Self-supervised End-to-end ToF Imaging Based on RGB-D Cross-modal Dependency

Weihang Wang[1] , **Jun Wang**[2] , **Fei Wen**[2,*]
[1]Soochow University
[2]Shanghai Jiao Tong University
whwang@suda.edu.cn, wangjeffrey1994@gmail.com, wenfei@sjtu.edu.cn

Abstract

Time-of-Flight (ToF) imaging systems are susceptible to various noise and degradation, which can severely affect image quality. Traditional sequential imaging pipelines often suffer from error accumulation due to separate multi-stage processing. Existing end-to-end methods typically rely on noisy-clean depth image pairs for supervised learning. However, acquiring ground-truth is challenging in real-world scenarios due to factors such as Multi-Path Interference (MPI), phase wrapping, and complex noise patterns. In this paper, we propose a self-supervised learning framework for end-to-end ToF imaging, which does not require any noisy-clean pairs yet generalizes well across various off-the-shelf cameras. Our framework leverages the cross-modal dependencies between RGB and depth data as implicit supervision to effectively suppress noise and maintain image fidelity. Additionally, the loss function integrates the statistical characteristics of raw measurement data, enhancing robustness against noise and artifacts. Extensive experiments on both synthetic and real-world data demonstrate that our approach achieves performance comparable to supervised methods, without requiring paired noisy-clean data for training. Furthermore, our method consistently delivers strong performance across all evaluated cameras, highlighting its generalization capabilities. The code is available at https://github.com/WeihangWANG/RGBD_imaging.

1 Introduction

Depth sensing techniques are playing an increasingly important role in many fields, such as 3D reconstruction, understanding, and interaction. Over the past decade, various depth sensing techniques have emerged, including stereo vision, structured light and Time-of-Flight (ToF). Due to low cost, high accuracy, and lightweight design, ToF cameras have attracted increasing attention in commercial applications. However, ToF cameras typically suffer from a variety of noise, including phase wrapping, multi-path interference, motion artifact, and shot noise. Therefore, the ToF imaging method for producing high-quality depth images is of significant value.

To improve the ToF imaging quality, some methods use a multi-stage pipeline and solve these problems in isolation, e.g., phase unwrapping [Järemo Lawin *et al.*, 2016; Wang *et al.*, 2021], multi-path removal [Marco *et al.*, 2017; Agresti and Zanuttigh, 2018], and motion artifact suppression [Chen *et al.*, 2020]. However, the cascaded imaging method suffers from error accumulation. Moreover, most of these methods rely on prior assumptions or pre-trained models, which usually deviate from real-world data. The domain gap between prior assumptions and real-world data results in limited performance.

To mitigate these issues, end-to-end learning frameworks have been proposed to recover the depth from raw correlation measurements directly in [Su *et al.*, 2018; Zheng *et al.*, 2021; Wang *et al.*, 2023a; Wang *et al.*, 2023b], which jointly realize multi-path interference removal, phase unwrapping, and de-noising. However, such frameworks also encounters the problem that the collection of real-world raw data paired with ground-truth depth images is difficult and even impractical. Supervision signals containing reconstruction error and misalignment also lead to poor imaging performance.

To address these challenges, this paper proposes a self-supervised end-to-end learning method which does not require any noisy-clean depth image pairs for training. We exploit the cross-modal dependency between the RGB and depth modalities as supervision information, which is inspired by the nature that RGB images also contain rich geometric information of the scene.

In summary, the contributions are as follows:

- We propose a self-supervised learning framework for end-to-end ToF imaging without the requirement of any noisy-clean depth image pairs. It exploits the cross-modal dependency between the RGB and depth modalities as implicit supervision to suppress noise and preserve fidelity.
- We design a hybrid loss function incorporating the statistical characteristics of raw measurement data, which enhances robustness against noise and artifacts.
- We conduct extensive real-world end-to-end ToF imag-

[*]Corresponding author

ing experiments on four different off-the-shelf ToF cameras. The results demonstrate that the proposed method achieves competitive performance compared to supervised methods. Furthermore, our method consistently delivers strong performance across all evaluated cameras, highlighting its generalization capabilities.

2 Related Work

2.1 Cascaded ToF Imaging Processing Methods

Phase Unwrapping. Phase wrapping is a fundamental problem in many applications, such as synthetic aperture radar (SAR) [Pritt, 1996; Chen and Zebker, 2002], as well as in ToF imaging. Classical phase unwrapping methods estimate the phase wrap numbers with only a single depth image. These methods follow man-made assumptions, including the relationship between amplitude and depth [McClure *et al.*, 2010; Cho *et al.*, 2012], and the phase jump between adjacent pixels [Frey *et al.*, 2001; Droeschel *et al.*, 2010b].

Multi-shot methods take two depth images at two different frequencies to solve depth ambiguity. The common one is Chinese remainder theorem, that is effective in the easy cases with the absence of other errors. To address the hard cases with the presence of MPI or noise, multiple image priors are proposed, such as amplitude constraints [Järemo Lawin *et al.*, 2016; Wang *et al.*, 2021], and frequency constraints [Droeschel *et al.*, 2010a].

Multi-path Interference Removal. Due to the complexity of multi-path effect, traditional methods usually take multiple measurements at different modulation frequencies. These methods recast the multi-path as a sparse estimation problem and solve from different perspectives, including optimization [Bhandari *et al.*, 2014; Freedman *et al.*, 2014], spectral estimation [Feigin *et al.*, 2015; Kirmani *et al.*, 2013], and compressive sensing (CS) [Xuan *et al.*, 2016]. Hardware-based methods introduce an extra projector [Naik *et al.*, 2015] or custom coding [Kadambi *et al.*, 2013].

Recently, learning based methods for MPI removal have been proposed [Son *et al.*, 2016; Marco *et al.*, 2017; Agresti and Zanuttigh, 2018; Qiu *et al.*, 2019; Gutierrez-Barragan *et al.*, 2021]. However, these learning based methods have an unsatisfactory performance in real-world data.

2.2 End-to-end ToF Imaging Processing Methods

The multi-stage pipeline often suffers from cumulative errors and information loss. To address these issues, the end-to-end framework is widely adopted, leveraging noisy-clean image pairs for ToF imaging processing in [Su *et al.*, 2018; Yan *et al.*, 2020; Zheng *et al.*, 2021; Gao *et al.*, 2021; Jung *et al.*, 2021; Li *et al.*, 2022; Meng *et al.*, 2024; Tang *et al.*, 2024]. Su *et al.* [Su *et al.*, 2018] firstly propose a GAN-based architecture for ToF imaging from dual-frequency, raw correlation measurements. Zheng *et al.* [Zheng *et al.*, 2021] design a multi-stage iterative CNN for I-ToF depth error removal. Jung *et al.* [Jung *et al.*, 2021] take an fusion of RGB and raw correlation data as input and focus on real-world scenarios with strong ambient light and far distances.

However, existing methods rely on either 3D reconstruction or other commercial depth cameras as ground truth, which is not sufficiently reliable. Zheng *et al.* [Zheng *et al.*, 2021] propose to use 3D reconstruction results as clean depth image, but this method suffers from reconstruction errors and misalignment. Su *et al.* [Su *et al.*, 2018] propose to generate noisy data by adding synthesized noise and degradation to clean data. However, real-world noise and degradation are complicated and difficult to simulate accurately. Therefore, the performance of these methods is limited due to the domain gap between real-world data and generated noisy-clean pairs.

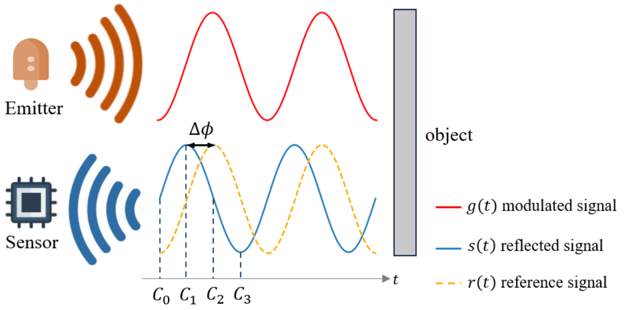


Figure 1: The illustration of ToF imaging principles. C_{0-3} are four sampling points.

2.3 Unsupervised Learning Methods for Image Enhancement

To address the challenges in collecting clean image for supervision, unsupervised and self-supervised learning methods without noisy-clean pairs have recently gained increasing popularity in image enhancement. For example, recent methods use noisy-noisy pairs [Lehtinen *et al.*, 2018], noisier-noisy pairs [Moran *et al.*, 2020; Krull *et al.*, 2019], noisy images only [Batson and Royer, 2019; Xu *et al.*, 2020; Wang *et al.*, 2023a] instead of noisy-clean pairs for image denoising. Most of these methods present a specific prior model on the noise signal, which may lead to unsatisfactory in real-world scenarios. Wang *et al.* [Wang *et al.*, 2023b] propose an optimal transport based method that shows remarkable superiority in raw depth image denoising without any prior models, but it cannot handle other error sources. Similarly, Agresti *et al.* [Agresti *et al.*, 2019] propose an unsupervised method for multipath interference removal based on pixel-level domain adaptation, while it still needs clean-noisy pairs for training.

3 Background

Typical ToF imaging systems consist of two main components, namely an active illumination module (emitter) and an image sensor (receiver), as shown in Figure 1. The distance d is derived from the phase shift $\Delta\phi$ between emitted signal $g(t)$ and reflected signal $s(t)$, which can be written as

$$d = \frac{\Delta\phi \cdot \lambda}{4\pi} = \frac{\Delta\phi \cdot c}{4\pi f_m}. \tag{1}$$

where $\Delta\phi$ is the phase shift, λ is the wavelength, f_m is the modulation frequency, and c is the speed of light. The problem is to solve for the phase shift.

The image sensor cannot measure the phase shift $\Delta\phi$ directly, but measures the integral of the raw correlation function instead as

$$h(\tau) = s(t) \otimes r(t), \tag{2}$$

with

$$g(t) = cos(f_m \cdot t), \tag{3}$$

$$s(t) = \alpha \cdot g\left(t - \Delta\phi\right) + \beta. \tag{4}$$

where $r(t)$ is the reference signal usually equal to the emitted signal $g(t)$, α is the amplitude attenuation coefficient, and β represents the intensity of the ambient light. Then, the raw correlation measurement in (2) can be written as

$$\begin{aligned} h(\tau) &= \frac{1}{T}\lim_{T\to\infty}\int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)r(t+\tau)dt, \\ &= \frac{\alpha}{2}cos(f_m\tau - \Delta\phi) + \beta. \end{aligned} \tag{5}$$

In practice, with four sampling points of $i\pi/2$ for $i \in \{0, 1, 2, 3\}$ in a period, we can obtain four raw correlation measurements h_0 to h_3 and calculate the phase shift $\Delta\phi$ with these four raw correlation measurements, further the depth d and amplitude value α as

$$\Delta\phi = \arctan\left(\frac{h_1 - h_3}{h_0 - h_2}\right), \tag{6}$$

$$d = \arctan\left(\frac{h_1 - h_3}{h_0 - h_2}\right) \cdot \frac{c}{4\pi f_m}, \tag{7}$$

$$\alpha = \sqrt{\left(\frac{h_1 - h_3}{2}\right)^2 + \left(\frac{h_0 - h_2}{2}\right)^2}. \tag{8}$$

The above method is most commonly used in ToF cameras. However, due to the imperfection of ToF measurements, some geometric information would be lost, while undesired noise may be introduced, resulting in inaccurate depth estimation, such as those caused by multi-path interference, phase wrapping, etc.

4 Methodology

Grounded in information theory, we derive a self-supervised formulation and implement it using WGAN-based adversarial learning, combined with MAE fidelity and smoothness losses, to guide training without the need for clean GT pairs. By maximizing the cross-modal mutual information, the model learns to generate high-quality depth from raw-data input.

4.1 Theoretical Analysis

Taking raw correlation measurements C as input, the ToF imaging model can be denoted as

$$\hat{X} := f(C), \tag{9}$$

where $\hat{X}$ is the estimated depth image of the model.

It is relatively easy to train an imaging model with the ground truth depth image X as supervision. However, the challenge lies in collecting noisy-clean pairs $(\hat{X}, X)$. To overcome this challenge, we use the mutual information between the RGB image R and the depth image X, i.e. $I(X; R)$ to supervise the learning of an imaging model. To achieve this, we consider the following formulation

$$\min_{f} \mathbb{D}\left(p_{\hat{X},R} \| p_{X,R}\right), \tag{10}$$

where $d(\cdot,\cdot)$ measures the distance between distributions, such as the KL divergence or Wasserstein distance. $p_{\hat{X},R}$ and $p_{X,R}$ are the joint distributions of $(\hat{X}, R)$ and (X, R), respectively.

According to information theory, mutual information $I(\hat{X}; R)$ is equivalent to the KL divergence $\mathbb{D}_{KL}(p_{\hat{X},R} \| p_{\hat{X}}p_R)$, represented as

$$\begin{aligned} I(\hat{X}; R) &= \int_{\hat{\mathcal{X}}\times\mathcal{R}} p_{\hat{X},R}\log\frac{p_{\hat{X},R}}{p_{\hat{X}}p_R}d\hat{X}dR, \\ &= \mathbb{D}_{KL}\left(p_{\hat{X},R} \| p_{\hat{X}}p_R\right). \end{aligned} \tag{11}$$

Generally, transition from an actual joint distribution to a joint distribution under the assumption of independence will either remain or increase the KL divergence. In other words, the KL divergence between $p_{\hat{X},R}$ and $p_{\hat{X}}p_R$ is no less than that between $p_{\hat{X},R}$ and $p_{X,R}$, which can be written as

$$\mathbb{D}_{KL}\left(p_{\hat{X},R} \| p_{X,R}\right) \leq \mathbb{D}_{KL}\left(p_{\hat{X},R} \| p_{\hat{X}}p_R\right) = I(\hat{X}; R). \tag{12}$$

The mutual information $I(\hat{X}; R)$ is upper bounded by $I(X; R)$ based on information theory, denoted as

$$I(\hat{X}; R) \leq I(X; R). \tag{13}$$

Therefore, minimizing the divergence between the two joint distributions $p_{\hat{X},R}$ and $p_{X,R}$ in (10) would maximize the mutual information $I(\hat{X}; R)$. The above analysis motivates us to design a self-supervised end-to-end framework leveraging RGB-D cross-modal dependency as supervision for training a ToF imaging model.

4.2 Network Architecture

The proposed formulation (10) is implemented based on a WGAN architecture, as shown in Figure 2. The generator G estimates the de-noised depth image while preserving the geometry information of raw measurements. The discriminator D forces the model to learn the common RGB-D cross-modal dependency between $\hat{X}$ and R, which can further suppress noise and improve fidelity.

The proposed generator consists of a multi-branch feature extraction module and a U-Net module. The feature extraction module includes a raw correlation branch, a depth branch, and a depth-amplitude branch. We use the Residual Channel Attention Block (RCAB) [Zhang *et al.*, 2018] as backbone, which can focus on more informative features. The features from the three branches are concatenated as input to the U-Net architecture for depth image generation.

The discriminator is designed to determine whether the generated depth image and the paired RGB image $(\hat{X}, R)$

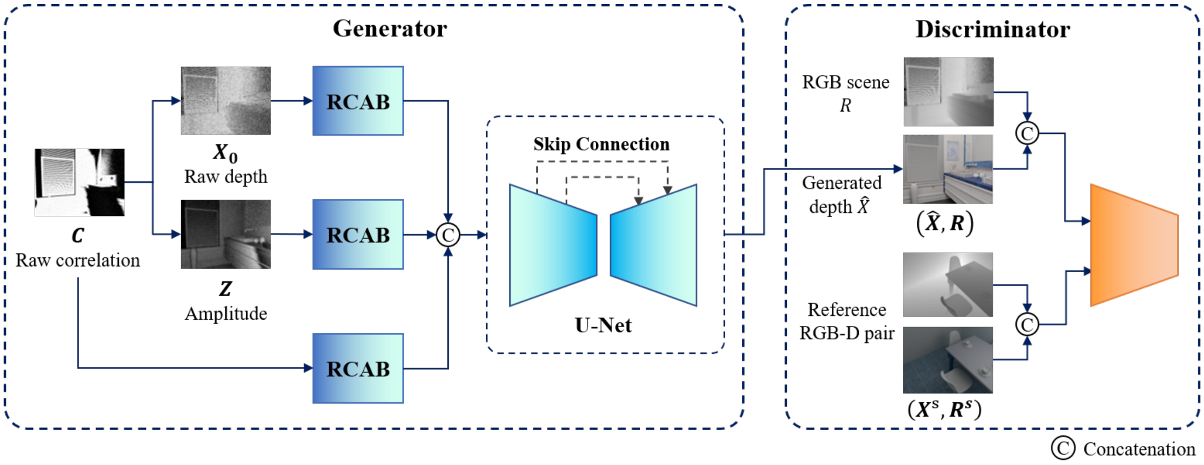


Figure 2: Overview of the proposed self-supervised end-to-end ToF imaging framework. The network architecture is based on WGAN. In particular, the discriminator leverages the RGB-D cross-modal dependency as implicit supervision, guiding the generator to suppress noise and maintain geometric fidelity. Additionally, the loss function integrates the statistical characteristics of raw measurement data, enhancing robustness against noise and artifacts.

maintain consistent cross-modal dependencies with the synthesized RGB-D image pair (X^s, R^s), which guides the model to generate a depth image with correct geometry and high fidelity relative to the paired RGB image. Note that the generated depth image $\hat{X}$ and the synthesized depth image X^s in the discriminator are not paired, which relaxes the requirement for paired noise-clean depth images in supervision.

4.3 Loss Functions

As shown in Figure 2, the proposed framework is implemented based on WGAN and we propose a hybrid loss to train the end-to-end ToF imaging model as

$$\mathcal{L}_G = E[-D(G(C), R)] + \lambda_1 \mathcal{L}_{dep} + \lambda_2 \mathcal{L}_{corr} + \lambda_3 \mathcal{L}_{smooth}, \tag{14}$$

$$\mathcal{L}_D = E[D(G(C), R)] - E[D(X^s, R^s)], \tag{15}$$

where λ_1, λ_2, and λ_3 are hyperparameters.

In (14), the first term represents cross-modal adversarial learning, which guides the model to suppress noise and improve fidelity based on the geometric information implied in the RGB image. Besides, the loss function takes into account the statistical characteristics of both depth data and raw correlation data. The last three terms correspond to three branches of the generator. The details are given as follows.

$\mathcal{L}_{dep}$ minimizes the mean absolute error between the output generated depth image $\hat{X}$ and the original depth image X_0 calculated from raw correlation measurements C as

$$\mathcal{L}_{dep} = \|G(C) - X_0\|_1. \tag{16}$$

$\mathcal{L}_{corr}$ minimizes the mean absolute error between the reconstructed correlation measurements $\hat{C}$ and input raw correlation measurements C as

$$\mathcal{L}_{corr} = \|\hat{C} - C\|_1, \tag{17}$$

where $\hat{C}$ is reconstructed based on $\hat{X}$ and amplitude Z at two modulation frequencies f_1 and f_2 in the following form as

$$\hat{C} = Z * \left[sin(\frac{\hat{X}}{c/4\pi f_1}), cos(\frac{\hat{X}}{c/4\pi f_1}), sin(\frac{\hat{X}}{c/4\pi f_2}), cos(\frac{\hat{X}}{c/4\pi f_2}) \right]. \tag{18}$$

$\mathcal{L}_{smooth}$ is used to ensure the local smoothness of generated depth image. Especially, we use a total variation loss [Su *et al.*, 2018] can be written as

$$\mathcal{L}_{smooth} = |\partial_x \hat{X}| e^{-|\partial_x Z|} + |\partial_y \hat{X}| e^{-|\partial_y Z|}. \tag{19}$$

5 Experimental Results

In this section, we evaluate the proposed method on both synthetic and real-world ToF raw data captured with four different off-the-shelf ToF cameras to evaluated the performance and generalization capabilities.

5.1 Experimental Settings

In our method, the discriminator is trained by paired RGB-D data $\{(\hat{x}_i, r_i)\}$ and $\{(x_i^s, r_i^s)\}$, while the generator takes raw correlation data c_i, directly calculated noisy depth data $\hat{x_i}$ and amplitude data z_i as input. The U-Net module of generator consists of two down-sampling CNN layers in the encoder and two up-sampling CNN layers in the decoder. The synthesized RGB-D pairs $\{(x_i^s, r_i^s)\}$ for discriminator are from the synthetic dataset [Zheng *et al.*, 2021], which is generated based on simulating the imaging system of the LUCID Helios camera. Table 1 shows the hyperparameter tuning procedure for the loss function. Using PSNR as the evaluation metric, we set $\lambda_1 = 10$ and $\lambda_3 = 20$. λ_2 is empirically set to 5. The initial learning rate is 0.1, and the optimizer is RMSprop. The maximum number of epochs is set to 200. The code is implemented in PyTorch and run on Nvidia 3090Ti. The comparison of computational complexity is shown in Table 2, which shows that our method has medium parameter amount and computation cost.

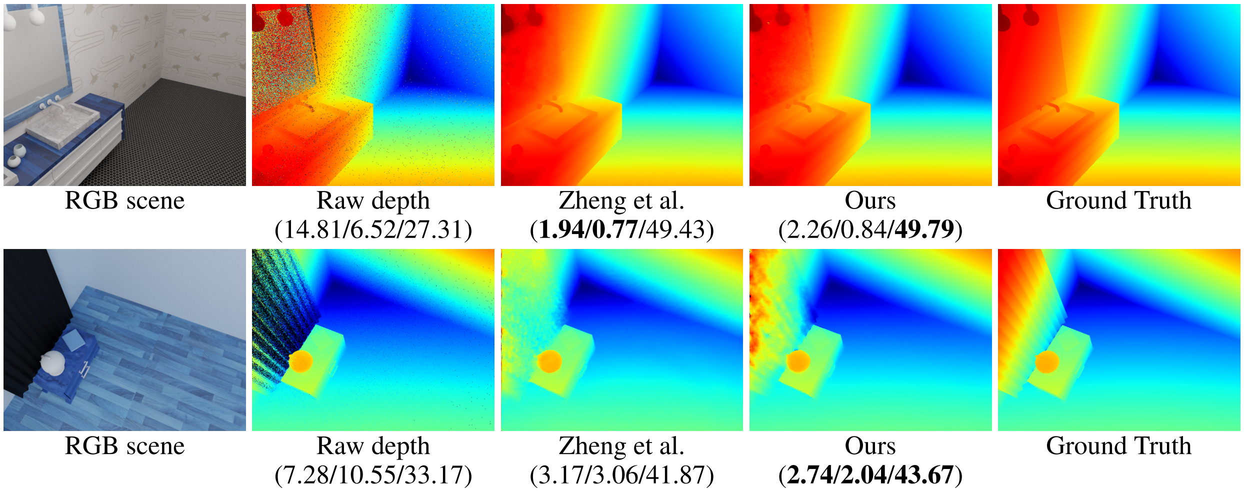


Figure 3: Visual quality and quantitative score comparison on the synthetic data. The quantitative scores are mean absolute error (MAE), relative error (Rela.), and PSNR respectively.

λ_1/λ_3	1	10	20	30
1	34.57	35.25	34.37	35.93
10	39.76	40.24	**42.87**	39.64
20	37.96	40.14	41.10	40.67
30	39.94	40.82	40.01	41.80

Table 1: Tuning procedure for hyperparameters in the loss function.

Method	Parameter Amount	FLOPs
[Su *et al.*, 2018]	16.6M	40.3G
[Zheng *et al.*, 2021]	2.1M	5.3G
Ours	4.8M	12.7G

Table 2: The comparison of computational complexity.

	MAE (cm)	Rela. (%)	PSNR (dB)
Raw measurement	9.73	3.47	30.26
[Su *et al.*, 2018]	4.23	1.78	-
[Zheng *et al.*, 2021]	**2.03**	**0.85**	**44.19**
Ours	2.56	1.12	42.53

Table 3: Quantitative comparison on the synthetic dataset.

5.2 Experiments on Synthetic Datasets

In this experiment, we evaluate the proposed method on a synthetic dataset [Zheng *et al.*, 2021] under the scenario with combined corruptions. Table 3 shows the quantitative results of the compared methods in terms of mean absolute error (MAE), relative error (Rela.), and PSNR. Without using noisy-clean pairs as supervision, our self-supervised method achieves competitive performance compared to the supervised method [Zheng *et al.*, 2021] which is trained using noisy-clean pairs.

ToF camera	Modulation Frequency (MHz)	Resolution	Sensor
Lucid	75 & 100	640×480	Sony
TI	40 & 70	320×240	TI
TCS	80 & 100	640×480	Sony
TCE	12 & 24	320×240	EPC660

Table 4: Parameters of the evaluated ToF cameras.

Moreover, Figure 3 compares the visual quality along with quantitative scores on samples from the synthetic dataset. The depth images are shown in pseudo-color, with a red-to-blue color scheme representing distances from near to far. The first one is a bathroom case with high reflectivity mirror and a corner. In the second case, the main degradation is low-intensity noise caused by a black curtain. It can be observed that generated depth images of ours can suppress noise while preserving geometry information, which achieve competitive performance with supervised method [Zheng *et al.*, 2021].

5.3 Experiments on Real-world Datasets

In this experiment, we evaluate the proposed method on real-world data captured by four off-the-shelf ToF depth cameras, respectively LUCID, TI, TCS, and TCE. The basic parameters of four cameras are listed in Table 4. The first two datasets are captured by [Zheng *et al.*, 2021] and [Su *et al.*, 2018]. We test on the data of two additional cameras to further prove the generalization of our proposed method.

Comparison on real-world data captured with TI. Figure 4 illustrates two cases of the daily scenes from the real-world dataset [Su *et al.*, 2018]. The raw measurements are captured by a TI OPT8241-CDK-EVM camera, including noise, invalid pixels, and complex reflectance. It can be observed that the supervised method [Zheng *et al.*, 2021] does not generalize well to the real-world data captured by a different depth sensor. Compared with [Su *et al.*, 2018], our proposed method can preserve detail information while ef-

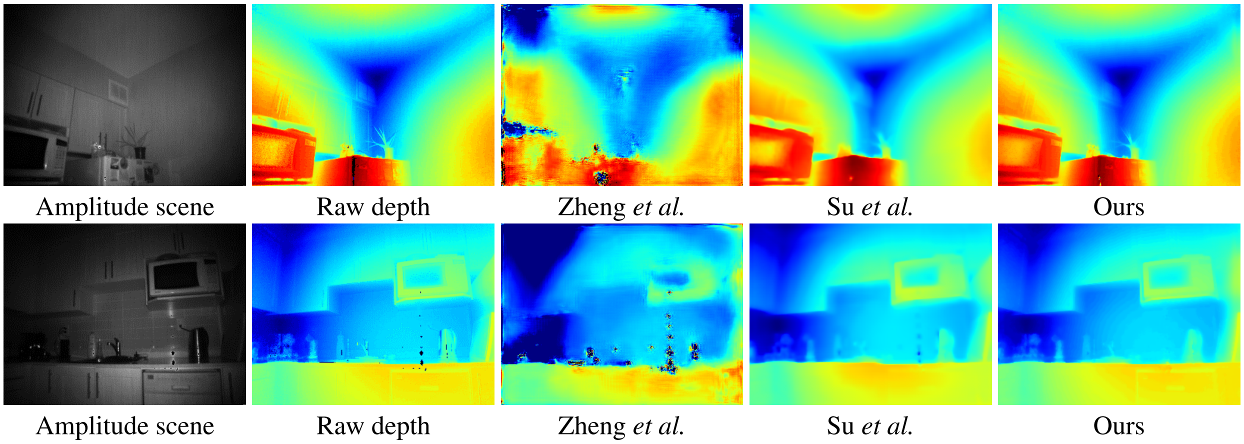


Figure 4: Visual quality comparison on the imaging outputs from the raw data captured by the TI camera.

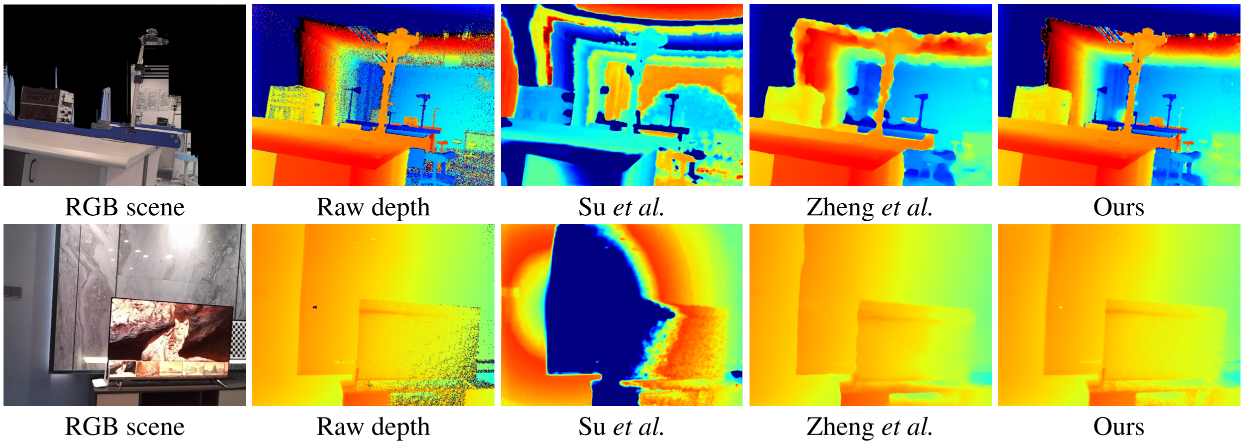


Figure 5: Visual quality comparison on the imaging outputs from the raw data captured by the LUCID camera.

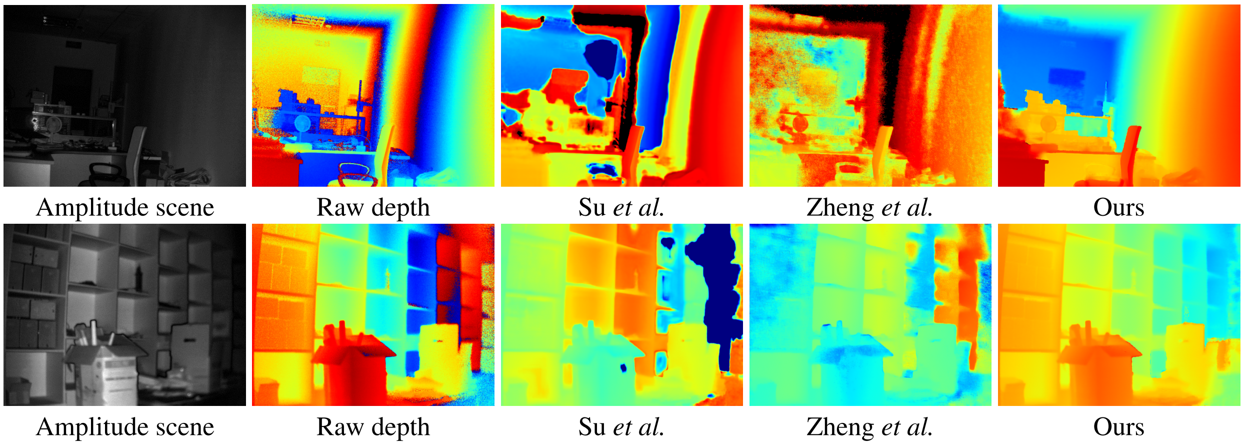
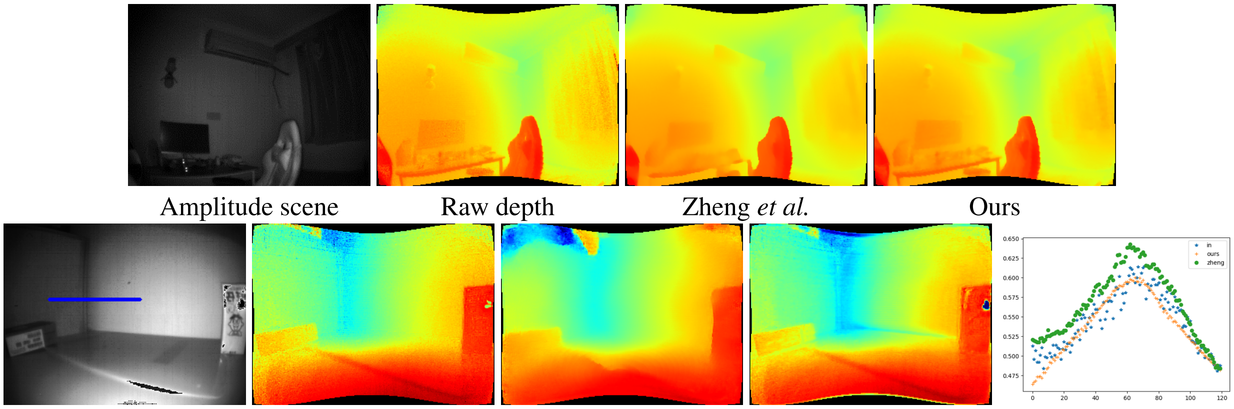


Figure 6: Visual quality comparison on the imaging outputs from the raw data captured by the TCS camera.


Amplitude scene | Raw depth | Zheng *et al.* | Ours

Amplitude scene | Raw depth | Zheng *et al.* | Ours | Scanline

Figure 7: Visual quality comparison on the imaging outputs from the raw data captured by the TCE camera.

fectively removing noise. For example, in the first sample, the potted leaves are blended into the background in [Su *et al.*, 2018], but our method retains the details.

Comparison on real-world data captured with LUCID. Figure 5 compares the visual quality on two samples from the real-world dataset captured by Lucid. It can be observed that the method [Su *et al.*, 2018] does not perform well in this dataset, due to the difference of the noise model. The supervised method [Zheng *et al.*, 2021] recovers smooth depth image, but it is trained using data captured by the LUCID camera. However, the method [Zheng *et al.*, 2021] changes the shape of the object, such as the distorted edges in the second scene. In comparison, our proposed method can recover a cleaner depth image with more reasonable 3D geometry.

Comparison on real-world data captured with TCS. Figure 6 illustrates two cases of the daily scenes in the office, captured with TCS. The first scene is a far-distance scene with a maximum range of 7 meters, while the second one is a close-range scene with complex reflectance and shape. The raw ToF measurements are captured at two high frequencies, respectively $80M$ and $100M$. Therefore, the depth images suffer from phase unwrapping and the pseudo-color visualization of the raw depth image exhibits periodic repetition. As shown in Figure 6, the complex noise leads to failure of [Su *et al.*, 2018] and [Zheng *et al.*, 2021]. Compared with these two methods, the proposed method can restore the depth image with explicit edges and less noise.

Comparison on real-world data captured with TCE. Figure 7 illustrates two cases of the daily scenes in the room. The first case is a full-view of room. The second one is a corner case. Both of the methods can remove random noise, but the restored depth image by [Zheng *et al.*, 2021] is over-smoothed, and errors occur at the boundary of the image. For an intuitive comparison of multi-path removal, we randomly select a scanline (blue line in Amplitude). As shown in Figure 7, the scanline of the raw depth and restored depth by [Zheng *et al.*, 2021] is far from the ideal structure of the corner, as the scanline is curved and messy. Note that Su *et al.* requires dual-frequency raw ToF correlation measurements as input to recover depth. However, the TCE camera provides only single-frequency measurements. Therefore, the results of Su *et al.* are not included in Fig. 7.

$\mathcal{L}_{dep}$	$\mathcal{L}_{corr}$	$\mathcal{L}_{smooth}$	MAE (cm)	Rela. (%)	PSNR (dB)
✓	-	-	4.8	1.60	39.52
✓	✓	-	2.92	1.35	41.75
✓	-	✓	3.13	1.46	40.65
✓	✓	✓	**2.62**	**1.14**	**42.41**

Table 5: Ablation study on the synthetic dataset.

5.4 Ablation Experiments

In this section, we provide ablation study to evaluate the components of the proposed loss functions. The ablation experiments are conduct on a small dataset which contains only a quarter of the full synthetic dataset [Zheng *et al.*, 2021].

As shown in Table 5, the model performs best when all the components are used as the loss function, which demonstrate that the hybrid loss function incorporating statistic characteristics of raw measurements can improve the performance of the imaging model.

6 Conclusion

In this paper, we propose a self-supervised learning framework for end-to-end ToF imaging, which does not require any noisy-clean pairs yet generalizes well across various off-the-shelf cameras. The proposed self-supervised framework utilizes the cross-modal dependency between RGB and depth data as implicit supervision to suppress noise and preserve fidelity. The loss function incorporates the statistical characteristics of raw measurement data, improving robustness against noise and artifacts. Experimental results on both synthetic and real-world data demonstrate that our proposed method can suppress noise and preserve fidelity, which achieves competitive performance with supervised methods. Furthermore, our method consistently delivers strong performance across all evaluated cameras, highlighting its generalization capabilities across various scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62271314.

References

[Agresti and Zanuttigh, 2018] Gianluca Agresti and Pietro Zanuttigh. Deep learning for multi-path error removal in tof sensors. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Agresti *et al.*, 2019] Gianluca Agresti, Henrik Schaefer, Piergiorgio Sartor, and Pietro Zanuttigh. Unsupervised domain adaptation for tof data denoising with adversarial learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5586, 2019.

[Batson and Royer, 2019] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019.

[Bhandari *et al.*, 2014] Ayush Bhandari, Achuta Kadambi, Refael Whyte, Christopher Barsi, Micha Feigin, Adrian Dorrington, and Ramesh Raskar. Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *Optics Letters*, 39(6):1705–1708, 2014.

[Chen and Zebker, 2002] Curtis W Chen and Howard A Zebker. Phase unwrapping for large sar interferograms: Statistical segmentation and generalized network models. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8):1709–1719, 2002.

[Chen *et al.*, 2020] Zhuo Chen, Peilin Liu, Fei Wen, Jun Wang, and Rendong Ying. Restoration of motion blur in time-of-flight depth image using data alignment. In *2020 International Conference on 3D Vision (3DV)*, pages 820–828, 2020.

[Cho *et al.*, 2012] Shung Han Cho, Kwanghyuk Bae, Kyu-Min Kyung, and Tae-Chan Kim. Fast and efficient method to suppress depth ambiguity for time-of-flight sensors. In *IEEE Global Conference on Consumer Electronics 2012*, pages 432–434. IEEE, 2012.

[Droeschel *et al.*, 2010a] David Droeschel, Dirk Holz, and Sven Behnke. Multi-frequency phase unwrapping for time-of-flight cameras. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1463–1469. IEEE, 2010.

[Droeschel *et al.*, 2010b] David Droeschel, Dirk Holz, and Sven Behnke. Probabilistic phase unwrapping for time-of-flight cameras. In *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, pages 1–7. VDE, 2010.

[Feigin *et al.*, 2015] Micha Feigin, Ayush Bhandari, Shahram Izadi, Christoph Rhemann, Mirko Schmidt, and Ramesh Raskar. Resolving multipath interference in kinect: An inverse problem approach. *IEEE Sensors Journal*, 16(10):3419–3427, 2015.

[Freedman *et al.*, 2014] Daniel Freedman, Yoni Smolin, Eyal Krupka, Ido Leichter, and Mirko Schmidt. Sra: Fast removal of general multipath for tof sensors. In *European Conference on Computer Vision*, pages 234–249. Springer, 2014.

[Frey *et al.*, 2001] Brendan J Frey, Ralf Koetter, and Nemanja Petrovic. Very loopy belief propagation for unwrapping phase images. *Advances in Neural Information Processing Systems*, 14, 2001.

[Gao *et al.*, 2021] Rongrong Gao, Na Fan, Changlin Li, Wentao Liu, and Qifeng Chen. Joint depth and normal estimation from real-world time-of-flight raw data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 71–78. IEEE, 2021.

[Gutierrez-Barragan *et al.*, 2021] Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu. itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging*, 7:1205–1214, 2021.

[Järemo Lawin *et al.*, 2016] Felix Järemo Lawin, Per-Erik Forssén, and Hannes Ovrén. Efficient multi-frequency phase unwrapping using kernel density estimation. In *European Conference on Computer Vision*, pages 170–185. Springer, 2016.

[Jung *et al.*, 2021] HyunJun Jung, Nikolas Brasch, Aleš Leonardis, Nassir Navab, and Benjamin Busam. Wild tofu: Improving range and quality of indirect time-of-flight depth with rgb fusion in challenging environments. In *2021 International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2021.

[Kadambi *et al.*, 2013] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):1–10, 2013.

[Kirmani *et al.*, 2013] Ahmed Kirmani, Arrigo Benedetti, and Philip A Chou. Spumic: Simultaneous phase unwrapping and multipath interference cancellation in time-of-flight cameras using spectral methods. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

[Krull *et al.*, 2019] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2124–2132, 2019.

[Lehtinen *et al.*, 2018] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pages 2965–2974, 2018.

[Li *et al.*, 2022] Jiaqu Li, Tao Yue, Sijie Zhao, and Xuemei Hu. Fisher information guidance for learned time-of-flight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16334–16343, June 2022.

[Marco *et al.*, 2017] Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6):1–12, 2017.

[McClure *et al.*, 2010] Shane H McClure, Michael J Cree, Adrian A Dorrington, and Andrew D Payne. Resolving depth-measurement ambiguity with commercially available range imaging cameras. In *Image Processing: Machine Vision Applications III*, volume 7538, pages 159–170. SPIE, 2010.

[Meng *et al.*, 2024] Yu Meng, Zhou Xue, Xu Chang, Xuemei Hu, and Tao Yue. itof-flow-based high frame rate depth imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4938, June 2024.

[Moran *et al.*, 2020] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020.

[Naik *et al.*, 2015] Nikhil Naik, Achuta Kadambi, Christoph Rhemann, Shahram Izadi, Ramesh Raskar, and Sing Bing Kang. 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*, pages 73–81, 2015.

[Pritt, 1996] Mark D Pritt. Phase unwrapping by means of multigrid techniques for interferometric sar. *IEEE Transactions on Geoscience and Remote Sensing*, 34(3):728–738, 1996.

[Qiu *et al.*, 2019] Di Qiu, Jiahao Pang, Wenxiu Sun, and Chengxi Yang. Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9994–10003, 2019.

[Son *et al.*, 2016] Kilho Son, Ming-Yu Liu, and Yuichi Taguchi. Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3390–3397. IEEE, 2016.

[Su *et al.*, 2018] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang 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.

[Tang *et al.*, 2024] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9763–9772, June 2024.

[Wang *et al.*, 2021] Jun Wang, Peilin Liu, Fei Wen, Rendong Ying, and Weihang Wang. Phase unwrapping for time-of-flight sensor based on image segmentation. *IEEE Sensors Journal*, 21(19):21600–21611, 2021.

[Wang *et al.*, 2023a] Jun Wang, Peilin Liu, and Fei Wen. Self-supervised learning for rgb-guided depth enhancement by exploiting the dependency between rgb and depth. *IEEE Transactions on Image Processing*, 32:159–174, 2023.

[Wang *et al.*, 2023b] Wei Wang, Fei Wen, Zeyu Yan, and Peilin Liu. Optimal transport for unsupervised denoising learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2104–2118, 2023.

[Xu *et al.*, 2020] Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, 29:9316–9329, 2020.

[Xuan *et al.*, 2016] Vinh Nguyen Xuan, Klaus Hartmann, Wolfgang Weihs, and Otmar Loffeld. Multi-target super-resolution using compressive sensing arguments for multipath interference recovery. In *2016 4th International Workshop on Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa)*, pages 148–152. IEEE, 2016.

[Yan *et al.*, 2020] Chen Yan, Ren Jimmy, Cheng Xuanye, Qian Keyuan, Wang Luyang, and Jinwei Gu. Very power efficient neural time-of-flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2257–2266, 2020.

[Zhang *et al.*, 2018] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[Zheng *et al.*, 2021] Zhuolin Zheng, Yinzhang Ding, Xiaotian Tang, Yu Cai, Dongxiao Li, Ming Zhang, Hongyang Xie, and Xuanfu Li. Iterative error removal for time-of-flight depth imaging. In *International Conference on Artificial Neural Networks*, pages 92–105. Springer, 2021.