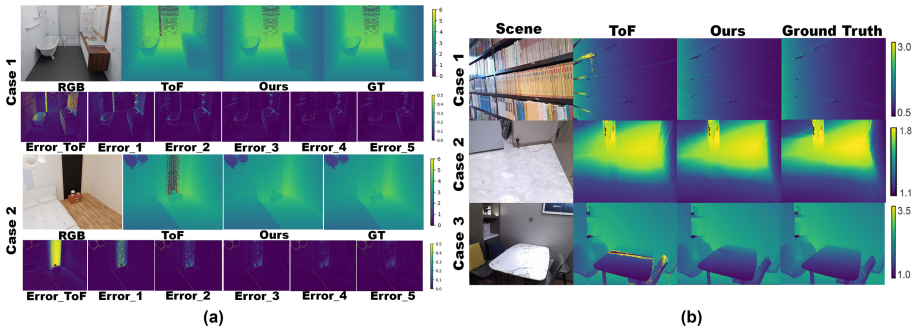

Fig. 3. (a) The results of ToF depth map error elimination in different stages, including RGB image (For visualize only, not the input of our method), ToF depth map, Result of our method, Ground Truth depth map, Absolute Error of the ToF depth map, Absolute Error of the depth map from module 1–5. (b) Results of our method on RealWorld dataset.

The six error maps below each case show the process of error removal from the ToF depth map by our proposed CNN. It can be seen from these error maps, that most of the imaging noise has been removed after the ToF depth map passes through the first module, and the relatively simple local MPI occurs in the bathtub and corner is reduced, too; while errors from other sources are still there. When the depth map passes through the second and third modules, various MPI begin to be eliminated. Also, with cascading of the modules, the feature map of each module captures higher and higher levels of information, so the understanding of the scene steadily increases. As a result, low intensity caused noise is gradually recovered, wrong depths are being corrected and missing depths are being filled. The last module refine the depth map from the antecedent module and is tend to convergence. Table 1 shows a quantitative evaluation that with iteration of modules, the performance gradually improves, and reach convergence on last two modules. This is the reason that we use five modules in total.

We selected several ToF depth maps with representative error patterns in different challenging scenes, which are a bookstore, kitchen, and dining room, respectively. From Fig. 3(b), it can be seen that our CNN still shows fine performance under the real-world dataset. In case 1, the depth on the lower reflectivity part of the bookshelf gap is completed. The general local MPI in case 2 makes the depth of the ToF camera notably larger around the corners, and our CNN outputs a result precisely removed the error. In case 3, the depth of the marble tabletop, which is affected by serious MPI due to the high specular reflectivity of the surface, is also corrected.

4.2 Compared to State-of-the-Art Methods

Table 2. Quantitative testing results of our method and other state-of-the-art methods on our dataset. Results with the best performance are marked in bold.

	Synthetic dataset		RealWorld dataset	
	MAE (cm)	Rela. MAE (%)	MAE (cm)	Rela. MAE (%)
ToF	9.73	3.47	4.17	2.51
Su et al. [22]	4.23	1.78	3.32	2.21
Agresti et al. [2]	1.87	0.81	2.27	1.38
Ours	**1.46**	**0.67**	**1.79**	**1.12**

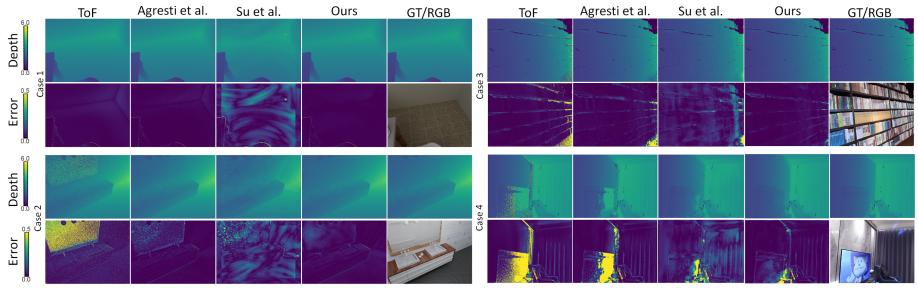

Fig. 4. Comparison between our method and other state-of-the-art methods. Each case shows the depth map and error map obtained by different methods, and provides the RGB image of the scene for reference.

On Our Datasets. We conducted a comprehensive evaluation to prove that our method is better than the state-of-the-art method. Specifically, we compare our work with Su et al. [22] and Agresti et al. [2], because they are representative and state-of-the-art work that focused on ToF depth error removal. The former takes only the raw correlation measurements at two modulation frequencies as input to predict a depth map in an end-to-end fashion. The latter input the depth map and intensity map at three modulation frequencies and make one residual prediction. We retrained them on our Synthetic and Real-World datasets to work with our data type using the same training procedure. Models are trained on Synthetic Dataset, and are fine tuned on RealWorld Dataset. These training procedures are all convergent to ensure reaching their optimal performance. Table 2 shows that our work surpassed other methods by a large margin, achieving the best performance on both datasets. Figure 4 shows a more

intuitive and detailed comparison, with case 1 and case 2 on Synthetic dataset, case 3 and case 4 on our Real-World dataset. Our approach gracefully removed most of the error mentioned above, including local MPI in case 1, imaging noise and mirror reflection in case 2 and case 4, and typical low intensity in case 3, and gave a clean smooth depth map and error map close to zero.

These two methods, one overly relies on the fitting ability of CNN in terms of input and output, leading to artifacts and performance degradation in some area [22]; the other only performs a single error prediction via a rather small-scale network, thus is unable to detect and correct all kinds of error, especially the part that needs the understanding of scenes [2]. This also proves the advanced nature of iterative error correction from the negative side since it avoids these issues.

Table 3. Quantitative testing results of our method and other state-of-the-art methods on datasets where they were proposed. Results with the best performance are marked in bold. (*) is the result on its MPI removal network. Median and Interquartile Range (IQR) are calculated on error (not absolute error).

Dataset	Datatype	Proposed approach on this dataset	Ours
Marco et al. [16]	Depth map	First quartile: 4 cm Second quartile: 9 cm Third quartile: 17 cm	**First quartile: 2.24 cm** **Second quartile: 4.89 cm** **Third quartile: 8.68 cm**
Su et al. [22]	Correlation maps in 40 MHz and 70 MHz	MAE: 2.9 cm SSIM: 0.9631	**MAE: 2.28 cm** **SSIM: 0.9906**
Agresti et al. [2]	Depth, Amplitude and Intensity map in 20 MHz, 50 MHz and 60 MHz	(Synthetic) MAE: 7.49 cm (RealWorld) MAE: 3.19 cm	**(Synthetic) MAE: 5.80 cm** **(RealWorld) MAE: 2.95 cm**
Guo et al. (FLAT) [9]	Correlation maps in 10 MHz, 50 MHz, 75 MHz	**Median: −0.01 cm*** IQR: 2.63 cm Percentile 90th: 4.16 cm	Median: −0.07 cm **IQR: 1.75 cm** **Percentile 90th: 2.60 cm**

On Other Datasets. We also compared our approach on the datasets on which previous methods are proposed. We collected, to our best knowledge, all publicly available datasets in the recent years proposed for ToF error correction. We then trained the proposed CNN on these datasets, with the number of modules and Resblocks slightly adjusted. By adjusting the early fusion of our CNN, we enable their data to fit in our network, and compare our method with the native methods on these datasets with the same quantitative evaluation. The results are shown in Table 3, which clearly shows the excellence of our method since our approach outperforms the previous methods proposed on its dataset in most of the indicators. Through testing on different datasets, feeding different inputs, and comparing with different evaluation criteria, we demonstrated the superiority and robustness of our method.

5 Conclusion and Future Work

In this paper, we proposed a new method for ToF depth error removal. We have shown that an iteratively error removing CNN can produce a better depth map for ToF imaging, because it conforms to the nature of the ToF depth map and can avoid the inherent defects of CNN, theoretically and experimentally. We achieved this by analyzing ToF error removal principles, designing a CNN framework for this problem, introducing two new large datasets, and experimenting on different datasets. In future work, we plan to perform super-resolution to the depth map of the ToF camera so that it can be aligned with an RGB image.

References

1. 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)
2. Agresti, G., Zanuttigh, P.: Deep learning for multi-path error removal in ToF sensors. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
3. Chen, Y., Ren, J., Cheng, X., Qian, K., Wang, L., Gu, J.: Very power efficient neural time-of-flight. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 2257–2266 (2020)
4. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM Trans. Graph. (ToG) **36**(4), 1 (2017)
5. Dorrington, A.A., Godbaz, J.P., Cree, M.J., Payne, A.D., Streeter, L.V.: Separating true range measurements from multi-path and scattering interference in commercial range cameras. In: Three-Dimensional Imaging, Interaction, and Measurement, vol. 7864, p. 786404. International Society for Optics and Photonics (2011)
6. 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
7. Fuchs, S.: Multipath interference compensation in time-of-flight camera images. In: 2010 20th International Conference on Pattern Recognition, pp. 3583–3586. IEEE (2010)
8. Godbaz, J.P., Cree, M.J., Dorrington, A.A.: Closed-form inverses for the mixed pixel/multipath interference problem in AMCW lidar. In: Computational Imaging X, vol. 8296, p. 829618. International Society for Optics and Photonics (2012)
9. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3D ToF artifacts through learning and the flat dataset. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 368–383 (2018)
10. 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), 1–18 (2015)
11. Hansard, M., Lee, S., Choi, O., Horaud, R.P.: Time-of-flight cameras: principles, methods and applications. Springer Science and Business Media (2012). https://doi.org/10.1007/978-1-4471-4658-2

12. He, Y., Liang, B., Zou, Y., He, J., Yang, J.: Depth errors analysis and correction for time-of-flight (ToF) cameras. Sensors **17**(1), 92 (2017)
13. Izadi, S., et al.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 559–568 (2011)
14. Jarabo, A., Marco, J., Muñoz, A., Buisan, R., Jarosz, W., Gutierrez, D.: A framework for transient rendering. ACM Trans. Graph. (ToG) **33**(6), 1–10 (2014)
15. Jongenelen, A.P., Bailey, D.G., Payne, A.D., Dorrington, A.A., Carnegie, D.A.: Analysis of errors in ToF range imaging with dual-frequency modulation. IEEE Trans. Instrum. Meas. **60**(5), 1861–1868 (2011)
16. Marco, J., et al.: DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Trans. Graph. (ToG) **36**(6), 1–12 (2017)
17. Naik, N., Kadambi, A., Rhemann, C., Izadi, S., Raskar, R., Bing Kang, S.: A light transport model for mitigating multipath interference in time-of-flight sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 73–81 (2015)
18. 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)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Sarbolandi, H., Lefloch, D., Kolb, A.: Kinect range sensing: structured-light versus time-of-flight Kinect. Comput. Vis. Image Underst. **139**, 1–20 (2015)
21. Song, S., Shim, H.: Depth reconstruction of translucent objects from a single time-of-flight camera using deep residual networks. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11365, pp. 641–657. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20873-8_41
22. 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)
23. Wang, Z., Chen, J., Hoi, Steven C.H.: Deep learning for image super-resolution: a survey. IEEE Trans. Patt. Anal. Mach. Intell., 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.2982166