Virtual Scanning: Unsupervised Non-line-of-sight Imaging from Irregularly Undersampled Transients

Xingyu Cui[1] **Huanjing Yue**[1] **Song Li**[2,3] **Xiangjun Yin**[1]
Yusen Hou[1] **Yun Meng**[2,3] **Kai Zou**[2,3] **Xiaolong Hu**[2,3] **Jingyu Yang**[1,*]
[1]School of Electrical and Information Engineering, Tianjin University, China
[2]School of Precision Instrument and Optoelectronic Engineering, Tianjin University, China
[3]Key Lab. of Optoelectronic Information Science and Technology, Ministry of Education, China

Abstract

Non-line-of-sight (NLOS) imaging allows for seeing hidden scenes around corners through active sensing. Most previous algorithms for NLOS reconstruction require dense transients acquired through regular scans over a large relay surface, which limits their applicability in realistic scenarios with irregular relay surfaces. In this paper, we propose an unsupervised learning-based framework for NLOS imaging from irregularly undersampled transients (IUT). Our method learns implicit priors from noisy irregularly undersampled transients without requiring paired data, which is difficult and expensive to acquire and align. To overcome the ambiguity of the measurement consistency constraint in inferring the albedo volume, we design a virtual scanning process that enables the network to learn within both range space and null space for high-quality reconstruction. We devise a physics-guided SURE-based denoiser to enhance robustness to ubiquitous noise in low-photon imaging conditions. Extensive experiments on both simulated and real-world data validate the performance and generalization of our method. Compared with the state-of-the-art (SOTA) method, our method achieves higher fidelity, greater robustness, and remarkably faster inference times by orders of magnitude. The code and model are available at https://github.com/XingyuCuii/Virtual-Scanning-NLOS.

1 Introduction

Non-line-of-sight (NLOS) imaging aims to reconstruct hidden scenes beyond the direct line of sight of the detector, garnering interest across various fields such as robot vision, autonomous driving, rescue operations, remote sensing, and medical imaging [1–5]. In a typical active confocal NLOS imaging system, as depicted in Fig. 1, a laser source and a detector are both focused on the same point on a relay surface. Pulses emitted by the laser reflect off the surface to illuminate the hidden scene. The detector captures photons bouncing back from the scene toward the relay surface, referred to as transients, from which the hidden scene can be recovered using elaborately designed algorithms.

While existing works have achieved remarkable breakthroughs, they also face significant limitations that hinder their practical applicability. These methods assume dense and regular scanning of a large relay surface, which may not be feasible in realistic scenarios with irregular relay surfaces such as latticed windows or fences. Transients obtained through irregular undersampling can result in severe ill-posedness, leading to artifacts or reconstruction failure. This raises the challenging task of *NLOS imaging from irregularly undersampled transients (IUT)*. To address this, Liu et al. [6] proposed introducing manually designed strong regularization terms under a functional optimization framework. However, this approach is hindered by the need for lengthy numerical iterative computations. Recent

*Corresponding author: yjy@tju.edu.cn.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

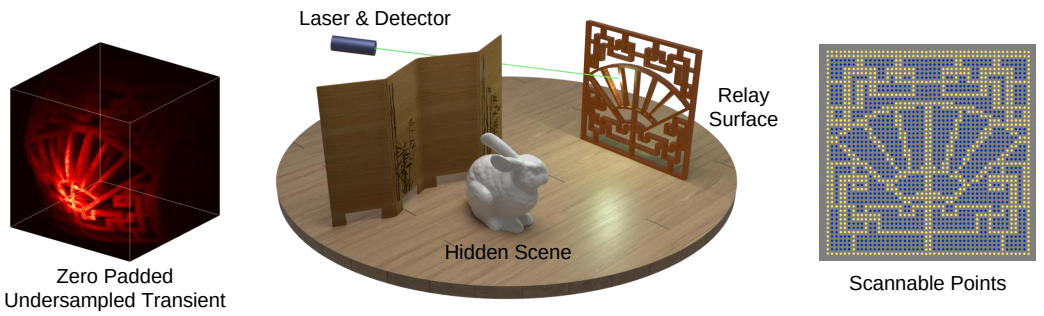

Figure 1: Illustration of the active confocal NLOS imaging with an irregular relay surface. Yellow points indicate scannable points, while blue points indicate non-scannable points.

learning-based NLOS imaging methods [7–10] enable quick inference, but they require supervision from a large amount of paired data, including well-aligned ground-truth albedos, which are difficult and expensive to acquire. Therefore, exploring NLOS imaging under unsupervised learning is worthwhile to eliminate the heavy reliance on paired data.

In this paper, we propose a novel unsupervised framework capable of learning implicit priors from noisy IUT. This framework comprises two components: a virtual scanning reconstruction network (VSRnet) that learns high-quality measurement-to-albedo mapping beyond the range space induced by the measurement consistency, and a SURE-based denoiser that enhances our method's robustness to measurement noise by incorporating the physics model of the low-photon time-resolved detector. We conduct extensive experiments on simulated data, publicly available real data, and data acquired from our self-built NLOS system. Our method outperforms existing algorithms, particularly in providing robustness for real data and diverse irregular relay surfaces.

Our main contributions are summarized as follows:

• We propose an unsupervised NLOS imaging framework capable of learning implicit priors from noisy IUT, effectively overcoming the dependency on paired data that is difficult to acquire and align.

• We introduce a virtual scanning process that enables the network to learn within both range and null spaces for high-quality reconstruction from IUT, extending NLOS imaging to realistic scenarios with irregular relay surfaces.

• We propose a SURE-based denoiser, an unsupervised physics-guided module that incorporates the low-photon time-resolved detector's physics model, to enhance our method's robustness to noise.

• We evaluate our method on simulated, publicly available, and self-captured real data, demonstrating its superior reconstruction quality and significantly faster inference compared to the SOTA method.

2 Related work

2.1 Model-based NLOS reconstruction

Model-based NLOS algorithms have achieved significant advances in recent years. Direct reconstruction algorithms [11–15] offer rapid implementations for NLOS imaging, while iterative algorithms [16–18, 5] leverage more accurate physical models for higher reconstruction quality. Recent efforts aim to extend NLOS imaging to challenging real-world scenarios. For instance, Manna et al. [19] and Gu et al. [20] addressed NLOS imaging with dynamic and non-planar relay surfaces, respectively. However, these methods still rely on large relay surfaces and dense measurements. Another set of algorithms [21–25] aims to achieve high spatial resolution reconstruction using sparse sampling measurements to significantly reduce acquisition time. Yet, they are constrained to specific scanning patterns, such as regular or Hadamard patterns. To address NLOS imaging from irregularly undersampled transients, Liu et al. [6] introduced a reconstruction model using Confocal Complemented Signal-Object Collaborative Regularization (CC-SOCR). Despite its effectiveness, CC-SOCR suffers from long inference times due to its iterative nature.

2.2 Learning-based NLOS reconstruction

Recently, deep learning has gained attention for NLOS imaging due to its learning capabilities and fast inference. Chopite et al. [7] first employed deep learning for NLOS reconstruction, but their method fell short compared to model-based approaches due to the lack of physical guidance. Physics-guided methods [8, 26, 9, 10] have since emerged to enhance reconstruction quality, particularly for real-world data. Recent works focus on NLOS imaging from regularly undersampled transients [27, 28], but they require large datasets of paired data and fail to reconstruct from IUT. Furthermore, these supervised algorithms still have significant room for improvement in robustness on real data due to the gap between simulated datasets used for training and real-world data.

3 Problem formulation and motivation

3.1 NLOS Imaging from IUT

Fig. 1 illustrates confocal NLOS imaging with time-resolved systems. By scanning a set of points $P = \{(x_i, y_i, 0) \mid i = 1, 2, \ldots, s,\ x_i, y_i \in \mathbb{R}\}$ on the relay surface, the forward model of NLOS imaging can be modeled as

$$\tau(p,t) = \int_Q \frac{\kappa(q)}{\|p-q\|^4} \cdot \delta(2\|p-q\| - tc) dq \tag{1}$$

where τ is the spatial-temporal measurement, $p = (x, y, 0)$ denotes the scanning point, $\kappa(q)$ denotes the albedo value at point q in the 3D hidden scene Q, c is the speed of light. The distance $\|p-q\|$ is related to the time of flight t through the Dirac delta function $\delta(\cdot)$. For compact presentation and analysis, we employed a discretized version of the above forward model. Let $u \in \mathbb{R}^{st}$ and $\rho \in \mathbb{R}^{l^2 z}$ represent the vectorized measurements and albedos of the hidden object, respectively, where l denotes the size of the vertical and horizontal dimensions, and z denotes the depth dimension. We denote the forward operator, also known as the light transport matrix, by $H \in \mathbb{R}^{st \times l^2 z}$. The forward processing can be described by the following linear model:

$$u = H\rho. \tag{2}$$

Notably, the forward operator H is related not only to the optical-electronic characteristics of the NLOS system but also to the scannable region on the relay surface.

3.2 Motivation

Deep learning-based algorithms have demonstrated significant potential to enhance NLOS imaging performance compared to model-based approaches. Most prior work has adopted a supervised paradigm, which requires substantial amounts of high-quality paired data. However, it is prohibitively expensive or even infeasible to acquire ground-truth 3D albedo volumes precisely aligned with the spatio-temporal measurements. This limitation motivates us to develop an unsupervised NLOS reconstruction framework that avoids this dependency and further enhances the generalization of learning-based methods to real-world data.

The standard approach in unsupervised learning is to train the reconstruction mapping f_θ by minimizing the measurement consistency (MC) loss:

$$\mathbb{E}_u \|H f_\theta(u) - u\|_2^2. \tag{3}$$

Nevertheless, solely enforcing the MC loss without ground truth supervision does not guarantee high-quality reconstruction. This can be analyzed from the perspective of the range-null decomposition [29]. Let $\mathcal{N}_H = \{v \in \mathbb{R}^{l^2 z} \mid Hv = 0\}$ be the null space of the operator H. Its complementary space is the range space of $H^\top$, denoted by $\mathcal{R}_H = \{H^\top u, u \in \mathbb{R}^{st}\}$, such that $\mathbb{R}^{l^2 z} = \mathcal{R}_H \oplus \mathcal{N}_H$.

Any albedo volume ρ can be decomposed into a range-space component and a null-space component.

$$\rho = \underbrace{H^\dagger H\rho}_{\text{range-space component}} + \underbrace{(I - H^\dagger H)\rho}_{\text{null-space component}} \tag{4}$$

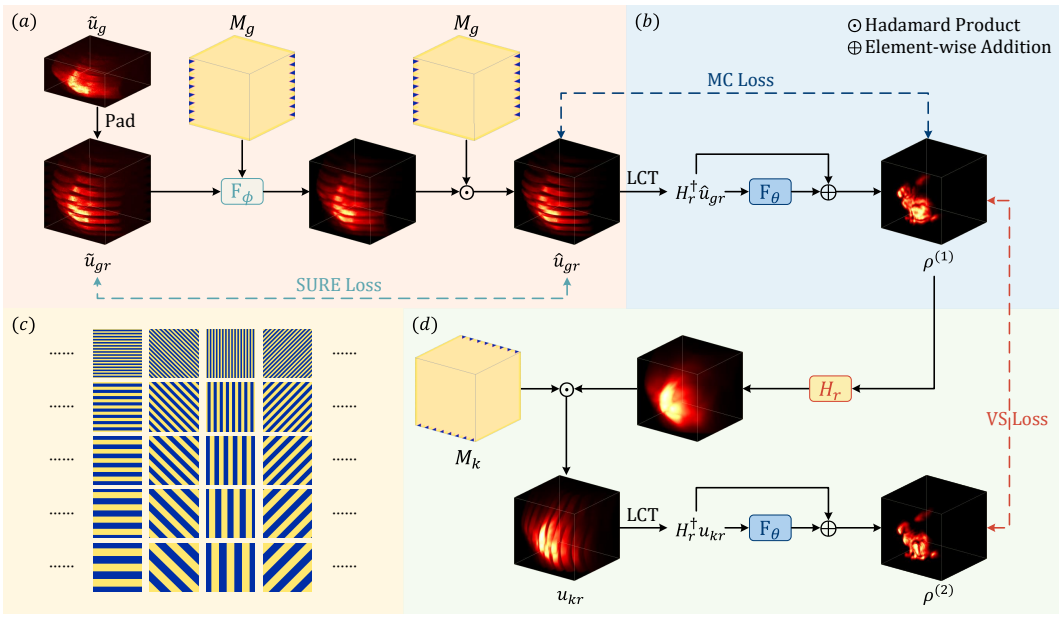

Figure 2: The pipeline of our unsupervised framework: (a) The SURE-based denoiser, which consists of an encoder-decoder network designed for IUT and is trained by minimizing SURE loss in the first stage; (b) The virtual scanning reconstruction network (VSRnet), which consists of a Unet-like network and is trained by MC loss and VS loss in the second stage; (c) Relay surfaces used for training (yellow indicates scanning areas, and their corresponding 3D binary masks are used in implementation); (d) The virtual scanning process, which involves virtually observing $\rho^{(1)}$ with a relay surface M_k that is distinct from M_g and enforcing consistency between $\rho^{(1)}$ and $\rho^{(2)}$.

where $H^\dagger \in \mathbb{R}^{l^2z \times st}$ is the pseudo-inverse of H satisfying $HH^\dagger H = H$. The operator $H^\dagger H$ projects the sample ρ into the range space $\mathcal{R}_H$: $\mathcal{D}_r(\rho) = H^\dagger H\rho$, whereas its complementary operator $(I - H^\dagger H)$ projects ρ into the null-space $\mathcal{N}_H$: $\mathcal{D}_n(\rho) = (I - H^\dagger H)\rho$. As long as the trained network f_θ reconstructs from input u as $f_\theta(u) = \mathcal{D}_r(\rho) + v_n$ for $\forall\ v_n \in \mathcal{N}_H$, the reconstructed volume $f_\theta(u)$ would fully meet the MC requirement: $H(\mathcal{D}_r(\rho) + v_n) = HH^\dagger H\rho + Hv_n = u$ since we have $HH^\dagger H\rho = u$ and $Hv_n = 0$. This suggests that the MC constraint only locates the albedo volume in a broad subspace surrounding the range-space projection $\mathcal{D}_r(\rho)$, and the inference of the component $\mathcal{N}_H$ is ad-hoc without guidance.

This necessitates an unsupervised framework capable of learning beyond the range space. We note that model-based algorithms [18, 6, 25, 23] manually design regularization terms to learn beyond range space, but suffer from long inference times. In other computational imaging tasks, supervised methods [30, 31, 29, 32] and unsupervised methods [33–37] have been proposed to recover null-space components of the reconstructions. Along this avenue, we propose an effective unsupervised framework capable of learning in both range-null spaces for NLOS imaging from IUT.

4 Method

4.1 Unsupervised framework

Fig. 2 shows the proposed unsupervised framework via virtual scanning for NLOS imaging from IUT. The framework consists of two components: 1) a virtual scanning reconstruction network (VSRnet) to recover the 3D albedo volume from both range and null space, and 2) a SURE-based denoiser to enhance the robustness to ubiquitous noise in transients. Given a set of G noisy irregularly undersampled transients $\mathcal{U} = \{\tilde{u}_g \in \mathbb{R}^{st} |\ g = 1, \ldots, G\}$ and the set of their corresponding forward operators $\mathcal{H} = \{H_g \in \mathbb{R}^{st \times l^2 z} |\ g = 1, \ldots, G\}$, our goal is to train a deep neural mapping without labeled supervision to reconstruct the 3D albedo volume ρ from the noisy IUT $\tilde{u}$. As discussed in

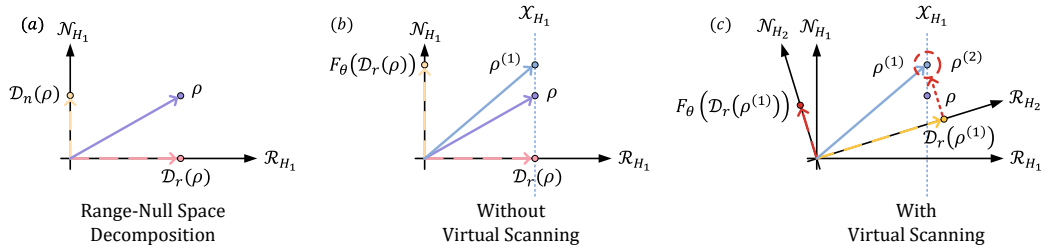

Figure 3: Toy visualization of NLOS reconstruction from the perspective of range-null space decomposition (RNSD). (a) illustrates the RNSD of ρ observed by H_1. (b) illustrates that ρ cannot be accurately recovered with only the measurement consistency loss. (c) shows that the proposed virtual scanning promotes the acquisition of null-space components.

Sec. 3.1, due to the one-to-one correspondence between the forward operator and the relay surface, we will loosely use H_g to refer to different relay surfaces without ambiguity in the following sections.

We first briefly sketch the workflow of the inference stage, as shown in Fig. 2(a) and 2(b). The noisy IUT $\tilde{u}_g$, observed through an irregular relay surface H_g, is first zero-padded to $\tilde{u}_{gr}$ on a full scanning grid. This is then passed through the SURE-based denoiser F_ϕ to remove noise. Subsequently, the albedo volume ρ is reconstructed using VSRnet f_θ, which incorporates the physical prior of NLOS imaging (LCT [11]) and a learned reconstruction mapping F_θ. During the training stage, the denoiser is regularized by the SURE loss, while the reconstruction network is regularized by the MC loss. Additionally, we introduce the virtual scanning process (Fig. 2(d)) to capture the measurement details within the null space of observation operators. In this process, the reconstructed $\rho^{(1)}$ is projected into the measurement space as u_{kr} by virtually scanning with a different relay surface $H_k \in \mathcal{H}$, which is distinct from H_g. The virtual scanned measurement u_{kr} is then projected back into the reconstruction space as $\rho^{(2)}$ by the reconstruction network. We impose a virtual scanning (VS) loss between the two reconstructed volumes, $\rho^{(1)}$ and $\rho^{(2)}$, to promote null-space learning (Sec. 4.2). The modules of our framework and the loss functions utilized for training are detailed in the following subsections. The structures of F_ϕ and F_θ are detailed in the Supplementary Material (SM).

4.2 Virtual Scanning

Strategy The training strategy of VSRnet is depicted in Fig. 2(b) and 2(d). Specifically, at the reconstruction module (Fig. 2(b)), the denoised sample $\hat{u}_{gr}$ is initially transformed to the albedo domain using the inverse operator of LCT $H_r^\dagger$. The resulting $H_r^\dagger \hat{u}_{gr}$ is then mapped to the albedo volume $\rho^{(1)}$ by the reconstruction network F_θ with a global residual connection. At the virtual scanning module, the reconstructed albedo volume $\rho^{(1)}$ is projected into a virtual undersampled measurement u_{kr} using another forward operator $H_k \in \mathcal{H}$ $(H_k \neq H_g)$. To achieve efficient implementation of forward operators, we decouple H_k into $M_k \odot H_r$, which can be efficiently computed using Hadamard product and the fast Fourier transform. In practice, $M_k \in \mathbb{R}^{l \times l \times z}$ is the 3D binary mask associated with the relay surface, constructed by repeating the 2D sampling pattern t times along the time dimension. H_r represents the forward operator of LCT. Following the same reconstruction pipeline, an albedo volume, denoted by $\rho^{(2)}$, is obtained from the virtual measurement u_{kr}. The processes of obtaining $\rho^{(1)}$ and $\rho^{(2)}$ can be formalized as:

$$\begin{aligned} \rho^{(1)} &= f_\theta(F_\phi(\mathrm{Pad}(\tilde{u}_g), M_g) \odot M_g), \\ \rho^{(2)} &= f_\theta(H_r \rho^{(1)} \odot M_k). \end{aligned} \tag{5}$$

The two reconstructed albedos, $\rho^{(1)}$ and $\rho^{(2)}$, should be identical if the learned mapping F_θ enables perfect reconstruction. This motivates us to impose a proximity constraint between $\rho^{(1)}$ and $\rho^{(2)}$ in the albedo domain, named the virtual scanning loss. The virtual scanning process in the training pipeline facilitates the learning of the null-space component, thereby providing a promising prior that complements the range space, as analyzed below.

Analysis Fig. 3 provides a more intuitive understanding of how the proposed virtual scanning facilitates the network in learning the null-space component. Let H_1 and H_2 be two observation

operators associated with two different relay surfaces. We observe ρ using the forward operator H_1, obtaining the measurement u_1. As illustrated in Fig. 3(a), ρ is projected into the range-space $\mathcal{R}_{H_1}$ of H_1. When we only impose the MC constraint, the resulting output $\rho^{(1)}$ will belong to the following set $\mathcal{X}_{H_1}$:

$$\mathcal{X}_{H_1} = \{v \mid H_1^{\dagger} H_1 v = \mathcal{D}_r(\rho), \mathcal{D}_r(\rho) \in \mathcal{R}_{H_1}\}. \tag{6}$$

As depicted by the blue dashed line in Fig. 3(b), there exist multiple outputs that satisfy the measurement consistency. If, for instance, we obtain inaccurate estimated results denoted as $\rho^{(1)}$ through $\rho^{(1)} = \mathcal{D}_r(\rho) + F_\theta(\mathcal{D}_r(\rho))$, we then utilize H_2 to virtually scan $\rho^{(1)}$ and project it into the range space $\mathcal{R}_{H_2}$. Subject to the constraint $\rho^{(1)} = \rho^{(2)}$, $F_\theta(\mathcal{D}_r(\rho^{(1)}))$ converges to $\mathcal{D}_n(\rho^{(1)})$ due to the relationship $\rho^{(1)} = \mathcal{D}_r(\rho^{(1)}) + \mathcal{D}_n(\rho^{(1)})$. Following this iteration, the network can learn within $\mathcal{R}_{H_1}$ and $\mathcal{N}_{H_2}$. The entire process is illustrated in Fig. 3(c). Similarly, by altering the order of operators H_1 and H_2, F_θ will also observe within $\mathcal{R}_{H_2}$ and $\mathcal{N}_{H_1}$. In practice, a set of operators $\mathcal{H}$ is provided to enhance the network's generalization across various operators with different relay surfaces.

4.3 SURE-based denoiser

NLOS imaging inherently operates under photon-limited conditions, and noise in transient measurements can lead to severe background artifacts in the albedo space. This hinders the network's ability to learn implicit priors from noisy IUT, especially in unsupervised learning. Inspired by previous studies [38–42], we introduce a physics-guided unsupervised denoiser to suppress measurement noise. Specifically, we leverage Stein's Unbiased Risk Estimator (SURE) [43] to derive an unsupervised learning loss function that considers the physical model of the time-resolved detector in NLOS imaging systems, which can be modeled as

$$\begin{aligned} \tilde{u} &\sim \text{Poisson}(u + b), \\ u &= H\rho, \end{aligned} \tag{7}$$

where b represents the dark counts of the detector along with the background photons. The parameterized denoiser F_ϕ learns to map the noisy measurement $\tilde{u}$ to its clean version u via minimization of the SURE loss function given in Eq. (8).

4.4 Loss function

Given a measurement set $\{\tilde{u}_{i,g}, i = 1, 2, \ldots, I, g = 1, 2, \ldots, G\}$ collected by observing I hidden scenes, each with G relay surfaces, the training of our framework involves three loss functions: the SURE loss $\mathcal{L}_{\text{SURE}}$, the MC loss $\mathcal{L}_{\text{MC}}$, and the VS loss $\mathcal{L}_{\text{VS}}$. The SURE loss is an unbiased estimation of the mean squared error (MSE) under the Poisson noise model taking dark count consideration:

$$\begin{aligned} \mathcal{L}_{\text{SURE}} = \mathbb{E}_{\{i,g\}}\{ &\frac{1}{st}\|\tilde{u}_{i,g} - F_\phi(\tilde{u}_{i,g})\|_2^2 - \frac{1}{st}(\mathbf{1} + b)^\top \tilde{u}_{i,g} \\ &+ \frac{2}{st} b^\top F_\phi(\tilde{u}_{i,g}) + \frac{2}{st\varepsilon}(e_{i,g} \odot \tilde{u}_{i,g})^\top \big(F_\phi(\tilde{u}_{i,g} + \varepsilon e_{i,g}) - F_\phi(\tilde{u}_{i,g})\big)\} \end{aligned} \tag{8}$$

where ε is a positive number, $e_{i,g} \in \{-1, 1\}^{st}$ is a binary vector, whose elements follow a Bernoulli distribution with equal probability [40], and $\odot$ denotes element-wise multiplication. Detailed derivation of the SURE loss is given in supp. material.

We adopted the mean squared error for $\mathcal{L}_{\text{MC}}$ and $\mathcal{L}_{\text{VS}}$:

$$\begin{aligned} \mathcal{L}_{\text{MC}} &= \mathbb{E}_{\{i,g\}}\|\hat{u}_{i,g} - H_g(f_\theta(\hat{u}_{i,g}))\|_2^2, \\ \mathcal{L}_{\text{VS}} &= \mathbb{E}_{\{i,g\}}\|f_\theta(\hat{u}_{i,g}) - f_\theta(H_k(f_\theta(\hat{u}_{i,g}))\|_2^2. \end{aligned} \tag{9}$$

We first trained the denoiser F_ϕ using the SURE loss. Next, we froze the SURE-based denoiser and trained the network F_θ with a combined loss function, $\mathcal{L}(\theta) = \mathcal{L}_{MC}(\theta) + \beta\mathcal{L}_{VS}(\theta)$, where β is a trade-off parameter (see SM for training details).

5 Experiment

5.1 Experiment setup

Dataset We generated 8,000 transients using the transient rasterizer from [8] with default parameters. The transients have a spatial-temporal resolution of $128 \times 128 \times 512$ with a bin width of 33 ps. The

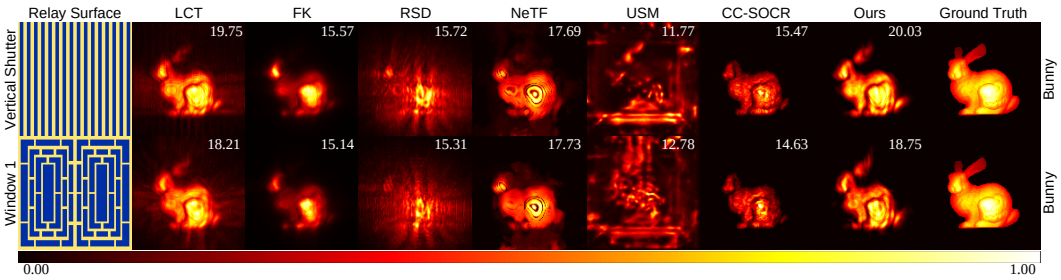

Figure 4: Reconstruction results of the "bunny" with different relay surfaces. The intensity images are normalized to the range of 0 to 1 through maximum value normalization. PSNR (dB) for each intensity image is displayed in the top right corner.

dataset contains 111 objects from the alphanumerics dataset [9], covering lowercase and uppercase letters from the English and Greek alphabets, and numerals from 0 to 9. We also rendered a sample "bunny" for quantitative comparison. To assess our method's generalization, we tested it on real-world data acquired by three different systems [10–12] and a self-built system (see SM for system details). The hidden scenes feature various reflective materials, depth ranges, and geometric shapes, and were acquired under different conditions including scanned areas, spatial resolutions, bin widths, and integration times. For a fair comparison with CC-SOCR [6] within a manageable time frame, we resized the full-sampled transients [10, 12] to $128 \times 128 \times 256$ and the full-sampled transients [11] to $64 \times 64 \times 256$. We used these resized transients for all methods.

Relay surface To simulate the irregularly undersampled process, we extracted signals from the full-sampled transients according to various irregular relay surfaces. For training, we sampled five horizontal shutter patterns with intervals of [4, 8, 12, 16, 20] and 40 uniform rotations from 0 to 180 degrees, resulting in 200 sampling patterns (see Fig. 2(c)). For testing, we included more realistic irregular relay surfaces to evaluate our method's generalization capability in real-world scenarios.

Compared methods We compare our method with three traditional direct reconstruction algorithms (LCT [11], FK [12], RSD [13]), two learning-based algorithms (Unsupervised NeTF [26], Supervised USM [28]), and one iterative algorithm (CC-SOCR [6]). Since LCT, FK, RSD, and USM are designed for raster scanning, we use zero-padded versions of the IUT as input, similar to our method. NeTF and CC-SOCR accept irregularly undersampled transients. For a fair comparison, we also applied our SURE-based denoiser to pre-denoise the transients for all compared methods. However, we did not apply the pre-denoising step for CC-SOCR, as the results reconstructed by CC-SOCR did not show significant improvements. This is because the strong CC-SOCR regularization inherently performs some level of denoising. We compute the peak signal-to-noise ratio (PSNR) to quantitatively evaluate the reconstruction results on the simulated dataset for all methods.

5.2 Results

Simulated data Fig. 4 shows comparison results on the "bunny". LCT, FK, NeTF, CC-SOCR, and our method can recover the main body. However, FK and CC-SOCR lose structures around the ear, while LCT and RSD exhibit aliasing artifacts due to irregular undersampling. Among the learning-based methods, NeTF produces a blurry object with diffused artifacts and loss of structures around the ear. USM struggles to adapt to irregularly undersampled transients and fails to reconstruct the hidden object. In contrast, our method successfully recovers most geometric structures of the bunny without aliasing artifacts, achieving the best quantitative results.

Real-world data We first tested our method on publicly available real datasets [10–12] (Fig.5), and then on self-captured real-world data (Fig.6). Without proper regularization, the direct reconstruction methods, *i. e.*, LCT and RSD, exhibit severe aliasing artifacts due to undersampling. FK is depth-sensitive and fails to recover far-end structures in the hidden scene. NeTF tends to produce blurry shapes due to its struggle with utilizing limited information in IUT. USM produces results that are nearly overwhelmed by aliasing artifacts in irregularly undersampled cases, but can achieve acceptable results in regularly undersampled cases (Fig.6). CC-SOCR can recover main objects for

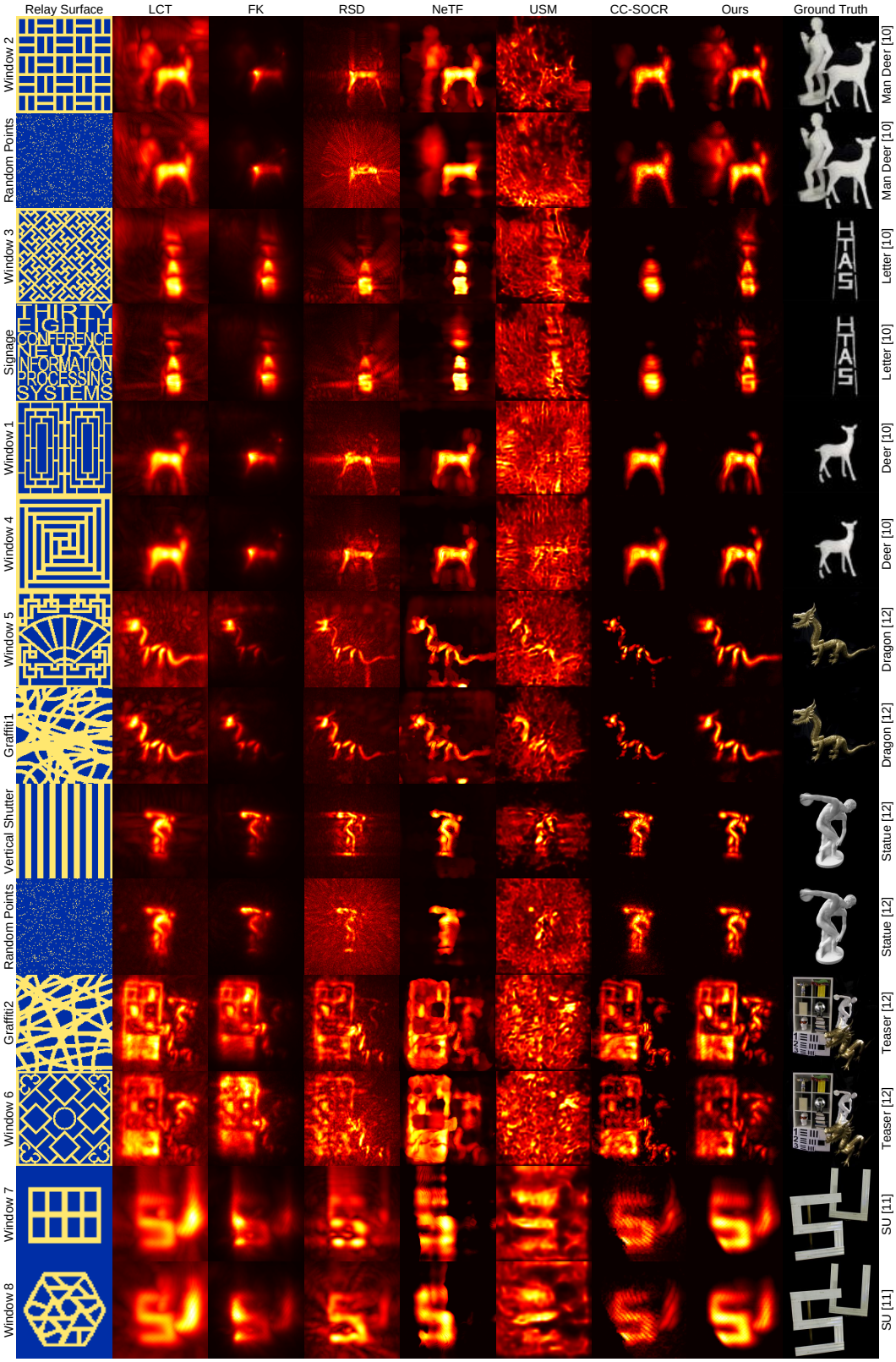

Figure 5: Reconstruction results of publicly available real-world dataset [10–12] with different relay surfaces.

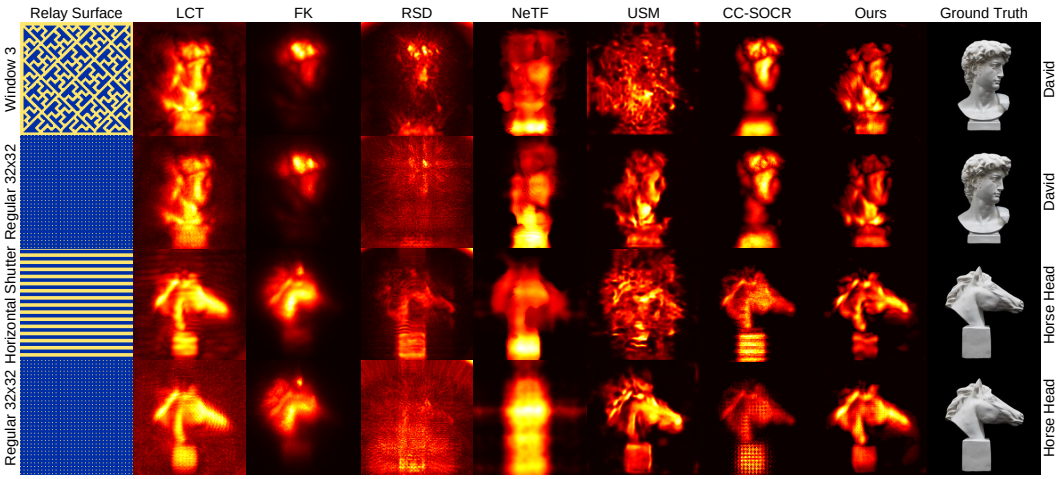

Figure 6: Reconstruction results of self-captured real-world dataset with different relay surfaces.

all the test cases, but loses fine structures at distant depth layers ("Man Deer", "Letter", "Deer" and "Teaser") and underperforms on glossy ("Dragon") or retroreflective ("SU") objects. Our method achieves the best quality for various objects and sampling patterns. Our method generates stable results on real-world data with diverse attributes and relay surfaces despite being trained only on a simple alphanumeric dataset and shutter-like relay surfaces. It notably outperforms the range-space solver LCT, successfully removing aliasing artifacts while preserving structure. The promising generalization highlights the effectiveness of our method in learning beyond the range space.

5.3 Inference time

Table 1: Inference time of various methods. The values in the table represent the average inference time across 16 IUT with size of $128 \times 128 \times 256$.

Method	LCT	FK	RSD	NeTF	USM	CC-SOCR	Ours
Runtime (CPU)	0.81 s	1.52 s	0.94 s	N/A	2.34 s	7.73 h	2.24 s
Runtime (GPU)	0.09 s	0.15 s	0.12 s	0.69 h	0.24 s	N/A	0.18 s

We compare the inference time of various methods on an Intel(R) Xeon(R) Platinum 8369B 2.90GHz CPU with 32 cores and an NVIDIA 3090 GPU, respectively. The inference times of NeTF and CC-SOCR vary with different IUT. Therefore, we average the inference times across 16 IUT shown in Fig. 5 and Fig. 6 each with a size of $128 \times 128 \times 256$. As shown in Tab. 1, the direct reconstruction methods, LCT, FK, and RSD, are faster than the other methods. Unlike these one-step approaches, CC-SOCR iteratively solves the functional model by a series of alternative sub-problems, requiring nearly eight hours to reconstruct an albedo. NeTF stands on the per-scene rendering framework and thus requires significantly longer inference times than the other two learning-based methods, USM, and our method. Note that both our method and CC-SOCR achieve more accurate reconstructions than the other methods. However, our method is 12,000$\times$ faster than CC-SOCR in CPU mode.

5.4 Ablation study

We validate the effectiveness of the two core components of our method: the virtual scanning process (VS) and the SURE-based denoiser. For quantitative evaluation, we simulated a dataset of 1,000 transients by rendering objects with complex geometries, such as chairs, clocks, guitars, sofas, and motorcycles. Our method and its two variants were tested on the simulated transients sampling with 15 different relay surfaces. For qualitative evaluation, we assess the two components using real-world datasets, "Teaser" and "Dragon". "Teaser" features complex structures, which helps evaluate VS's ability to recover details. "Dragon", with its glossy material, exhibits a lower signal-to-noise ratio in its acquired transient, which helps validate the SURE-based denoiser's effectiveness.

Table 2: Ablation study for the virtual scanning process and the SURE-based denoiser.

SURE-based denoiser	Virtual Scanning Process	PSNR (dB)
×	✓	18.69
✓	×	19.63
✓	✓	20.52

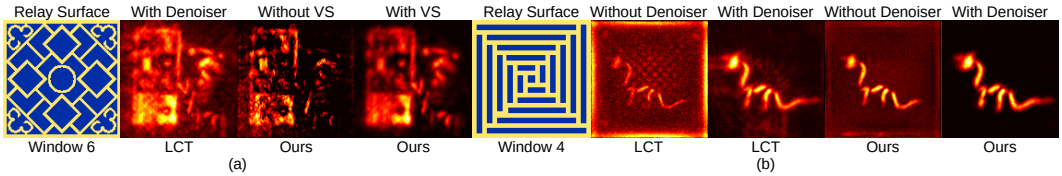

Figure 7: Ablation studies on the virtual scanning process (VS) (a) and the SURE-based denoiser (b).

Virtual Scanning As shown in Tab. 2, the virtual scanning process resulted in an average improvement of 0.89 dB. Fig. 7 demonstrates that without the VS component, our method can only recover a limited portion of the scene structure, which is impaired by aliasing artifacts. In contrast, our full model reconstructs more detailed and cleaner structures, confirming the effectiveness of VS component in recovering null-space components.

SURE-based denoiser As evident from the quantitative results, the SURE-denoiser achieved an average performance improvement of 1.83 dB, which is more significant than that of VS. This result aligns with expectations because background noise affects the entire reconstruction volume, while aliasing artifacts caused by irregular undersampling disrupt the main structure. For both LCT and our method, as illustrated in Fig. 7(b), the denoiser effectively suppresses noise artifacts. In Sec. 5.2, we apply the denoiser to competing methods to enhance their robustness to noise for a fair comparison. This demonstrates the versatility of the denoiser as a plug-and-play module in other NLOS algorithms.

Additional ablation studies on the hyperparameters within the loss functions and the relay surfaces used for training are provided in the supplementary materials.

6 Conclusion and limitations

Conclusion In this paper, we propose an unsupervised learning-based framework for NLOS imaging from irregularly undersampled transients (IUT). By introducing a virtual scanning process and a SURE-based denoiser, our framework achieves high-quality and fast NLOS reconstructions from IUT. Furthermore, it can be trained solely from noisy IUT, enabling future work on direct learning from real-world datasets to bridge the gap between simulated and real datasets. Our method outperforms the state-of-the-art method on both simulated and real-world data with various relay surfaces. In future work, we will extend our method to non-confocal imaging systems for more practical applications.

Limitations Our method faces two main limitations. Firstly, it is constrained to the confocal system due to the lack of a high-speed, low-memory non-confocal forward operator for deep learning training. However, the framework is theoretically extendable to general forward operators, indicating future research on new forward operators in NLOS imaging may not require special consideration for irregular undersampling. Secondly, while our method is entirely unsupervised, we only present the results of our method which is trained using simulated transients due to the time-consuming nature of collecting real transient datasets. Nonetheless, this approach partially mitigates the data gap introduced by simulated ground truth, and the results demonstrate excellent performance. Transitioning to deep learning that exclusively uses real transients is one of our future goals.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grants 62231018, 62071322, 62072331.

References

[1] D. Faccio, A. Velten, and G. Wetzstein, "Non-line-of-sight imaging," *Nature Reviews Physics*, vol. 2, no. 6, pp. 318–327, 2020.

[2] S. Chan, R. E. Warburton, G. Gariepy, J. Leach, and D. Faccio, "Non-line-of-sight tracking of people at long range," *Optics express*, vol. 25, no. 9, pp. 10 109–10 117, 2017.

[3] M. Isogawa, Y. Yuan, M. O'Toole, and K. M. Kitani, "Optical non-line-of-sight physics-based 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7013–7022.

[4] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. Dickmann, K. Dietmayer, B. Sick *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2068–2077.

[5] C. Wu, J. Liu, X. Huang, Z.-P. Li, C. Yu, J.-T. Ye, J. Zhang, Q. Zhang, X. Dou, V. K. Goyal *et al.*, "Non-line-of-sight imaging over 1.43 km," *Proceedings of the National Academy of Sciences*, vol. 118, no. 10, p. e2024468118, 2021.

[6] X. Liu, J. Wang, L. Xiao, Z. Shi, X. Fu, and L. Qiu, "Non-line-of-sight imaging with arbitrary illumination and detection pattern," *Nature Communications*, vol. 14, no. 1, p. 3230, 2023.

[7] J. Grau Chopite, M. B. Hullin, M. Wand, and J. Iseringhausen, "Deep non-line-of-sight reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 960–969.

[8] W. Chen, F. Wei, K. N. Kutulakos, S. Rusinkiewicz, and F. Heide, "Learned feature embeddings for non-line-of-sight imaging and recognition," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 6, pp. 1–18, 2020.

[9] F. Mu, S. Mo, J. Peng, X. Liu, J. H. Nam, S. Raghavan, A. Velten, and Y. Li, "Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] Y. Li, J. Peng, J. Ye, Y. Zhang, F. Xu, and Z. Xiong, "Nlost: Non-line-of-sight imaging with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 313–13 322.

[11] 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, 2018.

[12] D. B. Lindell, G. Wetzstein, and M. O'Toole, "Wave-based non-line-of-sight imaging using fast fk migration," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–13, 2019.

[13] X. Liu, I. Guillén, M. La Manna, J. H. Nam, S. A. Reza, T. Huu Le, A. Jarabo, D. Gutierrez, and A. Velten, "Non-line-of-sight imaging using phasor-field virtual wave optics," *Nature*, vol. 572, no. 7771, pp. 620–623, 2019.

[14] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6800–6809.

[15] S. I. Young, D. B. Lindell, B. Girod, D. Taubman, and G. Wetzstein, "Non-line-of-sight surface reconstruction using the directional light-cone transform," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1407–1416.

[16] M. La Manna, F. Kine, E. Breitbach, J. Jackson, T. Sultan, and A. Velten, "Error backprojection algorithms for non-line-of-sight imaging," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1615–1626, 2018.

[17] B. Ahn, A. Dave, A. Veeraraghavan, I. Gkioulekas, and A. C. Sankaranarayanan, "Convolutional approximations to the general non-line-of-sight imaging operator," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7889–7899.

[18] X. Liu, J. Wang, Z. Li, Z. Shi, X. Fu, and L. Qiu, "Non-line-of-sight reconstruction with signal–object collaborative regularization," *Light: Science & Applications*, vol. 10, no. 1, p. 198, 2021.

[19] M. La Manna, J.-H. Nam, S. A. Reza, and A. Velten, "Non-line-of-sight-imaging using dynamic relay surfaces," *Optics express*, vol. 28, no. 4, pp. 5331–5339, 2020.

[20] C. Gu, T. Sultan, K. Masumnia-Bisheh, L. Waller, and A. Velten, "Fast non-line-of-sight imaging with non-planar relay surfaces," in *2023 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2023, pp. 1–12.

[21] C. A. Metzler, D. B. Lindell, and G. Wetzstein, "Keyhole imaging: non-line-of-sight imaging and tracking of moving objects along a single optical path," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1–12, 2020.

[22] M. Isogawa, D. Chan, Y. Yuan, K. Kitani, and M. O'Toole, "Efficient non-line-of-sight imaging from transient sinograms," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Springer International Publishing, 2020, pp. 193–208.

[23] J.-T. Ye, X. Huang, Z.-P. Li, and F. Xu, "Compressed sensing for active non-line-of-sight imaging," *Optics Express*, vol. 29, no. 2, pp. 1749–1763, 2021.

[24] W. Yang, C. Zhang, W. Jiang, Z. Zhang, and B. Sun, "None-line-of-sight imaging enhanced with spatial multiplexing," *Optics Express*, vol. 30, no. 4, pp. 5855–5867, 2022.

[25] X. Liu, J. Wang, L. Xiao, X. Fu, L. Qiu, and Z. Shi, "Few-shot non-line-of-sight imaging with signal-surface collaborative regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 303–13 312.

[26] S. Shen, Z. Wang, P. Liu, Z. Pan, R. Li, T. Gao, S. Li, and J. Yu, "Non-line-of-sight imaging via neural transient fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2257–2268, 2021.

[27] J. Wang, X. Liu, L. Xiao, Z. Shi, L. Qiu, and X. Fu, "Non-line-of-sight imaging with signal superresolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 420–17 429.

[28] Y. Li, Y. Zhang, J. Ye, F. Xu, and Z. Xiong, "Deep non-line-of-sight imaging from under-scanning measurements," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] J. Schwab, S. Antholzer, and M. Haltmeier, "Deep null space learning for inverse problems: convergence analysis and rates," *Inverse Problems*, vol. 35, no. 2, p. 025008, 2019.

[30] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016.

[31] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly *et al.*, "Deep generative adversarial networks for compressed sensing automates mri," *arXiv preprint arXiv:1706.00051*, 2017.

[32] D. Chen and M. E. Davies, "Deep decomposition learning for inverse imaging problems," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 510–526.

[33] A. Pajot, E. De Bézenac, and P. Gallinari, "Unsupervised adversarial image reconstruction," in *International conference on learning representations*, 2018.

[34] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "Rare: Image reconstruction using deep priors learned without groundtruth," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1088–1099, 2020.

[35] B. Yaman, S. A. H. Hosseini, S. Moeller, J. Ellermann, K. Uğurbil, and M. Akçakaya, "Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data," *Magnetic resonance in medicine*, vol. 84, no. 6, pp. 3172–3191, 2020.

[36] J. Tachella, D. Chen, and M. Davies, "Unsupervised learning from incomplete measurements for inverse problems," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4983–4995, 2022.

[37] A. Bora, E. Price, and A. G. Dimakis, "Ambientgan: Generative models from lossy measurements," in *International conference on learning representations*, 2018.

[38] Y. C. Eldar, "Generalized sure for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, 2008.

[39] F. Luisier, T. Blu, and M. Unser, "Image denoising in mixed poisson–gaussian noise," *IEEE Transactions on Image Processing*, p. 696–708, Mar 2011. [Online]. Available: http://dx.doi.org/10.1109/tip.2010.2073477

[40] Y. Le Montagner, E. D. Angelini, and J.-C. Olivo-Marin, "An unbiased risk estimator for image denoising in the presence of mixed poisson–gaussian noise," *IEEE Transactions on Image processing*, vol. 23, no. 3, pp. 1255–1268, 2014.

[41] D. Chen, J. Tachella, and M. E. Davies, "Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5647–5656.

[42] S. Soltanayev and S. Y. Chun, "Training deep learning based denoisers without ground truth data," *Advances in neural information processing systems*, vol. 31, 2018.

[43] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[44] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[45] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[46] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

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

[49] S. Ramani, T. Blu, and M. Unser, "Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms," *IEEE Transactions on image processing*, vol. 17, no. 9, pp. 1540–1554, 2008.

[50] C. A. Metzler, A. Mousavi, R. Heckel, and R. G. Baraniuk, "Unsupervised learning with stein's unbiased risk estimator," *arXiv preprint arXiv:1805.10531*, 2018.

[51] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of ai," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 088–30 095, 2020.

A Network architecture

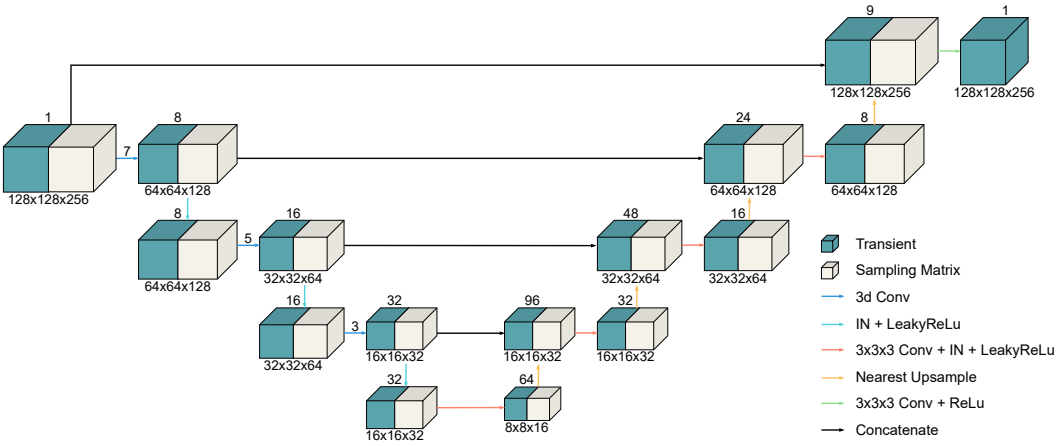

Figure 8: Overview of our P-Unet with 3D partial convolution for denoising irregularly undersampled transients. The spatial resolution of transients features in each layer is depicted below the blocks, and their respective channel counts are presented above the blocks. The resolutions and channels of the sampling matrices are identical to the transients features in the corresponding layer. The numbers on the blue arrows indicate the kernel size of the 3D convolution.

We begin by presenting the architecture of the network F_ϕ. Since the vanilla convolution treats all voxels of the zero-padded IUT as valid, it leads to the distortion of information in undersampled transients during feature propagation. As illustrated in Fig. 8, we present an encoder-decoder network (P-Unet) based on 3D partial convolution [44]. Partial convolution utilizes a sampling matrix to differentiate between valid and invalid voxels, effectively capturing spatial information from the transients. Additionally, we incorporate the instance normalization (IN) layer [45], which is insensitive to input distribution. This layer enhances the network's generalization capability to diverse zero-padded IUT with varying distributions from different sampling rates.

Given that the performance of F_θ in VSRnet mainly depends on acquiring null space information from the training forward operators, we adopt a 3D residual network featuring an attention-gated network [46]. While attention mechanisms might not significantly improve reconstruction quality, they help speed up the network's training process.

B Training details

Our method is implemented using PyTorch [47], and we employ the Adam optimizer [48] with a weight decay of 10^{-8}. In the first stage, the SURE-based denoiser model F_ϕ is trained with a batch size of 4 for 40 epochs. We set the initial learning rate to 1×10^{-3} and reduce it by a factor of 0.1 at epoch 30. Subsequently, in the second stage, the VSRnet model F_θ is trained with a batch size of 2 for 20 epochs, utilizing an initial learning rate of 5×10^{-4} and a reduction by a factor of 0.1 at epoch 10. In each epoch, we randomly select 40 complete simulated transients for each relay surface and extract signals from them to generate irregularly undersampled transients for training. All models were trained on 2 NVIDIA 3090 GPUs, taking nearly 40 hours in total. Regarding the loss function, the hyperparameters ε, β and b are set to 0.1, 0.001 and 4, respectively.

C Details of proposed SURE-based denoiser

C.1 Results of compared methods using our denoiser

We applied the proposed SURE-based denoiser to the compared methods and selected relatively better results for comparative experiments. As depicted in Fig. 9, the results of LCT, FK, and RSD exhibit reduced diffuse background noise when the denoiser is applied. For learning-based algorithms, our denoiser effectively removes cluster artifacts caused by measurement noise, thereby enhancing the

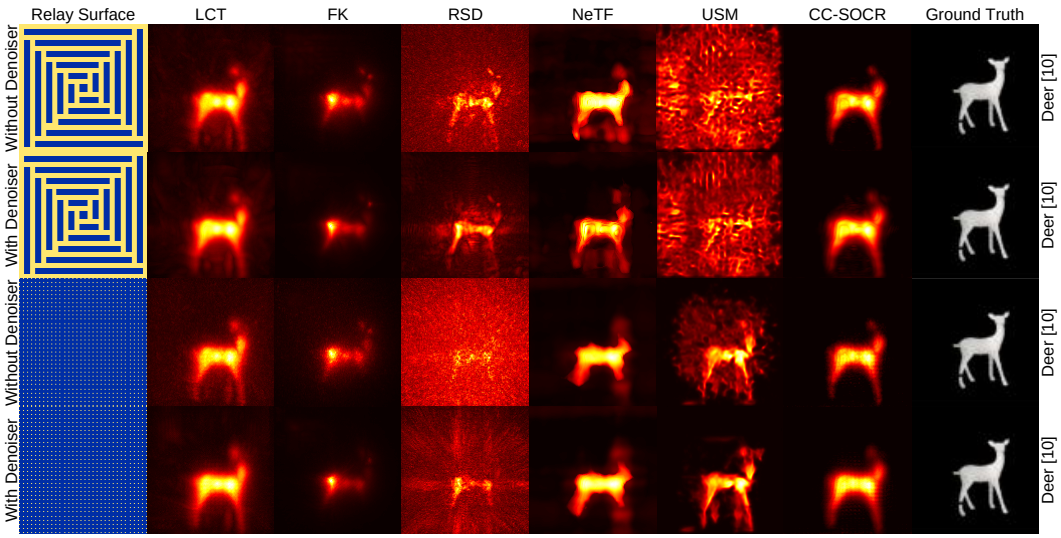

Figure 9: Qualitative results of compared methods with and without the SURE-based denoiser.

clarity of primary structures. However, in the case of CC-SOCR, which already incorporates strong regularization, the improvement of our denoiser is limited. This is likely due to the algorithm's sub-problem already including a denoising effect.

C.2 Derivation of the SURE loss

Stein's Unbiased Risk Estimation (SURE) framework has been developed and used in prior signal denoising works [38–42, 49]. These works did not consider the dark current of sensor outputs that might be removed in some image signal pipelines. Therefore, we extend the SURE denoising model to NLOS imaging by incorporating more practical noise characteristics of the detector. In this section, we provide the derivation of the SURE loss mentioned in the main text, following the notations in [40, 41].

For compact presentation, we simply denote the training set of transients by $\{\tilde{u}_j \in \mathbb{R}^{st}, j = 1, 2, \ldots, I \times G\}$ instead of the two-subscript version $\{\tilde{u}_{i,g}, i = 1, 2, \ldots, I, g = 1, 2, \ldots, G\}$. I is the number of observed hidden scenes, and G is the number of forward operators associated with relay surfaces. s and t are the number of scanning points and the number of time bins in each histogram of the transient, respectively.

In the supervised learning, given a paired dataset of noisy transients $\{\tilde{u}_j\}$ and its clean version $\{u_j\}$, the model F_ϕ can be trained by minimizing the mean squared error (MSE) loss function:

$$\mathbb{E}_{\{\tilde{u},u\}}\Big\{\sum_{j=1}^{J}\frac{1}{st}\|u_j - F_\phi(\tilde{u}_j)\|^2\Big\} = \mathbb{E}_u\Big\{\sum_{j=1}^{J}\frac{1}{st}\mathbb{E}_{\tilde{u}|u}\|u_j - F_\phi(\tilde{u}_j)\|^2\Big\}. \tag{10}$$

However, the clean transient set $\{u_j\}$ is not available. To achieve unsupervised learning only with noisy transients, our goal is to obtain an unbiased estimator of the mean squared error (MSE) standing on the SURE framework. To this end, we can further decompose the inner expectation in Eq. 10 as

$$\mathbb{E}_{\tilde{u}|u}\|u_j - F_\phi(\tilde{u}_j)\|^2 = \mathbb{E}_{\tilde{u}|u}\big\{\|u_j\|^2\big\} + \mathbb{E}_{\tilde{u}|u}\big\{\|F_\phi(\tilde{u}_j)\|^2\big\} - 2\mathbb{E}_{\tilde{u}|u}\big\{u_j^\top F_\phi(\tilde{u}_j)\big\}. \tag{11}$$

The unbiased estimator of the second term in Eq. 11 is $\|F_\phi(\tilde{u}_j)\|^2$, which can be obtained without the clean transient u_j. The first term in Eq. 11 can be rewritten as $\mathbb{E}_{\tilde{u}|u}\big\{u_j^\top \tilde{u}_j\big\}$ because $\mathbb{E}_{\tilde{u}|u}\big\{\tilde{u}_j\big\} = u_j$. The first and third terms depend on u. Thus, we need to obtain unbiased estimators for them by considering the characteristics of the noise model .

In the case of NLOS imaging, the noisy transients can be modeled as $\tilde{u} \sim \text{Poisson}(u + b)$, where b denotes the dark counts. The unsupervised expressions of $\mathbb{E}_{\tilde{u}|u}\big\{u_j^\top \tilde{u}_j\big\}$ and $\mathbb{E}_{\tilde{u}|u}\big\{u_j^\top F_\phi(\tilde{u}_j)\big\}$ can be obtained using the following lemma.

***Lemma* 1 (Lemma 1.2 in [40])** *Let $v \in \mathbb{R}^{st}$ such that $v \sim Poisson(u)$ be an independent random variables and let $\Phi : \mathbb{R}^{st} \to \mathbb{R}^{st}$ be a function such that $\mathbb{E}_{v|u}\{|\Phi_m(v)|\} < +\infty$ for all m.*

$$\mathbb{E}_{v|u}\{u^\top \Phi(v)\} = \mathbb{E}_{v|u}\{v^\top \Phi^{[-1]}(v)\}. \tag{12}$$

Let $\Phi(v) = F_\phi(\tilde{u})$, then $\mathbb{E}_{\tilde{u}|u}\{u_j^\top F_\phi(\tilde{u}_j)\}$ can be expressed as:

$$\begin{aligned}
&\mathbb{E}_{\tilde{u}|u}\{u_j^\top F_\phi(\tilde{u}_j)\} \\
&= \mathbb{E}_{v|u}\{u_j^\top \Phi(v_j)\} \\
&= \mathbb{E}_{v|u}\{(u_j + b)^\top \Phi(v_j)\} - \mathbb{E}_{v|u}\{b^\top \Phi(v_j)\} \\
&= \mathbb{E}_{v|u}\{v_j^\top \Phi^{[-1]}(v_j)\} - \mathbb{E}_{v|u}\{b^\top \Phi(v_j)\} \\
&= \mathbb{E}_{\tilde{u}|u}\{\tilde{u}_j^\top F_\phi^{[-1]}(\tilde{u}_j)\} - \mathbb{E}_{\tilde{u}|u}\{b^\top F_\phi(\tilde{u}_j)\}.
\end{aligned} \tag{13}$$

Here, we use the first-order Taylor approximation $F_\phi^{[-1]}(\tilde{u}_j) \approx F_\phi(\tilde{u}_j) - \partial F_\phi(\tilde{u}_j)$ to simplify Eq. 13 into:

$$\mathbb{E}_{\tilde{u}|u}\{\tilde{u}_j^\top F_\phi(\tilde{u}_j)\} - \mathbb{E}_{\tilde{u}|u}\{\tilde{u}_j^\top \partial F_\phi(\tilde{u}_j)\} - \mathbb{E}_{\tilde{u}|u}\{b^\top F_\phi(\tilde{u}_j)\}. \tag{14}$$

Then, the unbiased estimator of Eq. 14 is

$$\tilde{u}_j^\top F_\phi(\tilde{u}_j) - \tilde{u}_j^\top \partial F_\phi(\tilde{u}_j) - b^\top F_\phi(\tilde{u}_j). \tag{15}$$

Note that it is difficult to obtain an analytic form of $\partial F_\phi(\tilde{u}_j)$ when $F_\phi(\cdot)$ is a neural network. We adopt the Monte-Carlo approach [49] to obtain an estimate of the divergence of $F_\phi(\tilde{u})$ with the following approximation:

$$div_{\tilde{u}}\{F_\phi(\tilde{u})\} \approx \frac{e^\top}{\varepsilon}(F_\phi(\tilde{u} + \varepsilon e) - F_\phi(\tilde{u})), \tag{16}$$

where ε is a small positive number, and $e \in \{-1, 1\}^{st}$ is a binary vector whose entities follow a Bernoulli distribution with equal probability [40].

Similarly, let $\Phi(v)$ be an identity function such that $\Phi(v) = v$, and $v = \tilde{u}$. Using the first-order Taylor approximation, $\mathbb{E}_{\tilde{u}|u}\{u_j^\top \tilde{u}_j\}$ can be expressed as

$$\begin{aligned}
&\mathbb{E}_{\tilde{u}|u}\{u_j^\top \tilde{u}_j\} \\
&= \mathbb{E}_{v|u}\{v_j^\top \Phi^{[-1]}(v_j)\} - \mathbb{E}_{v|u}\{b^\top \Phi(v_j)\} \\
&\approx \mathbb{E}_{v|u}\{v_j^\top (v_j - \partial v_j)\} - \mathbb{E}_{v|u}\{b^\top \Phi(v_j)\} \\
&= \mathbb{E}_{\tilde{u}|u}\{\tilde{u}_j^\top (\tilde{u}_j - \partial \tilde{u}_j)\} - \mathbb{E}_{\tilde{u}|u}\{b^\top \tilde{u}_j\}.
\end{aligned} \tag{17}$$

We obtain the unbiased estimator of Eq. 17:

$$\tilde{u}_j^\top (\tilde{u}_j - \mathbf{1}) - b^\top \tilde{u}_j, \tag{18}$$

where $\mathbf{1}$ is a vector consisting of st ones. With the above derivation, the total unbiased estimator of the MSE is given by:

$$\begin{aligned}
&\sum_{j=1}^{J} \frac{1}{st}\{\|\tilde{u}_j\|^2 - 2\tilde{u}_j^\top F_\phi(\tilde{u}_j) + \|F_\phi(\tilde{u}_j)\|^2 - \mathbf{1}^\top \tilde{u}_j - b^\top \tilde{u}_j \\
&+ 2b^\top F_\phi(\tilde{u}_j) + \frac{2}{\varepsilon}(e_j \odot \tilde{u}_j)^\top (F_\phi(\tilde{u}_j + \varepsilon e_j) - F_\phi(\tilde{u}_j))\}.
\end{aligned} \tag{19}$$

After rearrangement, Eq. 19 can be expressed as:

$$\begin{aligned}
&\sum_{j=1}^{J} \frac{1}{st}\{\|\tilde{u}_j - F_\phi(\tilde{u}_j)\|^2 - (\mathbf{1} + b)^\top \tilde{u}_j \\
&+ 2b^\top F_\phi(\tilde{u}_j) + \frac{2}{\varepsilon}(e_j \odot \tilde{u}_j)^\top (F_\phi(\tilde{u}_j + \varepsilon e_j) - F_\phi(\tilde{u}_j))\}.
\end{aligned} \tag{20}$$

Thus, we get the SURE loss function as described in the main text:

$$\begin{aligned}
\mathcal{L}_{\text{SURE}} = \mathbb{E}_{\{i,g\}}\{&\frac{1}{st}\|\tilde{u}_{i,g} - F_\phi(\tilde{u}_{i,g})\|_2^2 - \frac{1}{st}(\mathbf{1} + b)^\top \tilde{u}_{i,g} \\
&+ \frac{2}{st}b^\top F_\phi(\tilde{u}_{i,g}) + \frac{2}{st\varepsilon}(e_{i,g} \odot \tilde{u}_{i,g})^\top \big(F_\phi(\tilde{u}_{i,g} + \varepsilon e_{i,g}) - F_\phi(\tilde{u}_{i,g})\big)\}.
\end{aligned} \tag{21}$$

During the training process, we use zero-padded irregularly undersampled transients as input, but only consider valid values in the calculation of the SURE loss.

D Details of our NLOS system

We used a femtosecond fiber laser as a light source with a central wavelength of 1560 nm and a pulse repetition rate of 82 MHz. The light was split by a $99:1$ fiber coupler. The channel with 1% of the light was attenuated and sent into a fast photodetector, functioning as the start signal for the time-to-amplitude convertor (TAC). The other channel, with 99% of the light, was connected to a collimator by a piece of single mode fiber (SMF) and then travelled through a hole in the mirror. The beam was diffusively reflected by the relay surface, and the raster scanning was controlled by the beam-steering mirror. After diffusive reflections by the object and again by the relay surface, the echo light travelled back along the same optical path and was reflected to a second collimator. Finally, the echo light was detected by a fractal superconducting nanowire single-photon detector (SNSPD) coupled with SMF. The output voltage pulses from the fractal SNSPD were amplified by the RF amplifiers and input to the TAC as a stop signal for the coincidence counting. The transients were captured over a $0.8 \times 0.8\ m^2$ scanning area, with a size of $128 \times 128 \times 512$ and a bin width of 8 ps.

E Discussion

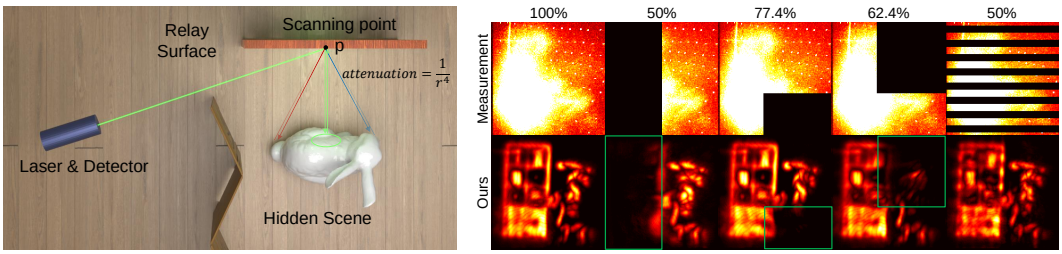

Figure 10: (a) The top view of the NLOS imaging system. The green oval region on the object indicates the area from which the scanning point can capture information. Light with different optical distances is represented by three colors. (b) The first row shows the maximum intensity projections of the transient along the time dimension, with the corresponding sampling rates indicated at the top.

In this section, we discuss how different relay surfaces affect reconstruction quality. Generally, a relay surface with a large missing area leads to result in poor reconstruction quality due to information loss. To verify this, we removed signals from the fully-sampled transient according to the relay surfaces and reconstructed them using our method. It is clear that parts of the hidden scene facing large missing areas are nearly impossible to reconstruct, as illustrated in the second, third, and fourth columns of Fig 10 (b). However, even with the same sampling rate, a more uniform distribution of scanning points produces better results (e.g., comparing the fifth column to the second).

Therefore, in the challenging task of NLOS imaging from IUT, the sampling rate is not the sole determinant of reconstruction quality. When the detector focuses a point on the relay surface, only the parts of the hidden scene that are close to that point can be observed. As shown in Fig 10 (a), the light attenuation term $1/r^4$ causes that a histogram $(1 \times 1 \times T)$ captured from scanning point p primarily contains information from hidden areas close to that point, as indicated by the green oval area. This results in poor reconstruction when large areas of the relay surface are invalid for scanning. However, determining the acceptable proportion of the invalid area relative to the entire surface for successful reconstruction is theoretically challenging, as it depends on various factors, including the relative depth between the hidden scene and the relay surface, the scene's reflectivity and normals, detector efficiency, and laser power. We consider this an important area for future work.

F Additional ablation study

In this section, we extend our ablation study further to explore the optimal values and sensitivity of the hyperparameters within the loss functions. Additionally, we examine the effect of relay surfaces used for training. To accommodate variations in testing data, we evaluated our method on different IUT of the "bunny", sampled according to 15 distinct relay surfaces mentioned in the main text.

F.1 Ablation study on hyperparameters

Table 3: Effect of the hyperparameters ε on the denoising performance in terms of PSNR (mean value and variance).

ε	0.0001	0.001	0.01
PSNR	39.11±1.99	40.19±2.91	41.16±2.61
ε	**0.1**	1	10
PSNR	**41.53 ± 1.25**	36.97±7.35	34.92±0.17

Effect of the positive number ε In our experiments, the transient is normalized to the range $[0, 100]$. Following the recommendation in [50], where the value of ε is suggested to be set around $\max(u)/1000$, we set ε to 0.1. As shown in Tab. 3, our SURE-based denoiser exhibits optimal performance when $\varepsilon = 0.1$.

Table 4: Effect of the hyperparameters β on the reconstruction performance in terms of PSNR (mean value and variance).

β	0	0.0001	0.0005
PSNR	16.19±0.21	17.36±0.45	18.25±0.77
β	**0.001**	0.005	0.01
PSNR	**19.03 ± 0.64**	18.39±0.81	18.03±0.86

Effect of the hyperparameter β The hyperparameter β acts as a weight to balance the MC loss and the VS loss. As shown in Tab. 4, our method achieves optimal performance when $\beta = 0.001$. The effectiveness of our method decreases notably as β decreases. When the VS loss is completely disabled ($\beta = 0$), the model fails to learn beyond the range space, highlighting the significance of virtual scanning in our approach.

F.2 Ablation study on relay surfaces for training

It's widely recognized that the robustness of neural networks depends heavily on the diversity of the training dataset. As discussed in [51], some neural networks struggle to adapt to changes in the sampling rate. They perform best when the sampling rate of input matches that of training dataset. Any deviation, whether a decrease or increase in the testing data's sampling rate, can lead to a decline in performance. To address this challenge, a common strategy is to augment the diversity of the training data's sampling rate.

Table 5: Effect of the intervals of relay surfaces for training on the reconstruction performance in terms of PSNR (mean value and variance).

Interval	[4]	[12]	[20]
PSNR	17.63±0.64	17.77±0.95	18.16±0.76
Interval	[8,16]	[4,12,20]	**[4,8,12,16,20]**
PSNR	18.36±1.05	18.85±0.88	**19.03 ± 0.64**

Interval In the context of NLOS imaging from irregularly undersampled transients, we can modify the shape of the relay surface to adjust the sampling rate of transients. In theory, light reflecting from the hidden scene spreads across the entire relay surface. However, due to the radiometric fall-off $\|p-q\|^4$, only a limited portion of the reflected light, concentrated in a small area of the relay surface, can be effectively captured by the time-resolved detector. As a result, we manipulate the interval of the shutter-like surfaces to modify the local sampling rate of undersampled transients.

As shown in Tab. 5, the optimal combination of interval values for relay surfaces is $[4, 8, 12, 16, 20]$. Our method's performance shows a decline with a reduction in the variety of interval types. This suggests that using undersampled transients with a more diverse range of sampling rates for training can enhance the generalization capability of our method.

Table 6: Effect of the rotations of relay surfaces for training on the reconstruction performance in terms of PSNR (mean value and variance).

Rotation	10	20	30
PSNR	18.08±0.65	18.33±1.05	18.74±1.32
Rotation	**40**	50	60
PSNR	**19.03 ± 0.64**	18.85±0.94	18.76±0.75

Rotation Similarly, we can modulate the diversity of relay surfaces by varying the number of rotations. As shown in Tab. 6, the optimal number of rotations is 40 when the count of transients using for training is 8,000 in total. When the number of rotations falls below 40, the diversity of relay surfaces becomes inadequate. Conversely, when it exceeds 40, the allocated transients for each relay surfaces becomes too limited.

G Additional results

We present additional results on publicly available real data [10, 12] with more irregular relay surfaces, as shown in Fig 11 and Fig 12. The experiment setup for data processing and the compared methods remains consistent with the description in the main text.

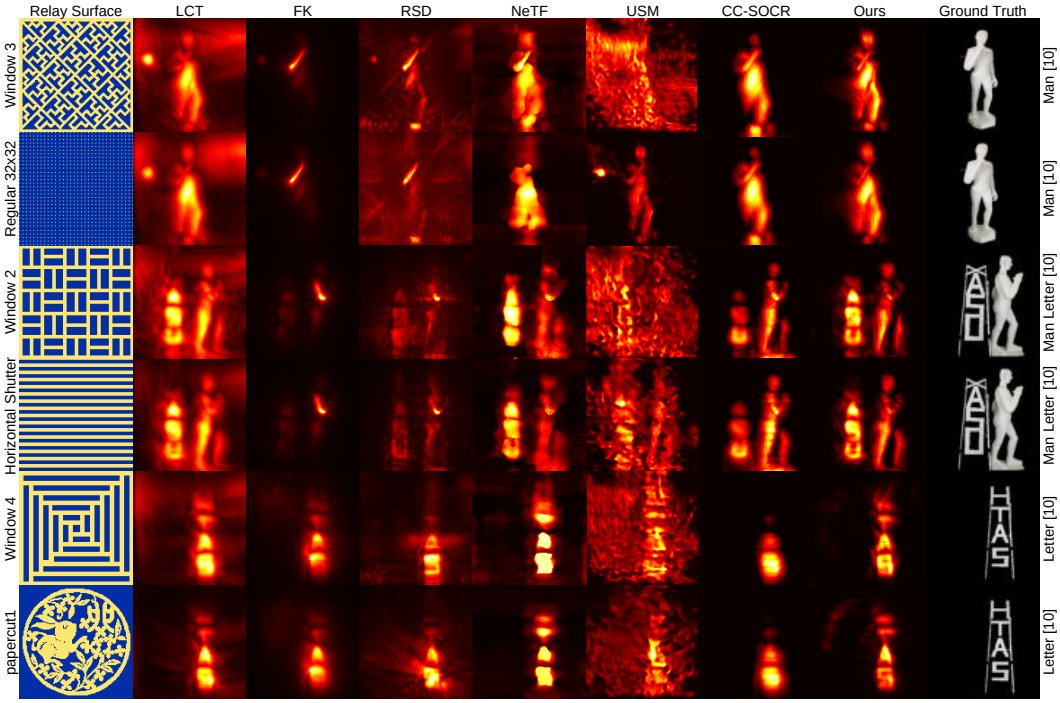

Figure 11: Reconstruction results of publicly available real-world dataset [10] with different relay surfaces.

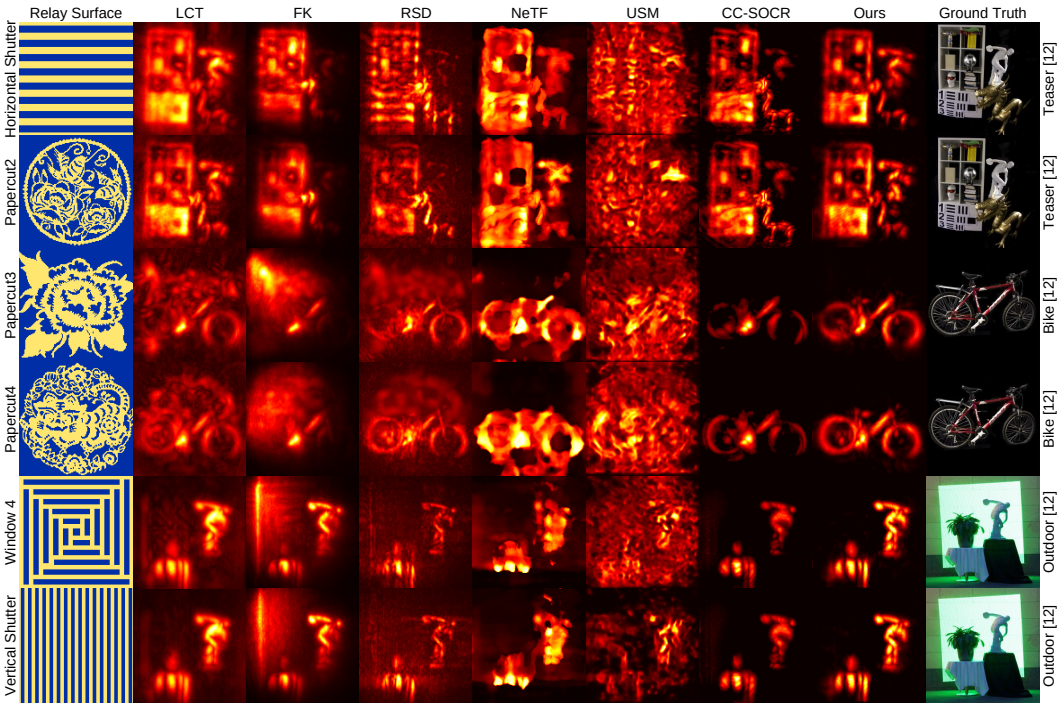

Figure 12: Reconstruction results of publicly available real-world dataset [12] with different relay surfaces.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make the main claims in introduction in Sec. 1. The claims matches theoretical results in Sec. 3.2 and experimental results in Sec. 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our paper discusses the limitations in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the analysis of the motivation in Sec. 3.2 and the virtual scanning strategy in Sec. 4.2. And we provide the derivation of the proposed SURE-based loss in Sec. C.2.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe proposed framework in Sec. 4. We present proposed network architecture in Sec. A and provide training details in Sec. B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The data and code will be released after publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and testing details in Sec. 5, Sec. B and Sec. F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The computational resources require for error bars are too high.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information on the computer resources in Sec. 5.3 and Sec. B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work on non-line-of-sight imaging does not have any apparent societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original paper that provided code package and dataset in references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The new assets, including the dataset and code, will be released with documentation after publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.