PROGRESSIVE PASSIVE NON-LINE-OF-SIGHT IMAGING WITH LARGE MODEL PRIORS

Xiaolong Du[1], Ruixu Geng[1], Jiarui Zhang[1], Yan Chen[1], Yang Hu[1]

[1]School of Electronic Engineering and Information Science
University of Science and Technology of China, China

ABSTRACT

Passive non-line-of-sight (NLOS) imaging has developed rapidly in recent years. However, existing models generally suffer from low-quality reconstruction due to the severe loss of information during the projection process. This paper proposes a two-stage passive NLOS imaging approach, aimed at reconstructing high-quality complicated hidden scenes. In the first stage, we train a coarse reconstruction network based on the optimal transport principle and using vector quantization to learn discrete priors for projection image encoding. This network generates a coarse reconstruction of the hidden image that seems blurry but contains the overall structure of the hidden image. In the second stage, we leverage a large, pre-trained text-to-image diffusion model to augment the coarse reconstruction and recover the image details. We elaborately design the controller modules and the loss functions of this fine reconstruction network to ensure the consistency between the generated image and the coarse reconstruction image. Comprehensive experiments on a large-scale passive NLOS dataset demonstrates the superiority of the proposed method.

***Index Terms*—** non-line-of-sight imaging, large model prior, diffusion model, vector quantization

1. INTRODUCTION

Non-line-of-sight (NLOS) imaging aims to image objects hidden in obstructed view by analyzing scattered light on a relay wall, as shown in Fig. 1. With the trait of seeing hidden objects, NLOS imaging has broad application prospects in many fields such as autonomous vehicles, robot vision, and remote sensing. Depending on whether a controllable light source is used, NLOS imaging can be divided into active imaging and passive imaging. Active imaging uses an ultrafast laser light to illuminate the relay surface area, and a high-resolution time-resolved detector to capture the transient response of three-bounce light. Active imaging enables 3D reconstruction, but it relies on expensive scanning equipment. In contrast, passive NLOS imaging uses an ordinary camera to capture the scattering of light on the relay surface, thereby eliminating the need for controllable illumination and complex detectors.

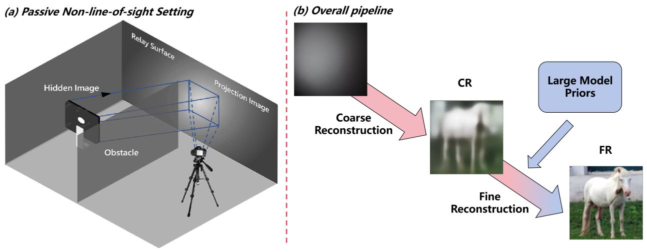

Fig. 1. (a) Passive NLOS imaging setting. Light emitted from the hidden image projects on the relay surface and then captured by a camera. (b) Our two-stage pipeline. First we perform a coarse reconstruction, and then supplement the details leveraging large model priors.

This work focuses on passive NLOS imaging, which can be seen as a special image restoration problem. However, in passive NLOS imaging, the degradation in the projection image is more complicated and severe, which makes it an extremely challenging problem. Directly applying existing image restoration models to passive NLOS imaging usually cannot achieve satisfactory reconstruction results. Recently, deep learning-based methods [1, 2, 3] attemped to establish the mapping from projection images to hidden images with neural networks. However, although some recent work such as [12] have achieved reasonable reconstruction for relatively simple datasets whose hidden images are clean synthetic ones from a single category, they fail to get satisfactory results for more complex datasets with natural hidden images coming from multiple classes. Humans can understand a dramatically degraded image by associating it with similar ones in memory to supplement the missing information. This inspires us that the lost information can be reasonably recovered by introducing large model priors of natural images.

Recently, text-to-image diffusion models [14, 15] show remarkable ability to generate high-quality (HQ) images based on user-provided prompts. This provides the possibility of leveraging the image generation ability of these models to assist challenging image restoration such as the passive NLOS imaging problem. To doing so, a big challenge is to keep the consistency between the image generated by the

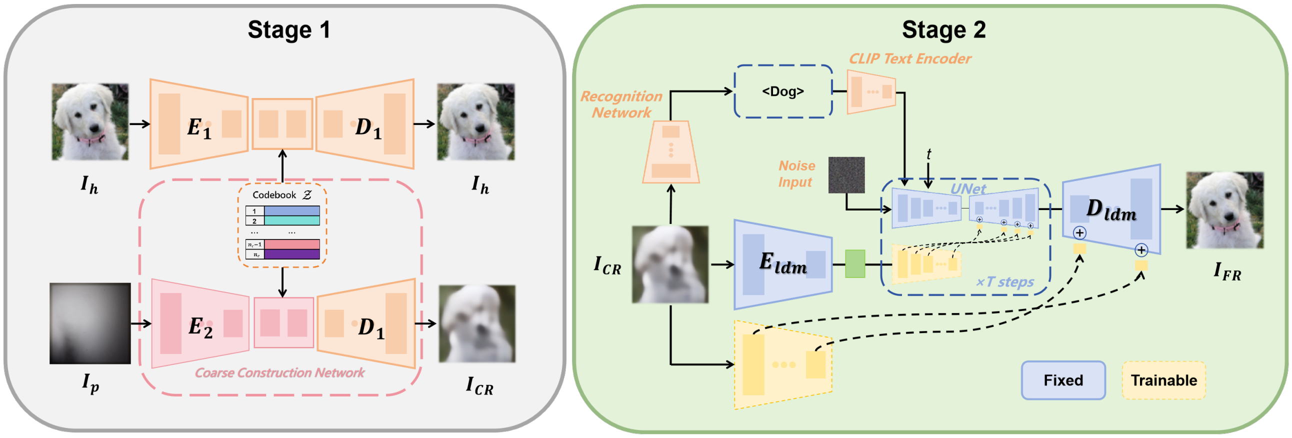

Fig. 2. Our effective two-stage reconstruction pipeline. We first use the CRN to get a rough reconstruction image I_{CR}. The optimal transport principle and the VQ technique have been employed in this network. We also use a parallel encoder design to improve the fidelity of I_{CR}. In the second stage, we leverage the power of a large, pre-trained text-to-image diffusion model to augment I_{CR} and generate high quality reconstruction I_{FR}.

pre-trained generative model and the given projection image. However, due to the severe degradation of the projection image, it is difficult to directly use it to effectively control the generative model to achieve this. We therefore first get a rough reconstruction of the hidden image and then use it to control the pre-trained diffusion model. This stratage naturally decomposes the passive NLOS imaging task into sub-problems with progressive goals which is a reasonable design for this challeging problem.

In this work we introduce a novel passive NLOS imaging method, which exploits the idea of progressive reconstruction and leveraging large model priors to fill in the image details lost during the light transportation process. Specifically, we propose a two-stage hidden image reconstruction method. In the first stage, we generate a rough reconstruction of the hidden image with a coarse reconstruction network (CRN). We first learn discrete latent representations of the hidden images with an autoencoder combined with vector quantization in the latent space. Then a parallel encoder is learned to map the projection images to discrete latent representations same as their corresponding hidden images. The coarse reconstruction is conducted with the decoder of the autoencoder. In the second stage, we refine the results of the first stage through a fine reconstruction network (FRN). In this stage, we leverage a large, pre-trained text-to-image diffusion model to recover the details missing in the coarse reconstruction (CR) images of stage one. We first use a recognition network to identify the category of the objects in the CR images so as to get the textual prompts for the diffusion model. Then we use the CR image as an additional condition to control the pre-trained diffusion model to generate final reconstruction image. We carefully design the added network controllers and the loss functions to ensure the consistence of the spatial structures and the region colors between the generated images the the corresponding CR images.

Our contributions can be summarized as follows:

- We propose a novel progressive passive NLOS imaging method that leveraging generative large model priors. By using large model priors, we effectively solve the problem of LQ reconstruction caused by the severe degradation of NLOS projection image.

- We propose a two-stage NLOS reconstruction pipeline, including using a coarse reconstruction network to obtain a coarse reconstruction image, and a fine reconstruction network that use the CR image to control a large, pre-trained diffusion model to generate a high quality reconstruction of the hidden image.

- Extensive experiments have been conducted on a large-scale passive NLOS dataset. The results show that the proposed method is superior to existing passive NLOS methods and several state-of-the-art image restoration methods.

2. RELATED WORKS

2.1. Passive Non-Line-of-Sight Imaging

Our work focuses on the 2D reconstruction problem in passive NLOS imaging. Existing methods mainly include placing partial occluders, using polarizers and applying deep learning [1, 2, 3]. Among them, deep learning-based passive NLOS imaging is attractive because the superior representation ability of deep neural networks can greatly improve the reconstruction resolution. Notably, Tancik et al. [11] used a variational autoencoder (VAE) for NLOS imaging. However, the model is limited to reconstructing a single specific object.

Geng et al. [12] developed NLOS-OT, a new framework that uses manifold embedding and optimal transport to map projection images to hidden images in latent space. In addition, they also established the first public large-scale passive NLOS dataset NLOS-Passive, which facilitates research in this field. However, due to the severe loss of information during the projection process, NLOS-OT still suffers from low-quality (LQ) reconstruction.

2.2. Image restoration and image prior

Image restoration aims to restore HQ images from degraded LQ versions. Existing work has extensively studied degradation modes such as noise [4], blur [5] and severe weather conditions [6]. Broadly speaking, if the projection process from the hidden image to the projection image is regarded as a kind of degradation, then the passive NLOS imaging problem can be regarded as a special image restoration problem. However, this problem is extremely different from traditional image restoration problems. The hidden images are distorted more severely in passive NLOS imaging than in other problems. Nevertheless, the method of image restoration problem provides a useful reference for solving passive NLOS imaging. Existing image restoration methods are mainly improved in two aspects: data utilization and image prior incorporation. The first focuses on increasing data diversity or improving model pipelines. The second type focuses on the use of image priors. While the "learning-from-scratch" approaches require large amounts of data and computational resources, using pre-trained generative models with rich texture priors has become a practical and efficient approach. Many studies [4, 26] utilize pre-trained Generative Adversarial Networks (GANs) to improve the image restoration process. Basically, they use a generator network to reconstruct the desired HQ image from a degraded LQ input, and a discriminator network to judge whether the HQ output is perceptually realistic. However, due to the inherent limitations of GANs, these methods occasionally produce unrealistic textures. Therefore, in recent research, there is increasing interest in using more advanced pre-trained generative models, such as denoising diffusion models [14, 15, 17, 20].

3. METHODOLOGY

In this work, we aim to exploit powerful large model priors to solve the passive NLOS problem. Our proposed framework employs an efficient and flexible two-stage pipeline. We adopt a conservative but feasible solution by first removing most of the degradation in the projection image and then using subsequent fine reconstruction network to reproduce the lost information. This design promotes the latent diffusion model to focus more on texture/detail generation without interference from degradation, and achieve more realistic/clear results, as shown in Fig. 5. Figure 2 shows the architecture of our framework. Specifically, given a projection image I_p, our goal is to reconstruct the hidden image I_h. In stage 1, the CRN is used to obtain a CR image I_{CR}. In stage 2, in order to introduce large model priors to supplement the details of I_{CR}, we use an image recognition network E_c to identify the content of I_{CR} and obtain the textual prompt $p = E_c(I_{CR})$. Then, a fine-tuned Stable Diffusion is used to generate a FR image I_{FR} guided by I_{CR} and p.

3.1. Coarse Reconstruction Network

The CRN is dedicated to generating the coarse reconstruction image I_{CR}. Due to the severe information loss during the NLOS projection process, it is still very challenging even only recovering the rough structure of I_h. NLOS-OT proposes to establish a mapping between the codings of I_h and I_p, which is based on the optimal transport (OT) theory. In this process, the model first establishes the latent space of I_h by autoencoding pre-training. Then, an encoder is trained using I_p, I_h pairs to map I_p to the representation of I_h in the latent space. By transforming the reconstruction problem into a high-dimensional to low-dimensional mapping problem, NLOS-OT achieves performance exceeding existing NLOS reconstruction methods. However, due to the lack of pixel-level constraints, it tends to obtain low-fidelity results, as shown in Fig. 5(a). In addition, ignoring noise interference (eg. light reflected from irrelevant objects and noise caused by low-exposure shooting) increases the difficulty of coding mapping, thus exacerbating this problem. Therefore, our CRN makes two effective improvements based on NLOS-OT, including vector quantization to resist noise and a parallel encoder to improve reconstruction fidelity.

Network Architecture. The CRN mainly consists of three parts, namely encoders E_1, E_2 and decoder D_1, as shown in Fig. 2. E_1 and D_1 form an autoencoder, which is pre-trained to achieve auto-encoding of the hidden image I_h and then frozen. E_2 is used to encode the projection image I_p. During training, E_1 is discarded, while E_2 and D_1 form a new network for coarse reconstruction.

Vector Quantization. The CRN uses vector quantization (VQ) which was first introduced by VQVAE [16] to learn discrete priors to encode images. During the encoding process, the elements of the spatial latent representation $l = E(x) \in R^{h \times w \times n_d}$ of image x will be replaced by the most similar code in a codebook, as shown in Fig. 2. Specifically, let $\mathcal{Z} = \{z_i\}_{i=1}^{n_c}, z_i \in R^{n_d}$ be a codebook, where n_c is the size of the codebook and n_d is the dimentionality of each code. We quantizing l_{ij} into one of the code in $\mathcal{Z}$ by performing nearest neighbor look-up, which can be formulated as:

$$z_q = \mathbf{q}(l_{ij}) = \arg\min_{z_i \in \mathcal{Z}} \left\| l_{ij} - z_i \right\|_1 . \qquad (1)$$

The HQ codebook is obtained by auto-encoding pre-training on the hidden image dataset, as shown in Fig. 3(a). The pre-trained loss function is comprised of three terms,

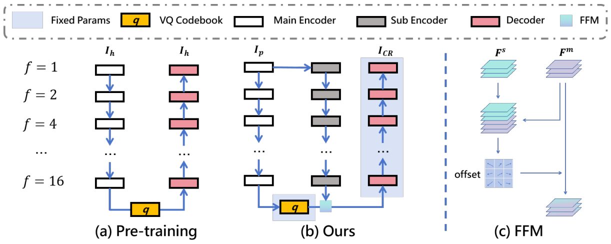

Fig. 3. Illustration of the architecture variants. (a) The structure used in our pre-training. (b) The proposed parallel encoder. (c) The feature fusion module (FFM). f is the compression patch size.

each serving a different purpose, as delineated in Eq. (2). The first component is the reconstruction loss, where $sg[\cdot]$ represents the stop-gradient operation. The codebook $\mathcal{Z}$ undergoes updates via the second term. Meanwhile, the third term is the commitment loss, which ensures that the encoder consistently commits to a specific codebook entry. The weights for these components are given by α and β, respectively.

$$\begin{aligned}\mathcal{L}_{VQ} = \|I'_h - I_h\|_1 + \alpha \left\| sg\left[l\right] - z_q \right\|_2^2 \\ + \beta \left\| sg\left[z_q\right] - l \right\|_2^2\end{aligned} \quad (2)$$

Parallel Encoder. The pre-trained codebook provides the information of I_h for coarse reconstruction, but it also brings the problem of fidelity variation. In order to extract the fidelity information in I_p without ”contaminating” the clean details generated by the codebook, we set up E_2 as a parallel encoder consisting of a main encoder E_m and a sub encoder E_s, as shown in Fig. 3(b). E_m uses OT loss [12] while vector quantizing to promote the feature F_m to be aligned to $E_1(I_h)$. F_m is obtained from the codebook, its fidelity may deviate from the hidden image. E_s performs conventional encoding on the projection image. Without the constraint of OT loss and the VQ step, the semantic information in the feature F_s extracted by E_s is supposed to be more consistent with I_p. By fusing the two features, we can obtain a better coarse reconstruction image. Specifically, we adopt deformable convolution [18] to distort F_s towards F_t, as shown in Fig. 3(c). We first concatenate these two features to generate offsets. The offsets are then used in deformable convolutions to distort F_m to match the fidelity of the input. The whole process can be formulated as:

$$I_{CR} = D_1(\mathcal{F}\left[E_s(I_p), \mathbf{q}(E_m(I_p))\right]), \quad (3)$$

where $\mathcal{F}\left[\cdot\right]$ is the feature fusion operation.

The loss function for coarse reconstruction is a combination of the reconstruction loss and the OT loss:

$$\begin{aligned}\mathcal{L}_{ot} &= \left\|\mathbf{q}(E_m(I_p)) - \mathbf{q}(E_1(I_h))\right\|_1, \\ \mathcal{L}_{reg} &= \mathcal{L}_1(I_h, I_{CR}) + \mathcal{L}_{per}(I_h, I_{CR}), \\ \mathcal{L}_{CRN} &= \mathcal{L}_{ot} + \lambda \mathcal{L}_{reg}.\end{aligned} \quad (4)$$

Here, L_{per} is a simple MSE Loss but measured by the difference between I_h and I_{CR} on VGG features. λ is a weighting factor.

The CRN obtains better reconstruction results than NLOS-OT and most image restoration models, which alleviates the difficulty of the fine reconstruction.

3.2. Fine Reconstruction Network

Stable Diffusion. Due to the severe information loss during the NLOS projection process, the CRN can only generate a rough reconstruction of I_h, which is of low quality and lacks image details. We therefore continue with a second stage reconstruction that resorts to generative model prior to complement the missing details. By introducing large model priors, FRN performs a fine reconstruction based on I_{CR} and a prompt to obtain I_{FR}. The prompt is generated by a recognition network. The FRN is built based on the large-scale text-to-image latent diffusion model - Stable Diffusion. In order to improve efficiency and training stability, Stable Diffusion trains an autoencoder, compresses the image x into a latent encoding z with encoder E_{ldm}, and reconstructs it with decoder D_{ldm}. The diffusion and denoising processes are performed in the latent space through an UNet. Gaussian noise with variance $\beta_t \in (0, 1)$ at time t is added to the latent $z = E_{ldm}(x)$ for producing the noisy latent. In denoising process, UNet is learned by predicting the noise ϵ conditioned on c (i.e., the text prompt) at a randomly chosen time stage t.

Fine-tuning Strategy. We use a set of I_{CR}, I_h pairs to fine-tune the pre-trained Stable Diffusion model with I_{CR} as the conditional control. By creating parallel modules, we fine-tune both the autoencoder and the UNet module in Stable Diffusion. During training, only the parallel modules are fine-tuned for our reconstruction task. This strategy effectively alleviates the overfitting problem of small training datasets and retains the HQ generation capabilities of Stable Diffusion. Furthermore, compared with ControlNet [17], our fine-tuning strategy is more effective for image reconstruction tasks. ControlNet only adds additional network structures to UNet to adjust the distribution of data in the latent space. Therefore, ControlNet can only control high-level information such as the spatial structure and semantic information, yet cannot achieve pixel-level control. In experiments, directly using ControlNet for image reconstruction results in severe color shifts, as shown in Fig. 5(d). Our FRN adds a parallel module to autoencoder, thereby ensuring the consistency of I_{FR} in low-level information such as color and texture.

Specifically, to guide the pre-trained Stable Diffusion model to generate the desired hidden image corresponding to the given projection image, we first use a lightweight image recognition network EfficientnNet-L2 [25] fine-tuned with a set of I_{CR} and their categorical information to recognize the semantic category of the object in I_{CR}. We then inject the

Table 1. Quantitative comparison on NLOS Passive dataset.

Method	FID↓	DISTS↓	LPIPS↓	CLIP-Score↑
NAFNet	264.96	0.4721	0.5896	0.5816
SwinIR	217.04	0.4531	0.5693	0.6170
NLOS-OT	280.72	0.4355	0.5068	0.6266
Uformer	211.45	0.4599	0.5488	0.6188
SR3	159.50	0.2731	0.3635	0.7823
DiffBIR	146.50	0.2649	0.3556	0.7403
Ours(stage1)	193.95	0.4391	0.5314	0.6475
Ours(stage2)	**121.13**	**0.2430**	**0.3107**	**0.8135**

identified semantic information into the FRN in the form of textual prompts. Then, we create parallel modules of the autoencoder and the UNet (indicated in yellow in Fig. 2), which contain the same structured network blocks as them. We initialize the added parameters with pre-trained parameters. The outputs of the parallel modules are added to the decoders of the autoencoder and the UNet respectively. Additionally, a 1 × 1 convolutional layer is applied before the addition operation at each scale.

During training, the original modules retains large model priors due to their frozen parameters. Therefore, their encoder part generate HQ but low-fidelity features. The parallel modules adjust their parameters based on the training I_{CR}, I_h pairs. The features generated by the original modules and the parallel modules are fused at different scales, and their influence weight on the fused features is controlled through a 1 × 1 convolutional layer. The decoders of the autoencoder and the UNet partially decode the fusion features to generate I_{FR}. Our training goal is to minimize the following loss function:

$$\begin{aligned}
\mathcal{L}_{Diff} &= E_{z,c,t,\epsilon,E_{ldm}(I_{CR})}\left[\left\|\epsilon - \epsilon_\theta(z_t, c, t, E_{ldm}(I_{CR})\right\|_2^2\right], \\
\mathcal{L}_{reg} &= \left\|I_{FR} - I_h\right\|_1 + \mu \left\|HSV(I_{FR}) - HSV(I_h)\right\|_1, \\
\mathcal{L}_{FRN} &= \mathcal{L}_{Diff} + \nu \mathcal{L}_{reg}.
\end{aligned} \tag{5}$$

Here $HSV(\cdot)$ means converting the image from RGB domain to HSV domain and retaining the last two dimensions. This loss function makes the network more sensitive to color. μ and ν are weight factors.

4. EXPERIMENTS

4.1. Experiment Settings

Dataset. We use the large-scale passive NLOS dataset NLOS-Passive [12] to evaluate the performance of our model. NLOS-Passive captures projection images under various light transport conditions by changing the distance between hidden images and relay surfaces, the camera angle, the ambient illumination, and the material of the relay surface. NLOS-Passive uses four different types of images as hidden images, namely MNIST [7], Style-GAN generated supermodel face dataset [8], anime face data DANBOORU2019 [9], and STL-10 [10]. Considering the relative simplicity of the first three data types and the already satisfactory reconstruction by NLOS-OT, we focus on the STL-10 dataset to assess our model's performance due to its complexity and the unsatisfactory reconstruction obtained by existing methods.

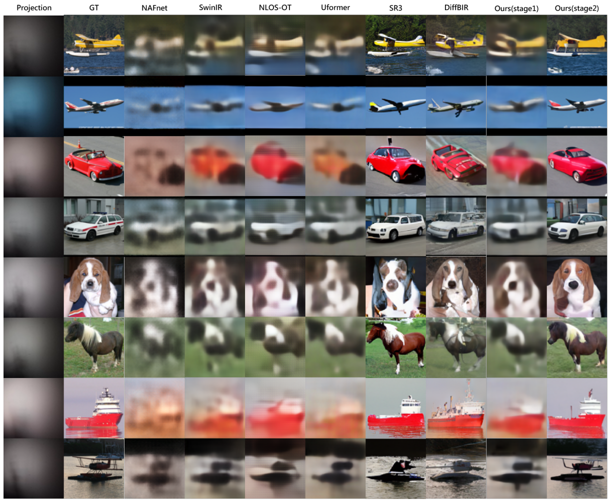

Fig. 4. The visual comparison on NLOS Passive dataset.

Baselines. We compare the model with state-of-the-art passive NLOS-OT imaging method [12] and five state-of-the-art image restoration methods, NAFNet [13], Uformer [6], SwinIR [4], SR3 [20], DiffBIR [19]. Considering the severe loss of information during optical transport, there will be inevitable detail differences between I_{FR} and I_h. We employ four perceptual metrics: FID [21], DISTS [22], LPIPS [23] and CLIP-Score [24]. FID, DISTS, and LPIPS measure perceptual distance, while CLIP-Score estimates semantic accuracy by evaluating the score between I_{FR} and I_h. We provide pixel-level image quality evaluations such as PSNR and SSIM in the ablation experiments of coarse reconstruction. Prior studies [22, 23] have shown that they are weakly correlated with human perception of image quality in real-world environments.

4.2. Result

We provide the quantitative comparison on NLOS Passive dataset in Table. 1. It is observed that our method achieves optimal or suboptimal results on most metrics. The visual comparison results are presented in Fig. 4. It can be observed that our method is able to restore the image more naturally, while other methods tend to distort the image or produce blurry output. It is worth noting that only using the CRN in stage 1 can obtain better results than most models, which shows that the CRN can reasonably establish the mapping of I_p to I_h. In addition, our method can also generate realistic details for

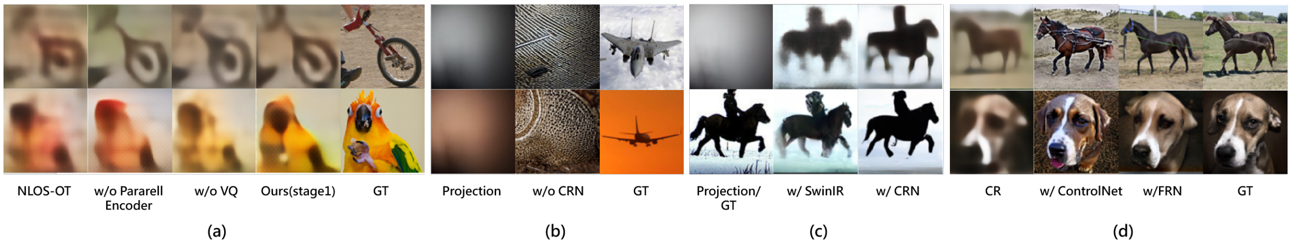

Fig. 5. Visual comparison of ablation studies. (a) w/o VQ cause artifacts and noise in I_{CR}, while w/o parallel encoder lead to low-fidelity results. (b) w/o the CRN incorrectly regards the shadow of I_p as semantic information, resulting in reconstruction collapse; The third and fourth columns show that two baselines based on diffusion model cannot obtain reasonable results due to the ill-posed nature of I_p. (c) w/ SwinIR, replace the CRN with SwinIR to reconstruct I_{CR} in stage 1. The first row is I_{CR}, and the second row is I_{FR}. SwinIR's incorrect reconstruction of shape and color results in a poor subsequent reconstruction; (d) ControlNet has a color shift problem, which can be solved by our fine-tuning strategy.

Table 2. Quantitative comparison of ablation study.

Exp.	**PSNR↑**	**SSIM↑**
NLOS-OT	17.46	0.5072
(a) without VQ	18.89	0.5206
(a) without Parallel Encoder	18.33	0.5158
Ours(stage1)	**19.54**	**0.5271**

Exp.	**FID↓**	**DISTS↓**	**LPIPS↓**	**CLIP-Score↑**
(b) without CRN	352.74	0.6445	0.7138	0.5553
(c) replace CRN with SwinIR	147.02	0.2703	0.3842	0.7758
(d) replace FRN with ControlNet	169.45	0.3233	0.5381	0.7409
Ours(stage2)	**121.13**	**0.2430**	**0.3107**	**0.8135**

natural images, while other methods generate semantically incorrect textures or inconsistent colors.

4.3. Ablation Study

The Importance of VQ and Parallel Encoders. We first study the effectiveness of our proposed new modules in CRN for coarse reconstruction. We remove VQ and parallel encoder respectively, and then perform the coarse reconstruction. The quantitative comparison of performance is shown in Table 2, and the visual comparison is shown in Figure 5(a). Both results demonstrate that by introducing VQ and the parallel encoder, the quality of the reconstructed image is evidently improved.

The Importance of Coarse Reconstruction Network. We then study the implications of CRN to our proposed two-stage pipeline. Here, we respectively remove the CRN and replace it with SwinIR. Removing or replacing the CRN resulted in a significant degradation in performance on the dataset, as shown in Table. 2. The visual comparison is shown in Fig. 5(b, c). As can be seen from the first example, fine-tuning a diffusion model directly with I_p causes the model (w/o OT) to incorrectly treat degradation as semantic information. In the second example, we replace the CRN with SwinIR. It can be seen that SwinIR cannot effectively eliminate the degradation of the projection image, thus affecting subsequent reconstruction. Two examples indicate that CRN is indispensable in degradation removal.

The Importance of Fine Reconstruction Network. Finally, we verify the effectiveness of our proposed FRN. Here, we compare with ControlNet, which also add controls to the pretrained large diffusion model. As shown in Fig. 5(d), ControlNet tends to output results with color shifts due to the lack of regularization for color consistency during training. This indicates that FRN can better control low-level information to improve the fidelity of results.

5. CONCLUSION

In this paper, we propose a progressive reconstruction method that leverages large model priors to achieve high quality passive NLOS imaging. Due to the great difficulty of this restoration problem, we employ an effective two-stage reconstruction pipeline. We first use a coarse reconstruction network to get a rough reconstruction of the hidden image. The optimal transport principle and the vector quantization technique have been employed in this network. We also use a parallel encoder design to improve the fidelity of the reconstruction. In the second stage, we leverage the power of a large, pretrained text-to-image diffusion model to augment the coarse reconstruction and generate high quality reconstruction of the hidden image. Experiments on a large passive NLOS dataset demonstrate the superiority

6. REFERENCES

[1] M. Aittala et al. . Computational mirrors: Blind inverse light transport by deep matrix factorization. In NeurIPS,

2019.

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

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

[4] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In ICCV, pages 1833–1844, 2021.

[5] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In arXiv preprint arXiv:2201.11793, 2022.

[6] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In CVPR, pages 17683–17693, 2022.

[7] Gwern Branwen and A Gokaslan. Danbooru2019: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. 2019.

[8] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics. In JMLR Workshop and Conference Proceedings, 215–223, 2011.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR. 4401–4410, 2019.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradientbased learning applied to document recognition. In Proc. IEEE 86, 11, 2278–2324, 1998.

[11] Matthew Tancik, Guy Satat, and Ramesh Raskar. 2018. Flash photography for data-driven hidden scene recovery. In arXiv preprint arXiv:1810.11710, 2018.

[12] Ruixu Geng, Yang Hu, Zhi Lu, Cong Yu, Houqiang Li, Hengyu Zhang, and Yan Chen. Passive non-line-of-sight imaging using optimal transport. In TIP, 110–124, 2021.

[13] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In ECCV, 2022.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684-10695, 2022.

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS, 35:36479–36494, 2022.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In arXiv preprint arXiv:1807.03748, 2018.

[17] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In arXiv preprint arXiv:2302.05543, 2023.

[18] Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In CVPR. pp. 9308–9316, 2019.

[19] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, Chao Dong. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. In arXiv:2308.15070, 2023.

[20] Saharia, Chitwan, et al. Image Super-Resolution via Iterative Refinement. In IPTMA, vol. 45, no. 4, 2022, pp. 1–14, 2023.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 30, 2017.

[22] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. In TPAMI, 44(5):2567–2581, 2020.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pages 586–595, 2018.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, pages 8748–8763. PMLR, 2021.

[25] Foret, Pierre, et al. ”Sharpness-aware minimization for efficiently improving generalization.” In arXiv preprint arXiv:2010.01412, 2020.

[26] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In TPAMI, 44(11):7474–7489, 2021.