TransiT: Transient Transformer for Non-line-of-sight Videography

Ruiqian Li* Siyuan Shen* Suan Xia* Ziheng Wang Xingyue Peng
Chengxuan Song Yingsheng Zhu Tao Wu Shiying Li† Jingyi Yu†
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{lirq1, shensy2023, xiasa2022, wangzh1, pengxy2023, songchx2022, zhuys2022,
wutao, lishy1, and yujingyi}@shanghaitech.edu.cn

Abstract

High quality and high speed videography using Non-Line-of-Sight (NLOS) imaging benefit autonomous navigation, collision prevention, and post-disaster search and rescue tasks. Current solutions have to balance between the frame rate and image quality. High frame rates, for example, can be achieved by reducing either per-point scanning time or scanning density, but at the cost of lowering the information density at individual frames. Fast scanning process further reduces the signal-to-noise ratio and different scanning systems exhibit different distortion characteristics. In this work, we design and employ a new Transient Transformer architecture called TransiT to achieve real-time NLOS recovery under fast scans. TransiT directly compresses the temporal dimension of input transients to extract features, reducing computation costs and meeting high frame rate requirements. It further adopts a feature fusion mechanism as well as employs a spatial-temporal Transformer to help capture features of NLOS transient videos. Moreover, TransiT applies transfer learning to bridge the gap between synthetic and real-measured data. In real experiments, TransiT manages to reconstruct from sparse transients of 16×16 *measured at an exposure time of 0.4 ms per point to NLOS videos at a* 64×64 *resolution at 10 frames per second. Our code, demo, and dataset are available at* https://wangzh1.github.io/TransiT/.

1. Introduction

Human and machine vision capture dynamic information from the surrounding environment and interpret these variations to avoid hazards and make timely decisions. However, a moving object does not necessarily lie within the line-of-sight (LOS), e.g., a pedestrian at corners may be occluded by obstacles. The capabilities to detect and recognize hidden moving targets in scenarios as such are crucial in domain specific applications, ranging from robotics and autonomous driving to post-disaster rescue efforts.

We observe that hidden targets can potentially be captured using a classic non-line-of-sight (NLOS) setup, where an intermediate surface, either naturally existing or intentionally placed, serves as a relay wall [9, 10]. Fig. 1 illustrates a typical NLOS scenario under a confocal setting, which simplifies the classic setup from a five-dimensional (5D) configuration into 3D [27]. After a laser beam illuminates a point on the relay wall, a portion of the photons bounce off and scatter in a spherical wavefront onto a hidden object. A single-photon detector then records a transient, including the number of photons that arrive back at the same point over a specific time interval. Although this scheme is originally designed to recover static objects, it can be extended to reconstruct dynamic scenes by applying frame-by-frame methods. Lindell et al. [17] pioneer the reconstruction of NLOS videos at a resolution of 32×32 whereas Liu et al. [19] and their follow-up work [26] use a higher resolution scan of 181×131. These methods require a high sampling rate to ensure high resolution results, which significantly reduces the frame per second (FPS), leading to artifacts such as motion blur and discontinuity.

SPAD arrays allow multiple-pixel detection in a single shot and can improve the FPS; however, they face issues such as crosstalk, as well as high computation and memory costs [26, 29, 42]. Alternatively, several attempts aim to reduce per point scanning time by using a single-pixel SPAD-based setup [5, 24, 28]. Lindell et al. [17] raster scan a uniform grid of 64×64 at 2 FPS on a full relay wall of 2 m $\times$ 2 m. Isogawa et al. [14] enhance the FPS by adopting a circular scanning strategy within a small scanning range. Ye et al. [39] exploit transients of 16×16 and recover an NLOS video at 4 FPS by reducing the sampling density and incorporating plug-and-play (PnP) regularization and compressed sensing priors. To improve spatial resolution, recent works resort to learning-based approaches [16, 33] to upsample transients from 8×8 and 16×16 to 32×32. They

*These authors contributed equally to this work.
†Corresponding authors

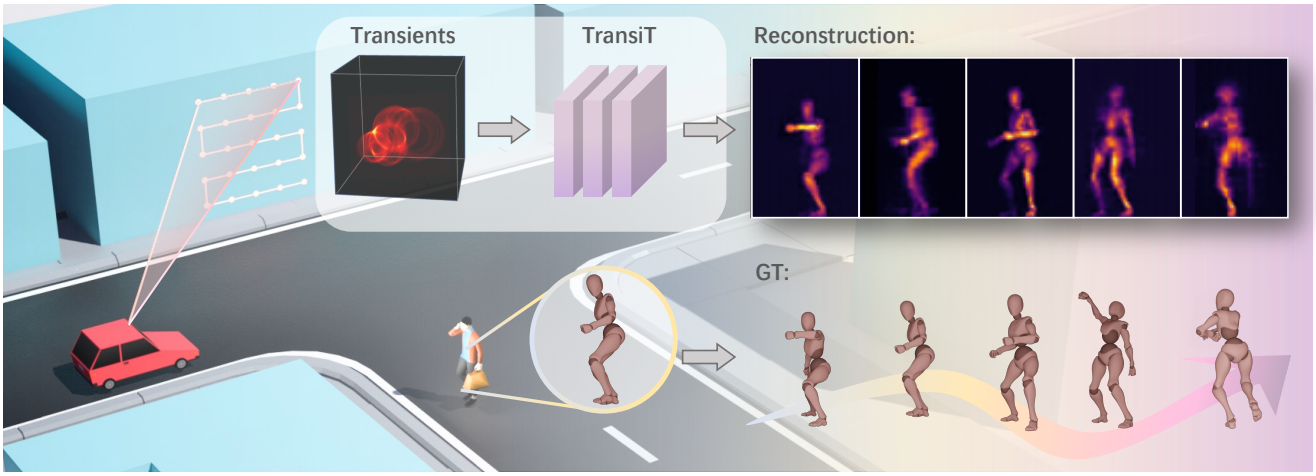

Figure 1. **NLOS videography.** An NLOS imaging system captures transients of a moving hidden object (e.g., a walking person) via a relay surface. TransiT is capable of reconstructing a high quality NLOS video of the person at 10 FPS using fast and sparse scanning.

are designed primarily for static NLOS reconstruction, and shows limited performance in dynamic scenarios.

We observe that videos captured by an NLOS system in the end are still videos. Recent video Transformers have demonstrated strong capabilities in understanding video contents and conducting feature extractions [3, 4, 8, 12, 22, 34]. Yet, unlike traditional videos, NLOS videos capture the transients of light that do not readily represent meaningful contents. In this paper, we aim to achieve NLOS videography by reconstructing dynamic frames with an initial scanning resolution of 16 $\times$ 16. We propose a Transient Transformer technique called TransiT to achieve real-time NLOS recovery under fast scans. Prior methods upsample sparse transients into virtual dense measurements for high resolution reconstruction [16, 33]. In contrast, TransiT directly compresses the temporal dimension of input transients to extract features, reducing computation costs and meeting high FPS requirements. However, this scheme introduces an issue for recovering fine details from sparse spatial-temporal signals. Therefore, TransiT designs a feature fusion mechanism to combine each frame with the features learned from the previous frame and then employs spatial-temporal attention via a Transformer to help capture the spatial-temporal features of NLOS transient videos.

To deploy our solution to real systems, we observe NLOS transients measured from a SPAD-based NLOS imaging system are also influenced by the image processing pipeline and hardware specifications, e.g., the laser, the galvanometer, and other devices. As a result, the measurements are often convolved with complex and heavy distortions, in particular in per point fast scanning. We thus formulate a new distortion model and construct a large-scale synthetic dataset of dynamic NLOS scenes. We select a variety of over 2K motion sequences from a public dataset [1] as well as include different motion styles, such as translation, rotation, body movements, etc. Our synthetic dataset contains nearly 200k frames of dynamic NLOS scenes. To bridge the gap between real and synthetic data, it is essential to align the features from both datasets. We design a maximum mean discrepancy (MMD)-based transfer learning method to fine-tune TransiT on real measurements, further enhancing its performance. To validate our technique, we conduct extensive experiments on both synthetic and real-measured datasets. The results demonstrate that TransiT enables us to achieve high quality NLOS videography, converting fast-scan frames at a raw spatial resolution of 16 $\times$ 16 to 64 $\times$ 64 at 10 FPS.

2. Related Work

Pioneered by the seminar works [11, 15, 32], time-resolved NLOS imaging has achieved remarkable progress. Back-projection (BP) based algorithms, such as filtered BP [5, 32] and fast BP [2], project each transient onto voxels of an NLOS scene and solve the inverse problem based on the ellipsoidal forward model. O'toole et al. [27] present the confocal setting to simplify the forward model from 5D into 3D and the light-cone transform (LCT) algorithm to solve the inverse problem as a 3D deconvolution process for fast reconstruction. The confocal setting and LCT become standard solution for NLOS imaging and benefit recent NLOS research [14, 16, 25, 30, 38, 43]. Because the signals detected are weak and are convolved into heavy noise of NLOS imaging system and process, the majority of approaches, including optimization-based [13, 31], wave-based [17–19], Fermat flow [38], and learning-based [6, 7, 30, 44], require dense transient measurements as input to reconstruct high resolution images of an NLOS scene. By treating each frame of a moving hidden object as static, these approaches are capable of recovering a complete NLOS video frame by frame [17, 26]. This video may be reconstructed in a low frame rate and without related information between adjacent frames, resulting in artifacts, such as motion blur and discontinuity. From either dense or sparse measurements, our framework enables us to learn related information of moving objects from neighboring frames and reconstruct spatial details of a single frame

and alleviate motion blur in NLOS videos.

For dynamic NLOS imaging, many attempts aim to accelerate transient measurement. Streak cameras can scan a line of area on the relay wall by a single shot with short exposure time [36, 37], but they are expensive for most practical applications. Owing to good balance in sensitivity and cost, SPAD cameras are the tool of choice to simultaneously record transients of multiple pixels [26, 29, 45]. However, SPAD cameras are currently under development and only support non-confocal configuration and limited pixels. Alternatively, single SPAD-based NLOS imaging systems can raster scan the relay wall in arbitrary sampling numbers and patterns [14, 17, 28]. For fast measurement, many works reduce scanning density [39], per-point exposure time [17], or scanning area [14]. Recently, Ye et al. [41] reach 4 FPS for complex dynamic real-life scenes. Through a keyhole, Metzler et al. [24] acquire transients of a moving object from a single light path and estimate the trajectory of the object. We exploit a tailored scanning strategy to reduce per point capture time and scanning density and achieve approximately 102 ms per frame of 16×16.

From sparse transients measured in per point scanning, recent optimization-based [20, 21, 40] and learning-based [16, 33] approaches attempt to reconstruct high resolution images of hidden objects. Using a superresolution network, Wang et al. [33] recover dense virtual measurements from sparse input and exploit existing algorithms for fast reconstruction. Their technique is potentially able to recover an NLOS video frame by frame. Ye et al. [39] reconstruct NLOS videos of moving hidden objects and filter image and motion blurring effects as noise. In contrast to these methods, we train TransiT on a large-scale synthetic dataset, which composes a variety of dynamic NLOS scenes, and leverage relevant information of diverse moving hidden objects for high quality NLOS videography.

3. Distortion Model Under Fast Scanning

To achieve NLOS videography at 10 FPS, we use a confocal imaging system [27] and scan a 16×16 grid on the relay wall by employing a serpentine scanning pattern, with approximately 0.4 ms scanning time per point.

Following the forward model [27], ideal transients $\tau(\bar{\mathbf{x}}_n, t)$ in Fig. 2(a) can be formulated as:

$$\tau(\bar{\mathbf{x}}_n, t) = \frac{1}{r^4} \iiint_{\Omega} \rho(\mathbf{x}) \cdot \delta(2\|\bar{\mathbf{x}}_n - \mathbf{x}\| - tc)\, d\mathbf{x}, \quad (1)$$

where Ω represents the NLOS space and ρ the albedo of the hidden scene at any point $\mathbf{x} = (x, y, z)$. The Dirac delta function δ converts the distance to time $t = 2\|\bar{\mathbf{x}}_n - \mathbf{x}\|/c$ from the illumination point to the hidden scene and back to the detection point, with c the speed of light.

Ideally, we assume that the laser beam moves instantly from point to point inside the scanning grid and can neglect

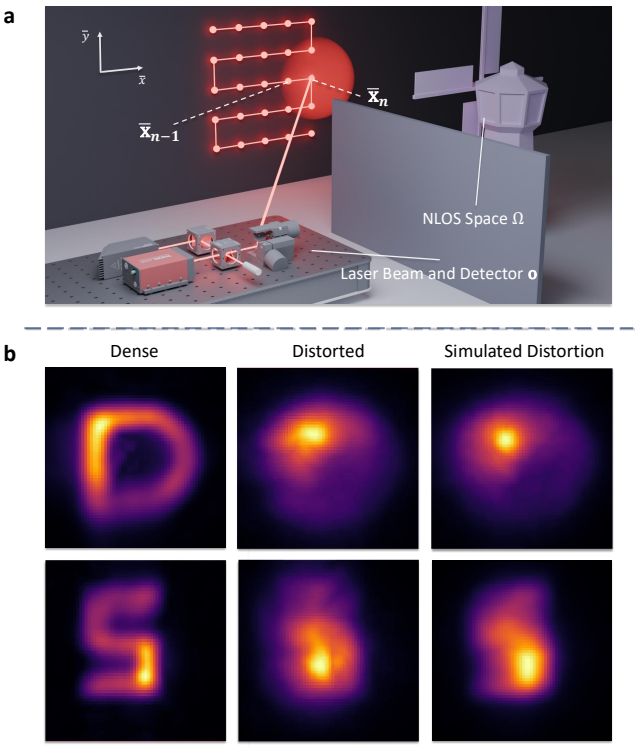

Figure 2. **Distortion model under fast scanning.** (a) System configuration for distortions under fast scanning. Due to the limited galvanometer's speed, illumination and detection scan on the relay wall in a linear form rather than at a single point. (b) Images reconstructed using f-k [17] from three different transients as input. Dense: 64×64 grid generated by applying our distortion model with 2 ms per point, distortion-free. Distorted: 16×16 grid real-measured with 0.4 ms per point. Simulated Distortion: 16×16 grid picked up from the Dense.

the time of its movement between scanning points. However, the direction of the laser beam is guided by a scanning galvanometer system, which has a minimum response time of ∼0.4 ms from when it receives a signal to when it completes the movement. In contrast to the ideal per-point scanning, this response time creates a continuous scanning path. Ideal transients $\hat{\tau}(\bar{\mathbf{x}}_n, t)$ recorded under fast scanning are therefore an integral of photon events illuminated along the path between adjacent points. Denote $\bar{\mathbf{x}}_n^m$ as the point on the wall located at $\bar{\mathbf{x}}_n$ being shifted a distance m towards the previous scanning point, we have:

$$\hat{\tau}(\bar{\mathbf{x}}_n, t) = \frac{1}{\|S\|} \int_S \tau\left(\bar{\mathbf{x}}_n^{\mathbf{s}}, t\right)\, d\mathbf{s}, \quad (2)$$

where $S = \bar{\mathbf{x}}_n - \bar{\mathbf{x}}_{n-1}$ represents the one-dimensional path between adjacent scanning points, $\frac{1}{\|S\|}$ the normalization term. We assume $\|S\| \neq 0$, otherwise $\hat{\tau}(\bar{\mathbf{x}}_n, t) = \tau(\bar{\mathbf{x}}_n, t)$. The serpentine scanning pattern restricts adjacent points to change along either $\bar{x}$- or $\bar{y}$-axis on the relay wall, allowing us to simplify this path integral to a single dimension.

Real measurement records two additional photon propagation paths compared to the ideal transients in Eq. 1: One from the laser to the relay wall and one from the relay wall to the detector. Treating the laser and detector

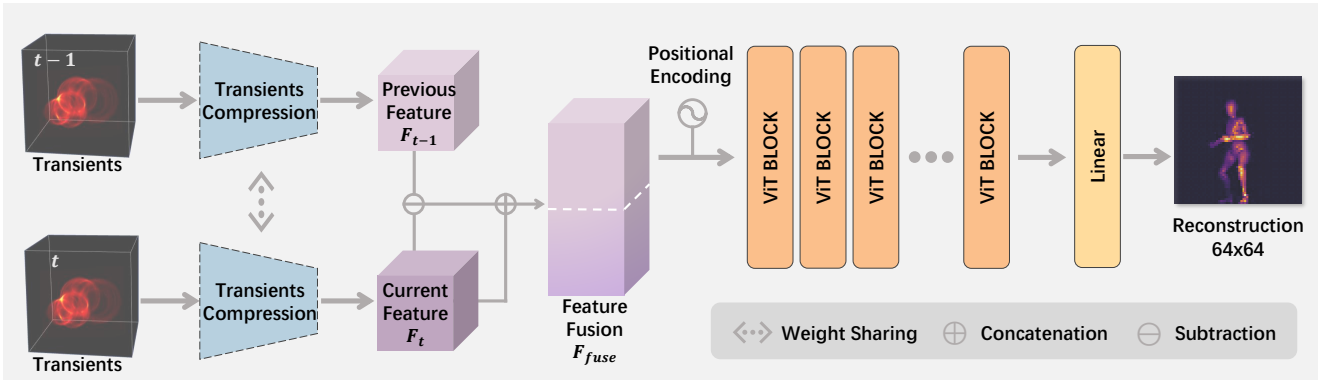

Figure 3. **Pipeline of TransiT.** TransiT is a Transformer-based architecture with the transients of current and previous frames as input. Transient compression extracts features by compressing the input transients along the temporal axis. Feature fusion combines the current frame's features with the difference between the current and previous frame features. ViT blocks with spatial-temporal attention then process the fused features. Followed by a linear layer, TransiT outputs a high resolution reconstruction.

as being at the same location $\mathbf{o} = (x_o, y_o, z_o)$, defining $\mathrm{d}_n^m = \|\bar{\mathbf{x}}_n^m - \mathbf{o}\|$ and $\mathrm{d}_n = \mathrm{d}_n^{\mathbf{0}}$, the real measurement is:

$$\tau_*(\bar{\mathbf{x}}_n, t) = \frac{1}{{\mathrm{d}_n}^2} \cdot \tau\left(\bar{\mathbf{x}}_n, t - \frac{2\mathrm{d}_n}{c}\right). \tag{3}$$

where $1/{\mathrm{d}_n}^2$ is the attenuation from diffuse reflection after photons arrive at the detection point. Combining Eq. 2 and Eq. 3, the real measurement under fast-scan scenario $\hat{\tau}_*(\bar{\mathbf{x}}_n, t)$ can be derived as:

$$\hat{\tau}_*(\bar{\mathbf{x}}_n, t) = \frac{1}{\|S\|} \int_S \left(\frac{\mathrm{d}_n^{\mathbf{s}}}{\mathrm{d}_n}\right)^2 \cdot \tau_*\left(\bar{\mathbf{x}}_n^{\mathbf{s}}, t + \frac{2(\mathrm{d}_n^{\mathbf{s}} - \mathrm{d}_n)}{c}\right) d\mathbf{s}, \tag{4}$$

where $\bar{\mathbf{x}}_n^{\mathbf{s}}$ represents a specific point on the path between $\bar{\mathbf{x}}_n$ and $\bar{\mathbf{x}}_{n-1}$, while $\mathrm{d}_n^{\mathbf{s}}$ is the directed line between $\bar{\mathbf{x}}_n^{\mathbf{s}}$ and the detector. We need to convert the real measurement to a scanning grid, Eq. 4 thus reveals the distortion introduced by the mismatching between the focal point and the scanning grid. Rewriting Eq. 4 using Eq. 3, the distortion in ideal transients with real measurement error becomes:

$$\hat{\tau}(\bar{\mathbf{x}}_n, t) = \frac{1}{\|S\|} \int_S \left(\frac{\mathrm{d}_n}{\mathrm{d}_n^{\mathbf{s}}}\right)^2 \cdot \tau\left(\bar{\mathbf{x}}_n^{\mathbf{s}}, t + \frac{2(\mathrm{d}_n - \mathrm{d}_n^{\mathbf{s}})}{c}\right) d\mathbf{s}. \tag{5}$$

This integral can then be discretized to:

$$\hat{\tau}(\bar{\mathbf{x}}_n, t) \approx \frac{1}{\|S\|} \sum_{i=1}^{M} \left(\frac{\mathrm{d}_n}{\mathrm{d}_n^{i\Delta\mathbf{s}}}\right)^2 \cdot \tau\left(\bar{\mathbf{x}}_n^{i\Delta\mathbf{s}}, t + \frac{2(\mathrm{d}_n - \mathrm{d}_n^{i\Delta\mathbf{s}})}{c}\right) \Delta\mathbf{s}, \tag{6}$$

where M and Δs represents the number of sampled points and distance between two adjacent scanning points (see Supplementary Materials for detailed derivations). Fig. 2 (b) shows the images reconstructed from distortion-free, real-world and simulated distorted data.

Eqs. 5 and 6 show the distortion model mapping from per-point scanned dense ideal transients to fast-scan sparse ideal transients. However, the inverse problem in this context is difficult to solve mathematically, i.e., we cannot recover undistorted transients from distorted. Therefore, when training TransiT, we generate synthetic data with high spatial resolution (e.g., 64×64) and intentionally introduce fast-scan distortions by applying Eq. 5, allowing TransiT to reconstruct the hidden scene from distorted data.

4. TransiT Architecture

Our TransiT tackles the challenges of high frame rate NLOS video reconstruction of dynamic scenes. The core innovation of TransiT lies in its ability to utilize the generative capabilities inherent in Transformer, enabling the recovery of NLOS video from low-resolution data with fast-scan distortion. Fig. 3 shows the overall pipeline of TransiT.

Transients compression. A crucial challenge in high frame rate NLOS Video reconstruction is balancing the need for high temporal resolution while preserving enough spatial detail for effective reconstruction. Unlike previous methods that rely on upsampling techniques to increase the resolution of input transients, our approach directly compresses the input transients using a linear layer to extract transients feature $\mathcal{F}$. Specifically, we compress the temporal axis of the input transients (represented as histograms) into a lower-dimensional feature space, significantly reducing the input complexity. Despite the original NLOS histograms spanning hundreds or even thousands of time bins, experiments demonstrate that compressing them into a 32-dimensional feature vector does not degrade performance. Intuitively, in dynamic and rapidly scanned NLOS scenes, the most valuable information from each histogram is often concentrated in one or a few peak positions and amplitudes.

Feature fusion. Another challenge in NLOS video reconstruction arises from the difficulty in capturing fine details of dynamic scenes from sparse data. To address this, we introduce a feature fusion technique that emphasizes temporal differences between consecutive frames, highlighting critical dynamic changes in the scene [35]. Specifically, we compute the difference between the features of the current frame transients $\mathcal{F}_t$ and the previous frame $\mathcal{F}_{t-1}$ and concatenate it with the current frame's features:

$$\mathcal{F}_{fuse} = concat(\mathcal{F}_t, \mathcal{F}_t - \mathcal{F}_{t-1}). \tag{7}$$

This fused feature $\mathcal{F}_{fuse}$ combines both the current state of the scene and the observed temporal changes, providing a richer representation for reconstruction.

Transformer with spatiotemporal attention. To model

the complex spatiotemporal dependencies in an NLOS video, we utilize Vision Transformer (ViT) blocks with spatial and temporal positional encodings. The spatial positional encoding $PE^{spatial}$ is two-dimensional, representing the x and y coordinates of the sampling points. Temporal positional encoding PE^{time}, on the other hand, is one-dimensional distinct frames of the transients sequence. Spatial and temporal positional encodings are combined via element-wise summation and subsequently applied to the fused feature $\mathcal{F}_{fuse}$. The resulting encoded feature is then passed through the Transformer blocks $\mathcal{B}$ with a simple linear layer to generate the final image I of current frame:

$$I = \mathcal{B}(\mathcal{F}_{fuse} + PE^{spatial} + PE^{time}). \tag{8}$$

Training. We use a two-stage training approach to optimize the model. In the first stage, we minimize the mean squared error (MSE) between the ground truth image I_{GT} and the predicted images from the network:

$$\mathcal{L}_{imaging} = \|I_{GT} - I\|^2. \tag{9}$$

A key challenge in NLOS video reconstruction is the distribution gap between synthetic and real-measured data, as system responses and noise levels vary across hardware platforms, and external factors can distort the data. To address this, we propose a second training stage that employs Maximum Mean Discrepancy (MMD) Loss which is used in transfer learning to measure the discrepancy between the distributions of synthetic and real-measured data. Specifically, we extract fused features from both synthetic and real-measured data to calculate the MMD loss between them:

$$\mathcal{L}_{MMD} = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi(\mathcal{F}^i_{real}) - \frac{1}{m}\sum_{j=1}^{m}\phi(\mathcal{F}^j_{syn}) \right\|^2, \tag{10}$$

where n and m represent the number of feature samples extracted from the real-measured and synthetic datasets, respectively. $\phi(\cdot)$ is the feature map corresponding to a Gaussian kernel. The total training loss for stage two combines MMD loss with image reconstruction loss:

$$\mathcal{L}_{total} = \mathcal{L}_{imaging} + \lambda\mathcal{L}_{MMD}, \tag{11}$$

where λ is a hyperparameter that controls the trade-off between reconstruction accuracy and domain alignment.

While this training strategy provides an effective solution to mitigate domain gaps, it remains an optional strategy rather than a mandatory requirement. In practical NLOS scenarios, discrepancies between synthetic and real-world settings may arise due to material properties, hardware characteristics, and noise artifacts. For instance, most simulations assume that both the relay surface and the hidden object exhibit ideal diffuse reflectance, whereas real-world conditions may introduce specular reflections. Additionally, factors such as laser source variations and SPAD-specific noise introduce distortions that are sometimes difficult to model precisely in simulation.

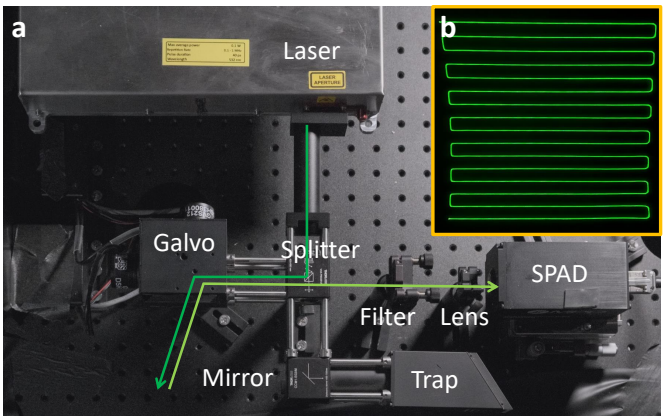

Figure 4. **System setup.** (a) Our NLOS imaging system, and (b) The fast scanning pattern.

A key advantage of MMD Loss in such cases is that it's self-supervised, requiring only real-world measurement without the need for corresponding ground truth, which is usually difficult to obtain. Moreover, MMD-based domain adaptation can achieve good performance with significantly fewer real-world samples than synthetic data, further enhancing its practical applicability. In cases where synthetic and real-world conditions are already well-aligned, this two-stage training strategy may be unnecessary. Further training details are provided in the next section.

5. Experiments

5.1. Experimental Setup

Synthetic dataset. Based on our distortion model, we construct a large scale synthetic dataset, which includes 10,000 motion sequences and 100,000 frames rendered from a diversity of objects, e.g., letters, windmills, propellers, and full-body and half-body of humans. The object sizes range from 1 m $\times$ 1 m to nearly 2 m $\times$ 2 m. The motion styles vary from simple translation and rotation to complex movements. Each motion sequence spans a duration of 5 seconds, with 10 FPS, yielding 50 consecutive frames. The motion speed is approximately 40 cm/s across motions. We exploit the motion sequences of full-body and half-body of humans from Mixamo [1], a human motion dataset with over 2,000 motion sequences. We generate the synthetic transients at spatial resolution of 64 $\times$ 64 and at temporal resolution of 20 ps by scanning on a 2 m $\times$ 2 m relay wall.

Real-measured data. We construct an NLOS imaging system, as illustrated in Fig. 4, under a confocal configuration. Light from the laser (Katana 05-HP, 532 nm, repetition rate 1 MHz) is directed by a 2D galvanometer (GVS012) and scan on the relay wall. A fast-gated SPAD with a 50 mm lens is coupled with a PicoHarp 300 and a delayer (PSD-MOD) to record transients at a time bin of 4 ps. The hardware is positioned 2 m away from the relay wall. The system is calibrated with a time jitter of 72 ps using the on-

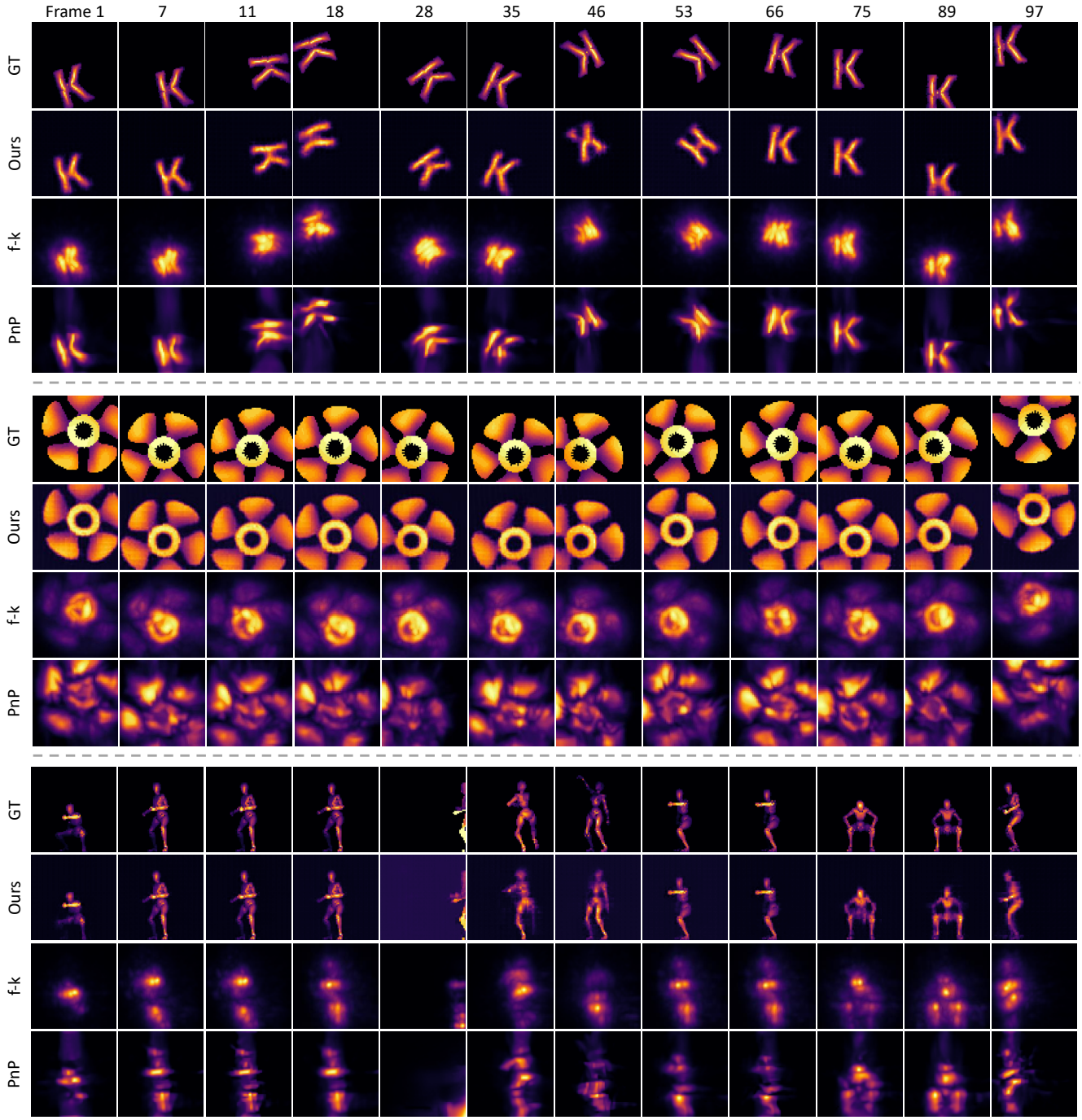

Figure 5. **Comparison of synthetic results.** From top to bottom: Ground truth, ours, f-k and PnP. The results are reconstructed across multiple frames of 16×16 of noisy synthetic data for different objects — a character 'K', a propeller, and a human.

site method [28]. Due to the mechanical speed limitations of the galvanometer, we adopt a serpentine scanning pattern (Fig. 4(b)) to minimize rapid large-angle deflections by optimizing traversal paths between adjacent scanning points. We scan a 16×16 grid to cover an area of 1 m $\times$ 1 m, with each point being scanned for 0.4 ms.

Implementation. We implement TransiT using PyTorch, employing the AdamW optimizer [23] with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay set to 0.01. The learning rate follows a cosine decay schedule, starting from 5×10^{-3} to 1×10^{-4}, with a linear warmup over the first 10 epochs. The input transients, originally of size $16 \times 16 \times T$, are first compressed to a 32-dimensional latent representation on T. Generally, higher-dimensional embedding enables richer representations and better performance. The choice of feature dimension was determined through balancing reconstruction accuracy and computational constraints. For training, we use a batch size of 64 per GPU and distribute the process across 24 NVIDIA A800 GPUs. The total number of training epochs is set to 1000, and it re-

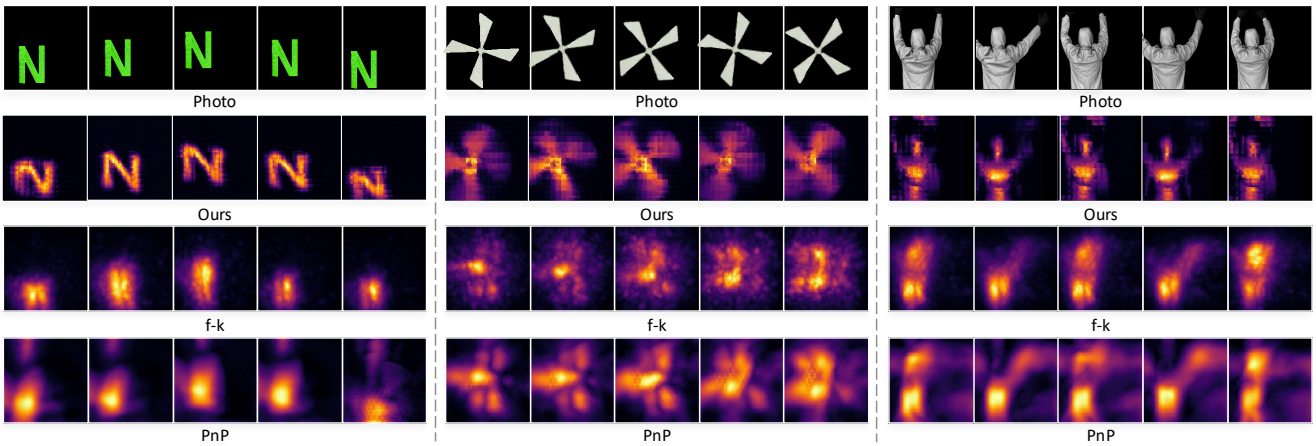

Figure 6. **Comparison of real-measured results.** From top to bottom: Ground truth, ours, f-k, and PnP. The results are reconstructed across multiple frames of 16×16 of real-measured data for different objects — a character 'N', a windmill, and a human.

Table 1. Comparison results. We compare the reconstruction performance of different moving objects using different algorithms (f-k, PnP, and ours), evaluated by Euclidean Distance (ED), Cosine Similarity (CS), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR).

Object	Method	ED↓	CS↑	SSIM↑	PSNR↑
Character	f-k	0.1286	0.7876	0.6677	17.86
	PnP	0.0923	0.8575	0.6764	20.77
	Ours	**0.0520**	**0.9418**	**0.9227**	**25.87**
Propeller	f-k	0.3180	0.6854	0.1902	10.01
	PnP	0.2707	0.7800	0.2818	11.39
	Ours	**0.0904**	**0.9781**	**0.8211**	**20.91**
Human	f-k	0.1136	0.6265	0.5791	19.00
	PnP	0.1018	0.6939	0.6706	19.92
	Ours	**0.0415**	**0.9272**	**0.8041**	**29.45**

quires approximately 24 hours to optimize with Eq. 9. For real experiments, we fine-tune TransiT for an additional 100 epochs using 200 real NLOS transient frames on 8 NVIDIA A800 GPUs and it requires approximately 2 hours to optimize with Eq. 11 with $\lambda = 0.01, n = 32, m = 64$. Since the domain gap varies with different systems, the optimal λ is setup-dependent. Our goal is to provide an adaptation strategy, and we encourage practitioners to tune λ based on their hardware. Training is conducted with mixed precision using PyTorch AMP to accelerate computations. For inference, we deploy the model on a single NVIDIA RTX 3090 GPU. Total per-frame latency is $\sim$106 ms ($\sim$10 FPS), including 102 ms for acquisition, $\sim$3.7 ms for arranging photon events, and $\sim$0.6 ms for reconstruction.

Comparison methods. We carry out experiments on the datasets using three approaches: our TransiT, f-k [17], and PnP [39]. To ensure a fair comparison, we standardize the input size to 16×16 and the output size to 64×64, and apply the identical distortion model across all methods. For the synthetic data, we apply our distortion model to convert distortion-free transients at 64×64 into distorted transients at 16×16. For real-measured data, we directly use our captured transients at 16×16. For f-k, we upsample the 16×16 input to 64×64 via interpolation, producing an output of 64×64. For PnP, we follow the preprocessing steps from their public code, padding the 16×16 spatial resolution to 64×64 before applying the algorithm.

5.2. Synthetic Results

Fig. 5 showcases the multi-frame reconstruction results of our synthetic data. From top to bottom, the sequences depict the character 'K', a propeller, and a full-body of human. Specifically, the motion of the character 'K' involves a combination of translation and rotation, while the propeller is randomly sampled from frames along different rotational sequences. The full-body human is similarly sampled from frames across various motion sequences.

Due to fast-scan distortions, f-k produces highly blurred reconstructions and fails to distinguish the object although it can track the dynamic position. PnP performs better for the character and full-body human, where shape and position are more discernible, but shows less satisfactory performance in the propeller. Our method, in contrast, achieves superior reconstructions by accounting for the system noise and fast-scan distortions during training, while leveraging the strong learning and representation capabilities of the Transformer. To quantitatively analyze the performance, we compare the results from different methods across various objects in terms of ED, CS, SSIM, and PSNR. As shown in Table 1, TransiT outperforms f-k and PnP. Additional results are included in Supplementary Materials.

5.3. Real-measured Results

Fig. 6 presents reconstruction results for three different real objects: a character 'N', a windmill, and a half-body human. The size of the planar character is 50 cm $\times$ 50 cm, with a movement speed of approximately 20 cm/s. The windmill has a diameter of around 80 cm, with a rotational speed of approximately 15°/s. In Fig. 6, TransiT demonstrates accurate shapes and positions of three moving objects. Due to the distortions, f-k and PnP produce highly blurred shapes of the objects although they can track the dynamic positions. Compared to the other two objects, the

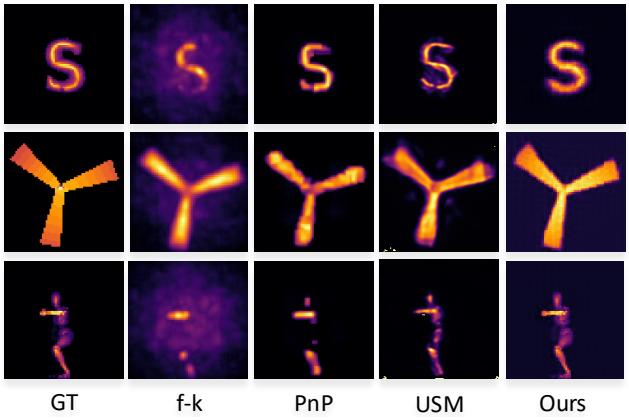

Figure 7. Ablation study. The results are reconstructed from transients of three static objects under sparse scanning.

windmill yields noticeably blurrier results mainly because of its higher rotational speed.

The results on both synthetic and real-measured datasets demonstrate that TransiT enables us to robustly handle complex and heavy distortions. We further conduct experiments on the real-measured data from f-k [17] and show comparison results in Supplementary Materials. Notably, some reconstructions exhibit block-like artifacts, which we attribute to material disparities between real objects and synthetic data. While our training dataset consists of diffuse surfaces, real-world objects often demonstrate retro-reflective properties. Addressing this limitation, potentially through integration of larger NLOS datasets with diverse materials or advanced NLOS foundation models, remains an avenue for future work.

5.4. Ablation Study

To assess performance of TransiT in the absence of distortions, we carry out an ablation study. We fine-tune TransiT on distortion-free data for 100 epochs and test it on transients of three static objects: A character 'S', a propeller, and a human. These transients are captured under sparse scanning. Fig. 7 shows the comparison results from four methods. f-k yields noisy results with noticeable artifacts whereas PnP obtains clearer results by incorporating a band-pass filtering and a video denoising network. USM [16] generates relatively clear reconstruction results in some cases, owing to its transient recovery network, whereas the reconstruction of the letter 'S' exhibits slight blur. In contrast, TransiT offers superior performance under the scenario without fast-scan distortions.

We further provide quantitative comparisons of four methods in terms of ED, CS, SSIM, and PSNR as shown in Table 2. Both qualitative and quantitative results demonstrate that our method outperforms the others in terms of metrics and achieves high quality reconstruction for either static or dynamic NLOS scenes.

Table 2. Quantitative results of ablation study. We compare the performance of four methods (f-k, PnP, USM, and ours) for different static objects in terms of ED, CS, SSIM, and PSNR.

Object	Method	ED↓	CS↑	SSIM↑	PSNR↑
Character	f-k	0.1352	0.6657	0.1224	17.37
	PnP	0.0763	0.8639	0.8852	22.34
	USM	0.0548	0.9317	0.9163	25.19
	Ours	**0.0531**	**0.9567**	**0.9261**	**25.49**
Propeller	f-k	0.1487	0.8456	0.2365	16.55
	PnP	0.1204	0.8926	0.7817	18.38
	USM	**0.1081**	0.9336	0.8016	**21.92**
	Ours	0.1174	**0.9556**	**0.8308**	18.61
Human	f-k	0.1525	0.4828	0.0957	16.34
	PnP	0.0754	0.6925	0.8515	22.49
	USM	0.0483	0.8641	0.9227	24.24
	Ours	**0.0241**	**0.9652**	**0.9361**	**26.39**

6. Discussion and Conclusion

We have presented TransiT, a new Transformer-based technique, to reconstruct high frame rate, high quality non-line-of-sight videos. This new solution allows us to capture sparse transients (16×16) at the exposure time of 0.4 ms per point using a single-pixel SPAD. Resultant transients involve distortions, which are convolved during the imaging and fast scanning processes. We have formulated a distortion model and presented effective solutions to practically deploy TransiT to real NLOS imaging systems. Comprehensive experiments show that TransiT outperforms the state-of-the-art on both reconstruction quality and frame rates. While our system currently operates at 10 FPS, we could theoretically achieve even higher frame rates without modifying the hardware or network structure. However, we selected 10 FPS to balance reconstruction quality and temporal consistency.

While TransiT has made significant progress, there remain a number of challenges in NLOS videography. First, fast scanning reduces per point exposure time and the scanning density, resulting in sparse and distorted transients. Existing techniques are difficult to denoise these distortions. A potential solution is to employ continuous scans as inputs for the reconstruction process, rather than sparse and grid-based data. This would require TransiT to be fine-tuned to cope with the new scanning process and data types. Second, for complex NLOS scenes, single-pixel SPADs require prolonged acquisition time to capture dense measurements with rich information, which limits the achievable frame rates. Although SPAD arrays are the tool in the future, they are currently constrained by the resolution and the sensitivity. Our TransiT was initially designed to single-point scanning but could be potentially extended to SPAD arrays for achieving even higher frame rates. These are our immediate future work and hopefully a major step towards real-time NLOS video reconstruction in practical applications.

Acknowledgments

The authors appreciate the anonymous reviewers and area chairs for their valuable comments. This work was supported in part by the National Natural Science Foundation of China under Grants W2431046 and 61977047, and by the MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence, and the HPC Platform of ShanghaiTech University.

References

[1] Mixamo, https://www.mixamo.com/#/. 2, 5

[2] Victor Arellano, Diego Gutierrez, and Adrian Jarabo. Fast back-projection for non-line of sight reconstruction. Optics Express, 25(10):11574–11583, 2017. 2

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6836–6846, 2021. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In ICML, page 4, 2021. 2

[5] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. Optics express, 23(16):20997–21011, 2015. 1, 2

[6] Wenzheng Chen, Fangyin Wei, Kiriakos N Kutulakos, Szymon Rusinkiewicz, and Felix Heide. Learned feature embeddings for non-line-of-sight imaging and recognition. ACM Transactions on Graphics, 39(6):1–18, 2020. 2

[7] Javier Grau Chopite, Matthias B Hullin, Michael Wand, and Julian Iseringhausen. Deep non-line-of-sight reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 960–969, 2020. 2

[8] Fan Deng-Ping, Ji Ge-Peng, Cheng Ming-Ming, Sakaridis Christos, and Van Gool Luc. Advances in deep concealed scene understanding. Visual Intelligence, 1(16), 2023. 2

[9] Daniele Faccio, Andreas Velten, and Gordon Wetzstein. Non-line-of-sight imaging. Nature Reviews Physics, 2(6): 318–327, 2020. 1

[10] Ruixu Geng, Yang Hu, Yan Chen, et al. Recent advances on non-line-of-sight imaging: Conventional physical models, deep learning, and new scenes. APSIPA Transactions on Signal and Information Processing, 11(1), 2021. 1

[11] Otkrist Gupta, Thomas Willwacher, Andreas Velten, Ashok Veeraraghavan, and Ramesh Raskar. Reconstruction of hidden 3d shapes using diffuse reflections. Optics express, 20 (17):19096–19108, 2012. 2

[12] Guan He, Song Chunfeng, and Zhang Zhaoxiang. Lidar-camera cooperative semantic segmentation. Machine Intelligence Research, 2025. 2

[13] Felix Heide, Matthew O'Toole, Kai Zang, David B Lindell, Steven Diamond, and Gordon Wetzstein. Non-line-of-sight imaging with partial occluders and surface normals. ACM Transactions on Graphics, 38(3):1–10, 2019. 2

[14] Mariko Isogawa, Dorian Chan, Ye Yuan, Kris Kitani, and Matthew O'Toole. Efficient non-line-of-sight imaging from transient sinograms. In European conference on computer vision, pages 193–208. Springer, 2020. 1, 2, 3

[15] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using transient imaging. In IEEE 12th International Conference on Computer Vision, pages 159–166. IEEE, 2009. 2

[16] Yue Li, Yueyi Zhang, Juntian Ye, Feihu Xu, and Zhiwei Xiong. Deep non-line-of-sight imaging from under-scanning measurements. In Advances in Neural Information Processing Systems, pages 1–12, 2023. 1, 2, 3, 8

[17] David B Lindell, Gordon Wetzstein, and Matthew O'Toole. Wave-based non-line-of-sight imaging using fast fk migration. ACM Transactions on Graphics (ToG), 38(4):1–13, 2019. 1, 2, 3, 7, 8

[18] Xiaochun Liu, Ibón Guillén, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. Non-line-of-sight imaging using phasor-field virtual wave optics. Nature, 572(7771): 620–623, 2019.

[19] Xiaochun Liu, Sebastian Bauer, and Andreas Velten. Phasor filed diffraction based reconstruction for fast non-line-of-sight imaging systems. Nature communications, 11:1645, 2020. 1, 2

[20] Xintong Liu, Jianyu Wang, Leping Xiao, Xing Fu, Lingyun Qiu, and Zuoqiang 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, pages 13303–13312, 2023. 3

[21] Xintong Liu, Jianyu Wang, Leping Xiao, Zuoqiang Shi, Xing Fu, and Lingyun Qiu. Non-line-of-sight imaging with arbitrary illumination and detection pattern. Nature Communications, 14(1):3230, 2023. 3

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3202–3211, 2022. 2

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 6

[24] Christopher A. Metzler, David B. Lindell, and Wetzstein Gordon. Keyhole imaging: Non-line-of-sight imaging and tracking of moving objects along a single optical path. IEEE Transactions on Computational Imaging, 7:1–12, 2021. 1, 3

[25] Fangzhou Mu, Sicheng Mo, Jiayong Peng, Xiaochun Liu, Ji Hyun Nam, Siddeshwar Raghavan, Andreas Velten, and Yin Li. Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022. 2

[26] Ji Hyun Nam, Eric Brandt, Sebastian Bauer, Xiaochun Liu, Marco Renna, Alberto Tosi, Eftychios Sifakis, and Andreas Velten. Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. Nature communications, 12:6526, 2021. 1, 2, 3

[27] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. Nature, 555(7696):338–341, 2018. 1, 2, 3

[28] Zhengqing Pan, Ruiqian Li, Tian Gao, Zi Wang, Siyuan Shen, Ping Liu, Tao Wu, Jingyi Yu, and Shiying Li. On-site non-line-of-sight imaging via online calibration. IEEE Photonics Journal, 14(5):1–11, 2022. 1, 3, 6

[29] Chengquan Pei, Anke Zhang, Yue Deng, Feihu Xu, Jiamin Wu, U David, Lei Li, Hui Qiao, Lu Fang, and Qionghai Dai. Dynamic non-line-of-sight imaging system based on the optimization of point spread functions. Optics Express, 29(20): 32349–32364, 2021. 1, 3

[30] Siyuan Shen, Zi Wang, Ping Liu, Zhengqing Pan, Ruiqian Li, Tian Gao, Shiying Li, and Jingyi Yu. Non-line-of-sight imaging via neural transient fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7):2257–2268, 2021. 2

[31] Chia-Yin Tsai, Aswin C Sankaranarayanan, and Ioannis Gkioulekas. Beyond volumetric albedo–a surface optimization framework for non-line-of-sight imaging. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1545–1555, 2019. 2

[32] Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. Nature communications, 3 (1):745, 2012. 2

[33] Jianyu Wang, Xintong Liu, Leping Xiao, Zuoqiang Shi, Lingyun Qiu, and Xing Fu. Non-line-of-sight imaging with signal superresolution network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17420–17429, 2023. 1, 2, 3

[34] Lishun Wang, Miao Cao, Yong Zhong, and Xin Yuan. Spatial-temporal transformer for video snapshot compressive imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7):9072–9089, 2022. 2

[35] Yihao Wang, Zhigang Wang, Bin Zhao, Dong Wang, Mulin Chen, and Xuelong Li. Propagate and calibrate: real-time passive non-line-of-sight tracking. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 972–981, 2023. 4

[36] Feng Xiaohua and Gao Liang. Toward non-line-of-sight videography. Optics and Photonics News, 32(11):22–29, 2021. 3

[37] Feng Xiaohua and Gao Liang. Ultrafast light field tomography for snapshot transient and non-line-of-sight imaging. Nature Communications, 12:2179, 2021. 3

[38] Shumian Xin, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis 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, pages 6800–6809, 2019. 2

[39] Juntian Ye, Yu Hong, Xiongfei Su, Xin Yuan, and Feihu Xu. Plug-and-play algorithms for dynamic non-line-of-sight imaging. ACM Transactions on Graphics, 43(5):1–12, 2024. 1, 3, 7

[40] Jun-Tian Ye, Xin Huang, Zheng-Ping Li, and Feihu Xu. Compressed sensing for active non-line-of-sight imaging. Optics Express, 29(2):1749–1763, 2021. 3

[41] Jun-Tian Ye, Yi Sun, Wenwen Li, Jian-Wei Zeng, Yu Hong, Zheng-Ping Li, Xin Huang, Xianghui Xue, Xin Yuan, Feihu Xu, Xiankang Dou, and Jian-Wei Pan. Real-time non-line-of-sight computational imaging using spectrum filtering and motion compensation. Nature Computational Science, 4: 920–927, 2024. 3

[42] Aaron Young, Nevindu M Batagoda, Harry Zhang, Akshat Dave, Adithya Pediredla, Dan Negrut, and Ramesh Raskar. Enhancing autonomous navigation by imaging hidden objects using single-photon lidar. arXiv preprint arXiv:2410.03555, 2024. 1

[43] Sean I Young, David B Lindell, Bernd Girod, David Taubman, and Gordon 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, pages 1407–1416, 2020. 2

[44] Yanhua Yu, Siyuan Shen, Zi Wang, Binbin Huang, Yuehan Wang, Xingyue Peng, Suan Xia, Ping Liu, Ruiqian Li, and Shiying Li. Enhancing non-line-of-sight imaging via learnable inverse kernel and attention mechanisms. In IEEE/CVF International Conference on Computer Vision, pages 10563–10573, 2023. 2

[45] Wenjun Zhang, Enlai Guo, Shuo Zhu, Chenyang Huang, Lijia Chen, Lingfeng Liu, Lianfa Bai, Edmund Y. Lam, and Jing Han. Real-time scan-free non-line-of-sight imaging. APL Photonics, 9(12), 2024. 3