Passive Non-Line-of-Sight Imaging Using Optimal Transport

Ruixu Geng, *Graduate Student Member, IEEE*, Yang Hu, Zhi Lu, *Graduate Student Member, IEEE*, Cong Yu, *Graduate Student Member, IEEE*, Houqiang Li, *Fellow, IEEE*, Hengyu Zhang, and Yan Chen, *Senior Member, IEEE*

***Abstract*— Passive non-line-of-sight (NLOS) imaging has drawn great attention in recent years. However, all existing methods are in common limited to simple hidden scenes, low-quality reconstruction, and small-scale datasets. In this paper, we propose NLOS-OT, a novel passive NLOS imaging framework based on manifold embedding and optimal transport, to reconstruct high-quality complicated hidden scenes. NLOS-OT converts the high-dimensional reconstruction task to a low-dimensional manifold mapping through optimal transport, alleviating the ill-posedness in passive NLOS imaging. Besides, we create the first large-scale passive NLOS imaging dataset, NLOS-Passive, which includes 50 groups and more than 3,200,000 images. NLOS-Passive collects target images with different distributions and their corresponding observed projections under various conditions, which can be used to evaluate the performance of passive NLOS imaging algorithms. It is shown that the proposed NLOS-OT framework achieves much better performance than the state-of-the-art methods on NLOS-Passive. We believe that the NLOS-OT framework together with the NLOS-Passive dataset is a big step and can inspire many ideas towards the development of learning-based passive NLOS imaging. Codes and dataset are publicly available (https://github.com/ruixv/NLOS-OT).**

***Index Terms*— Non-line-of-sight imaging, optimal transport, autoencoder, manifold embedding.**

I. Introduction

NON-LINE-OF-SIGHT (NLOS) imaging enables hidden objects to be seen when occluded from direct view by analyzing the scattered light on a relay wall. With the trait of seeing hidden objects, NLOS imaging has numerous potential applications in autonomous vehicles, robotic vision and remote sensing. This inherent characteristic of greatly expanding the field of view and improving the observation capability has made NLOS imaging arouse great attention in recent years.

Depending on whether a controllable light source is used, NLOS imaging can be divided into active imaging [1]–[7] and passive imaging [8]–[12]. Among them, active NLOS imaging uses an ultrafast laser to illuminate the area on the relay surface, and exploits a high resolution time-resolved detector to capture the transient response of three-bounce light. Exploiting the controllable light source, active imaging can obtain photon responses at different moments and positions, and reconstruct the three-dimensional hidden scene with high quality. In this paper, we focus on passive imaging methods without controllable light sources to complete NLOS reconstruction using an ordinary camera, as shown in Fig. 1.

Passive NLOS imaging is an extremely challenging problem because of uncontrollable probe illumination [10]. Specifically, the close contribution between pixels due to isotropic diffuse reflection makes the condition number of light transport matrix in passive NLOS imaging very large, causing it difficult to obtain good reconstructions from the observations. To alleviate this problem, many methods have been proposed, including placing a partial occluder [9], [11], using polarizers [10] and applying deep learning [12]–[14]. Among them, deep learning-based passive NLOS imaging [12]–[15] is attractive since the superior representation ability of deep neural networks can greatly improve the reconstruction resolution. However, there are still several challenges when applying deep learning for passive NLOS imaging. Firstly, existing methods utilize the U-Net [16], a mature network structure that has been verified to be effective in not very ill-posed tasks including image segmentation and deblurring, as the basic network structure, which is however not quite effective since the distributions of the input and output of the passive NLOS imaging are extremely different. Secondly, there is no large-scale dataset for the passive NLOS imaging, due to which the advantages of deep learning cannot be fully explored. Recent work [12] simulates the forward propagation process based on the Phong model [17] to produce datasets, which however cannot be used for practical lighting conditions due to the ideal assumptions in the model.

In this paper, we are committed to addressing the above challenges. Particularly, we propose a network architecture based on the optimal transport (OT) theory [18], NLOS-OT,

Manuscript received July 1, 2021; revised October 11, 2021 and November 3, 2021; accepted November 4, 2021. Date of publication November 22, 2021; date of current version November 30, 2021. This work was supported by the National Natural Science Foundation of China under Grant 62172381. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Salerno. *(Corresponding authors: Yang Hu; Hengyu Zhang; Yan Chen.)*

Ruixu Geng, Zhi Lu, and Cong Yu are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: gengruixu@std.uestc.edu.cn; zhilu@std.uestc.edu.cn; congyu@std.uestc.edu.cn).

Yang Hu and Houqiang Li are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: eeyhu@ustc.edu.cn; lihq@ustc.edu.cn).

Hengyu Zhang is with the Department of Cardiology, West China Hospital, Sichuan University, Chengdu 610041, China (e-mail: zhanghe8635_cn@sina.com).

Yan Chen is with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: eecyan@ustc.edu.cn).

This article has supplementary downloadable material available at https://10.1109/TIP.2021.3128312, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3128312

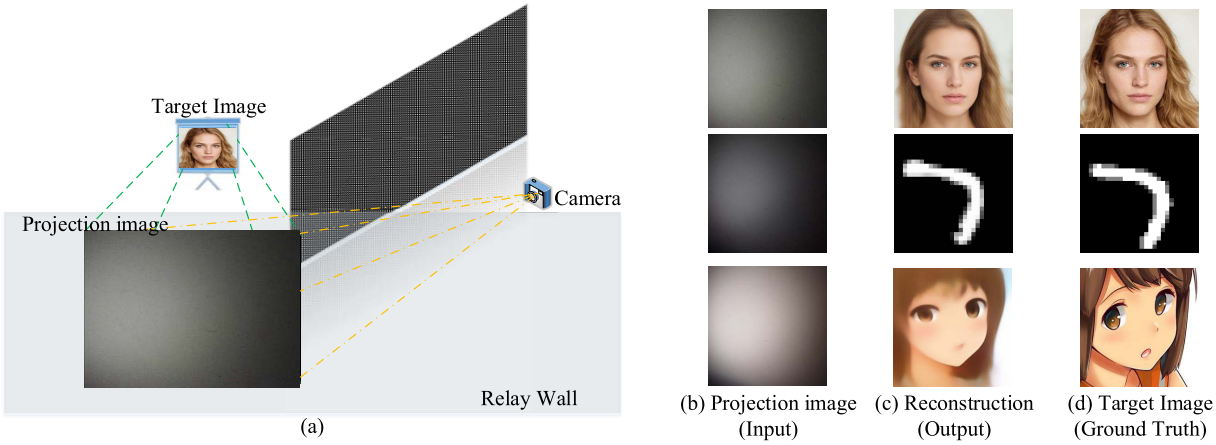


Fig. 1. **Passive NLOS Imaging. (a)** Experimental setup, where the LCD screen displays target images, the corresponding light projects onto the relay wall, and the camera captures the projection image on the wall. **(b)** Projection image on the relay wall captured by the camera. **(c)** Reconstructed image by the proposed NLOS-OT. **(d)** Target image displayed on the LCD screen.

for the passive NLOS imaging to resolve the unbalance distribution challenge between the input and output. The proposed NLOS-OT first obtains the latent code for the target image through an autoencoder, and then maps the projection image to the latent code through another encoder by the optimal transport. It is worth noting that the OT, in theory, has the capability to map a projection image with limited information to the latent space of the target image. In addition, we create the first large-scale dataset, NLOS-Passive, for passive NLOS imaging utilizing a common LCD and a mobile phone camera. The NLOS-Passive includes 50 groups[1] and more than 3,200,000 projection images, where the projections of the MNIST [19], public face data [20], animation face data [21], [22], and STL-10 [23] on different relay walls are captured. NLOS-Passive contains data under different optical transport conditions, such as brightness, camera angle, hidden object position, etc. Thus, NLOS-Passive can be used to not only study the performance of different algorithms, but also compare the performances of a specific algorithm in different optical conditions. We believe that the NLOS-OT framework together with the NLOS-Passive dataset is a big step and can inspire many ideas towards the development of learning-based passive NLOS imaging. Our primary contributions are summarized as follows:

1) We propose a novel framework named NLOS-OT, for the challenging passive NLOS imaging. The NLOS-OT enables passive NLOS imaging in complex scenes through manifold embedding and optimal transport. Through experiments, we have demonstrated that NLOS-OT performs significantly better than existed end-to-end training framework.
2) We verify through experiments that the diffuse projection image, even captured by cell phone camera, under unknown partial occlusion contains enough information about the hidden scene. When using widely distributed data with unknown occlusion for training, NLOS-OT can complete the test on a completely different dataset, which shows that NLOS-OT has strong generalization and further demonstrates the feasibility of passive NLOS imaging tasks.
3) We build NLOS-Passive, a large-scale passive NLOS dataset containing more than 50 groups of data and 3,200,000 samples. NLOS-Passive includes data of different hidden scenes collected under different light transport conditions, which can be used to evaluate the performance of passive NLOS imaging algorithms. To the best of our knowledge, NLOS-Passive is the first public large-scale passive NLOS dataset.

The remainder of this work is organized as follows. Section II presents related works on NLOS imaging, optimal transport, generative models and optical eavesdropping. Section III describes our proposed NLOS-OT model in detail. The experimental setups and results are provided in Sec. IV. Finally, discussions and conclusions are drawn in Sec. V and Sec. VI respectively.

II. Related Works

In this section, we introduce the related works on passive NLOS imaging, generative models, OT theory, and optical eavesdropping respectively.

A. Passive Non-Line-of-Sight Imaging

Depending on whether there is a controllable external light source, NLOS imaging has active methods and passive methods. Here, we focus on passive NLOS imaging.

Without a controllable light source, the passive methods collect mostly intensity information and depend on incoherent ambient light for illumination [24] or directly reconstruct the two-dimensional image displayed on a screen, which has also received much attention in recent years. Torralba and Freeman observed for the first time that the surrounding environment can be used as an "accidental" camera to recover hidden

[1]Each "group" refers to the data of the same dataset collected under the same optical conditions (such as distance, angle, illumination, reflective surface). The size of the group depends on the size of the image dataset. For example, it is 70,000 for MNIST.

objects [25]. Passive methods usually have a much lower cost and faster data collection speed than active methods. However, since passive NLOS imaging only collects the intensity of reflected light, which lacks information such as time and phase in active methods, most existing works [9], [12] only focused perform 2D reconstruction or localization while a few recent works can estimate both hidden shape and depth with partial occluder [26], [27]. Specifically, for passive reconstruction tasks, intensity information is the most commonly utilized information [9]–[11], [13], [26]. In such a case, two-dimensional reconstruction can be completed, such as the computational periscopy [9] and computational mirrors [13]. Besides, by introducing depth information into the forward model, the distance of a simple scene can be roughly estimated with partial occlusion [26], [27]. If the goal is localization, all we need is to get the distance between several target points and the relay wall. Because several target points can reflect or emit coherent light independently and move over time, coherence-based information (such as spatial coherence [8], optical ToF [28]) and space-time information [29] can be exploited to complete passive sensing and localization.

In this paper, we mainly focus on 2D reconstruction in passive NLOS scenes, i.e., passive NLOS imaging as shown in Fig. 1. As described in Section I, the existing methods usually include physical-based methods (i.e., placing partial occlusion [9], [11], using polarizers [10] or exploiting optical memory effect [30]) and deep learning-based methods [12]–[15]. However, physical-based methods can only complete the rough reconstruction of simple scenes, while the methods based on deep learning have challenge on generalization ability and the reconstruction performance greatly depends on the similarity between the training set and the test set. Therefore, it is meaningful to develop a new passive NLOS imaging model that can not only achieve extremely high-quality reconstruction on specific datasets, but also have great generalization capabilities on large-scale datasets, which are the features of the proposed NLOS-OT.

B. Optimal Transport and Generative Models

1) Optimal Transport: Optimal transport(OT) theory studies the transmission problem of different distributions and has been successfully used in domain adaptation [31], image processing and other fields [32]. From a geometric view, OT can measure the distribution difference between two manifolds embedded in a high-dimensional space, which is similar to "earth mover's distance" (EMD) [33] used in WGAN [34]. Aude et al. used the stochastic gradient descent method to solve the OT problem [35], while Lei et al. applied OT to deep learning through convex optimization [36]. For extremely challenging image restoration tasks, using OT to map the input to the latent code of target space can effectively exploit the information in the input image [37], [38]. Nevertheless, to the best of our knowledge, there is no research on applying OT to passive NLOS imaging tasks.

2) Generative Models: Numerous generative models have been successfully applied to image restoration tasks in recent years. Encoder-Decoder based models(AEs) [16], [39]–[42] and Generative Adversarial Networks (GANs) [34], [43]–[47] are two of the most dominant approaches since they can generate high-quality and realistic results. However, due to limited model interpretation, injected noise, and element-wise noise, AE-based models often produce blurry images [40]. On the other hand, GAN is difficult to train and prone to mode collapse/mixture problems since the transport map is discontinuous while DNNs can only represent continuous maps [48]. In this paper, different from AEs and GANs, NLOS-OT first obtains the latent code through manifold embedding and then employs optimal transport to map the input data space to the latent space. The reduction of dimensionality makes passive NLOS problem more practical.

C. Optical Eavesdropping

Techniques for utilizing optical compromising emanations to obtain user privacy have a rich history [49]–[53]. In these tasks, the eavesdroppers used sensors near the user's display screen to monitor the information on the screen. As an early work, Kuhn [49] exploited photosensor to spy CRT (cathode-ray tube) computer monitors. Due to the raster scan of CRTs, time-resolved sensors can separate different pixels, and then complete pixel-level reconstruction. For NLOS scenes, because the response time of the diffuse reflection is short enough, [49] has completed the reconstruction of the CRTs screen reflected by a diffuse wall. However, for flat-panel displays (FPDs, e.g., LCD monitors and plasma screens), it is tough to resolve temporal information, which greatly increases the difficulty of NLOS reconstruction [49], [50]. References [51] and [52] used a relay surface with specular material (such as eyeballs) and a high-power telescope to complete NLOS eavesdropping. It can be seen that most optical eavesdropping works essentially avoid the diffuse reflection on the wall – [49], [50] exploited the raster scan of CRTs, and [51], [52] adopted specular reflective materials to replace the diffuse wall. On the contrary, our work aims to recover the hidden scenes displayed on an ordinary FPD and reflected by a diffuse wall, which is a typical problem statement in passive NLOS imaging [9], [10], [12], [13], [54]. As a price, compared with optical eavesdropping, the existing passive NLOS imaging detection distance is very short, which makes it difficult to use in snooping scenes despite its vast potential.

III. Our Approach

Here, we propose NLOS-OT, a novel framework designed for passive NLOS tasks. In this section, we introduce our settings and explain the motivations of NLOS-OT, then discuss our network and loss function.

A. Problem Setup

As shown in Fig. 1 (a), the objective of passive NLOS imaging is to recover the target image, i.e., the hidden object, by processing the projection image, i.e., the observed information, on the diffuse reflection wall. Assuming that each pixel on the target image is an independent point light source,

then the corresponding measured projection on the wall can be written as

$$I(p_y) = \int\int_{p_f \in F} A(p_f, p_y) I(p_f) dp_f + n_{b+d}(p_y) \quad (1)$$

where $I(p_y)$ is the light intensity on the pixel p_y of the detected projection area, and $I(p_f)$ is the intensity on the pixel p_f of the hidden source display area. Besides, $A(p_f, p_y)$ is the optical transport from the point light source p_f to area p_y on the relay wall. F represents all pixels on the entire screen, which is a rectangular area with two spatial dimensions, corresponding to the two integrals in Eq. 1. $n_{b+d}(p_y)$ is the noise at the pixel p_y, generated by the background (b) light and the detector (d) itself. The model can be discretized as

$$\mathbf{y} = \mathbf{A}\mathbf{f} + \mathbf{n_{b+d}} \quad (2)$$

where $\mathbf{f} \in \mathbb{R}^{H_f W_f}$ is the vectorized scene intensities and $H_f \times W_f$ is the resolution of the display. $\mathbf{y} \in \mathbb{R}^{H_y W_y}$ is the vectorized observation, and $H_y \times W_y$ is the resolution of the measured projection image. $\mathbf{A} \in \mathbb{R}^{H_y W_y \times H_f W_f}$ is the light transport matrix. $\mathbf{n_{b+d}} \in \mathbb{R}^{H_y W_y}$ represents the vectorized noise. Considering that the optical transport matrix $\mathbf{A}$ is determined by the bidirectional reflectance distribution function (BRDF) of the wall μ, the range between the hidden object and the wall r, the position of the camera c, $\mathbf{A}$ can be denoted as $\mathbf{A}(\mu, r, c)$. Hence, Eq. 2 can be rewritten as

$$\mathbf{y} = \mathbf{A}(\mu, r, c)\mathbf{f} + \mathbf{n_{b+d}} \quad (3)$$

We further use the matrix form to represent the scene intensities and observation, where $T_{\mu,r,c,n}$ is the corresponding transformation from the target image f to the projection image y. Please note that the $T^{-1}_{\mu,r,c,n}$ here is not equivalent to the inverse of A^{-1}, but a comprehensive consideration of $\mathbf{A^{-1}}$, noise $\mathbf{n}$ and data distribution. Thus, the passive NLOS imaging problem can be written as

$$f = T^{-1}_{\mu,r,c,n}(y) \quad (4)$$

where the mapping $T^{-1}_{\mu,r,c,n}$ is the reconstruction process that the proposed NLOS-OT aims to do.

Due to the large condition number of the optical transport matrix $\mathbf{A}$, the passive NLOS reconstruction is very challenging. Existing methods alleviated this problem by exploiting the additional obstacles [9] and polarizers [10]. In this paper, we resolve this problem by utilizing the OT theory and exploiting the deep image prior with an autoencoder. For specific datasets, such as MNIST, NLOS-OT can obtain sharp and high-quality reconstruction results by learning data distribution, inverse transport process, and noise distribution; for widely distributed datasets, such as STL-10, NLOS-OT mainly learns inverse light transport to obtain a good result with solid generalization ability. Besides, NLOS-OT does not need to assume prior knowledge of $\mathbf{A}$, e.g., the BRDF of the wall.

B. Network Architecture of NLOS-OT Model

1) Motivation of NLOS-OT Network: Existing data-driven passive NLOS imaging methods [12]–[14] mainly use the U-Net [16], which however do not consider the characteristic of NLOS imaging task where the distribution of the input and output are very unbalanced. Such methods can only reconstruct simple scenes and will result in fuzzy artifacts on the complex scene as illustrated in our experimental results. Moreover, if changing the U-Net to the conditional GAN, mode collapse is prone to occur.

The main reason for the above problems is that the existing network structure cannot efficiently map the limited features in the projection image y to the target space. Particularly, existing methods have completed this task in the image space, which has too many modes to achieve good results. On the contrary, we hope to complete this task in the embedded latent space using OT. Hence, the proposed NLOS-OT framework consists of two steps: obtaining the latent code of the target image through an autoencoder in step 1; and mapping the projection image to the latent space through the OT theory in step 2.

Therefore, the difficult passive NLOS imaging task is decomposed into two simple tasks in the proposed NLOS-OT model, i.e., step 1 manifold embedding and step 2 optimal transport. Note that in step 1, because the input contains all the information of the output, the manifold embedding is easy to implement. In step 2, the OT is performed in the latent space $\mathcal{Z}$, which greatly reduces the difficulty. Moreover, the OT theory can alleviate the mode collapse problem that may occur during the mapping process. Both step 1 and step 2 are implemented by deep learning, of which the network structure and loss function are described thoroughly in the following.

2) Network of NLOS-OT: As shown in Fig. 2, NLOS-OT mainly consists of three parts, namely encoder E_1 and decoder D_1 for manifold embedding, and encoder E_2 for optimal transport.

Among them, E_1 and D_1 form an autoencoder to complete the first step of the manifold embedding task; E_2 is responsible for the second step of the optimal transport task, mapping the projection image to the latent space (L).

In these three networks, E_1 and E_2 have the same network structure, but obviously the weights are different. As a decoder, D_1 has a symmetrical structure with E_1. Their structure is based on IntroVAE [40] with some modifications including the change of the IntroVAE structure into an autoencoder and the use of Tanh at the end of decoder D_1 for activation and normalization. The specific network structures of Encoder E_1 and E_2 are illustrated in Fig. 3

Among them, BatchNorm is used to complete the normalization, and the activation function is LeakyReLU. Through the fully connected layer, the encoder outputs a vector of $1*512$, which is the latent code. The first step of NLOS-OT is to train the autoencoder E_1 and D_1 to get the transformation from target images to latent code; the second step is to train E_2 to get the code from projection images to latent code.

C. Loss Function of NLOS-OT

There are two steps in the training process, each of which has an independent loss function.

1) Step 1. Manifold Embedding Using Encoder E_1 and Decoder D_1: In step 1, the encoder E_1 and decoder D_1 are trained with the target image f as the input and output to

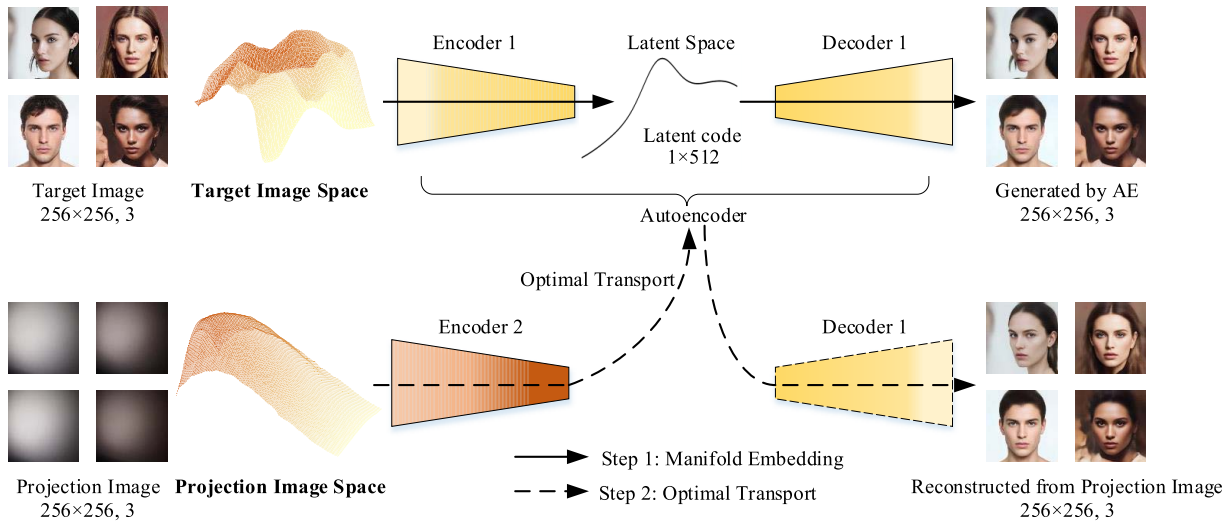


Fig. 2. Inspired by [48], NLOS-OT separates the manifold embedding and optimal transport of generative models. It first obtains the latent code from the target image space, and then completes the mapping from the projection image space to the latent space. The training phase is divided into two steps. The first step is to train the autoencoder composed of E_1 and D_1, and the second step is to train E_2 to get the optimal transport with fixed E_1 and D_1.

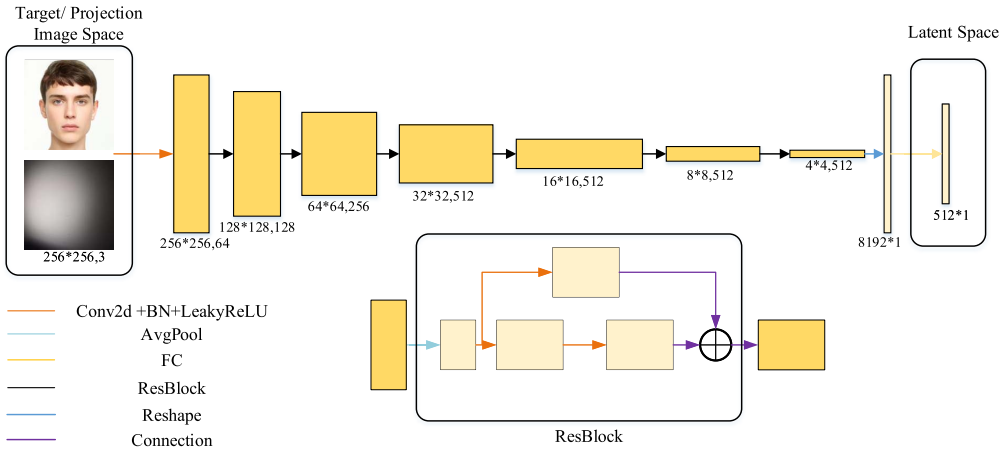


Fig. 3. **The network architecture of encoders in NLOS-OT.** The performance of the autoencoder determines the upper limit of NLOS-OT reconstruction, so the structure of the encoder cannot be too simple. Inspired by IntroVAE [40], we adopt this network structure, which can be used to encode multiple types of images.

obtain the latent code z. Here, E_1 represents the mapping from the target image space $\mathcal{T}$ to latent space $\mathcal{Z}$, while D_1 represents the mapping from the latent space $\mathcal{Z}$ to the target image space $\mathcal{T}$. The network architecture of the autoencoder is similar to those in IntroVAE [40] and PGGAN [55]. The size of the latent code is adjusted to 1×512 so that it is large enough to meet the requirements of the manifold embedding as well as small enough to reduce the difficulty of optimal transport. In addition, we add the activation function Tanh at the end of the decoder D_1 to make sure both input and output data are within [-1,1].

The objective of the manifold embedding is to find the latent code of the target image through optimizing E_1 and D_1. The corresponding loss function is divided into two parts, the pixel space distortion loss L_d and the feature space perceptual loss L_p

$$L_{AE} = \underbrace{\underbrace{L_p}_{\text{perceptual loss}} + \underbrace{\lambda \cdot L_d}_{\text{distortion loss}}}_{\text{loss of step1}} \tag{5}$$

Here, L_p is a simple MSE Loss but measured by the difference between f (the target hidden image) and $\hat{f}$ (the reconstructed hidden image) on VGG features [45], [56] as

$$L_p = \frac{1}{W \times H} \sum_{m=1}^{W} \sum_{n=1}^{H} \left(\phi_{3,3}(f)_{m,n} - \phi_{3,3}\left(\hat{f}\right)_{m,n} \right)^2$$

where $\phi_{3,3}(\cdot)$ represents the feature map obtained by $(3, 3)$ convolutional layer within the VGG19 network pretrained on ImageNet. W and H are the dimensions of feature maps. The L_d in Eq.(5) measures the distortion between target image f and $\hat{f}$ pixel-wisely using L_1-norm:

$$L_d = \left\| f - \hat{f} \right\|_1$$

Both the perceptual loss L_p and distortion loss L_d measure the difference between f and $\hat{f}$. We find that the perceptual loss L_p can well restore the high-level features and spatial structure, but cannot preserve the colors. Conversely, distortion loss L_d can restore the image color more accurately, but

with blurry results(See Appendix II in the supplementary for details).

2) Step 2. Optimal Transport Using Encoder E_2: In step 2, the encoder E_2 is trained with the parameters of E_1 and D_1 fixed. Here, E_2 is the mapping from the projection image space $\mathcal{P}$ to the latent space $\mathcal{Z}$. Ideally, given a projection image y, we hope that E_2 would map it to the latent code z that corresponds to the target image f. The network structure of E_2 is set to be identical with that of E_1.

The actual outputs we get from E_2 constitute another latent space $\hat{\mathcal{Z}}$. Calculating the distance loss between the two distributions $\mathcal{Z}$ and $\hat{\mathcal{Z}}$ and using the gradient descent method and backpropagation to force E_2 to encode the mapping from the projection image space $\mathcal{P}$ to latent space $\mathcal{Z}$ is the purpose of *step 2*.

To consider the geometry of the underlying spaces $\mathcal{Z}$ and $\hat{\mathcal{Z}}$, NLOS-OT employs the optimal transport (OT) metrics to assess the divergence between them. Let $\hat{\mu}$ and μ be two measures on latent spaces $\hat{\mathcal{Z}}$ and $\mathcal{Z}$ respectively, and $T : \hat{\mathcal{Z}} \rightarrow \mathcal{Z}$ is a measure preserving transport map, i.e. $\mu(B) = \hat{\mu}\left(T^{-1}(B)\right)$ for any μ-measurable set B, notated as $T_{\#\hat{\mu}} = \mu$. The cost function is denoted as $c : \hat{\mathcal{Z}} \times \mathcal{Z}$ where $c(\hat{z}, z)$ measures the cost to transport a unit mass from $\hat{z} \in \hat{\mathcal{Z}}$ to $z \in \mathcal{Z}$. The purpose of using OT in step 2 is to minimize the cost function c by training E_2, which is, according to Kantorovich's optimal transport theory [57], given $\hat{\mu} \in \mathcal{P}(\hat{\mathcal{Z}})$ and $\mu \in \mathcal{P}(\mathcal{Z})$

$$\min\ L_{OT} = \int_{\hat{\mathcal{Z}} \times \mathcal{Z}} c(\hat{z}, z) \mathrm{d}\pi(\hat{z}, z) \tag{6}$$

$$\text{s.t. } \pi(\hat{\mathcal{Z}}_i \times \mathcal{Z}) = \hat{\mu}(\hat{\mathcal{Z}}_i) \tag{7}$$

$$\pi(\mathcal{Z}_i \times \hat{\mathcal{Z}}) = \mu(\mathcal{Z}_i) \tag{8}$$

where $\hat{z} = E_{\theta_{E_2}}(y)$ is the output of E_2 with the input of projection image y. $\pi \in \mathcal{P}(\hat{\mathcal{Z}}, \mathcal{Z})$ is a measure that satisfies marginal constraints of $\hat{\mu}$ and μ in Eq.(7) and (8), i.e., the sum of mass removed from any measurable set $\hat{\mathcal{Z}}_i \in \hat{\mathcal{Z}}$ should be equal to $\hat{\mu}(\hat{\mathcal{Z}}_i)$, also same for $\mathcal{Z}$ and μ. Hence, $\pi(\hat{z}, z)$ is the amount of mass transported from $\hat{z}$ to z. Compared to map T used in Mong's OT theory [58] only transporting $\hat{z}$ to another z, $\pi(\hat{z}, z)$ allows mass in $\hat{z}$ to be mapped separately, which is a natural advantage.

In general, the cost function $c(\hat{z}, z)$ changes with z, such as $c(\hat{z}, z) = \|\hat{z} - z\|_1$ for original Monge's OT [58] and $c(\hat{z}, z) = \|\hat{z} - z\|_2$ for Wasserstein distance [34]. In step 2, however, because the indices of $\mathcal{Z}$ and $\hat{\mathcal{Z}}$ are already matched, i.e., $i - th$ latent code $\hat{z}_i \in \hat{\mathcal{Z}}$ and the $i - th$ latent code $z_i \in \mathcal{Z}$ are paired in each batch, the cost function $c(\hat{z}, z)$ can be simplified to a finite constant(e.g., 1) if and only if $\hat{z}$ and z have the same index and are generated by the same hidden image, else it would be $+\infty$. Additionally, for two latent codes $\hat{z}_i \in \hat{\mathcal{Z}}$ and $z_i \in \mathcal{Z}$, the amount of mass $\pi(\hat{z}, z)$ is set to the L_1-norm $\|\hat{z}_i - z_i\|_1$. Considering that z_i is known when $\hat{z}_i$ is given, $\pi(\hat{z}, z)$ conforms to the constraint in Eq.(7). In the same way, it also meets the constraint in Eq.(8). Therefore, we can optimize E_2 through gradient descent by minimizing L_{OT}.

The most crucial role of OT is to transform the original mapping from high-dimensional space to high-dimensional space into a mapping from high-dimensional space to low-dimensional embedded space. For the highly challenging task of passive NLOS imaging, other CNN-based end-to-end methods do not constrain the target space and directly complete the mapping from the projection images to the target images, making it difficult for the network to converge. In NLOS-OT, OT optimizes the mapping by measuring the difference between two low-dimensional vectors, thereby improving reconstruction quality. Following experiments show that according to the characteristics of the dataset, NLOS-OT can achieve extreme high-quality reconstruction or have strong generalization capabilities with reasonably good reconstruction.

IV. Experimental Results

In this section, we first introduce the experimental setup, including the used dataset NLOS-Passive, baseline methods and training details. After that, we evaluate the performance of NLOS-OT, in terms of reconstruction quality and generalization ability, through conducting experiments on the NLOS-Passive dataset. Specifically, we first fix the optical transport condition and evaluate the reconstruction quality of the NLOS-OT on different types of data. Then, we fix the type of data and evaluate how NLOS-OT works under various optical transport conditions. Finally, we train NLOS-OT with a broadly distributed dataset (STL-10) and test it under different conditions to evaluate the generalization ability and robustness to light conditions of NLOS-OT.

A. Experimental Setup

1) Dataset (NLOS-Passive): To effectively learn the encoders (E_1, E_2) and decoder (D_1) in NLOS-OT, we need a large-scale dataset. However, to our best knowledge, currently there is no existing large-scale passive NLOS imaging dataset available to use. Thus, we create a dataset, namely NLOS-Passive, with the experimental setting shown in Fig. 1(a), i.e., we display different target image f on the LCD and use a camera to capture the projection image y on the wall.

We use four different types of target images f, namely MNIST [19], the supermodel face dataset generated by StyleGAN [59], the animation face data DANBOORU2018 [21] and a widely distributed natural image dataset STL-10 [23].

Among them, MNIST, supermodel faces dataset and animation faces dataset are specific datasets with limited distribution, which is used to study the reconstruction effect of the model on special scenes. For each of the three target image datasets, we control four optical transport conditions: distance between the LCD and the wall D, the angle of the camera $\angle\alpha$, the ambient illumination L, and the material of the relay surface. Each of the four conditions has 2 values. Please see Fig. S4 in Appendix III in the supplementary for details. Therefore, for each of the three types of data (MNIST, supermodel faces, animation faces), we collected 16 groups respectively. These 16 sets of data are all without partial occlusion. It should be noted that in order to verify the influence of occlusions on the reconstruction methods, we collected a group of data for supermodel faces with partial occlusion as well, described in Appendix III in the supplementary for detail.

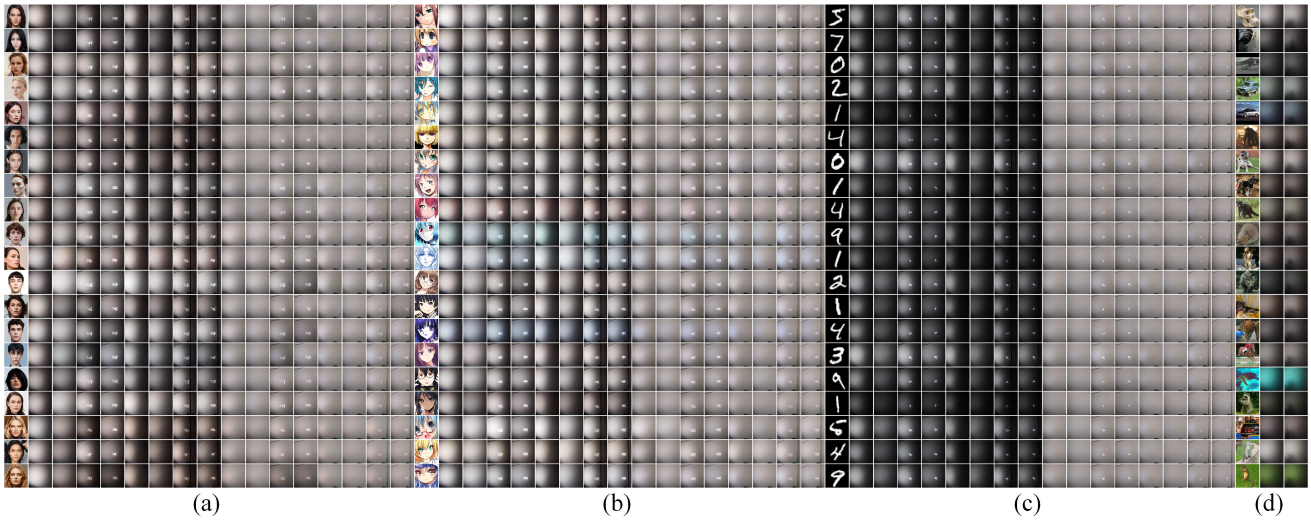


Fig. 4. **Samples from NLOS-Passive.** It contains 4 different kinds of target images, including **(a)** supermodel faces, **(b)** anime faces, **(c)** MNIST and **(d)** STL-10. (a)- (c) are narrow-distributed specific datasets. The 16 columns behind each target image represent projection images collected under 16 different optical transport conditions. (d) are broad-distributed data. The 2 columns behind each target image represent projection images collected with or without unknown partial occlusion. NLOS-Passive has a total of more than 50 groups and $3,200,000$ samples.

STL-10 is used to analyze the performance and generalization ability of the model on a broad dataset. Specifically, we collected two groups of STL-10 data without occlusion to verify the performance of NLOS-OT on a broader dataset. Besides, we also collected two groups of STL-10 data as well as some samples from other datasets with partial occlusion to demonstrate the generalization ability. Each group of STL-10 data contains $113,000$ samples.

During the acquisition process, various camera parameters, including white balance, focal length, exposure time, etc., remain unchanged to prevent changes in camera parameters from affecting the data. We totally collect more than 50 groups of data and 3,200,000 samples, as shown in Fig. 4. The details of NLOS-Passive and the hardware we used are covered in Appendix III and IV in the supplementary respectively.

2) Baseline Methods: Since there are only a few works using deep learning for the passive NLOS imaging [12]–[15], in this paper, we compare the proposed NLOS-OT with the following two baseline methods.

a) U-Net used in Phong [12]: The first baseline is the U-Net structure used in [12], [15].

b) Conditional GANs: U-Net can be regarded as an encoder-decoder based model, representing the state-of-the-art passive NLOS imaging methods. However, in recent years, GANs have also achieved superior results in areas including image deblurring and denoising. Thus, we also compare with GANs to verify the performance of NLOS-OT. Since there is no passive NLOS imaging method based on GANs, we establish a C-GAN reconstruction network according to Pix2Pix GAN [44] and DeblurGAN [45] as a baseline. See Appendix I in the supplementary for more details.

3) Training Details: All three methods, including two baselines and the proposed NLOS-OT, are implemented using PyTorch [60]. The training is performed on a single GeForce GTX 1080 Ti. The captured projection image has the size of 720×720, but all input images and output images are resized to 256×256.

The training process of NLOS-OT is divided into two steps, manifold embedding and optimal transport. In the first step, the network is an autoencoder containing only E_1 and D_1. The input and output of the network are both target images f. The learning rate is 0.0001 in the first 30 epochs, and exponentially decays to 0 in the subsequent epochs. We also adopt an early stopping strategy to prevent over-fitting and reduce the difficulty of optimal transport in step 2. Specifically, with the increase of training epoch in step 1, although the reconstruction result of step 1 is more realistic, it becomes harder to map the projection image space $\mathcal{P}$ to the latent space $\mathcal{Z}$ in step 2. We thus choose 100 epochs for MNIST and supermodel faces data, and 10 epochs for anime faces data and STL-10 data considering the trade-off between the accuracy of manifold embedding and the difficulty of step 2.

In the OT step, E_2 is added to the network, and the parameters of E_1 and D_1 are fixed based on step 1. In the training process of step 2, the input includes projection images y and target images f. E_1 extracts the features from f to get the latent code z, while E_2 is continuously optimized to complete the optimal mapping from y to the latent code $\hat{z}$.

In the testing phase, the network only has E_2 and D_1 where E_2 maps the projection images f to the latent code $\hat{z}$ and D_1 decodes $\hat{z}$ to reconstruct the hidden images.

B. Reconstruction of Complex Hidden Scenes

Existing passive NLOS imaging algorithms can only recover simple hidden scenes with blurry reconstruction. The proposed NLOS-OT reduces the reconstruction difficulty by simplifying the generation target from the image space $\mathcal{T}$ to the latent space $\mathcal{Z}$, and thus can be applied to more complex hidden scenes with good reconstruction quality.

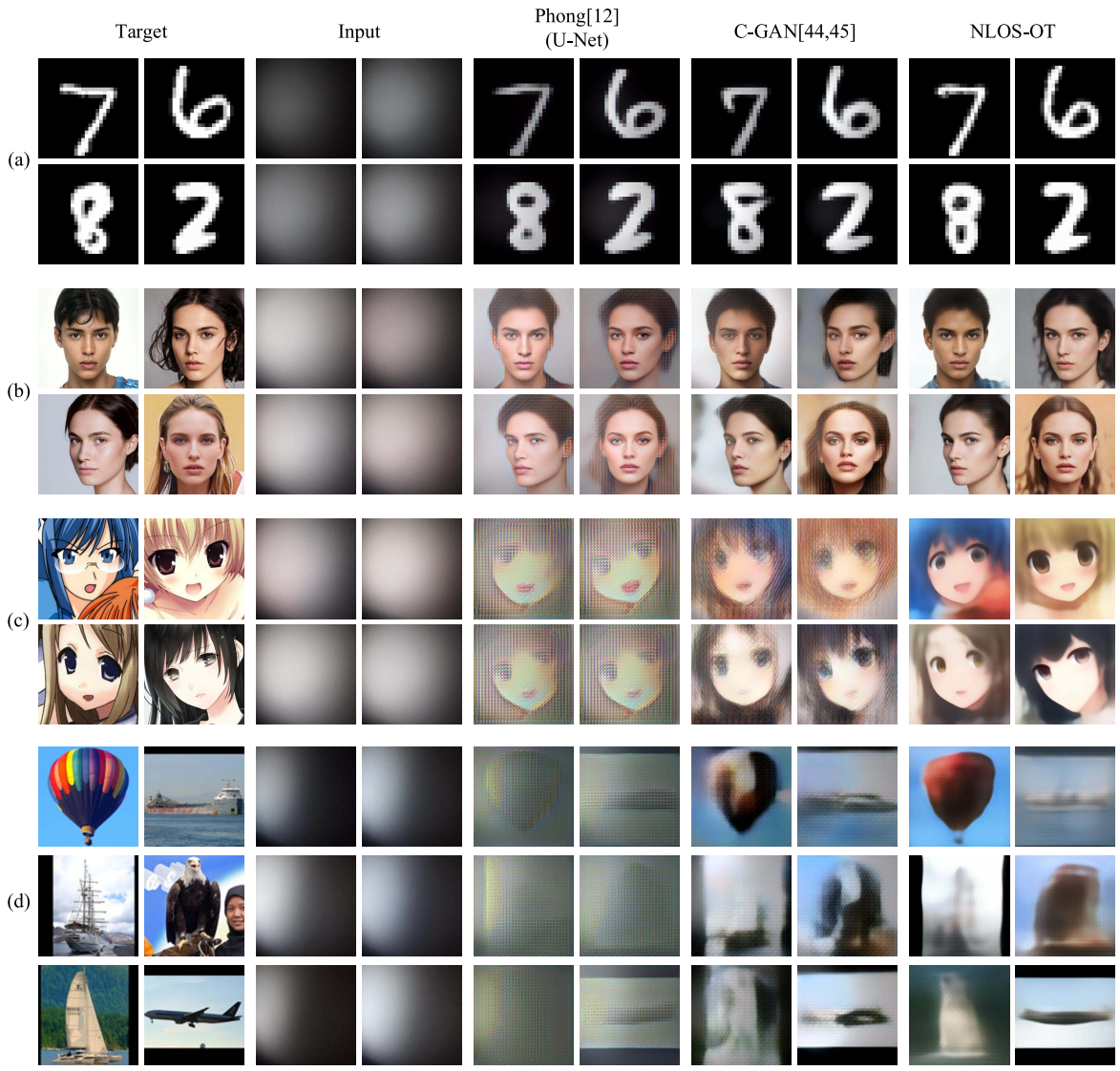


Fig. 5. **Results under different kinds of hidden images.** The reconstruction results of different methods when the target image is from MNIST **(a)**, supermodel faces **(b)**, anime faces **(c)** and STL-10 **(d)**. All data are collected on the wall under $D = 100$ cm, angle $\angle\alpha_1$ and dark light.

By keeping the distance D, the angle $\angle\alpha$, the illumination L and BRDFs unchanged, we fix the optical transport conditions to train and test on the four datasets respectively, as shown in Fig. 5. Among the four types of data, MNIST is the simplest so that U-Net, C-GAN and NLOS-OT can all achieve good reconstruction. When the data becomes more complex, the advantages of NLOS-OT become more apparent. For supermodel faces, the SSIM of the reconstruction obtained by NLOS-OT is higher than those obtained by the baseline methods. When the data is the anime face and STL-10 with more patterns, only NLOS-OT can obtain acceptable results, indicating that the proposed NLOS-OT has the superior ability on the passive NLOS imaging for complex hidden scenes.

Table I compares the quantitative results of NLOS-OT and the baseline methods. It can be seen that NLOS-OT achieves the best performance in terms of both SSIM [61] and PSNR, which verifies that by encoding the target image to latent code and then performing the optimal transport in the latent space, NLOS-OT can greatly improve the reconstruction of passive NLOS imaging. The SSIM and PSNR of traditional model-based methods are about 0.28 and 7.3dB respectively [9], [10], which is also exceeded by the proposed NLOS-OT.

TABLE I

QUANTITATIVE COMPARISON WITH SSIM AND PSNR. THE STATISTICAL RESULTS OF DIFFERENT METHODS ON THREE DATASETS, WHICH ARE OBTAINED UNDER TYPICAL OPTICAL TRANSPORT CONDITION (REFLECTED ON THE WALL, D = 100cm, ANGLE $\angle\alpha_1$, DARK LIGHT, WITHOUT OCCLUSION)

Method	MNIST		SuperModel Faces		STL-10	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Phong [12]	0.67	14.52dB	0.33	13.32dB	0.04	10.31dB
C-GAN	0.80	14.83dB	0.54	14.87dB	0.34	13.59dB
ours	**0.83**	**15.27dB**	**0.59**	**16.04dB**	**0.46**	**15.41dB**

Fig. 6. **Results under different optical transport conditions.** With a specified target image, **(a)** illustrate the results of different methods under the optical transport condition: dart environment, distance $D = 100$, angle $\angle\alpha_1$, and the wall reflection. **(b)-(e)** represent the reconstruction results obtained by different methods after changing one of the four optical transport conditions in (a), respectively.

C. Reconstruction Under Different Optical Transport Conditions

Existing passive NLOS imaging methods mainly work in a dark environment, and may fail to achieve good reconstruction with more ambient light. In addition to ambient light, other optical transport conditions can also affect the NLOS reconstruction. Therefore, besides the ability to reconstruct complex hidden scenes, the robustness to different optical transport conditions is also an important criterion.

To evaluate the robustness of the NLOS-OT under different optical transport conditions, we first select a certain target image set (e.g., the SuperModel Faces) and then determine a combination of optical transport conditions. The results are shown in Fig. 6, where (a) are the reconstructions under the optical transport condition: dark light environment, distance D $=$ 100, angle $\angle\alpha_1$, and the wall reflection. By changing one of the four optical transport conditions in (a), (b)-(e) represent the corresponding reconstruction results obtained by different methods. "Day Light" here refers to the measurement conditions with obvious ambient light. After calculation, the "day light" data is equivalent to adding 0.05 dB (SNR) shot noise that obeys the ambient light distribution to the dark light projection.

It can be seen that under different optical transport conditions, the reconstruction can be different for the same method. Overall, low illumination and high reflection relay surface have greater impact on the reconstructions, while the distance D and the angle $\angle\alpha$ have less influence on the results. In all cases, the proposed NLOS-OT achieves the best reconstruction under different conditions. These results show that the proposed NLOS-OT is robust to optical transport conditions.

D. Assessment of Generalization Ability

1) Generalization of Unseen Test Data: Generalization ability is an essential factor needed to be considered for the learning-based methods due to the implicit assumption that the distribution of the test data is similar to that of the training data.

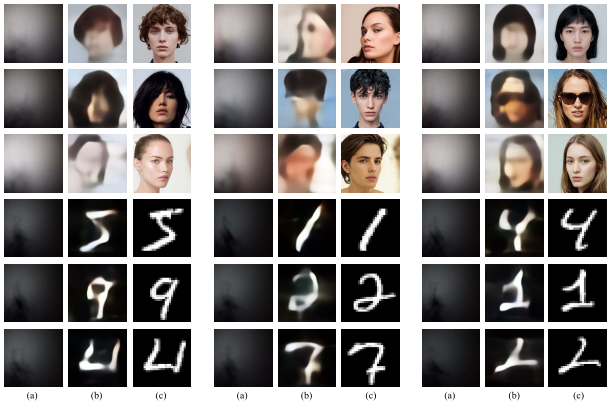


Fig. 7. **Generalization of NLOS-OT model.** When the dataset is complex, NLOS-OT has good generalization ability. The training set is STL-10, and the test set is MNIST and supermodel dataset. Left: input, middle: output, right: ground truth.

For passive NLOS imaging tasks, there are mainly three types of knowledge that NLOS-OT can learn: inverse transport process, noise distribution and data prior. The generalization ability will be poor when learning too much data prior, which is not expected. When the distribution of the training dataset is sufficiently broad, the difficulty of learning the data prior will increase, forcing the network to learn more inverse light transport knowledge, thus having a stronger generalization ability.

Therefore, we collected STL-10 data with an unknown partial occlusion (randomly placing a tripod in front of the screen, described in Appendix V in the supplementary) as the training set. On the other hand, some MNIST and supermodel face data were collected as the test set, keeping all conditions unchanged. Then, the generalization ability can be evaluated by training on the STL-10 dataset, and testing on the MNIST and supermodel face dataset. The results are shown in Fig. 7,

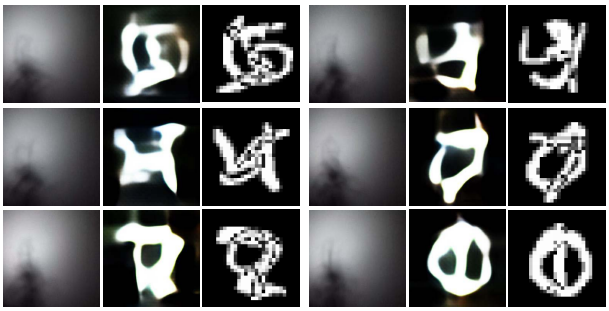


Fig. 8. **The results of adding two test images together as the input of the network.** Left: input image. Middle: output image. Right: ground truth. The network is trained by the projections of STL-10 dataset with an unknown occluder.

TABLE II

QUANTITATIVE COMPARISON BETWEEN NLOS-OT AND TRADITIONAL ALGORITHMS. ALL NLOS-OT DATA ARE COLLECTED UNDER (DARK, ANGLE B, D = 100, WALL CONDITIONS, AND ONLY TRAINED ON STL-10). QUANTITATIVE RESULTS OF TRADITIONAL ALGORITHMS ALL COME FROM [10]

	Test set / Methods	PSNR (dB)	SSIM
NLOS-OT Only trained by STL-10	STL-10	17.46	0.507
	MNIST	15.14	0.229
	Supermodel Faces	17.35	0.581
	Real Images	14.30	0.509
Traditional Methods	TV regularization	8.8	0.37
	Polarized NLOS	11.7	0.43

illustrating that NLOS-OT can be trained on STL-10, and tested on the MNIST and supermodel face datasets with good results. Besides, we also find that when two tested input images are added as the new input, the output is almost the added result of the corresponding target images (as shown in Fig. 8). Combined with Eq. 3, it can be seen that NLOS-OT has almost completed the theoretical linear transformation.

All the results show that when the training dataset is appropriate, NLOS-OT can indeed learn something very similar to the inverse optical transport matrix and have strong generalization capabilities, which also demonstrates that the projection data in Fig. 7-(a) does contain enough information to restore almost any hidden scene. Moreover, the performance is better than the previous methods. For more information about the generalization ability experiment, including the test results on the realistic pictures, and the comparison of the generalization ability with baseline methods [12], [45], please see Appendix VI in the supplementary.

2) Comparison With Traditional Physics-Based Algorithms Using Unseen Test Data: Besides using different types of data (MNIST and supermodel faces) in NLOS-Passive for generalization ability verification, we also used images from existing physics-based works to illustrate the generalization ability of NLOS-OT. Specifically, we used our screen to play the hidden images in [9] and [10], then tested them under our light transport settings (wall, D = 100cm, angle $\angle\alpha_1$, dark light). The test result is compared with the result in [9] and [10], as shown in Fig. 9.

It can be seen that for the hidden image in [9], the results obtained by NLOS-OT are not as good as [9]. Nevertheless, NLOS-OT does not need to measure the parameters of the scene and perform scaling calibration on different color channels, and can achieve better results when the target scene is similar to the training images. For (e)-(g) in Fig. 9, NLOS-OT can achieve similar and even better results than traditional algorithms in [10]. In fact, from quantitative comparison, the results of our method on the new test data are better than the traditional algorithm (the quantitative results of the traditional algorithms come from [10]), as shown in Tab. II. This can be explained by NLOS-OT's ability of not only learning the optical transport matrix, but also learning the data prior.

All these experiments can show that NLOS-OT has strong generalization capabilities and can be used for new distribution of images that have never been seen before.

3) The Adaptability of NLOS-OT to Different Data Collection Conditions: What is described above is the classic generalization ability, that is, the ability of NLOS-OT to adapt to different distributed data, which is essentially the ability of NLOS-OT to learn the inverse light transport process. Next, we will discuss another kind of "generalization ability", which measures the adaptability of NLOS-OT to different data collection conditions, and is essentially the ability of NLOS-OT to learn data prior. We believe that although in the usual sense, learning physical mapping is more important than learning data prior. However, for specific NLOS imaging tasks, it is also important to learn data priors to achieve high reconstruction quality and be applied to different light transport conditions. Specifically, we choose a narrow dataset (supermodel faces) and mix the data under 16 different optical transport conditions to form a new dataset for the training and testing. The results are shown in Fig. 12. It can be seen that NLOS-OT still performs best on this mixed dataset, which means that NLOS-OT has better "generalization ability" on different light conditions than existing methods based on U-Net [12], [15] and C-GAN.

The results in Fig. 7 and Fig. 12 illustrate the ability of NLOS-OT to learn data prior and to learn the process of inverse light transport, respectively. This shows that NLOS-OT can learn the inverse light transport process for a widely distributed dataset to obtain strong generalization ability, instead of difficult to converge as other existing end-to-end methods. Furthermore, NLOS-OT can get higher quality results and generalization ability to different light conditions for a specific dataset by learning the data prior.

E. Assessment of Robustness to Noise

As described in Section III, the noise of passive NLOS imaging is mainly background noise caused by different ambient light levels, which is shot noise based on the distribution of ambient light. Besides, some Gaussian noise would be added due to the camera. Here, we use the wall projection data collected under a dark environment, $D = 100$ and angle $\angle\alpha_1$ to train the network, and add these two types of noises at

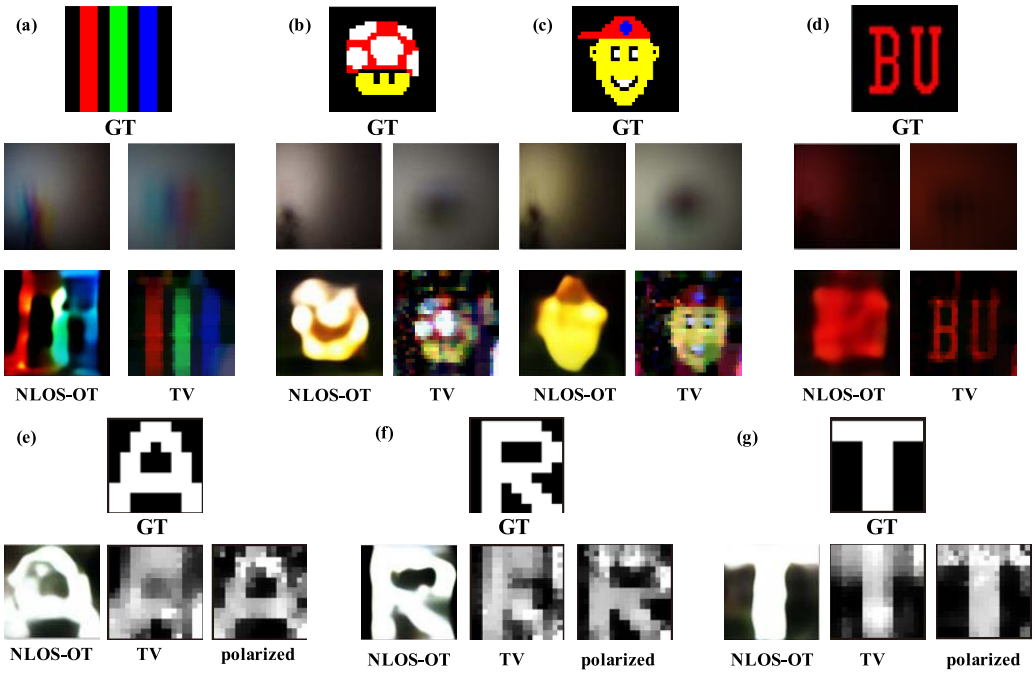


Fig. 9. **Comparison of NLOS-OT and traditional algorithms [9] and [10]**. The TV regularization results of (a)-(d) are from [9], and the traditional algorithms results of (e)-(g) are from [10]. NLOS-OT uses STL-10 data for training, and the collected projection data for testing.

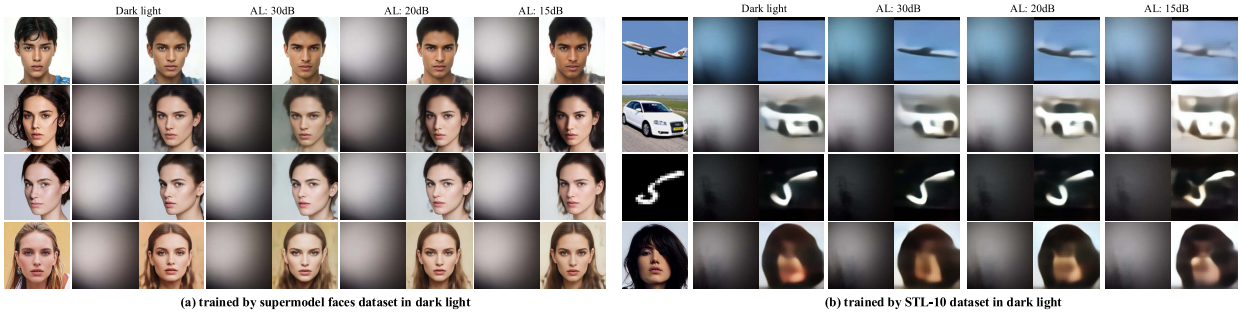


Fig. 10. **The test results on different levels of noise due to ambient light (AL)**. All the results in this figure are trained under dark light and tested after adding ambient light noise with different SNRs. **(a)** Test results on small-scale dataset. **(b)** Test results on large-scale dataset. Both (a) and (b) are tested under three different ambient noise levels (30dB, 20dB and 15 dB).

different levels to the projection images separately to illustrate the robustness of NLOS-OT to noise.

1) Shot Noise Caused by Ambient Light: We exploited a stable light source with adjustable brightness as the noise source, and collected test data under different SNRs (30dB, 20dB, 15dB). After that, we separately evaluated the robustness of the smaller-scale dataset (supermodel faces) and the larger-scale dataset (STL-10) to ambient light. The results are shown in Figure 10-(a) and (b) respectively.

It can be seen from Fig. 10 that when the noise continues to increase, the distribution of the input images will change significantly, resulting in a sharp drop in the quality of the reconstructed images. When the scale of the training set is small, the results under different SNRs change more significantly. On the contrary, when using a larger dataset, the ambient light noise has less influence on the reconstructions, which means a stronger robustness. For example, when using the STL-10 dataset for training, the number "5" in MNIST can be clearly recovered even at an SNR of 15dB. In terms of visual effects, the tolerance of NLOS-OT to noise generated by ambient light is about 10dB.

2) Gaussian Noise: To analyze the robustness of NLOS-OT to Gaussian noise, we added 10dB, 20dB, and 30dB Gaussian noise to each input image to test the reconstruction effect of the network on these noisy images.

Figure 11 shows the reconstruction results under different SNRs. It can be seen that NLOS-OT has strong robustness to noise. Considering that NLOS imaging is a very ill-conditioned problem, this robustness to noise is largely due to the contribution of the scene prior to NLOS-OT. With the noise around 10dB, NLOS-OT began to produce completely different reconstruction results.

Apart from noise, we also evaluated the robustness of NLOS-OT to other factors, including the distance from the display to the wall D, the camera angle, and the relay surface material. The results show that NLOS-OT has good robustness

Fig. 11. **Reconstruction results under Gaussian noise with different SNRs.**

to the distance D and relay surface materials, but is less robust to the camera angle and the position of the occluder. Please see Appendix VI in the supplementary for details.

V. Discussions

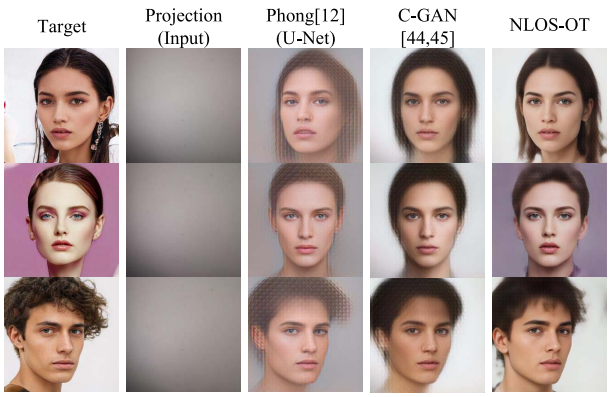
In this section, we first discuss what NLOS-OT has learned, the contribution to passive NLOS imaging tasks, and the challenge faced by NLOS-OT, then analyze the limitations of NLOS-Passive and our future work.

A. Discussion About Our Method: NLOS-OT

1) Transport Matrix vs Data Prior: As mentioned above, there are two main types of knowledge that can be exploited in passive NLOS imaging tasks: the inverse light transport process and data prior (including noise distribution). Traditional physics-based methods mainly exploit the light transport process by approximating the light transport matrix and using other ways (e.g., adding partial occlusions [9], [62], [63] and polarizers [10]) to decrease the condition number, hence obtaining strong generalization capabilities (not dependent on data distribution) but low-quality results. On the contrary, the existing end-to-end deep learning methods [12] mainly learn the data distribution of the data set (MNIST), hence obtaining higher quality but poor generalization ability results.

2) What Does NLOS-OT Learn: To some extent, NLOS-OT combines the strengths of the above two methods. That is, NLOS-OT can learn both the data prior and the inverse light transport process. For datasets with simple distribution, physics-based methods cannot effectively extract data priors, but NLOS-OT can effectively extract them and complete the state-of-the-art, high-quality reconstructions. On the other hand, for datasets with complex distributions, the existing deep learning-based methods are difficult to converge. Still, NLOS-OT is effective and learn the inverse transport process so that the generalization ability is equivalent to the traditional methods (any hidden scenes can be reconstructed), but with faster reconstruction speed (neural network only needs forward propagation during testing) and better quality (still learning a small amount of data prior) results.

Which one (data prior or inverse transport) is the predominant knowledge learned by NLOS-OT depends on the distribution of the data set. When the dataset is relatively small, the learning experience is very conducive to the decline of the loss function, and the network will tend to learn experience at this time. However, when the dataset is complex, the loss function will be challenging to decline through the learning experience. On the contrary, learning effective physical transformations will become a more efficient way. Hence, the network will learn the physical transformation behind passive NLOS imaging.



	Phong[12]	C-GAN	NLOS-OT
PSNR	12.87dB	14.85dB	**18.57dB**
SSIM	0.333	0.577	**0.658**

Fig. 12. **The adaptability of NLOS-OT to different optical transport conditions.** When all the data under different optical transport conditions are mixed, the proposed NLOS-OT has the best reconstruction, reflecting the better generalization ability to different light conditions.

To verify the conclusion, we conduct two experiments. In the first experiment, we keep the ambient light, angle, distance, and relay surface material unchanged to collect new data with partial occlusion. The results with and without occlusion are shown in Figure 13-(b) and (c), and the comparison of their loss functions on the validation set is shown in Figure 13-(d). In quantitative comparison, with occlusion, SSIM is **0.674**, while PSNR is **19.37dB**, both of which are higher than that without occlusion. Thus, it can be seen that NLOS-OT has indeed learned the effective physical mapping in passive NLOS even with narrow datasets. Another experiment fed the STL-10 dataset in Fig. 7 to NLOS-OT for training, and used the data measured in [9] for testing. Since what the NLOS-OT learn is mainly the optical transport matrix in this situation, and the STL-10 dataset we collected has a completely different optical transport matrix from the data in [9], the reconstruction cannot be completed theoretically, which is consistent with the experiment results. Please see Appendix V - (4) in the supplementary for details.

3) Challenges: Nevertheless, NLOS-OT still faces many limitations. First, due to the "black box" nature and the gradient backpropagation mechanism, the network will automatically learn knowledge that is easy to learn, making it extremely difficult to use small-scale datasets to obtain strong generalization results. In addition, because of the ill-posedness of passive NLOS imaging, it is tough to obtain extremely high-quality results on complex datasets. We believe two possible ways can alleviate these problems. The first is through transfer learning, enabling the data trained in one type of hidden scene to be quickly transferred to another type. The second is to

Fig. 13. **The results of with and without partial occlusion on supermodel face reconstruction. (a)** Hidden images. **(b)** Input and reconstruction results when there is no occlusion. **(c)** Input and reconstruction results when there is a partial occluder. In terms of quantitative comparison, the PSNRs of with and without occlusion are 16.04dB and 19.37dB, and SSIMs are 0.597 and 0.674, respectively. **(d)** Loss functions with/without partial occlusion. This shows that partial occlusion is helpful to NLOS-OT, proving that NLOS-OT has learned relevant knowledge about inverse light transport.

combine with model-based methods to force the network to learn only the information of the optical transport matrix, e.g., using matrix factorization to simulate the optical transport matrix [13]. We leave these to our future studies.

B. Discussion About the Dataset: NLOS-Passive

In NLOS imaging, due to many unavailable calibration processes, collecting data, especially a large-scale dataset, is very difficult. Therefore, most existing works generate simulated datasets through rendering models. In this work, we use different hidden images to collect data under different optical transport conditions, and made the NLOS-Passive, including more than 50 groups and 3,200,000 samples in total.

We believe that NLOS-Passive can be useful in various aspects. First, since it contains hidden scenes of different complexity, NLOS-Passive can be used to study the performance of passive NLOS imaging algorithms, whether conventional or data-driven methods. Secondly, with data captured under different light conditions, NLOS-Passive can be used to study the influence of lighting on the reconstruction algorithm and the optical transport matrix. Last but not least, NLOS-Passive is an experimentally collected dataset, so it can be used to train the optical transport process to help optimize the existing imaging model.

Despite its many potential applications, NLOS-Passive also has its limitations. First, because different optical transport conditions need to be manually controlled, we only collect data under limited optical conditions, as well as only use two relay materials, wall and whiteboard, which led to NLOS-Passive cannot represent enough data space. This limitation can be untangled by using NLOS-Passive as a constraint to improve the data rendering model. Besides, using RAW data is also an approach to enhance NLOS-Passive. They are left to our future studies.

VI. Conclusion

In this paper, we developed NLOS-OT, which enables passive NLOS imaging in complex scenes through manifold embedding and optimal transport. In addition, we created NLOS-Passive, a large-scale passive NLOS dataset totally containing more than 50 groups of data and 3,200,000 samples, which is, to the best of our knowledge, the first public large-scale passive NLOS dataset.

The proposed NLOS-OT resolves the unbalanced distribution problem by first performing manifold embedding to obtain the latent space, and then using optimal transport to map the projection image to the latent space. Such procedures greatly simplify the imaging since the dimension of the latent space is much lower than that of the target image space, as shown in Tab. I. We also show that NLOS-OT has the ability to reconstruct any hidden images, which means that NLOS-OT has a strong generalization ability. We anticipate that the NLOS-OT framework together with the NLOS-Passive dataset will accelerate the development of learning-based passive NLOS imaging research, and thereby enable real-time and high-quality passive NLOS imaging in the near future.

References

[1] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten, "Non-line-of-sight imaging using a time-gated single photon avalanche diode," *Opt. Exp.*, vol. 23, no. 16, p. 20997, 2015.
[2] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar, "Looking around the corner using ultrafast transient imaging," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 13–28, Oct. 2011.
[3] X. Liu *et al.*, "Non-line-of-sight imaging using phasor-field virtual wave optics," *Nature*, vol. 572, no. 7771, pp. 620–623, Aug. 2019.
[4] M. La Manna, F. Kine, E. Breitbach, J. Jackson, T. Sultan, and A. Velten, "Error backprojection algorithms for non-line-of-sight imaging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1615–1626, Jul. 2019.
[5] M. O'Toole, D. B. Lindell, and G. Wetzstein, "Confocal non-line-of-sight imaging based on the light-cone transform," *Nature*, vol. 555, no. 7696, pp. 338–341, Mar. 2018.
[6] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, "Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging," *Nature Commun.*, vol. 3, no. 1, pp. 1–8, Jan. 2012.
[7] S. Xin, S. Nousias, K. N. Kutulakos, A. C. Sankaranarayanan, S. G. Narasimhan, and I. Gkioulekas, "A theory of fermat paths for non-line-of-sight shape reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6800–6809.
[8] M. Batarseh, S. Sukhov, Z. Shen, H. Gemar, R. Rezvani, and A. Dogariu, "Passive sensing around the corner using spatial coherence," *Nature Commun.*, vol. 9, no. 1, pp. 3629–3634, Dec. 2018.
[9] C. Saunders, J. Murray-Bruce, and V. K. Goyal, "Computational periscopy with an ordinary digital camera," *Nature*, vol. 565, no. 7740, pp. 472–475, Jan. 2019.
[10] K. Tanaka, Y. Mukaigawa, and A. Kadambi, "Polarized non-line-of-sight imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2136–2145.

[11] A. B. Yedidia, M. Baradad, C. Thrampoulidis, W. T. Freeman, and G. W. Wornell, "Using unknown occluders to recover hidden scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12231–12239.

[12] C. Zhou, C.-Y. Wang, and Z. Liu, "Non-line-of-sight imaging off a phong surface through deep learning," 2020, *arXiv:2005.00007*.

[13] M. Aittala *et al.*, "Computational mirrors: Blind inverse light transport by deep matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14311–14321.

[14] M. Tancik, G. Satat, and R. Raskar, "Flash photography for data-driven hidden scene recovery," *CoRR*, vol. abs/1810.11710, pp. 1–11, Oct. 2018.

[15] T. Yu, M. Qiao, H. Liu, and S. Han, "Non-line-of-sight imaging through deep learning," *Acta Optica Sinica*, vol. 39, no. 7, 2019, Art. no. 0711002.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[17] B. T. Phong, "Illumination for computer generated pictures," *Commun. ACM*, vol. 18, no. 6, pp. 311–317, Jun. 1975.

[18] G. Peyré *et al.*, "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[20] S. Guo. (Dec. 2019). *Generated Seeprettyface Dataset*. Accessed: Sep. 24, 2020. [Online]. Available: https://github.com/a312863063/seeprettyface-generator-model

[21] Anonymous, D. Community, and G. Branwen. (2020). *Danbooru2019: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. [Online]. Available: https://www.gwern.net/Danbooru2019

[22] (Dec. 2019). *Anime Faces Dataset*. Accessed: Sep. 24, 2020. [Online]. Available: http://www.seeprettyface.com/mydataset_page3.html#anime

[23] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.

[24] A. Beckus, A. Tamasan, and G. K. Atia, "Multi-modal non-line-of-sight passive imaging," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3372–3382, Jul. 2019.

[25] A. Torralba and W. T. Freeman, "Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 374–381.

[26] C. Saunders, R. Bose, J. Murray-Bruce, and V. K. Goyal, "Multi-depth computational periscopy with an ordinary camera," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 9299–9305.

[27] S. W. Seidel, J. Murray-Bruce, Y. Ma, C. Yu, W. T. Freeman, and V. K Goyal, "Two-dimensional non-line-of-sight scene estimation from a single edge occluder," 2020, *arXiv:2006.09241*.

[28] J. Boger-Lombard and O. Katz, "Passive optical time-of-flight for non line-of-sight localization," *Nature Commun.*, vol. 10, no. 1, pp. 3343–3351, Dec. 2019.

[29] K. L. Bouman *et al.*, "Turning corners into cameras: Principles and methods," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2270–2278.

[30] O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," *Nature Photon.*, vol. 8, no. 10, pp. 784–790, Oct. 2014.

[31] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

[32] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.

[33] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[35] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, "Stochastic optimization for large-scale optimal transport," in *Proc. Adv. neural Inf. Process. Syst.*, 2016, pp. 3440–3448.

[36] N. Lei, K. Su, L. Cui, S.-T. Yau, and X. D. Gu, "A geometric view of optimal transportation and generative model," *Comput. Aided Geometric Design*, vol. 68, pp. 1–21, Jan. 2019.

[37] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.

[38] R. Webster, J. Rabin, L. Simon, and F. Jurie, "Detecting overfitting of deep generative networks via latent recovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11273–11282.

[39] Z. Ding *et al.*, "Guided variational autoencoder for disentanglement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7920–7929.

[40] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan, "IntroVAE: Introspective variational autoencoders for photographic image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 52–63.

[41] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[42] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19667–19679.

[43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[45] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[46] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[47] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1300–1309.

[48] D. An, Y. Guo, N. Lei, Z. Luo, S.-T. Yau, and X. Gu, "AE-OT: A new generative model based on extended semidiscrete optimal transport," in *Proc. ICLR*, 2020.

[49] M. G. Kuhn, "Optical time-domain eavesdropping risks of CRT displays," in *Proc. IEEE Symp. Secur. Privacy*, May 2002, pp. 3–18.

[50] M. G. Kuhn, "Compromising emanations: Eavesdropping risks of computer displays," Ph.D. dissertation, Comput. Lab., Univ. Cambridge, Cambridge, U.K., 2002.

[51] M. Backes, M. Dürmuth, and D. Unruh, "Compromising reflections-or-how to read LCD monitors around the corner," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 158–169.

[52] Y. Xu, J. Heinly, A. M. White, F. Monrose, and J.-M. Frahm, "Seeing double: Reconstructing obscured typed input from repeated compromising reflections," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2013, pp. 1063–1074.

[53] S. Chakraborty, W. Ouyang, and M. Srivastava, "LightSpy: Optical eavesdropping on displays using light sensors on mobile devices," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 2980–2989.

[54] T. Maeda, Y. Wang, R. Raskar, and A. Kadambi, "Thermal non-line-of-sight imaging," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2019, pp. 1052–1061.

[55] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[57] L. Kantorovich, "On the translocation of masses," in *Proc. Dokl. AN SSSR*, vol. 37, 1942, pp. 199–201.

[58] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, pp. 666–704, 1781.

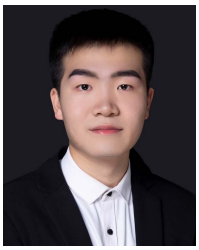
[59] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[60] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[62] J. Murray-Bruce, C. Saunders, and V. K. Goyal, "Occlusion-based computational periscopy with consumer cameras," in *Proc. Wavelets Sparsity*, 2019, Art. no. 111380X.

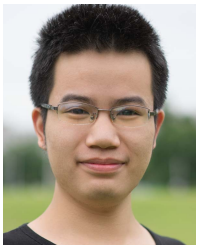
[63] J. Rapp *et al.*, "Seeing around corners with edge-resolved transient imaging," 2020, *arXiv:2002.07118*.



Ruixu Geng (Graduate Student Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, in 2019, where he is currently pursuing the M.S. degree with the School of Information and Communication Engineering. His research interests include computational imaging and computer vision.

Yang Hu received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. She was with the University of Maryland Institute for Advanced Computer Studies as a Research Associate from 2010 to 2015. She is currently an Associate Professor with the School of Information Science and Technology, University of Science and Technology of China. Her current research interests include computer vision, machine learning, and multimedia signal processing.

Zhi Lu (Graduate Student Member, IEEE) received the B.S. and M.A. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. His research interests are in data mining and multimedia.

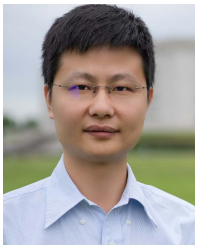


Cong Yu (Graduate Student Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering. His current research interests include wireless sensing, image synthesis, and adversarial learning.

Houqiang Li (Fellow, IEEE) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively.

He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He has authored and coauthored over 200 papers in journals and conferences. His research interests include image/video coding, image/video analysis, computer vision, and reinforcement learning. He is the Winner of National Science Funds (NSFC) for Distinguished Young Scientists, the Distinguished Professor of Changjiang Scholars Program of China, and the Leading Scientist of Ten Thousand Talent Program of China. He was a recipient of the National Technological Invention Award of China (second class) in 2019 and the National Natural Science Award of China (second class) in 2015. He was also a recipient of the Best Paper Award for VCIP 2012, ICIMCS 2012, and ACM MUM in 2011. He served as the General Co-Chair for ICME 2021 and the TPC Co-Chair for VCIP 2010. He is an Associate Editor (AE) of IEEE TRANSACTIONS ON MULTIMEDIA, and served as an AE for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013.

Hengyu Zhang received the master's and Ph.D. degrees from the West China Medicine School, Sichuan University, in 1996 and 2010, respectively. He was a Visiting Scientist with the School of Medicine, National University of Singapore, from February 2008 to February 2009. He has been the Director of the West China Syncope Center since 2021, where medical images are heavily involved. He is currently a Professor with the school and the Deputy Chief Physician of the West China Hospital, engaging cardiac pacing, electrophysiology, and pacemaker implantation with international leading level.



Yan Chen (Senior Member, IEEE) received the bachelor's degree from the University of Science and Technology of China in 2004, the M.Phil. degree from The Hong Kong University of Science and Technology in 2007, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2011.

He was with Origin Wireless Inc. as a Founding Principal Technologist. From September 2015 to February 2020, he was a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He is currently a Professor with the School of Cyber Science and Technology, University of Science and Technology of China. He has coauthored *Reciprocity, Evolution, and Decision Games in Network and Data Science* (Cambridge University Press, 2021) and *Behavior and Evolutionary Dynamics in Crowd Networks: An Evolutionary Game Approach* (Springer, 2020), and over 200 technical papers, including more than 100 IEEE journal articles. His research interests include multimodal sensing and imaging, multimedia signal processing, and wireless multimedia. He was a recipient of multiple honors and awards, including the Best Paper Award at the APSIPA ASC in 2020, the Best Student Paper Award at the PCM in 2017, the Best Student Paper Award at the IEEE ICASSP in 2016, and the Best Paper Award at the IEEE GLOBECOM in 2013. He is the Chair for the APSIPA Signal and Information Processing Theory and Methods (SIPTM) Technical Committee, and the Secretary-General for the CES Young Scientist Network Multimedia Technical Committee. He is the Organizing Co-Chair of PCM 2017, the Special Session Co-Chair of APSIPA ASC 2017, the 10K Best Paper Award Committee Member of ICME 2017, the Multimedia Communications Symposium Lead Chair of WCSP 2019, and the Area Chair for ACM Multimedia 2021. He is an Associate Editor for IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, and on the Editorial Board for MDPI *Sensors*. He is a Distinguished Lecturer of APSIPA.