Very Power Efficient Neural Time-of-Flight

Yan Chen[1,2]*, Jimmy Ren[1]*, Xuanye Cheng[1], Keyuan Qian[2], Luyang Wang[1], Jinwei Gu[1]

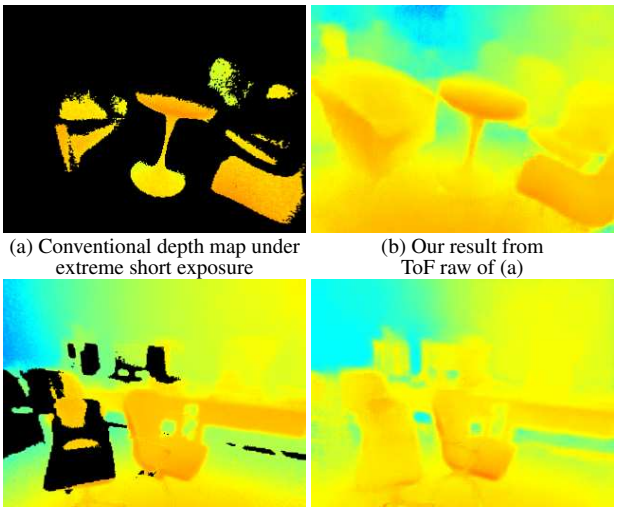
[1]SenseTime Research
{chenyan, rensijie, chengxuanye, wangluyang, gujinwei}@sensetime.com
[2]Tsinghua University
qianky@sz.tsinghua.edu.cn

Abstract

Time-of-Flight (ToF) cameras require active illumination to obtain depth information thus the power of illumination directly affects the performance of ToF cameras. Traditional ToF imaging algorithms are very sensitive to illumination and the depth accuracy degenerates rapidly with the power of it. Therefore, the design of a power efficient ToF camera always creates a painful dilemma for the illumination and the performance trade-off. In this paper, we show that despite the weak signals in many areas under extreme short exposure setting, these signals as a whole can be well utilized through a learning process which directly translates the weak and noisy ToF camera raw to depth map. This creates an opportunity to tackle the aforementioned dilemma and make a very power efficient ToF camera possible. To enable the learning, we collect a comprehensive dataset under a variety of scenes and photographic conditions by a specialized ToF camera. Experiments show that our method is able to robustly process ToF camera raw with the exposure time of one order of magnitude shorter than that used in conventional ToF cameras. In addition to evaluating our approach both quantitatively and qualitatively, we also discuss its implication to designing the next generation power efficient ToF cameras.

1. Introduction

Depth sensing is one of the core components of many computer vision tasks. Amplitude-modulated continuous-wave (AMCW) time-of-flight (ToF) has a brief and definite physical meaning in depth construction of scenes thus it attracts a lot of commercial attention, such as Kinect V2. It is also widely used in academic research of computer vision [13, 30], including human tracking [29], 3D scene reconstruction [18], robotics [17], object detection, gesture recognition [23, 33], and scene understanding [31, 15]. However, comparing with traditional RGB cameras, ToF cameras compute the depth by emitting a periodic amplitude modulated illumination signal and receive the demodulated signal reflected by the objects. Higher power of active illumination enables the ToF sensor to receive the signal with higher signal noise ratio (SNR) and higher level of confidence. Therefore, the power of illumination directly influences the performance of ToF cameras.



(a) Conventional depth map under extreme short exposure
(b) Our result from ToF raw of (a)
(c) Conventional depth map under regular exposure
(d) Our result from ToF raw of (c)

Figure 1. We propose an end-to-end pipeline to translates the weak and noisy ToF camera raw to high quality depth map. (a) Depth image produced by the ToF camera's default imaging pipeline with 200us exposure time. The quality is very poor. (b) Depth image produced by our method applied to the ToF camera raw from (a). (c) Depth image produced by the ToF camera's default imaging pipeline with regular exposure time. Some of the depth information is still lost due to objects with low reflectivity or long distances. (d) Depth image produced by our method applied to the ToF camera raw from (c).

*Indicates equal contribution.

Traditional ToF imaging algorithms are very sensitive to illumination and the depth accuracy degenerates rapidly with the decreasing illumination power. In order to obtain more accurate depth information, one way is to increase the intensity of the received active illumination signal. Other than increasing the illumination power, an alternative treatment to this issue is to increase the physical size of the pixels on the sensor to collect more light. However, this significantly decreases the depth map resolution. According to the inverse square law, one can also cut the depth sensing range of the camera. This obviously decreases the usability of the camera in many applications. Therefore, to make a ToF camera with satisfactory depth quality as well as reasonable resolution and sensing range, the painful dilemma for the illumination and the performance trade-off always troubles the designer of the camera if the conventional imaging pipeline is used.

Such dilemma can be tackled if there is a way to recover high quality depth information from weak signals. A number of recent studies show that it is plausible to recover low SNR natural images from very noisy data using deep learning. [27, 16, 5]. Chen et al. [5] showed impressive results on recovering high quality color image from camera Bayer pattern which is captured under extremely low light condition with short exposure. Inspired by these research, we show for the first time that for ToF cameras, despite the weak signals in many areas under the extreme short exposure setting, these signals as a whole can be well utilized through a learning process which directly translates the weak and noisy ToF camera raw to high quality depth map. This creates an opportunity to address the aforementioned dilemma and makes it possible to design a very power efficient ToF camera possibly with higher resolution and longer sensing range. To enable the learning, we collect a comprehensive dataset under a variety of scenes and photographic conditions via a specialized ToF camera. The dataset contains ToF raw measurements and depth maps collected under extreme short exposure settings and long exposure settings respectively. We show in the experiments that our proposed method is able to robustly process ToF raw measurements with an exposure time that is one order of magnitude shorter than that used in a conventional ToF camera.

The contributions of our work can be summarized as follows.

- We show for the first time that our proposed method is able to recover high quality depth information from very weak ToF raw data (one order of magnitude shorter exposure time).
- We introduce a real-world dataset used for training and validating the this learning tasks.
- We shed light on the design of the next generation ToF camera by providing an effective alternative to optimize the performance and power consumption trade-off.

2. Related Work

Depth reconstruction based on ToF cameras. ToF cameras face a lot of challenging problems when extracting depth from raw phase-shifted measurements with respect to emitted modulated infrared signal. Dorrington et al. [7] established a two-component, dual-frequency approach to resolving phase ambiguity, achieving significant improvements of the accuracy when distortion is caused by multi-path interference (MPI). Several methods were proposed to deal with MPI distortions, including adding or modifying hardware [34, 14, 26, 3], employing multiple modulation frequencies [7, 8, 4, 12] and estimating light transport via an approximation of depth [9, 10]. Marco et al. [22] correct MPI errors by a two-stage training strategy, training the encoder to represent MPI-corrupted depth images with captured dataset firstly and then use synthetic scenes to train the decoder to correct the depth. However, the above pipelines are based on the assumption that there is no cumulative error and information loss introduced in the previous stages, thus the final results of these methods are likely to contain cumulative errors of multiple stages.

Krishna et al. [35] filled the missing depth pixels by using a color-aware Gaussian-weighted averaging filter to estimate depth value. However, its performance is limited by the similarity between the neighborhood pixels and target pixels and the information of the target region is wasted. An end-to-end ToF image processing framework presented by Su et al. [32] can efficiently reduce noise, correct MPI and resolve phase ambiguity. However, the training data is not realistic. Therefore, depth reconstruction may fail when the scene contains low reflectivity materials and objects. To the best of our knowledge, none of existing depth reconstruction method is able to obtain high quality depth map from the weak and noisy ToF camera raw measurements.

Image enhancement under low light. For conventional RGB cameras, photography in low light is challenging. Several techniques have been proposed to increase the SNR of the recovered image [11, 25, 20, 21, 6]. Chen at el. [5] established a pipeline by training a fully convolutional neural network which directly translate the very noise and dark Bayer pattern camera raw to high quality color images. Though impressive results from the aforementioned studies, deep learning and data-driven approaches have not yet been adopted to recover high quality depth information from weak and noisy ToF raw. It remains unclear if such methodology is effective for ToF imaging. The aim of this paper is to disclose its feasibility.

Depth datasets. Although recently many datasets of depth maps are proposed, most of them consist of synthetic data, such as transient images generated via time-resolved rendering. A dataset of ToF measurements [22] is proposed

via simulating 25 different scenes with a physically-based, time-resolved renderer. Besides, Su et al. [32] offer a large-scale synthetic dataset of raw correlation time-of-flight with ground truth labels. However, the ToF raw with artificial distortions and Gaussian noise is not realistic enough to support the real life generalization especially when dealing with areas with large noise caused by low reflectivity. Only the raw RGB data, depth map and accelerometer data are provided in the NYU-Depth V2 dataset [24] but ToF raw measurements are missing. Thus, this dataset can not be used to train ToF raw to depth map conversion. Furthermore, most existing depth datasets concentrate on images captured under appropriate illumination or ideal environments, they are not suitable for evaluating imaging with low active illumination power or weak reflected signal. In this paper, we propose a comprehensive dataset to fill these gaps and enable the training and validation of our proposed model.

3. Method and Analysis

3.1. Imaging with Time-of-Flight Sensors

Distance measurement. The distance measurement mode of Time of Flight uses the on chip driver and the external LED/LD to provide modulated light on the target. Generally, the period of the modulation control signal is programmable. The modulator generates all signals to modulate the external LED/LD and simultaneously all demodulation signals to the pixel-field. We can describe the programmable modulation optical signal with angular frequency ω as

$$s(t) = \cos(\omega t), \tag{1}$$

where the amplitude is normalized. Once the signal is reflected by the object, the modulated optical signal goes back to the sensor with certain amplitude attenuation and certain phase shift, then the received signal can be expressed as

$$r(t) = \alpha \cos(\omega t - \varphi) + \delta, \tag{2}$$

where δ is the offset, α is the amplitude after attenuation, and φ is the phase shift. In order to achieve demodulation, the original emission signal needs to be used as a correlation signal and demodulated with the received signal as

$$\begin{aligned} \varphi_{sr} &= r(t) \otimes s(t) \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} r(t)s(t+\tau)\, dt \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} [\alpha \cos(\omega t - \varphi) + \delta][\cos(\omega t + \omega\tau)]\, dt \\ &= \frac{\alpha}{2} \cos(\omega t + \varphi), \end{aligned} \tag{3}$$

where the relevant signal is denoted as $C(\tau) = \varphi_{sg}(\tau)$. ToF cameras need to sample the correlation signal $C(\tau)$ four times in one cycle. That is, sampling is performed when $\omega\tau_0 = 0°, \omega\tau_{90} = 90°, \omega\tau_{180} = 180°, \omega\tau_{270} = 270°$. Considering the received signal is mainly superimposed on the background image, we also need to consider an offset here. Then the phase shift φ and the amplitude $\frac{\alpha}{2}$ can be obtained from the four sample values

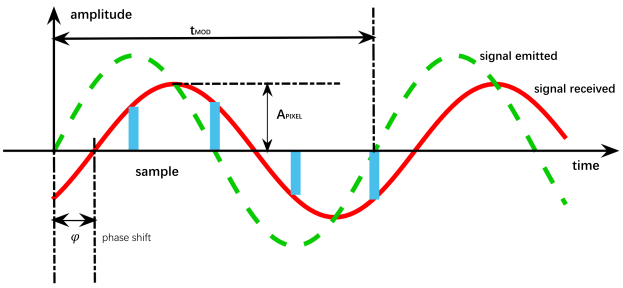


Figure 2. Sample of received signal per $\pi/4$.

$$\varphi = \arctan \frac{C(\tau_3) - C(\tau_1)}{C(\tau_0) - C(\tau_2)}, \tag{4}$$

$$\alpha = \frac{1}{2}\sqrt{[C(\tau_3) - C(\tau_1)]^2 + [C(\tau_0) - C(\tau_2)]^2}, \tag{5}$$

and we can find the distance value by the phase shift

$$d = \frac{c\varphi}{2\omega}, \tag{6}$$

where c is the speed of light.

Quality of the measurement result. Raw ToF measurements contain the distance information, as well as the quality and the validity (confidence level) of the received optical signal. A higher amplitude of the measured signal represents a more accurate distance measurement. The depth data for each pixel has its own validity and quality in ToF cameras. The amplitude of the modulated light received by the ToF sensor is the primary quality indicator for the measured distance data. It can be calculated as Eq. 5. However, excessive active illumination will make the amplitude of the raw measurements very large. This leads to errors in the depth value due to the problem of over-exposure of the ToF sensor.

Problems of Traditional Pipeline. In order to recover high-quality depth maps from imperfect ToF raw measurements, traditional methods of ToF camera imaging often require a series of specialized processing techniques, such as denoising, correction of multipath distortion and nonlinear compensation, etc. However, these components are independent to each other and often relies on the assumption of no cumulative error and information loss in the previous stages. In practice, this assumption is almost always not true. It may cause large errors in the final depth map.

3.2. Learning from imperfect ToF camera raw

In this section, our approach of depth reconstruction is presented in detail. We first describe the advantage of our

Network Architecture												
Name	D1	D2	D3	D4	U1	Conv1	U2	Conv2	U3	Conv3	U4	Conv4
Layer	conv+ Relu	conv+ Relu+ conv+ Relu	conv+ Relu+ conv+ Relu	conv+ Relu	conv+ Relu+ upsample	conv+ Relu+ conv+ Relu	conv+ Relu+ upsample	conv+ Relu+ conv+ Relu	conv+ Relu+ upsample	conv+ Relu+ conv+ Relu	conv+ Relu+ upsample	conv
Kernel	5×5	3×3	3×3	3×3	3×3	1×1/ 3×3	3×3	1×1/ 3×3	3×3	1×1/ 3×3	3×3	3×3
Stride	2	2/1	2/1	2	1	1	1	1	1	1	1	1
I/O	4/16	16/32	32/64	64/128	128/128	192/128	128/96	128/64	64/32	48/16	16/8	8/4
Input	ToF raw	D1	D2	D3	D4	D3+U1	Conv1	D2+U2	Conv2	D1+U3	Conv3	U4

Table 1. Architecture of our network. "conv" represents a convolutional layer. "upsample" means an up-sampling layer.

method of recovering high-quality depth images from weak and noisy ToF camera raw measurements compared to traditional ToF imaging methods. Then, we give a brief description of our whole pipeline to learn a mapping from ToF measurements acquired under low power illumination to corresponding high-quality depth map. We will introduce our carefully designed ToF loss and the network architecture of our method, as shown in Tab. 1, will be introduced. Finally, we present how we train the model and implementing details.

Comparison to traditional pipeline. The raw ToF measurements have a very low signal-to-noise ratio (SNR) and amplitude intensity, when the active illumination signal received by the ToF sensor is very low. In this case, conventional edge aware filtering methods such as bilateral filter tend to fail. Traditional method of ToF measurements denoising is based on arbitrary rules and assumptions, but these rules and assumptions often become invalid with changes in scenes and intensity of the received signals. This is particularly true for weak input signals. Therefore, it is very difficult to select the optimal parameters for all the image processing components to achieve good results for all scenarios. In contrast, the proposed method adopts the end-to-end learning and inference approach to translate the weak and noisy ToF camera raw to high quality depth map which avoids the highly complex parameter tuning for such noisy and weak input signals.

Network structure. To build intuition for this end-to-end approach, we have analyzed several previous work of image-to-image mapping. Most of them have adopted an encoder-decoder network with or without skip connections [28], which consists of down-sampling, residual blocks and up-sampling. The pixel value of ToF depth map is closely-related to camera settings, scene architecture and layout, compared with RGB images. Besides, the geometry and architecture of scene for both depth map and raw measurements are required to be consistent. And these specific characteristics of ToF raw measurements should be combined with the previous work of image translation, when designing network architecture. Finally, the depth map can be calculated with four-frame raw data by Eq. 4 and Eq. 6.

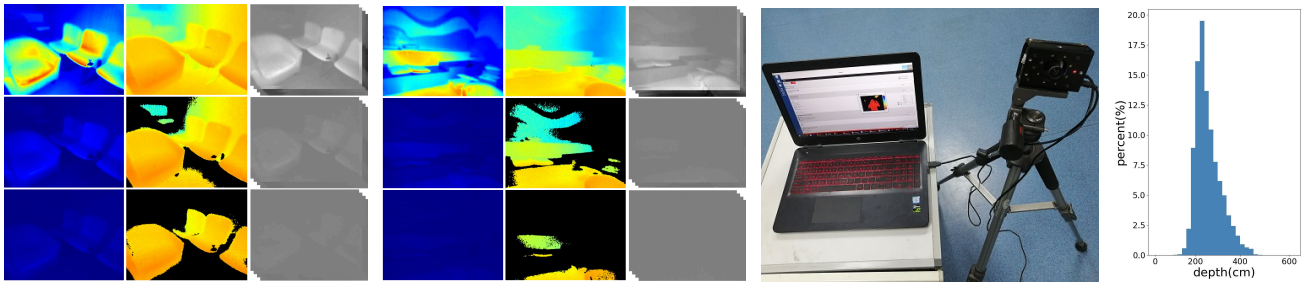
For the above considerations, we devise our network architecture based on the encoder-decoder construction with skip connections. The size of input is progressively decreasing in pace with going through the down-sampling layers for four times, until it reaches the layer U1. And after passing through four up-sampling layers, the size of input becomes larger and restored to its original size. The strided convolution layers combined with activation layers serve as decoder and the strided convolution layers combined with activation layers and up-sample layers are regarded as encoder. Moreover, we added the skip connections to the network between each pair of i D layer and n-i Conv layer following the U-net to enhance the accuracy of results.

ToF Loss. Different from traditional RGB camera imaging, both of the depth map and ToF raw measurements should be consistent with the underlying scene structure and geometry and can be influenced by spatial structures, modulation frequency, materials of target objects and many other factors. Thus, in order to improve the accuracy of depth reconstruction, we devise a method of mapping from raw ToF measurements captured under short exposure time to ToF raw measurements captured under long exposure and then calculate the depth value from generation ToF raw according to the physical meanings of ToF camera imaging. Considering the above, we propose a new loss named ToF Loss consisting of loss based on raw measurements and depth value for depth reconstruction.

In order to minimize the mean absolute error between each frame of raw measurements and its corresponding frame of ground truth, we introduce the L_{raw} loss as follow:

$$L_{raw} = \frac{1}{N}\sum_{i,j}^{N}[\left|r_{i,j}^{gt_0} - r_{i,j}^{pre_0}\right| + \left|r_{i,j}^{gt_1} - r_{i,j}^{pre_1}\right| + \left|r_{i,j}^{gt_2} - r_{i,j}^{pre_2}\right| + \left|r_{i,j}^{gt_3} - r_{i,j}^{pre_3}\right|], \tag{7}$$

where $r_{i,j}^{pre_n}$ (n=0, 1, 2, 3) represents ToF raw measurements generated by our network and $r_{i,j}^{gt_n}$ means ToF raw measurements captured under long exposure settings as ground truth.



(a) a group of measurements (b) a group of measurements ToF camera Depth statistics.

Figure 3. We use EPC660 ToF camera from ESPROS to collect a dataset of multiple pairs of short-exposure and corresponding long-exposure depth measurements. Diverse indoor scenes are collected in the dataset, including office room, restaurant, bedroom and Laboratory. The depth range is reasonable for indoor scenes of our dataset, for depth values range from 0cm to 591cm and has a mean of 236.88cm.

In order to ensure that high-quality depth maps can be calculated from raw measurements generated and minimize the mean absolute error between depth value calculated from ToF raw measurements and target depth value, the L_{depth} loss are introduced in Eq. 8.

$$L_{depth} = \frac{1}{N}\sum_{i,j}^{N}\left|d_{i,j}^{gt} - d_{i,j}^{pre}\right|, \tag{8}$$

$$d_{i,j}^{pre} = \frac{c}{2}\cdot\frac{1}{2\pi f_{LED}}\cdot[\pi + atan2(\frac{r_{i,j}^{pre_1} - r_{i,j}^{pre_1}}{r_{i,j}^{pre_2} - r_{i,j}^{pre_0}})], \tag{9}$$

where f_{LED}=6MHz, c=299792458 m/s. In this case, $d_{i,j}^{pre}$ does not represent depth value of the network output and its meaning has been changed to depth value calculated from ToF raw measurements produced by network according to Eq. 9. Thus, in order to applying this new ToF Loss to our network, we have to change the output of the network from depth maps to raw measurements and the output of last layer to four channels.

Thus, the ToF Loss consists of weighted L_{raw} and L_{depth},

$$L_{tof} = \alpha L_{depth} + \beta L_{raw}, \tag{10}$$

Experiments show that we can achieve better performance by adopting ToF Loss than just with the supervision of depth map.

Training details. Our networks are implemented in Pytorch. During training, inputs of the network are the ToF raw measurements captured under short exposure and the ground truth is the corresponding depth map captured under regular or long exposure. We randomly crop out 128×128 images on the original 320×240 images for data augmentation. This strategy effectively improves the robustness of the model. We train our network using the Adam optimizer [19] with an initial learning rate of 0.0002 for the first 200 epochs, before linearly decaying it to 0 over another 1800 epochs. In additon, our method can achieve best peformance with α = 1 and β = 0.1, and more details will be shown in supplementary materials. Experiments show that our network can process single-frame image on the GPU of chip on high-end mobile phones at a speed of 20ms. Besides, the NPU of Qualcomm with higher computing power can also support all the layers of our network, such that our algorithm can meet the requirements of frame rate in application.

4. Dataset

To enable the learning, we collect a comprehensive dataset under a variety of scenes and photographic conditions by a specialized ToF camera with raw data access. Due to the limitations of hardware devices, it is difficult to change the intensity of received signals by directly changing the physical size of the pixels on the ToF sensor or the power of the infrared LED illumination of the development kit. However, we can modify the intensity of received signal by changing the exposure time of the ToF camera, since the exposure time is directly proportional to the intensity of received signal.

We use EPC660 ToF camera from ESPROS to collect a dataset of multiple pairs of short-exposure and corresponding long-exposure depth measurements for training the proposed architecture. ToF raw measurements, amplitude image and depth map at 320×240 resolution are collected for each scene with an exposure time. We captured 5000 groups of measurements with 200us and 400us exposure time respectively and 5000 groups of corresponding long-exposure images from a variety of scenes with varying materials. During the experiments, we use 4500 groups for training and 500 groups for testing.

Diverse indoor scenes are collected in the dataset, including office room, restaurant, bedroom and laboratory. We adopt the ideal sinusoidal modulation functions to avoid the wiggling effect. The images are generally captured at night in rooms without infrared monitoring to avoid the influence of solar radiation and infrared light emitted by some particular machines. Note that a variety of hard cases such as distant objects, fine structures, irregular shapes and var-

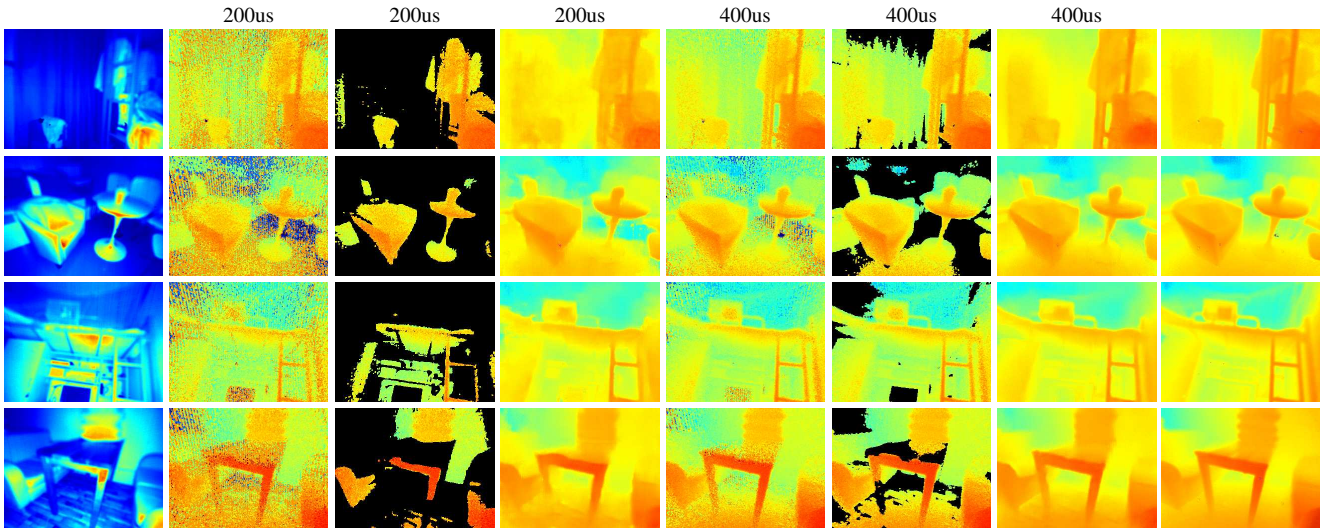


Figure 4. Experiment results. In order to verify the effectiveness of our proposed method, we validated our method on our test set for exposure times of 200us and 400us respectively. The results show that our method is able to robustly process ToF camera raw with the exposure time of one order of magnitude shorter than that of conventional ToF cameras.

ious materials including fabric, metals with low reflectivity and dark object with high absorptivity exist in our scenes.

We mount the ToF camera on a sturdy tripod to avoid camera shaking and other vibration when capturing. Due to continuous modulation, 6MHz was selected as modulation frequency for measuring depth in our scenes with range of 0-6 meters to prevent roll-over being observed. Then exposure time is adjusted to obtain high-quality raw data. After long-exposure ToF measurements captured, we decrease the exposure time to 200 us and 400 us respectively via software on computers to collect data without touching the cameras.

A mask to evaluate the quality of ToF measurements will be introduced into our dataset. Actually, the quality and validity of the received signal exists in raw data collected by ToF cameras. The signal amplitude as well as the ratio of ambient-light E_{BW} to the value of modulated light E_{ToF} (AMR) indicates the quality and validity of received signal. We combine these two features of received signals in a certain proportion to generate a quantitative criteria for evaluating the quality of each pixel in measurements. A threshold for criteria can be defined to produce a mask for each pixel. This mask can be adopted in network training and depth map generation. For instance, unconfident pixels in the labels can be ignored during the computation of error gradients in training.

Fig. 3 shows quantitative analysis of depth-range distribution of ToF measurements in our dataset. The depth range is reasonable for indoor scenes of our dataset, for depth values range from 0cm to 591cm and has a mean of 236.88cm. There are some regions with no depth value or much noise when short-exposure, due to few reflected photons detected. The ToF measurements is sufficient to serve as ground truth, though some noise still exists.

5. Experiments and Results

5.1. Quantitatively results

We first quantify depth error with the mean absolute error (MAE) and the structural similarity (SSIM) [36] of predicted depth map compared to the ground truth. At the same time, we will quantitatively analyze the variation of the error of our method at different detection distances. Then we will discuss the impact of our approach on the power consumption of ToF cameras. Finally, we also compare the denoise effect with traditonal denoising method.

	200us		400us	
	MAE	SSIM	MAE	SSIM
Traditional pipeline	56.41	0.2216	32.53	0.5180
BM3D	44.06	0.4658	30.12	0.7475
Ours	10.13	0.9156	7.94	0.9342

Table 2. This table reports the mean absolute error (MAE)(cm) and the structural similarity (SSIM)(%) of 200us exposure time and 400us exposure time. Traditional method can recover depth information only in the local position under the low exposure setting, so the overall error is very large.

Effect of exposure time. Our dataset contains raw data acquired under 200us and 400us exposure time and their corresponding depth maps collected under regular exposure

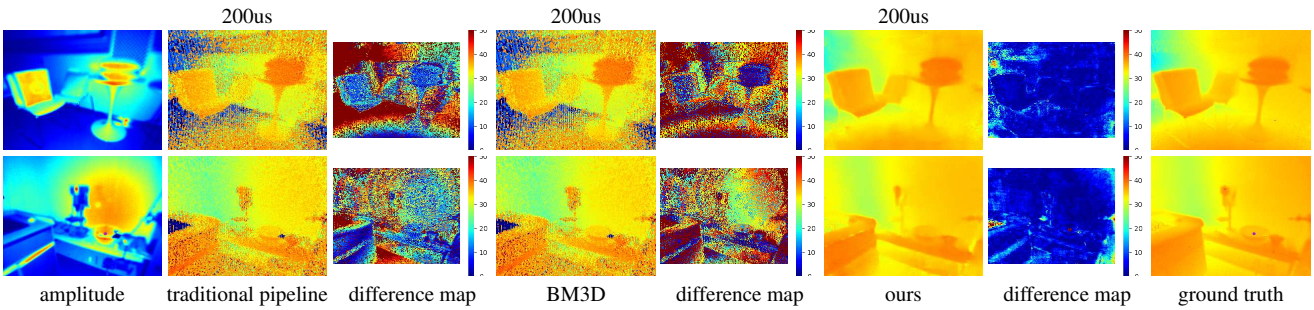


Figure 5. The performance of BM3D is compared to that of our method and our results are better in terms of details and precision in the depth map. We also provide a difference map that visually compares the difference between two processed results and ground truth respectively. Blue color indicates a small error in the difference map.

time. We have trained two models on the ToF raw measurements under 200us and 400us exposure respectively, and tested the accuracy of the two models with the corresponding test set. Then we calculate the mean absolute error (MAE) and the structural similarity (SSIM) [36] and compare the results of the traditional ToF camera pipeline with that of our proposed method on the test set. Note that the result is calculated over the whole test database in which the object distance varies between 0 to 591cm as indicated in the previous section.

As shown in Tab. 2, our results meet an overall 7.94cm depth error with raw captured under 400us exposure time and 10.13cm with raw captured under 200us exposure time. Although the accuracy of depth map produced by our method decreases with the reduction of exposure time, the experimental results of the two models both greatly exceed that of traditional pipeline method.

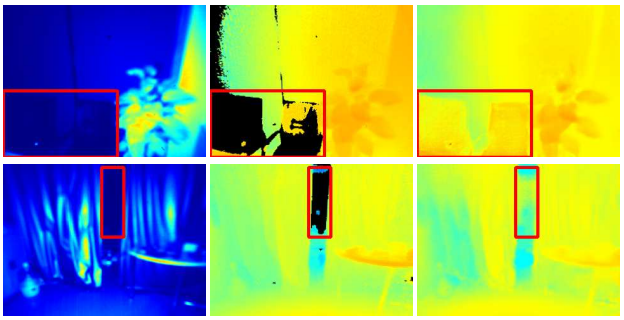
Effect of power consumption. It is necessary to analyze the power trade-offs of running a deep network vs. brighter illumination. We refer to the datasheet of the ToF camera EPC660 to learn that the power of eight Infrared LED - SFH4715s is 27200mw [2], thus reducing the exposure time by 20 times lowers the power to five percent of original power, which is about 1360mw. In addition, a mature AI chip Movidius has a power of less than 500mw [1]. Therefore, the total power consumption of chip and short exposure settings is less than 1860 mW, much smaller than that of long exposure settings, 272000 mw.

Comparison to denoising processing. In order to improve the accuracy of the depth map, a natural idea is to process the depth map with the existing denoising algorithm. The denoising performance of the BM3D algorithm outperforms most recent denoising models in natural images. Thus, we use the BM3D algorithm to denoise the depth map under short exposure. After that, we compare the results processed by BM3D with the depth map that we recovered from the corresponding raw data. As shown in Fig. 5, the depth map processed by BM3D loses more details and leaves perceptually significant noise compared to our results. Our results are also significantly better than that of the BM3D processing on difference map. As shown in Tab. 2, our results have better performance on both SSIM and MAE indicators, compared with BM3D.

5.2. Qualitative results on our dataset

Then, we present the results of our method and the traditional ToF camera imaging pipeline in extreme cases on our test dataset. In this section, we verify that our proposed end-to-end solution can still reconstruct accurate depth value in extreme case. Moreover, compared with the traditional method, if the exposure time is set to regular, our method is more robust to scenes with objects of high absorptivity or regions in distance. In addition, we also verify that our network is able to generate excellent results with ToF raw inputs collected under short exposure in the presence of infrared background light.

Qualitative results with different exposure time. We have shown that the amplitude value is an important indicator for evaluating the raw data quality of the ToF camera. Considering that the effect of the power level of the ToF camera active illumination system on the amplitude value in the amplitude map is equivalent to the effect of the length of exposure time, we simulate the power level of active illumination by controlling the length of the exposure time. In order to verify the effectiveness of our proposed method, we used the ToF raw measurements acquired under exposure time of 200us and 400us as the input of the network to predict the corresponding depth map. As shown in Fig. 4, our results have better performance, compared with the depth map generated by the traditional ToF pipeline. Experiments show that our method is able to robustly process ToF camera raw with the exposure time of one order of magnitude shorter than that used in conventional ToF cameras. However, though our method can still generate depth information when the confidence of reflected signal is too low, the accuracy will be influenced and this should be taken into account in application.



(a) Amplitude (b) ToF depth map with suitable exposure time (c) Our results from ToF raw of (b)

Figure 6. Traditional pipeline fails to recover the depth value of the regions marked out of the first group, since the black chair in the scene strongly absorbed signal emitted by ToF camera. For the regions marked out of the second group, too large distance results in few photons received by the ToF sensor. In contrast, our method is able to obtain high-quality depth maps for these two regions.

Robustness under regular exposure. Since there may exist some objects with low reflectivity or too large distance in the scene, choosing the appropriate exposure time or a strong power active illumination does not guarantee that the depth map of the entire scene is of high quality. However, our proposed method has better performance in the depth estimation of these objects, compared with traditional ToF process, due to the ability of translating the weak and noisy ToF camera raw to depth map directly. As shown in Fig. 6, we deliberately collected some scenes with dark objects and scenes with large distances(such as black stools and computer screens, glass doors with specular reflections, as well as objects with particularly large depth differences in the scene) to prove the robustness of our method in this case.

Robustness in the presence of ambient light. To verify the robustness of this method in the presence of infrared background light, we collected 500 groups of ToF measurements in an environment containing sunlight or infrared equipment to fine-tune our network. We discovered that if the camera uses the long exposure settings (e.g. 4000us) for ground truth collection, the influence of other infrared light is very small. This is the case for both outdoor and indoor scenes with sunlight. As shown in Fig. 7, by using such labels in training, the network is able to generate high-quality depth map with ToF raw inputs captured under short exposure settings.

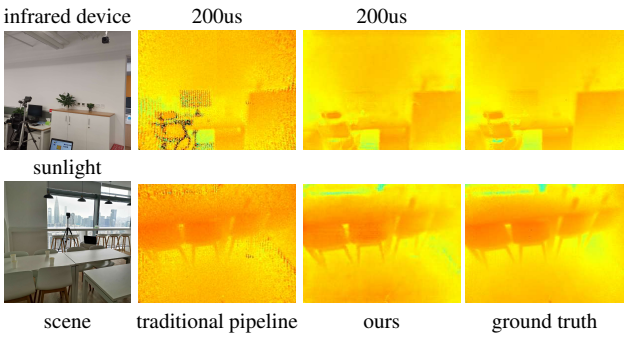


Figure 7. To verify the robustness of this method in the presence of infrared background light, we selected some scenes containing a monitoring device with infrared illumination and sunlight. The results show that even with some infrared signal interference, our method can still recover high-quality depth maps robustly.

6. Discussion and Conclusion

6.1. Implication to ToF camera design

Using neural network to robustly process ToF camera raw with very short exposure time (raw data with low SNR) is a novel alternative to optimize the power efficiency of the whole ToF system. Despite the involvement of neural network computation, the inception of many recent low power neural network hardware makes it a practical solution. Thus, our method can be very power efficent with high speed of neural network compution and low exposure settings for hardware. In addition to lowering the power consumption of ToF system, the results of this paper also provide a few extra design choices. First, higher depth frame rate may be achievable because the exposure time can be significantly reduced. Second, with the proposed method much smaller pixel size may be considered despite the SNR of the sensor raw could be low. Thus, higher depth resolution can thus be obtained with a reasonable power consumption. Such possibilities pave the way for new innovation in the ToF camera design.

6.2. Concluding remarks

In this paper, we discover that it is possible to devise a deep learning model to recover high quality depth information from very weak and noisy ToF raw measurements using deep learning. To realize the learning process, we collected a comprehensive dataset using a real-world ToF camera. We show in the experiments that our proposed method is able to robustly process ToF camera raw with the exposure time of one order of magnitude shorter than that used in conventional ToF cameras. While this neural network approach forms a key building block of a very power efficient ToF camera, it also shed new light on new innovations of the ToF camera design.

For future research, we will continue to improve the quality of our datasets. Specifically, we would adopt HDR imaging to improve the quality and precision of the ground truth depth map. Another opportunity for future work is to explicitly model the correction of the MPI error in an end-to-end trainable model to further enhance the accuracy of the results.

References

[1] Ai chip movidius: Myriad 2 vision processor. http://uploads.movidius.com/1441734401-Myriad-2-product-brief.pdf.

[2] Infrared led - sfh4715s. https://html.alldatasheet.com/html-pdf/1015980/OSRAM/SFH4715S/175/1/SFH4715S.html.

[3] S. Achar, J. R. Bartels, W. L. R. Whittaker, K. N. Kutulakos, and S. G. Narasimhan. Epipolar time-of-flight imaging. *Acm Transactions on Graphics*, 36(4):37, 2017.

[4] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar. Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *Optics letters*, 39(6):1705–1708, 2014.

[5] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. *arXiv preprint arXiv:1805.01934*, 2018.

[6] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu. Fast efficient algorithm for enhancement of low lighting video. IEEE, 2011.

[7] A. A. Dorrington, J. P. Godbaz, M. J. Cree, A. D. Payne, and L. V. Streeter. Separating true range measurements from multi-path and scattering interference in commercial range cameras. In *Three-Dimensional Imaging, Interaction, and Measurement*, volume 7864, page 786404. International Society for Optics and Photonics, 2011.

[8] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt. Sra: Fast removal of general multipath for tof sensors. In *European Conference on Computer Vision*, pages 234–249. Springer, 2014.

[9] S. Fuchs. Multipath interference compensation in time-of-flight camera images. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3583–3586, 2010.

[10] S. Fuchs, M. Suppa, and O. Hellwich. Compensation for multipath in tof camera measurements supported by photometric calibration and environment integration. In *International Conference on Computer Vision Systems*, pages 31–41, 2013.

[11] X. Guo, Y. Li, and H. Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017.

[12] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(5):156, 2015.

[13] F. Heide, W. Heidrich, M. Hullin, and G. Wetzstein. Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(4):36, 2015.

[14] F. Heide, M. B. Hullin, J. Gregson, and W. Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics (ToG)*, 32(4):45, 2013.

[15] S. Hickson, S. Birchfield, I. Essa, and H. Christensen. Efficient hierarchical graph-based segmentation of rgbd videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 344–351, 2014.

[16] Z. Hu, S. Cho, J. Wang, and M.-H. Yang. Deblurring low-light images with light streaks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3382–3389, 2014.

[17] S. Hussmann and T. Liepert. Robot vision system based on a 3d-tof camera. In *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*, pages 1–5, 2007.

[18] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, 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*, pages 559–568, 2011.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. volume 1, pages 541–551. MIT Press, 1989.

[21] H. Malm, M. Oskarsson, E. Warrant, P. Clarberg, J. Hasselgren, and C. Lejdfors. Adaptive enhancement and noise reduction in very low light-level video. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[22] J. Marco, Q. Hernandez, A. Munoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez. Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6):219, 2017.

[23] A. Memo and P. Zanuttigh. Head-mounted gesture controlled interface for human-computer interaction. *Multimedia Tools and Applications*, 77(1):27–53, 2018.

[24] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Transactions on Consumer Electronics*, 63(2):178–184, 2017.

[26] C. Peters, J. Klein, M. B. Hullin, and R. Klein. Solving trigonometric moment problems for fast transient imaging. *ACM Transactions on Graphics (TOG)*, 34(6):220, 2015.

[27] T. Remez, O. Litany, R. Giryes, and A. M. Bronstein. Deep convolutional denoising of low-light images. *arXiv preprint arXiv:1701.01687*, 2017.

[28] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[29] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, 30(3):217–226, 2012.

[30] S. Shrestha, F. Heide, W. Heidrich, and G. Wetzstein. Computational imaging with multi-camera time-of-flight systems. *ACM Transactions on Graphics (ToG)*, 35(4):33, 2016.

[31] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 567–576, 2015.

[32] S. Su, F. Heide, G. Wetzstein, and W. Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018.

[33] M. Van den Bergh and L. Van Gool. Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 66–72, 2011.

[34] A. Velten, D. Wu, A. Jarabo, B. Masia, C. Barsi, C. Joshi, E. Lawson, M. Bawendi, D. Gutierrez, and R. Raskar. Femto-photography: capturing and visualizing the propagation of light. *ACM Transactions on Graphics (ToG)*, 32(4):44, 2013.

[35] K. R. Vijayanagar, M. Loghman, and J. Kim. Refinement of depth maps generated by low-cost depth sensors. *IEEE International SoC Design Conference (ISOCC)*, pages 533–536, 2012.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.