DOI: 10.1111/cgf.14958
Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Multi-scale Iterative Model-guided Unfolding Network for NLOS Reconstruction

X. Su[1,4], Y. Hong[2], J. Ye[2], F. Xu[2,3] and X. Yuan[4]

[1]Zhe Jiang University, College of Computer Science and Technology, China
[2] University of Science and Technology of China, CAS Center for Excellence in Quantum Information and Quantum Physics, China
3 University of Science and Technology of China, Hefei National Laboratory for Physical Sciences at Microscale and Department of Modern Physics, China
4 Westlake University, School of Engineering, China

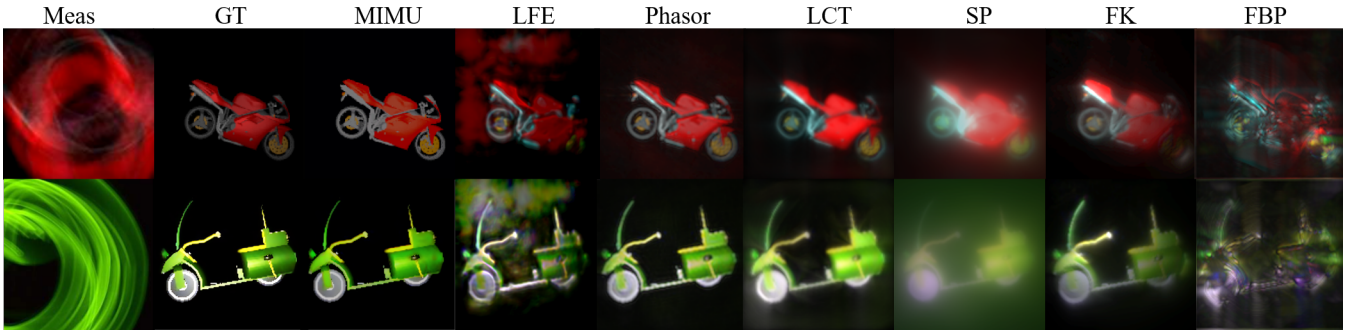

Figure 1: *Qualitative visualization of two color samples in* $256 \times 256 \times 512$ *size reconstructed by the proposed MIMU and previous strong baseline methods. "Meas","GT" stands for measurement and ground truth respectively. "MIMU" stands for our proposed method. LFE [CWK*20], Phasor [LGLM*19], LCT [OLW18], SP [WLH*21], FK [LWO19] and FBP [AGJ17] are baseline methods.*

Abstract

Non-line-of-sight (NLOS) imaging can reconstruct hidden objects by analyzing diffuse reflection of relay surfaces, and is potentially used in autonomous driving, medical imaging and national defense. Despite the challenges of low signal-to-noise ratio (SNR) and ill-conditioned problem, NLOS imaging has developed rapidly in recent years. While deep neural networks have achieved impressive success in NLOS imaging, most of them lack flexibility when dealing with multiple spatial-temporal resolution and multi-scene images in practical applications. To bridge the gap between learning methods and physical priors, we present a novel end-to-end Multi-scale Iterative Model-guided Unfolding (MIMU), with superior performance and strong flexibility. Furthermore, we overcome the lack of real training data with a general architecture that can be trained in simulation. Unlike existing encoder-decoder architectures and generative adversarial networks, the proposed method allows for only one trained model adaptive for various dimensions, such as various sampling time resolution, various spatial resolution and multiple channels for colorful scenes. Simulation and real-data experiments verify that the proposed method achieves better reconstruction results both in quality and quantity than existing methods.

CCS Concepts

• ***Computing methodologies*** → ***Computational photography;***

1. Introduction

Because light travels in straight lines, it is challenging to capture images of objects that are hidden around corners. To break this restriction, non-line-of-sight (NLOS) imaging analyzes the diffuse reflection from a relay wall to image hidden objects [FVW20, MSS*19], which has broad applications in many fields, such as medical imaging, autonomous driving, and robotic vision [LWK19, SMBG19, SOG18, XST*18]. With the rapid de-

velopment of photon-sensitive sensors and imaging algorithms, most current NLOS imaging techniques utilize inverse physical models [LZH*22, LWL*21] that are constructed with active or passive lighting and reconstruction algorithms to recover hidden scenes [BZT*15, GWV*12, LGLM*19, VWG*12, WZH*21, WLH*21]. Most recently, deep learning algorithms for NLOS imaging have received a lot of attention [GCHWI20, CWK*20].

In most cases, active NLOS imaging surpasses passive one, since the active methods can collect different kinds of information, including intensity, time, and coherence. In order to perform a high-resolution 3D reconstruction with active methods, a sensitive time-resolved detector is used to detect the light reflected on the relay surface, hidden objects, and relay surface in sequence. Then, the collected third-bounce light is analyzed by different optimized algorithms, e.g., backprojection [VWG*12], inverse methods [HXHH14, AGJ17], and SPIRAL-3D [WLH*21] to reconstruct the hidden scene. Learning based approaches [CWK*20, GCHWI20] integrated with physical models have demonstrated successful results by modeling the underlying physics of NLOS imaging, while leveraging learned scene priors useful for reconstructing familiar shapes and visual details. Unfortunately, these existing approaches produce unsatisfactory results in a fixed time-spatial resolution setup. In other problems, camera sensors are usually standard industrial products, the captured measurements thus are more easily reproduced than NLOS without concerning the setting for physical parameters such as wall size, time resolution and number of sampling points.

On one hand, due to the high-order loss with distance and environmental noise during the light transmitting, NLOS imaging is an ill-posed problem with low SNR [MSS*19], making high-quality reconstruction extremely difficult. Different hidden scenes may produce the same measurement, which deepens the ill-condition of the problem. On the other hand, the spatial resolution of the reconstruction result is limited by the size of the scanning area (wall size) and the system's temporal resolution. When the size of the hidden scene is large or complex, the spatial resolution will be limited by the computational complexity. To conquer these limitations, we introduce an unfolding architecture consisting of an inverse propagation module and a voxel mapping module into a deep neural network for multi-scale NLOS reconstruction. Specifically, our inverse propagation module adapts the iteration process of SPIRAL-3D [WLH*21] operator previously used for NLOS reconstruction with large Poisson noise, and our volume renderer takes the temporal resolution as an input and uses a condition module to transmit information about the input features to each stage, thus enabling our model to be trained under multiple time-spatial resolutions. Furthermore, to enable our model to handle various scenes more robustly, we adopt a more challenging training dataset with blur and expect it to serve as a universal dataset in future NLOS research. Through extensive experiments, we demonstrate that this design yields superior reconstruction quality for both synthetic data and real captures.

Specific contributions of this paper are as follows:

- First, we propose a flexible MIMU network for confocal NLOS (C-NLOS) reconstruction with physical model guided iterative optimization.
- Second, we develop a multi-scale strategy to promote network adaptability, which enables the proposed method to reconstruct C-NLOS with different spatial-temporal resolutions through a single trained model.
- Finally, extensive experiments demonstrate MIMU's robustness when handling various scenarios with state-of-the-art performance (including challenging large light spot synthetic and mainstream real-world data) and attractive running time.

2. Related Work

The objective of NLOS imaging is to retrieve hidden information about a target scene that cannot be directly observed by a camera. Abramson [Abr78] was the first to show a holographic capture system for transient imaging, and Kirmani [KHDR09] temporally resolved light transport measurements capturing short pulses of light before the global transport reaches steady state, which was initially proposed. Reconstruction algorithm can be classified into four main categories [MMP*22]: backprojection methods [VWG*12], wave propagation-based methods [HXHH14], iterative optimization methods [WLH*21], and geometry-based methods [XNK*19]. SPIRAL-3D [WLH*21] obtains an approximate solution of the inverse problem with LASSO-type optimization. The initial step of this method involves modeling the imaging forward model as a Poisson process and then discretizing the transient function. However, this process is time-consuming, requiring hundreds of iterations. Our objective is to enhance and expedite this iterative process by optimizing the update of a parametric representation.

It can be seen that the conventional methods based on physical imaging models, are still the mainstream of research in recent years [GCHWI20, HOZ*19, IH20, ICY*20, LYP*22, PDV19, TSG19, XNK*19, YHLX21, YLG*20]. In traditional approaches, NLOS imaging is constrained by three factors: i) the capabilities of the imaging setup and illumination (e.g., whether it is active or passive), ii) the accuracy of the forward imaging models, and iii) the effectiveness of the reconstruction algorithms. As an illustration, in [WLH*21], confocal settings and a dual-telescope setup were utilized to achieve an impressive NLOS imaging range of 1.43 km. Similarly, Liu et al. [LGLM*19] employed the phasor field to transform the NLOS imaging model into a LOS imaging model, enabling high-quality reconstruction of complex scenes. Another approach by O'Toole [OLW18] involved assuming a high-order transmission as a convolution with variable parameters, effectively converting the NLOS reconstruction problem into a 3D deconvolution problem.

However, recent advancements have capitalized on diverse learning-based approaches to overcome the theoretical and hardware limitations of NLOS reconstruction by leveraging statistical scene priors learned from extensive datasets. In terms of network design principles, deep learning methods applied in NLOS imaging can be categorized into two groups: end-to-end networks [GCHWI20] and physics-based networks [CWK*20]. LFE [CWK*20] presents a method that learns customized feature embeddings for NLOS reconstruction, as well as specific imaging, classification, and object detection tasks. Deep learning algorithms, in comparison to traditional algorithms [NBB*21, PZD*21], have

the capacity to comprehensively learn scene priors, automatically extract features, and successfully reconstruct hidden objects. Our proposed MIMU draws inspiration from this line of research [MJY20, MWZ22, WZM21, ZG18] and, for the first time, develops a **robust multi-scale iterative model-guided unfolding (MIMU) network for C-NLOS reconstruction**.

3. Physical Forward Model

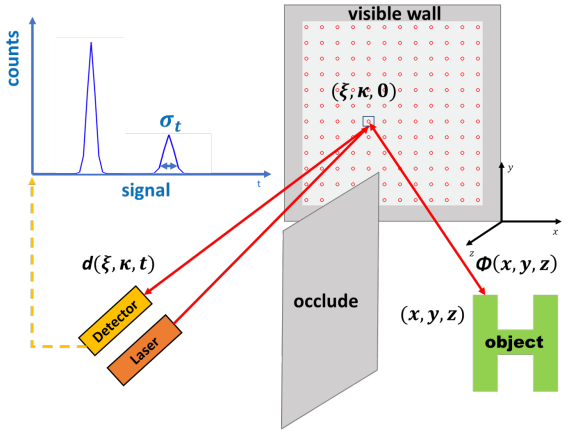

Figure 2: *A schematic diagram of the C-NLOS imaging system.*

The C-NLOS imaging system typically consists of a scanning pulsed laser and a single photon time-resolved detector, which focuses on the same points on a diffuse reflective wall. As shown in Figure 2, the directly illuminated points $(\xi, \kappa, 0)$ on the visible wall are considered as the sampling points. Then, the first diffuse reflection ray propagates to the points $(x, y, z) \in \Omega$ on the hidden object. $\boldsymbol{\phi}(x, y, z)$ is the second reflected wave from the object, and after the third reflections, of which reflective position $(\xi, \kappa, 0)$ is the same with the first one, a time resolved diffusive intensity $d(\xi, \kappa, t)$ is received by the detector. Here t represents the time of photon flight between the first reflection and the third reflection. Finally, a three dimensional (3D) light transient $\boldsymbol{d}(\xi, \kappa, t)$ is measured by an $m \times m$ array sampling. As derived in [OLW18], the 3D continuous signal is formulated as

$$\begin{aligned} &\boldsymbol{d}(\xi, \kappa, t) = \\ &\iiint_{\Omega} \frac{1}{\boldsymbol{r}^4(x-\xi, y-\kappa, z)} \boldsymbol{\phi}(x, y, z) \delta(2\boldsymbol{r} - ct) dx dy dz, \\ &\boldsymbol{r}(x-\xi, y-\kappa, z) = \sqrt{(x-\xi)^2 + (y-\kappa)^2 + z^2}, \end{aligned} \tag{1}$$

where the Dirac delta function δ models the light propagation, c is the speed of light. Note that $\boldsymbol{r}$ is the distance between the sampling points and the corresponding points on the surface of the hidden object. Combining all the detected photon arrival events into a single histogram results in a discrete inhomogeneous Poisson-distributed random variable as

$$\boldsymbol{y} \sim \text{Poisson}(\mathbf{A}\bar{\boldsymbol{\phi}} + \boldsymbol{b}), \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^{n_x n_y n_t}$ represents the discretized measurement by scanning point (n_x, n_y) with respect to the discretized time bin n_t. $\mathbf{A} \in \mathbb{R}^{n_x n_y n_t \times n_x n_y n_z}$ is the discretized version of the volumetric Albedo model in (1). $\bar{\boldsymbol{\phi}} \in \mathbb{R}^{n_x n_y n_z}$ represents the discretized Albedo of the hidden object. $\boldsymbol{b} \in \mathbb{R}^{n_x n_y n_t}$ denotes the dark count of the detector and background noise [BVT*16]. Since solving $\bar{\boldsymbol{\phi}}$ is an ill-posed problem, we reconstruct the hidden object by considering it as a regularized convex optimization problem.

4. Proposed MIMU Network

In this section, we begin by providing a comprehensive overview of the conventional iterative reconstruction algorithm SPIRAR-3D [WLH*21]. Subsequently, we shift our attention to the MIMU network, which serves as the fundamental component of the proposed framework. Lastly, we delve into specific implementation details.

4.1. SPIRAR-3D for C-NLOS

SPIRAR-3D [WLH*21] obtains an approximate solution of the inverse problem with LASSO optimization [Tib96] with various regularization. The proposed method first regards the imaging forward model as a Poisson process in (2) and subsequently obtains the probability function:

$$p\left(\boldsymbol{y} \mid \mathbf{A}\bar{\boldsymbol{\phi}} + \boldsymbol{b}\right) = \prod_{i=1} \frac{\left(\boldsymbol{e}_i^T \mathbf{A}\bar{\boldsymbol{\phi}} + \boldsymbol{\beta}\right)^{y_i}}{y_i!} exp(-\boldsymbol{e}_i^T \mathbf{A}\bar{\boldsymbol{\phi}} - \boldsymbol{\beta}), \tag{3}$$

where the background noise $\boldsymbol{b}$ is considered in the Poisson process with $\boldsymbol{\beta}$ [BVT*16]. $\boldsymbol{e}_i$ is the ith canonical basis unit vector. Denoting $\boldsymbol{f}$ as the estimation of $\bar{\boldsymbol{\phi}}$, the negative Poisson log-likelihood of (3) is given by

$$F(\boldsymbol{f}) = \mathbf{1}^T \mathbf{A}\boldsymbol{f} - \sum_{i=1}^{m} y_i \log\left(\boldsymbol{e}_i^T \mathbf{A}\boldsymbol{f} + \boldsymbol{\beta}\right), \tag{4}$$

$$\nabla F(\boldsymbol{f}) = \mathbf{A}^T \mathbf{1} - \sum_{i=1}^{m} \frac{y_i}{\boldsymbol{e}_i^T \mathbf{A}\boldsymbol{f} + \boldsymbol{\beta}} \mathbf{A}^T \boldsymbol{e}_i, \tag{5}$$

$$\nabla^2 F(\boldsymbol{f}) = \mathbf{A}^T \left[\sum_{i=1}^{m} \frac{y_i}{\left(\boldsymbol{e}_i^T \mathbf{A}\boldsymbol{f} + \boldsymbol{\beta}\right)^2} \boldsymbol{e}_i \boldsymbol{e}_i^T\right] \mathbf{A}, \tag{6}$$

where $\mathbf{1}$ is a vector of ones and $log(y_i!)$ is neglected. In a sequential quadratic approximation [HMW12], for iteration k, we compute a separable quadratic approximation to (4) using its second-order Taylor series approximation at $\boldsymbol{f}^k$. $\nabla F(\boldsymbol{f})$ is the gradient matrix and the Hessian Matrix $\nabla^2 F(\boldsymbol{f})$ will be approximated by an identity matrix $\alpha_k I$, with step length $\alpha_k > 0$ chosen by Barzilai-Borwein method [HMW12], similar to the second-order Taylor series approximation of (4) in [WNF09] and then yields

$$F^k(\boldsymbol{f}) = F(\boldsymbol{f}^k) + (\boldsymbol{f} - \boldsymbol{f}^k)^T \nabla F(\boldsymbol{f}^k) + \frac{\alpha_k}{2} \|\boldsymbol{f} - \boldsymbol{f}^k\|_2^2. \tag{7}$$

Then we can translate it to an optimized subproblem:

$$\boldsymbol{s}^k = \boldsymbol{f}^k - \frac{1}{\alpha_k} \nabla F(\boldsymbol{f}^k), \tag{8}$$

$$\boldsymbol{f}^{k+1} = \underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{argmin}} \frac{1}{2} \left\| \boldsymbol{f} - \boldsymbol{s}^k \right\|_2^2 + \frac{\tau}{\alpha_k} \text{pen}(\boldsymbol{f}) \; s.t. \boldsymbol{f} \geq 0, \tag{9}$$

where $\boldsymbol{s}^k$ is the gradient descent. The regularization penalty is considered as the penalty term $\text{pen}(\boldsymbol{f})$ with coefficient τ.

4.2. Unfolding the Iterative Algorithm

Due to the above optimized iteration progress, the method requires an interpretable forward and inverse model. Nevertheless, the real

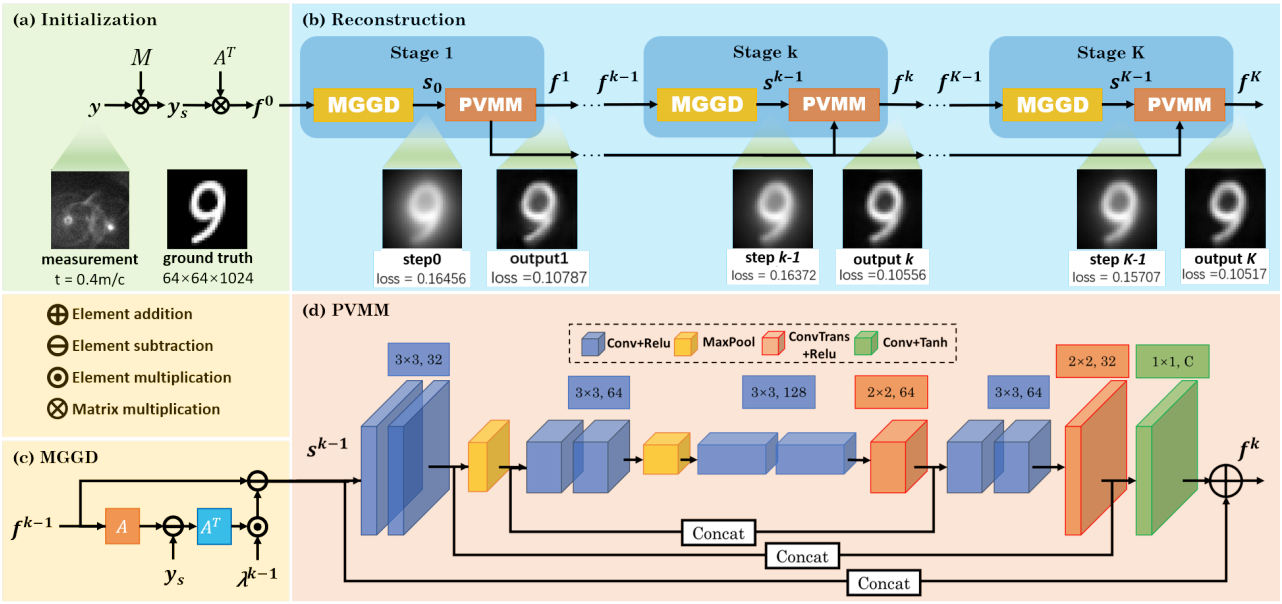

Figure 3: *Illustration of the proposed MIMU framework, which includes two parts:* ***(a)*** *Initialization and* ***(b)*** *Reconstruction.* ***(c)*** *MGGD module in each stage.* ***(d)*** *Each PVMM module composes of downsampling blocks and unsampling blocks.*

transient forward process is hardly modeled due to the uncertainty with the temporal and spatial variant distribution [HDY*21, SCZ*20]. Inspired by SPIRAR [HMW12] with a fixed 3D total variation denoiser, unfolding method [MJY20, MWZ22, WZM21, ZG18] trained jointly in all stages by optimizing the each denoiser at the same time. Specifically, we construct the unfolding network with n concatenated stages with respect to n iterations for optimizing $\boldsymbol{f}^k$, shown in Figure 3. The proposed network includes the following modules:

- Model-Guided Gradient Descent (MGGD) works as the gradient descent operation in (11).
- Pyramid Voxel Mapping Module (PVMM) is a trainable operator which works to improve the k-th 3D denoiser module $\mathcal{F}_{\mathbf{A}}^k$ in SPIRAR.
- Unlike various regularizer design in the primitive SPIRAR, we utilize the same scalable U-Net framework in all stages to generate the spatially shift-invariant filters. According to (12), we filter the $\boldsymbol{s}^k$ by the generated 3D filters for updating the $\boldsymbol{f}^{k+1}$.

In the resampling and initialization section, shown in Figure 3(a), the matrix $\mathbf{M}$ is from resampling operator proposed in LCT [OLW18] code, which means changing of variables in the integral by $z = \sqrt{u}$ and $v = (tc/2)^2$. C-NLOS forward model can be expressed as a 3D convolution $\boldsymbol{y}_s$ after resampling.

Subsequently we formulate the problem as a regularized Least-Square problem, of which nonlocal extension is beneficial to improving the incoherence between sampling matrix and $L1$ sparse dictionaries under the framework of model-based NLOS reconstruction. (9) can be specified as:

$$\boldsymbol{f}^{k+1} = \underset{\boldsymbol{f}\in\mathbf{R}^n}{\operatorname{argmin}} \tfrac{1}{2}\left\|\boldsymbol{f}-\boldsymbol{s}^k\right\|_2^2 + \tfrac{\tau}{\alpha_k}||\boldsymbol{f}||_1. \tag{10}$$

The proximal mapping in SPIRAR is derived as a soft thresholding function and in the proposed MIMU this step of iteration is updated by

$$\boldsymbol{s}^k = \boldsymbol{f}^k - \tfrac{1}{\alpha_k}\mathbf{A}^T(\boldsymbol{y}_s - \mathbf{A}\boldsymbol{f}^k), \tag{11}$$

$$\boldsymbol{f}^{k+1} = \mathcal{F}_{\mathbf{A}}^k(\boldsymbol{s}^k), \tag{12}$$

where $\mathcal{F}_{\mathbf{A}}^k(\cdot)$ is implemented by the proposed PVMM.

4.3. Network Structure

We propose a pyramid mapping method following the model-guided gradient descent process to allow one trained model applicable for various input sizes shown in Figure 3. Subsequently, we choose corresponding loss function to learn multiple tasks such as NLOS intensity image reconstruction and depth estimation. In order to intuitively illustrate how the proposed MIMU works, we plot the intermediary results in the bottom of Figure 3(b).

Model-Guided Gradient Descent (MGGD) In the discretized Poisson function (2), $\mathbf{A}$ is hard to estimate precisely due to the Gaussian distribution and jitter distribution in spatial and temporal domain respectively. In order to conquer the divergence in iterations, λ_k is set as a trainable parameter to control the step size in each stage, instead of the Barzilai-Borwein method used in [HMW12], which reduces amounts of auxiliary variable and memory occupation. A 3D interpolation layer with adaptive scaling factor is further adopted in each MGGD to increase the spatial-temporal resolutions of various feature map. The gradient descent in our proposed module can be expressed as:

$$\boldsymbol{s}^k = \boldsymbol{f}^k - \lambda_k\mathbf{A}^T(\boldsymbol{y}_s - \mathbf{A}\boldsymbol{f}^k). \tag{13}$$

Pyramid Voxel Mapping Module (PVMM) After caculating the optimized gradient, the goal of the next step is to make $\boldsymbol{f}^k$ closer to the desired voxel feature mapping $\boldsymbol{f}^{k+1}$. For updating the parameter

$\boldsymbol{f}^{k+1}$, we replace the previous TV-denoiser with the module shown in Figure 3(d). This lightweight U-Net based module includes two down-sampling blocks (encoding) and two up-sampling blocks (decoding), which deals with multi-scale feature maps due to redundancies of natural scenes. Each of them consists of two Convolutional layers (3×3 filters) or Trans-Conv layers (3×3 filters) with ReLU nonlinear activation function. The two Max Pooling layers are utilized to down-sample the voxel maps with a scaling factor of 2 and we noticed that the max pooling works better than average pooling in our problem, which ensures the flexibility of the whole framework and enhances the scalability for multi-scale input. The channel numbers of the voxel features in the PVMM are 32, 64, 128, 64 and 32, respectively. In order to alleviate the information loss in the down-sampling process, the skip connection (concatenation operation with the voxel feature maps in the same spatial dimension) between down-sampling blocks and up-sampling blocks are set as shown in Figure 3(d). The output voxel feature of is also added with the voxel feature map generated by MGGD module. Then the updated $\boldsymbol{f}^{k+1}$ is sent to the subsequent stage to refine the voxel feature maps by $\boldsymbol{f}^{k+1} = \mathcal{F}_{\mathbf{A}}^{k}(\boldsymbol{s}^{k})$. Ablation study in Section 5.5 shows that the PVMM module has advantages in adapting to various time resolution measurement, reducing dependence on the accuracy of the provided physical parameters.

The motivation of PVMM is to transform the voxel feature mapping to temporal measurement domain and shrink the value of each voxel rapidly with the help of ReLU and tanh nonlinear layers, which consumes hundreds of iterations to converge in the primitive SPIRAR method. Compared to original TV-based denoiser, PVMM performs a highway for data propagation without numerous loop computation, which makes it more efficient. We will verify the computing efficiency by comparing the running times later. We consider the PVMM module in each stage to play the role of mitigating the distortion introduced by the signal domain transformation model. Considering the gradient vanishing problem and intrinsic information loss such as edges and textures in MIMU framework, the long pathways connecting the encoder layers of different stages are designed to reuse multiple hidden states. In this way, voxel feature maps generated from each stage can preserve refined spatial information, leading to an spatially adaptive feature mapping operation.

5. Results

In this section, we first describe some implementation details of our proposed algorithm and then present the simulation and real data results to demonstrate the superiority of MIMU. At last, an ablation study is also conducted to evaluate each module.

5.1. Implementation Details

Training and testing datasets For the multi-scale C-NLOS reconstruction, our training dataset consists of 3000 generated measurement with corresponding intensity image and depth 256×256 pixel resolution 2D images, which involves various `bikes` 3D model. We also did data augmentation by scaling, shifting and rotation the 3D model among the samples. The real data include 6 scenes solely provided in [CWK*20]. We randomly split the synthetic dataset in training and testing pairs with 7 : 3.

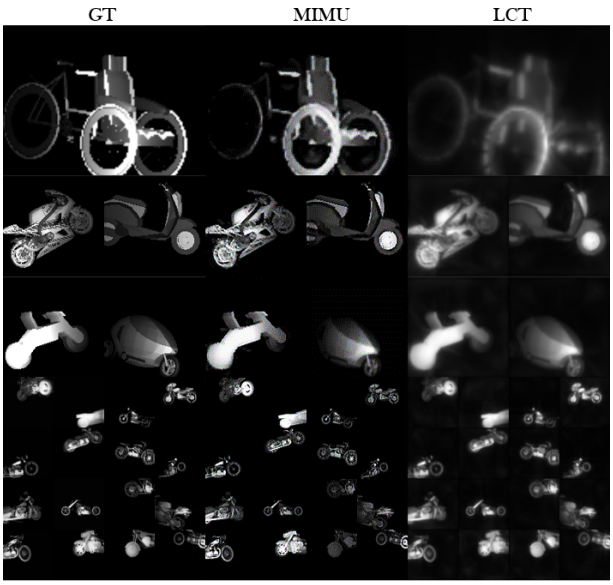

Figure 4: *Intensity image prediction in* $512 \times 512 \times 1024$*,* $256 \times 256 \times 512$ *and* $128 \times 128 \times 512$ *size.*

Loss Function Considering the albedo and depth information in each voxel of NLOS, the *loss function* is the combination of them, not like discrepancy and constraint loss used before [ZG18]. Specifically, the loss function is various norm distance between the ground truth and output of the last stage K

$$\mathcal{L} = \left\| \boldsymbol{f}^{*}_{albedo} - \boldsymbol{f}^{K}_{albedo} \right\|_{p} + \mu \left\| \boldsymbol{f}^{*}_{depth} - \boldsymbol{f}^{K}_{depth} \right\|_{p}, \tag{14}$$

where $\boldsymbol{f}^{*}_{albedo}$ and $\boldsymbol{f}^{*}_{depth}$ is the 2D ground truth of albedo and depth value generated from 3D voxel. $\boldsymbol{f}^{K}_{albedo}$ and $\boldsymbol{f}^{K}_{depth}$ are generated from the output 3D voxel of the last stage. p is selected as 1 or 2. Here we take the maximums for each pixel of the 3D voxel as the 2D albedo value, and their indexes multiplying a corresponding distance coefficient as the 2D depth value. We adopts mixed loss function, such as $L1$ plus $L2$ loss, but the convergence results remains a stabilization error especially in depth maps in our model when $p = 1$. We set $\mu = 1$ and $p = 2$ for the final training.

Training Strategy The proposed method is implemented by the pytorch 1.7. The optimizer is initialized with Adam and a digressive learning rate from 0.01 to 0.00001. We utilize a Nvidia GeForce RTX 3090 to train the proposed model consuming 2 days to make the training loss converge. We also try to add the batch normalization layer or dropout strategy to facilitate the training progress.

5.2. Results on Simulated Datasets

Gray Scale In order to compare the C-NLOS reconstruction results fairly in the same size with existing methods, we evaluate the performance in quality and quantity on the same test data in $256 \times 256 \times 512$ size, 32ps time resolution and 2m wall size. In addition to these large scale dataset in Figure 4, we further test on a small scale datasets with large light spot by rendering MNIST in

size of $64 \times 64 \times 1024$, 8ps time resolution and 0.7m wall size. The results are consistent with the common dataset.

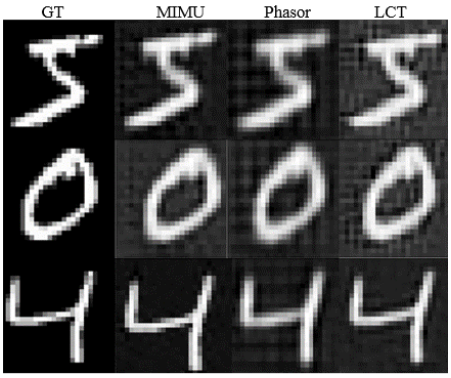

Figure 5: *Intensity image prediction on the blur MNIST dataset in* $64 \times 64 \times 1024$ *size.*

Methods	MIMU	LFE [CWK*20]	SP [WLH*21]	FBP [AGJ17]	LCT [OLW18]	FK [LWO19]
PSNR(dB)	**28.77**	27.54	18.34	17.63	19.02	23.51
SSIM	**0.92**	0.89	0.76	0.62	0.85	0.87
RMSE	**0.06**	0.08	0.65	0.73	0.59	0.54
TIME(S)	1.12	0.96	-	**0.65**	0.89	1.63

Table 1: *Quantitative results on the motorbikes testing dataset*

We quantitatively evaluate the results among these baseline approaches in Table 1, where we can see that the accuracy of the proposed MIMU exceeds existing methods, although FBP is the simplest and fastest method.

Challenging Blur Figure 5 shows the reconstruction results on the synthetic MNIST augmented by the spatial blur and temporal jitter, which is proposed in a long distance NLOS system [WLH*21]. The forward model (1) is updated to (15). The 3D convolution operation is shown in Figure 6, and more details in supplementary material (SM).

$$\begin{aligned} \boldsymbol{d}(\xi,\kappa,t) &= g_{xy} * {}_{xy} g_t *_t \\ &\iiint_\Omega \frac{1}{\boldsymbol{r}^4(x-\xi,y-\kappa,z)} \boldsymbol{\phi}(x,y,z)\delta(2\boldsymbol{r}-ct)dxdydz, \end{aligned} \tag{15}$$

where $g_t = \exp(-\frac{t^2}{2\sigma_t^2})$, $g_{xy} = \exp(-\frac{\xi^2}{2\sigma_x^2} - \frac{\kappa^2}{2\sigma_y^2})$ represent temporal and spatial distribution modeled by Gaussian function with respect to standard deviation $\sigma_t = 60ps$ and $\sigma_x = \sigma_y = 5/64m$ respectively.

In order to generate 3D dataset with depth from the 2D MNIST, we set every number as a plane randomly fixed in the space $0m$ to $1.2m$ from the reflective wall. Since there is no pretrained model

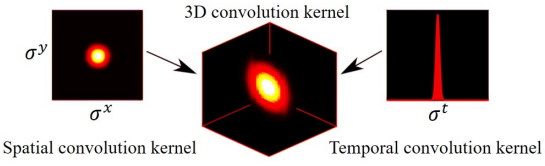

Figure 6: *The augmented operation on MNIST dataset.*

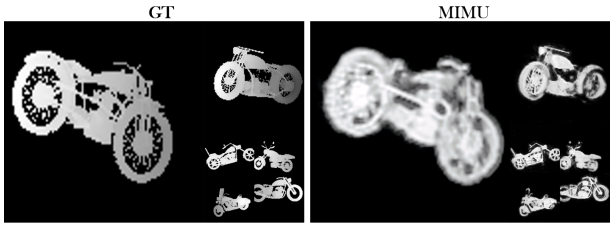

Figure 7: *Depth prediction in* $512 \times 512 \times 1024$, $256 \times 256 \times 512$ *and* $128 \times 128 \times 512$ *size. "GT" stands for depth ground truth. "MIMU" stands for our proposed method.*

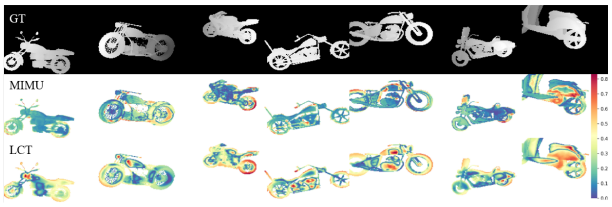

Figure 8: *Depth error heatmap of the public motorbike dataset.* ***1st row:*** *ground truth of depth map.* ***2nd row:*** *depth error map of MIMU.* ***3rd row:*** *depth error map of LCT.*

of LFE in this size, we compare with two main traditional methods. The proposed MIMU retains good quality consistently in both motorbike dataset without blur, shown in Figure 4 and blurred 3D MNIST dataset, shown in Figure 5.

Colorful Scale The proposed MIMU is able to retrieve not only the sharp shapes of the objects but also the colorful fine texture and details, as shown in Figure 1. All methods adopt the same physical parameters, specially in the iterative optimization method SPIRAL-3D, the iteration is set to 150 due to much more time cohnsuming than other methods. For the learning based method LFE, FK is adopted as its feature propagation module, which manually performs better than LCT and Phasor modules. Uniformly, the final results from all methods are normalized in the same way, as the SPIRAL-3D (SP) reconstruction appears brighter than the others. We note that the proposed MIMU, LFE and LCT are able to achieve decent geometric performance while LCT causes distortion in color. This indicates the proposed MIMU's scalability and robustness in multiple spectral channels. More importantly, the proposed MIMU retains sharper edges and finer texture on the motorbikes. Simultaneously, the proposed MIMU avoids noise and artifacts on the background that appear in other methods. Interestingly, FK misses the dark part of the reconstructed object due to the oversensitivity of the illumination condition.

RMSE Evaluation As shown in Figure 7, compared with the ground truth, the proposed method generates the least distance error and sharpest edges. The RMSE in Table 1 shows that it especially outperforms existing methods thanks to the albedo-depth joint training strategy.

Error Heatmap Besides RMSE used in [MMP*22], we also test the model with error map [CWK*20] in Figure 8 and SM. With

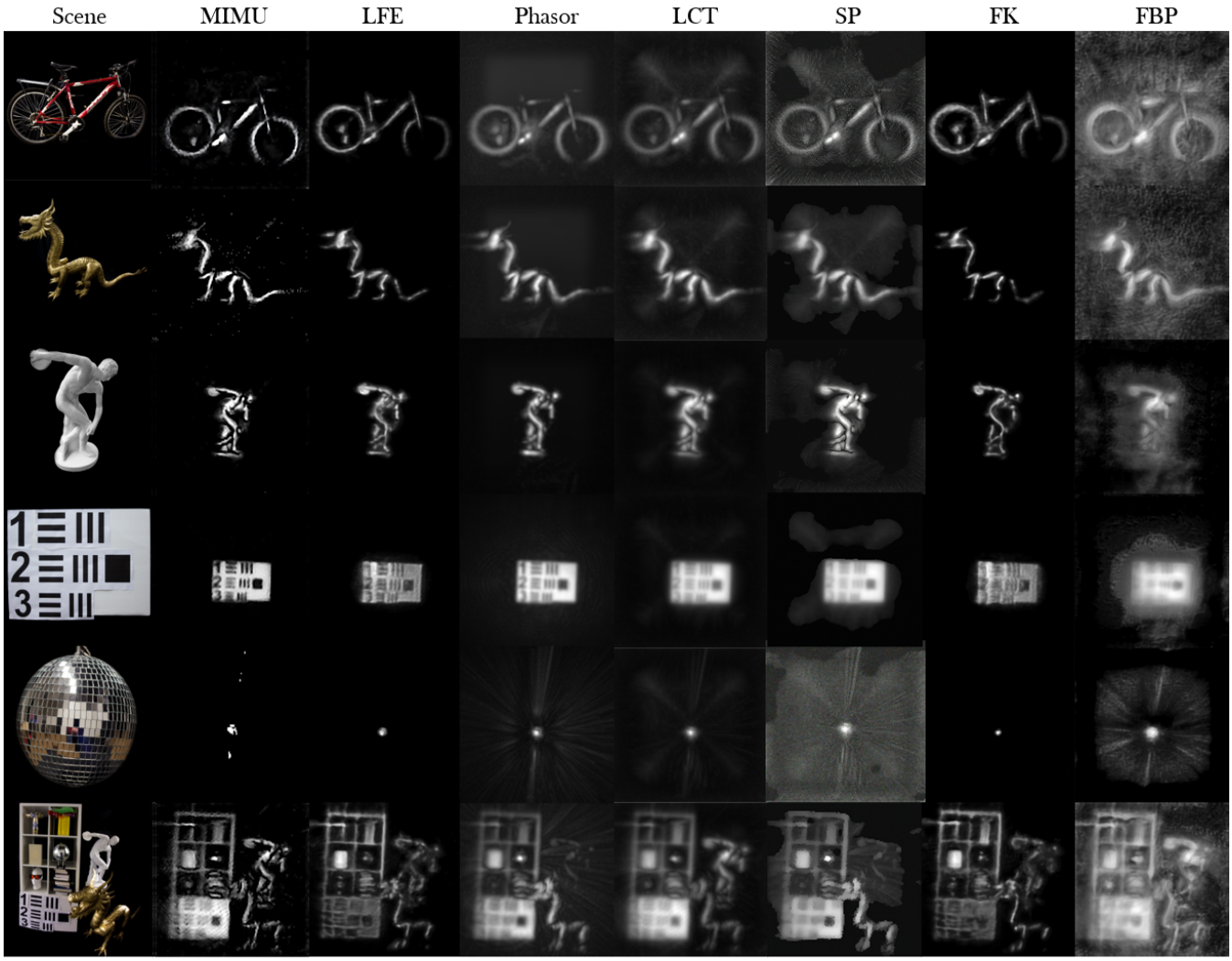

Figure 9: *Reconstructions from real measurements at* 32*ps time resolution. "MIMU" stands for proposed method. LFE [CWK*20], Phasor [LGLM*19], LCT [OLW18], FK [LWO19], FBP [AGJ17] and SP [WLH*21] are baseline methods.*

the help of the bottom color bar, we can see that the error distributions vary spatially and align with the edges and textures of the objects. Assisted by the long pathways connecting different stages' encoder layers, the entire network emphasizes the edges and textures. A quantitative assessment of testing simulation depth maps are shown in residual error map between reconstructed depth map and ground truth depth map. Compared with other methods, the proposed method generates the lowest distance error. Figure 8 demonstrates that it particularly surpasses existing methods in the background region, which can be attributed to the denoising module $\mathcal{F}\mathbf{A}^k(\cdot)$ in each stage. More details of the experiments are provided in the SM.

5.3. Results on Real Data

The qualitative results in Figure 9 is a common captured data provided in [LWO19] adopted by existing mainstream methods consisting of multiple distances, scenes, illumination and challenging mirror surfaces, which is sufficient to verify the practical performance.

The proposed MIMU is trained exclusively on synthetic bike dataset and performs well to the real measurements provided by the system [LWO19]. To ensure a fair comparison with existing methods, the 3D object model used for training, which was downloaded from a Google Drive link, is identical to that of LFE. Figure 9 validates that the proposed MIMU is able to restore hidden details, high-frequency texture information and weak illumination part, which achieves superior results in terms of both quality and quantity. In comparison to previous works, the proposed MIMU

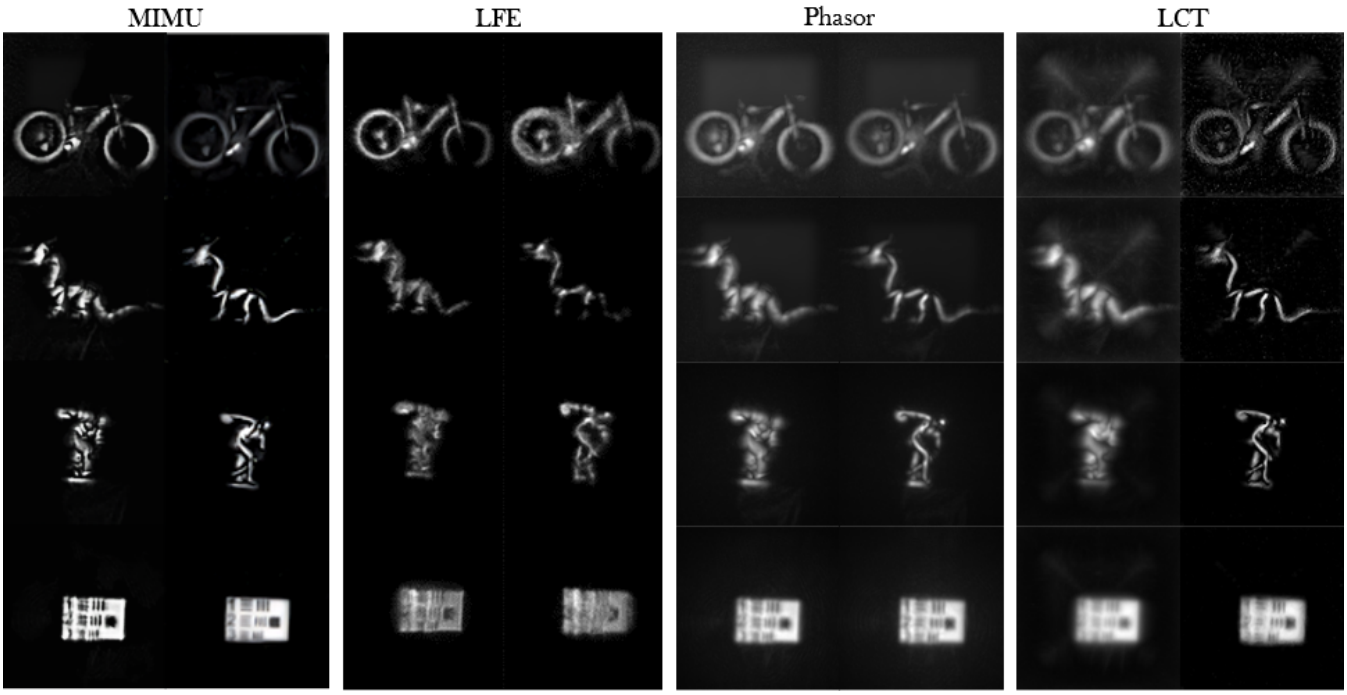

Figure 10: *Performance of reconstruction on sampling time resolution disturbance. In each method,* ***Left column:*** *data with time resolution* = 31*ps,* ***Right column:*** *data with time resolution* = 33*ps.*

network reconstructs the surface of the hidden object with sharper and cleaner details, particularly in scenes with multiple objects. It effectively eliminates the Gaussian noise from the background and the Poisson noise introduced from the detector, as demonstrated by the reconstructed resolution chart presented in the Figure 9. Moreover, the proposed MIMU method faithfully restores the bike front wheel, the dragon tail, and the adjacent blocks of the discoball, surpassing other learning-based or feed-forward methods. MIMU leverages the 3D spatial prior learned from the training dataset, enhancing the model's ability to interpret hidden information in complex scenes. For instance, in the bottom row of Figure 9, the statue is easily overlooked by existing methods due to its greater distance from other foreground objects. Our multi-stage structure takes advantage of dynamic weight adjustment in each stage, enhancing the perception of weak signals.

Additionally, it corrects the reflection error on object surfaces, ensuring that the white resolution chart appears brighter than the book on the bookshelf. When dealing with the challenging discoball, the proposed MIMU successfully distinguishes the specular blocks, demonstrating its robustness in handling diverse surfaces.

5.4. Robustness on Time Resolution

Considering the sensor errors in the real data, we evaluate the performance of the proposed MIMU at two different sampling time resolutions, which are different from the training settings. Figure 10 shows the results for four scenes. We compare the performance of two learning-based methods (MIMU, LFE) and two geometric methods (LCT, Phasor). The results demonstrate that the proposed MIMU achieves stable reconstruction even when the time resolution of the measurements differs from the parameters used during training, thereby validating its robustness against temporal interference. Another interesting insight is that the non learning-based method, such as LCT and Phasor, have fault tolerance for slight time resolution error caused by detector error. However, the performance of the LFE method significantly deteriorates as the time resolution varies under the same conditions. The MIMU, which leverages multiple stages including geometric propagation instead of relying on fixed learning priors, exhibits tolerance to parameter changes during information recovery.

5.5. Ablation Study

In order to investigate the behaviour of stage number and long pathway among multiple stages. We implement two groups of ablation studies to control variable. The top table shows that multiple stages of MIMU architecture improves performance in PSNR. As mentioned in the paper, the spatial resolution of the test dataset is 256×256.

Number of stages. The Table 2 shows how stages' number of MIMU method influences the performance. We can find that the performance increases with the larger number of stages, demonstrating the effectiveness of the proposed MIMU method. Our multiple stages structure takes advantages of the dynamic adjustment

Stages	2	3	4	8
PSNR(dB)	25.5036	28.7724	28.8753	**29.1893**

Table 2: *Ablation study of the stage number.*

Mode	Proposed	w/o Pathway	w/o Shortcut
PSNR(dB)	**28.7724**	28.0175	28.0387

Table 3: *Ablation study of different connecting mechanism.*

with corresponding weights in different stages to enhance the perception of the weak signal. By making a trade-off between performance and computational complexity, we employ 3 stages in our proposed network.

Pathway Design. As illustrated in Table 3, to verify the effectiveness of the long pathways connecting among different stages' encoder layers, we remove it from the proposed MIMU method, represented as "w/o Pathway". Additionally, we investigate the effectiveness of the shortcut in PVMM module in each stage by removing it, represented as "w/o Shortcut" in Table 3. The performance degradation demonstrates the positive effect of these designs, and it is evident that the long pathways contribute more to the final results.

Backbone Comparison. Since MGGD integrated with SPIRAR is irreplaceable, we test different backbones in PVMM in Table 4. We also replace PVMM with PnP framewor [VBW13] for comparison, which has been demonstrated in multiple image restoration tasks [ZLZ*22, ZZZ19, YLSD20].

Backbone	Proposed	VDSR [KLL15]	VGG16 [SZ15]	PnP [TDV20]	MAE [HCX*22]
PSNR(dB)	**28.7724**	27.8352	25.4581	24.0384	22.5920

Table 4: *Comparison of various light backbone in PVMM.*

5.6. Memory and Runtime

We implement all experiments on an NVIDIA GeForce RTX 3090 GPU. The inference process of the proposed MIMU consumes about 5GB of GPU memory, which is significantly lower than the feed-forward methods. The runtime of each method is listed in Table 1.

6. Discussion and Conclusion

We have proposed a robust MIMU architecture method to closely integrate the iterative optimization and learning prior for versatile NLOS reconstruction tasks, such as 2D image reconstruction, depth map estimation and colorful image recovery. Visual results and quantity evaluation demonstrate that the proposed MIMU consistently outperforms existing learning-based, feed-forward and iterative optimization baseline methods.

The main advancement of MIMU benefits from the multiple concatenated stages comparing existing deep learning method. Specifically, LFE [CWK*20] takes physical prior one time and Feed-Forward [GCHWI20] is an end-to-end network without physical model, which tends to generate unsatisfactory results with temporal disturbance due to a local optimized solution. Our MIMU method plays different roles with corresponding physical prior in different stages, which guides the reconstruction convergence closer to the ground truth step by step.

Simulator used to generate training dataset inevitably mismatches real physical process and affects the performance of learning-based methods on experimental data. In the future, we will design simulators closer to real physical process to bridge this gap.

References

[Abr78] ABRAMSON N.: Light-in-flight recording by holography. *Opt. Lett. 3*, 4 (Oct. 1978), 121–123. Publisher: Optica Publishing Group. URL: https://opg.optica.org/ol/abstract.cfm?URI=ol-3-4-121, doi:10.1364/OL.3.000121. 2

[AGJ17] ARELLANO V., GUTIERREZ D., JARABO A.: Fast back-projection for non-line of sight reconstruction. In *ACM SIGGRAPH 2017 Posters* (Los Angeles California, July 2017), ACM, pp. 1–2. URL: https://dl.acm.org/doi/10.1145/3102163.3102241, doi:10.1145/3102163.3102241. 1, 2, 6, 7

[BVT*16] BRONZI D., VILLA F., TISA S., TOSI A., ZAPPA F.: Spad figures of merit for photon-counting, photon-timing, and imaging applications: A review. *IEEE Sensors Journal 16*, 1 (2016), 3–12. doi:10.1109/JSEN.2015.2483565. 3

[BZT*15] BUTTAFAVA M., ZEMAN J., TOSI A., ELICEIRI K., VELTEN A.: Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express 23*, 16 (2015), 20997–21011. 2

[CWK*20] CHEN W., WEI F., KUTULAKOS K. N., RUSINKIEWICZ S., HEIDE F.: Learned feature embeddings for non-line-of-sight imaging and recognition. *ACM Transactions on Graphics 39*, 6 (Dec. 2020), 1–18. URL: https://dl.acm.org/doi/10.1145/3414685.3417825, doi:10.1145/3414685.3417825. 1, 2, 5, 6, 7, 9

[FVW20] FACCIO D., VELTEN A., WETZSTEIN G.: Non-line-of-sight imaging. *Nature Reviews Physics 2*, 6 (2020), 318–327. 1

[GCHWI20] GRAU CHOPITE J., HULLIN M. B., WAND M., ISERINGHAUSEN J.: Deep non-line-of-sight reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 960–969. 2, 9

[GWV*12] GUPTA O., WILLWACHER T., VELTEN A., VEERARAGHAVAN A., RASKAR R.: Reconstruction of hidden 3d shapes using diffuse reflections. *Optics express 20*, 17 (2012), 19096–19108. 2

[HCX*22] HE K., CHEN X., XIE S., LI Y., DOLLÁR P., GIRSHICK R.: Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 16000–16009. 9

[HDY*21] HUANG T., DONG W., YUAN X., WU J., SHI G.: Deep Gaussian Scale Mixture Prior for Spectral Compressive Imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, June 2021), IEEE, pp. 16211–16220. URL: https://ieeexplore.ieee.org/document/9578572/, doi:10.1109/CVPR46437.2021.01595. 4

[HMW12] HARMANY Z. T., MARCIA R. F., WILLETT R. M.: This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and Practice. *IEEE Transactions on Image Processing 21*, 3 (Mar. 2012), 1084–1096. Conference Name: IEEE Transactions on Image Processing. doi:10.1109/TIP.2011.2168410. 3, 4

[HOZ*19] HEIDE F., O'TOOLE M., ZANG K., LINDELL D. B., DIAMOND S., WETZSTEIN G.: Non-line-of-sight imaging with partial occluders and surface normals. *ACM Transactions on Graphics (ToG) 38*, 3 (2019), 1–10. 2

[HXHH14] HEIDE F., XIAO L., HEIDRICH W., HULLIN M. B.: Diffuse Mirrors: 3D Reconstruction from Diffuse Indirect Illumination Using Inexpensive Time-of-Flight Sensors. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (June 2014), pp. 3222–3229. ISSN: 1063-6919. doi:10.1109/CVPR.2014.418. 2

[ICY*20] ISOGAWA M., CHAN D., YUAN Y., KITANI K., O'TOOLE M.: Efficient non-line-of-sight imaging from transient sinograms. In *European Conference on Computer Vision* (2020), Springer, pp. 193–208. 2

[IH20] ISERINGHAUSEN J., HULLIN M. B.: Non-line-of-sight reconstruction using efficient transient rendering. *ACM Transactions on Graphics (ToG) 39*, 1 (2020), 1–14. 2

[KHDR09] KIRMANI A., HUTCHISON T., DAVIS J., RASKAR R.: Looking around the corner using transient imaging. In *2009 IEEE 12th International Conference on Computer Vision* (2009), IEEE, pp. 159–166. 2

[KLL15] KIM J., LEE J. K., LEE K. M.: Accurate image super-resolution using very deep convolutional networks. *CoRR abs/1511.04587* (2015). URL: http://arxiv.org/abs/1511.04587, arXiv:1511.04587. 9

[LGLM*19] LIU X., GUILLÉN I., LA MANNA M., NAM J. H., REZA S. A., HUU LE T., JARABO A., GUTIERREZ D., VELTEN A.: Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature 572*, 7771 (2019), 620–623. 1, 2, 7

[LWK19] LINDELL D. B., WETZSTEIN G., KOLTUN V.: Acoustic non-line-of-sight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6780–6789. 1

[LWL*21] LIU X., WANG J., LI Z., SHI Z., FU X., QIU L.: Non-line-of-sight reconstruction with signal–object collaborative regularization. *Light: Science & Applications 10*, 1 (2021), 1–20. 2

[LWO19] LINDELL D. B., WETZSTEIN G., O'TOOLE M.: Wave-based non-line-of-sight imaging using fast f-k migration. *ACM Transactions on Graphics 38*, 4 (July 2019), 116:1–116:13. URL: https://doi.org/10.1145/3306346.3322937, doi:10.1145/3306346.3322937. 1, 6, 7

[LYP*22] LIU P., YU Y., PAN Z., PENG X., LI R., WANG Y., YU J., LI S.: Hiddenpose: Non-line-of-sight 3d human pose estimation. In *2022 IEEE International Conference on Computational Photography (ICCP)* (2022), IEEE, pp. 1–12. 2

[LZH*22] LIU J., ZHOU Y., HUANG X., LI Z.-P., XU F.: Photon-efficient non-line-of-sight imaging. *IEEE Transactions on Computational Imaging 8* (2022), 639–650. 2

[MJY20] MENG Z., JALALI S., YUAN X.: GAP-net for Snapshot Compressive Imaging, Dec. 2020. arXiv:2012.08364 [eess]. URL: http://arxiv.org/abs/2012.08364. 3, 4

[MMP*22] MU F., MO S., PENG J., LIU X., NAM J. H., RAGHAVAN S., VELTEN A., LI Y.: Physics to the Rescue: Deep Non-line-of-sight Reconstruction for High-speed Imaging, Aug. 2022. arXiv:2205.01679 [cs, eess]. URL: http://arxiv.org/abs/2205.01679. 2, 6

[MSS*19] MAEDA T., SATAT G., SWEDISH T., SINHA L., RASKAR R.: Recent advances in imaging around corners. *arXiv preprint arXiv:1910.05613* (2019). 1, 2

[MWZ22] MOU C., WANG Q., ZHANG J.: Deep Generalized Unfolding Networks for Image Restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, June 2022), IEEE, pp. 17378–17389. URL: https://ieeexplore.ieee.org/document/9878586/, doi:10.1109/CVPR52688.2022.01688. 3, 4

[NBB*21] NAM J. H., BRANDT E., BAUER S., LIU X., RENNA M., TOSI A., SIFAKIS E., VELTEN A.: Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. *Nature communications 12*, 1 (2021), 1–10. 2

[OLW18] O'TOOLE M., LINDELL D. B., WETZSTEIN G.: Confocal non-line-of-sight imaging based on the light-cone transform. *Nature 555*, 7696 (Mar. 2018), 338–341. URL: http://www.nature.com/articles/nature25489, doi:10.1038/nature25489. 1, 2, 3, 4, 6, 7

[PDV19] PEDIREDLA A., DAVE A., VEERARAGHAVAN A.: Snlos: Non-line-of-sight scanning through temporal focusing. In *2019 IEEE International Conference on Computational Photography (ICCP)* (2019), IEEE, pp. 1–13. 2

[PZD*21] PEI C., ZHANG A., DENG Y., XU F., WU J., DAVID U., LI L., QIAO H., FANG L., DAI Q.: Dynamic non-line-of-sight imaging system based on the optimization of point spread functions. *Optics Express 29*, 20 (2021), 32349–32364. 2

[SCZ*20] SOLOMON O., COHEN R., ZHANG Y., YANG Y., HE Q., LUO J., VAN SLOUN R. J. G., ELDAR Y. C.: Deep Unfolded Robust PCA With Application to Clutter Suppression in Ultrasound. *IEEE Transactions on Medical Imaging 39*, 4 (Apr. 2020), 1051–1063. URL: https://ieeexplore.ieee.org/document/8836615/, doi:10.1109/TMI.2019.2941271. 4

[SMBG19] SAUNDERS C., MURRAY-BRUCE J., GOYAL V. K.: Computational periscopy with an ordinary digital camera. *Nature 565*, 7740 (2019), 472–475. 1

[SOG18] SMITH B. M., O'TOOLE M., GUPTA M.: Tracking multiple objects outside the line of sight using speckle imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6258–6266. 1

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)* (2015), Computational and Biological Learning Society, pp. 1–14. 9

[TDV20] TASSANO M., DELON J., VEIT T.: Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 1354–1363. 9

[Tib96] TIBSHIRANI R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288. 3

[TSG19] TSAI C.-Y., SANKARANARAYANAN A. C., GKIOULEKAS I.: Beyond volumetric albedo–a surface optimization framework for non-line-of-sight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1545–1555. 2

[VBW13] VENKATAKRISHNAN S. V., BOUMAN C. A., WOHLBERG B.: Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing* (2013), IEEE, pp. 945–948. 9

[VWG*12] VELTEN A., WILLWACHER T., GUPTA O., VEERARAGHAVAN A., BAWENDI M. G., RASKAR R.: Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications 3*, 1 (Jan. 2012), 745. URL: http://www.nature.com/articles/ncomms1747, doi:10.1038/ncomms1747. 2

[WLH*21] WU C., LIU J., HUANG X., LI Z.-P., YU C., YE J.-T., ZHANG J., ZHANG Q., DOU X., GOYAL V. K., XU F., PAN J.-W.: Non–line-of-sight imaging over 1.43 km. *Proceedings of the National Academy of Sciences 118*, 10 (Mar. 2021), e2024468118. URL: https://pnas.org/doi/full/10.1073/pnas.2024468118, doi:10.1073/pnas.2024468118. 1, 2, 3, 6, 7

[WNF09] WRIGHT S. J., NOWAK R. D., FIGUEIREDO M. A. T.: Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing 57*, 7 (2009), 2479–2493. doi:10.1109/TSP.2009.2016892. 3

[WZH*21] WANG B., ZHENG M.-Y., HAN J.-J., HUANG X., XIE X.-P., XU F., ZHANG Q., PAN J.-W.: Non-line-of-sight imaging with picosecond temporal resolution. *Physical Review Letters 127*, 5 (2021), 053602. 2

[WZM21] WT Z., ZHANGT J., MOU C.: Dense Deep Unfolding Network with 3D-CNN Prior for Snapshot Compressive Imaging, Oct. 2021. doi:10.1109/ICCV48922.2021.00485. 3, 4

[XNK*19] XIN S., NOUSIAS S., KUTULAKOS K. N., SANKARANARAYANAN A. C., NARASIMHAN S. G., GKIOULEKAS I.: A theory of fermat paths for non-line-of-sight shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6800–6809. 2

[XST*18] XU F., SHULKIND G., THRAMPOULIDIS C., SHAPIRO J. H., TORRALBA A., WONG F. N., WORNELL G. W.: Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *Optics express 26*, 8 (2018), 9945–9962. 1

[YHLX21] YE J.-T., HUANG X., LI Z.-P., XU F.: Compressed sensing for active non-line-of-sight imaging. *Optics Express 29*, 2 (2021), 1749–1763. 2

[YLG*20] YOUNG S. I., LINDELL D. B., GIROD B., TAUBMAN D., WETZSTEIN G.: Non-line-of-sight surface reconstruction using the directional light-cone transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 1407–1416. 2

[YLSD20] YUAN X., LIU Y., SUO J., DAI Q.: Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 1447–1457. 9

[ZG18] ZHANG J., GHANEM B.: ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, June 2018), IEEE, pp. 1828–1837. URL: https://ieeexplore.ieee.org/document/8578294/, doi:10.1109/CVPR.2018.00196. 3, 4, 5

[ZLZ*22] ZHANG K., LI Y., ZUO W., ZHANG L., VAN GOOL L., TIMOFTE R.: Plug-and-Play Image Restoration With Deep Denoiser Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 10 (Oct. 2022), 6360–6376. URL: https://ieeexplore.ieee.org/document/9454311/, doi:10.1109/TPAMI.2021.3088914. 9

[ZZZ19] ZHANG K., ZUO W., ZHANG L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1671–1681. 9