TriNLOS: Triplane Representations for Neural Non-Line-of-Sight Imaging

Bonmu Do[1] and Ji Hyun Nam[1⋆]

Department of Data Science, Inha University, Incheon, Republic of Korea
kmkdbh01@gmail.com, namjihyun@inha.ac.kr

Abstract. Non-line-of-sight (NLOS) imaging aims to reconstruct hidden scenes from time-resolved light transport, enabling vision around corners. While classical physics-based methods provide principled inversion, learning-based approaches often rely on dense 3D backbones with cubic computational complexity. We propose a hybrid deep learning framework that combines a physics-guided initialization with a triplane-based backbone for high-fidelity volumetric reconstruction. The initialization is provided by a learnable Enhanced Light-Cone Transform (ELCT), which produces a stable physics-consistent coarse volume, while the learned backbone replaces expensive $O(N^3)$ 3D processing with scalable $O(N^2)$ triplane feature extraction. ConvNeXt-style residual convolutions, Restormer attention, and axis-aware cross-attention jointly refine structure and recover missing geometry. Experiments on synthetic and real NLOS data demonstrate improved reconstruction fidelity compared to representative physics-based and learning-based baselines.

1 Introduction

Non-line-of-sight (NLOS) imaging seeks to recover the 3D structure of objects that are not directly visible by analyzing multiply scattered light measured on a visible relay surface. By illuminating the wall with a pulsed laser and capturing time-resolved reflections with a SPAD detector, it becomes possible to computationally "see around corners," enabling applications in autonomous navigation, search-and-rescue, surveillance, and medical diagnostics [9,12]. A NLOS configuration is shown in Fig. 1.

NLOS imaging systems are commonly categorized into *confocal* and *non-confocal* configurations. In confocal setups the illumination and detection points on the relay surface coincide, producing a symmetric light-cone geometry that admits efficient inversion via operators such as the light-cone transform (LCT) [35]. Non-confocal configurations allow spatially separated illumination and detection, supporting more general sampling patterns and wider fields-of-view [29, 46, 50]. In this work we focus on the confocal configuration. We present a hybrid physics–ML framework that combines a physics-guided volumetric initialization with a triplane architecture [4] adopted for NLOS reconstruction. Concretely, the pipeline first extracts low-level spatio-temporal features from raw

⋆ Corresponding author.

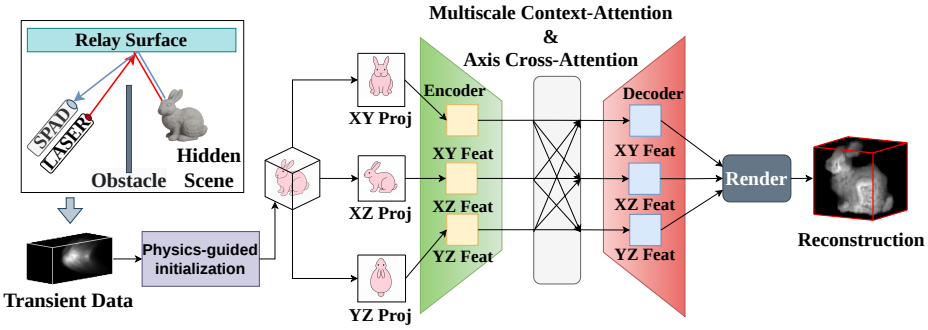


Fig. 1: Confocal NLOS setup and TRINLOS overview: A pulsed laser illuminates a relay wall, light scatters toward the hidden scene, and time-resolved returns are captured by a SPAD detector. The measured transient data are then processed through appropriate reconstruction procedures to form a 3D volumetric feature map, which is subsequently represented and refined using a triplane-based architecture.

transient measurements using a lightweight shallow CNN. These learned features are provided as inputs to an *Enhanced Light-Cone Transform* (ELCT) module which computes a coarse 3D estimate; ELCT extends the classical LCT [35] via depth re-parameterization, cone-aligned resampling, multi-scale frequency-domain enhancement, and depth-aware fusion. The resulting physically informed volume is projected onto three orthogonal planes and processed by lightweight 2D ConvNeXt-style backbones [31] augmented with multi-scale Restormer attention [52]. A dedicated axis cross-attention mechanism exchanges missing 3D cues between planes, and a gradient-blocked rendering head enables stable joint supervision of intensity and depth. An overview of the proposed framework is provided in Fig. 1, while Fig. 2 shows the detailed pipeline.

Experiments on synthetic and real NLOS datasets indicate that combining a physics-consistent initialization with a triplane refinement stage can produce high-quality reconstructions while reducing the effective volumetric complexity from cubic to quadratic in the backbone processing.

Our contributions are:

- **Enhanced Light-Cone Transform (ELCT):** a physics-guided initialization that combines depth remapping, cone-aligned resampling, and frequency-domain sharpening to produce a stable coarse volume for learned refinement.
- **Triplane backbone with cross-axis fusion:** a scalable triplane architecture that reduces volumetric complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ via 2D feature processing using ConvNeXt-style residual blocks [31] and Restormer attention [52], while restoring 3D spatial information lost in plane-wise projection through axis cross-attention and voxel-wise fusion, tailored to the geometric structure of NLOS transient measurements.
- **High reconstruction quality with memory efficiency:** The proposed model achieves quantitatively superior reconstruction performance over existing physics-based and learning-based methods, while offering favorable

memory–performance trade-offs among learning-based approaches, and maintains strong reconstruction quality on real-world measurements.

Code and dataset. Our implementation, dataset, and pretrained model are publicly available at https://github.com/CIL-Research/TriNLOS.

2 Related Work

Physics-based NLOS inversion. Classical NLOS reconstruction relies on analytic operators—such as delay-and-sum back-projection, and the light-cone transform (LCT)—to provide robust, training-free solutions [1,2,14,18,23,35,44,49]. These frameworks have been extended through wave-based optics, diffraction modeling, surface-normal estimation, and spherical-slice transforms to handle complex interference, surface geometry, and non-confocal layouts [25–27,42,50]. Recent refinements further incorporate advanced signal analysis and optimization-based regularization strategies to mitigate heavy noise, photon sparsity, and irregular sampling in transient measurements [24,28]. However, such purely physical models often assume idealized conditions, making them susceptible to the diverse variables and noise present in real-world environments [9,12].

Learning-based NLOS methods. To address physical model mismatches, data-driven approaches learn complex priors from transient measurements. Early attempts reconstructed the scene using U-Nets [13], and subsequent work has explored neural implicit representations, which achieve higher geometric fidelity [11,37]. Modern hybrid pipelines now integrate differentiable physics with deep architectures, leveraging advanced deep learning techniques to tackle challenging regimes such as under-scanning, high-speed acquisition, and dynamic scene reconstruction [6,20–22,33,40,41,51]. Furthermore, these learned representations demonstrate superior robustness, maintaining stable reconstruction performance even under extreme noise and low photon-count conditions.

Efficient NLOS imaging. Efforts to improve system efficiency span from hardware-level SPAD array optimizations [3,34,36] to algorithmic enhancements that reduce memory and computational overhead [38,53]. Operating directly in the transient sinogram domain and utilizing FFT-based implementations of physical operators allow for rapid processing compared to dense 3D volumetric methods [18,23,26,35]. Sparse acquisition strategies further reduce the required number of scanning points without compromising reconstruction quality [7,48]. These approaches collectively facilitate high-resolution imaging while keeping the computational footprint manageable for practical applications.

Structured Triplane Representations. Triplane representations factorize a 3D volume into three orthogonal axis-aligned 2D feature planes, reducing complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ [4], and have since been extended to vector–matrix tensor decompositions [5] and generalized d-dimensional planar factorizations [10]. This factorization has enabled a wide range of applications including diffusion-based 3D generative synthesis [39], digital avatar generation [45], fast text-to-3D generation [47], wavelet-refined neural fields [19], and hybrid

triplane–Gaussian single-view reconstruction [55]. Beyond computer vision, the same structured multi-plane factorization has demonstrated strong performance across diverse medical volumetric tasks—including cross-attention diffusion for 3D medical image synthesis [54], and zero-shot CT/MRI inverse problem solving [15]—demonstrating their generality across heterogeneous sensing domains.

Positioning of this work. We combine a lightweight learned preprocessing stage with a physics-guided ELCT initialization and a triplane refinement network augmented by cross-plane attention. The design is tailored to characteristics specific to NLOS imaging.

Although NLOS imaging that accounts for full 3D geometry with partial occlusions has been demonstrated [17], in practice it remains extremely challenging and limited in realistic settings. As a result, most practical NLOS scenarios focus on dominant-view or front-facing geometry, and the reconstruction problem is often closer to a 2.5D representation than a fully view-consistent 3D field.

Under this regime, the projection-induced information loss of triplane representations is less restrictive than in general 3D vision tasks. Building on this property, our framework exploits the triplane factorization to maintain high-quality reconstruction while reducing learned volumetric computation in the backbone from $\mathcal{O}(N^3)$ to approximately $\mathcal{O}(N^2)$. By combining physics-guided initialization, structured 2D factorization, and cross-plane interaction, the method balances reconstruction fidelity and computational efficiency in high-resolution NLOS scenarios.

3 Method

3.1 Problem Formulation

Given a transient tensor $T \in \mathbb{R}^{\tau \times H \times W}$, our goal is to reconstruct the hidden 3D volume $V \in \mathbb{R}^{D \times H \times W}$.

A core difficulty is the weak geometric correspondence between the transient domain and the volumetric target. Learning a direct map $T \to V$ forces the network to infer complex 3D structure from spatio-temporal measurements that are not arranged volumetrically; this mismatch is amplified for geometrically complex scenes with depth discontinuities and fine details.

An alternative is to insert a physics-based module that produces a coarse volume, which a 3D network refines. In practice, such volumes are dominated by diffuse background and noise while meaningful geometry occupies a small spatial fraction. Feeding this dense volume into a 3D backbone therefore wastes model capacity on irrelevant structure and incurs cubic cost $\mathcal{O}(N^3)$.

We mitigate these issues with a triplane factorization that reduces learned volumetric complexity to $\mathcal{O}(N^2)$. Crucially, axis-wise max-projection from 3D to 2D naturally suppresses low-magnitude background responses in favor of stronger foreground activations, reducing the influence of noise and focusing learning on geometrically salient regions. **Notation.** $[\cdot \,\|\, \cdot]$ denotes channel concatenation and "+" a residual addition.

3.2 Physics-Guided Initialization via ELCT

We begin with a shallow 3D convolutional extractor for mild denoising while preserving spatial resolution. A learnable stride-2 downsampling is applied along the temporal axis, followed by shallow residual refinement blocks [16].

To provide a physically grounded initialization, we introduce the *Enhanced Light-Cone Transform* (ELCT), a differentiable refinement of LCT. ELCT consists of two components.

Learnable depth re-mapping. The transient axis is re-parameterized with a learnable exponent α:

$$t' = \left(\frac{t}{\tau}\right)^{\alpha} \tau, \qquad T_\alpha(t,h,w) = T(t',h,w). \tag{1}$$

Frequency-domain structural sharpening. A Laplacian-of-Gaussian filter enhances structural components, while a soft cone-shaped mask suppresses off-cone diffuse responses. A learnable depth gate balances them:

$$V_0(z) = \gamma(z)\, V_{\text{mask}}(z) + \big(1-\gamma(z)\big)\, V_{\text{unmask}}(z), \quad \gamma(z) \quad = \big(1+\big(\tfrac{z}{s}\big)^{p}\big)^{-1}. \tag{2}$$

ELCT concludes with a lightweight inverse filter in the frequency domain:

$$\hat{V}(k) = \frac{\hat{T}(k)\,\hat{h}^*(k)}{|\hat{h}(k)|^2 + \epsilon}, \tag{3}$$

where $\hat{h}$ denotes the NLOS point-spread function.

With only three learnable parameters (α, s, p), ELCT provides a compact yet adaptive physics prior. The resulting coarse volume $V_0 = ELCT(T)$ serves as a depth-corrected and structurally consistent initialization for the triplane backbone.

Depth equalization. Because transient sampling is significantly denser along the depth axis, we apply depth-only downsampling. A max-pooling path and a lightweight learnable downsampler are fused through a normalization convolution to produce the balanced volume $\tilde{V}$ used for triplane projection.

3.3 Triplane Projection and 2D Feature Encoding

After obtaining the physics-guided coarse volume $\tilde{V}$, we extract features efficiently while matching the sparse structure of NLOS scenes. Instead of heavy 3D convolutions, $\tilde{V}$ is converted into three orthogonal 2D *triplane* views via axis-wise max-projection:

$$M_{ab}(u,v) = \max_{w \in c} [\, \tilde{V}(u,v,w) + E(u,v,w) \,], \tag{4}$$

where $(ab,c) \in \{(xy,z),(xz,y),(yz,x)\}$ and E is a learnable 3D positional embedding that provides cues for the 2D projection. This projection suppresses

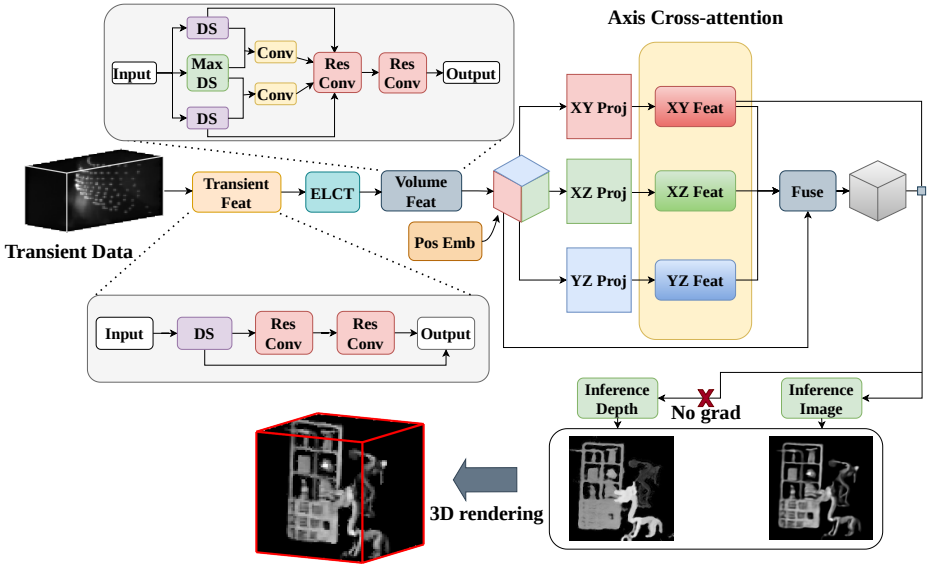


Fig. 2: Overview of the proposed TRINLOS framework. The transient input is first converted into a coarse 3D volume via ELCT, then projected into three orthogonal triplanes. These are processed by 2D FeatureNets with multi-scale attention, and a slice-wise axis cross-attention exchanges missing geometric cues. The fused volume is refined by a lightweight 3D head, and final intensity and depth maps are rendered using a stable depth branch. DS denotes a learnable downsampling layer, while Max DS is a max-pooling block with kernel size 2 and stride 2.

diffuse background and produces three image-like maps M_{xy}, M_{xz}, M_{yz} providing complementary views of the hidden geometry. Each triplane captures a partial scene view, and together they encode the full 3D structure. We process them independently with lightweight 2D feature encoders (see Fig. 3).

Local feature extraction. Each triplane passes through a ResConvNeXt block with ConvNeXt-style residual connections. This block is applied before the Restormer encoder, after its output, and again after axis cross-attention. Compared to Swin Transformer [30], it is more memory-efficient and reduces VRAM usage, since it avoids computing spatial attention maps, while still enhancing local detail and preserving mid-range context at a low computational cost.

Hierarchical Channel-Attentive Context Integration. Each triplane encoder includes a U-Net-style two-stage attention module built on Restormer blocks, computing channel-wise attention to refine input features X into X'.Unlike standard spatial self-attention [8, 43], which has $\mathcal{O}(N^2)$ complexity with respect to the number of tokens N, Restormer attention scales as $\mathcal{O}(HWC^2)$, making it computationally more efficient. It has also proven effective in image restoration tasks like denoising and deblurring.

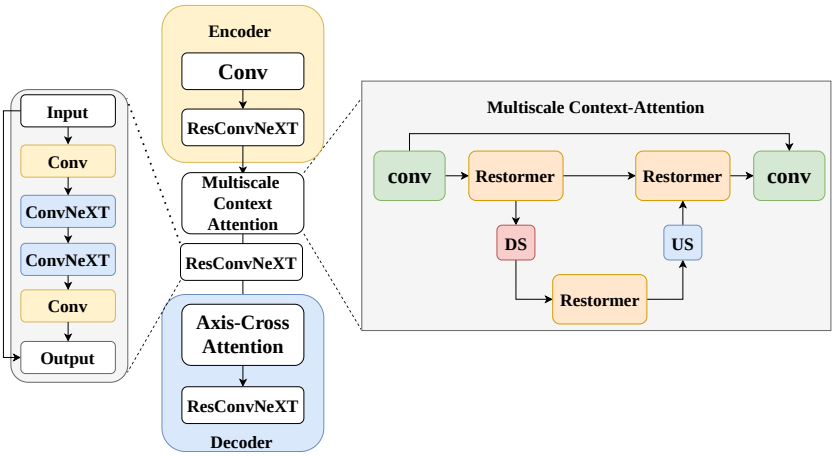


Fig. 3: Architecture of the encoder–decoder. ResConvNeXt and Restormer refine features, axis cross-attention compensates for missing axis cues. DS and US denote learnable down- and up-sampling, respectively.

3.4 Cross-Axis Aggregation and Volumetric Reconstruction

Triplane projection reduces computational cost, but each plane loses one spatial dimension and cannot fully capture 3D geometry alone. To recover information lost during max-projection, we employ an *axis cross-attention* mechanism that exchanges features along shared spatial axes.

Cross-axis reasoning. For any pair of triplanes that share an axis s, we extract slice features $\{F_i\}$ and $\{G_i\}$ indexed along s and update each slice using slice-wise cross-attention:

$$F_i \leftarrow F_i + \mathrm{softmax}\Big(\frac{(F_i W_Q)(G_i W_K)^\top}{\sqrt{d}}\Big)(G_i W_V), \tag{5}$$

where W_Q, W_K, and W_V are learned projections. This focuses attention exactly where geometric cues are aligned—along the shared axis— instead of performing more expensive full 2D or 3D attention. The three triplanes update one another cyclically so that each receives complementary viewpoints from its two partners.

3D Volume Fusion and Refinement. After cross-axis refinement, each triplane feature is lifted to 3D by broadcasting along the collapsed dimension. Since the three resulting volumes vary in coverage and reliability, we predict per-voxel confidence maps W_{xy}, W_{xz}, W_{yz} via a lightweight axis-attention module and directly produce the triplane-derived volume as

$$V_{\mathrm{tri}} = \mathrm{Linear}(W_{xy}F_{bxy} + W_{xz}F_{bxz} + W_{yz}F_{byz}), \quad V_{\mathrm{out}} = \mathrm{fusion}([\,V_{\mathrm{tri}} \parallel \tilde{V}\,]). \tag{6}$$

This consolidates the three complementary planar views into a unified volumetric estimate and subsequently fuses it with the physics-guided coarse volume $\tilde{V}$ through a shallow 3D refinement network, yielding the final output volume used for rendering.

3.5 Rendering and Training Objective

The final stage converts the reconstructed volume into 2D intensity and depth maps. Since many NLOS applications ultimately operate on 2D outputs, we attach a lightweight rendering head rather than predicting albedo and depth directly in 3D.

2D Rendering. We summarize the volume by an axial maximum and concatenate it with the final xy feature to form a compact 2D descriptor. Shallow 2D heads then predict intensity and depth:

$$\begin{aligned} F(x,y) &= \phi\big([\max_z V_{\text{out}}(x,y,z) \parallel F^{2D}_{xy}(x,y)]\big), \quad g(x,y) \quad = \arg\max_z V_{\text{out}}(x,y,z), \\ \hat{T}_t(x,y) &= \text{render}_t\big(F(x,y) \parallel g(x,y)\big), \quad t \in \{I, D\}. \end{aligned} \tag{7}$$

The rendered maps can be used to reconstruct an approximate 3D volume if needed.

Training objective. We supervise intensity $\hat{I}$ and depth $\hat{D}$ with an L_1 composite loss:

$$\mathcal{L} = \|I - \hat{I}\|_1 + \alpha \|D - \hat{D}\|_1, \qquad \alpha = 1. \tag{8}$$

No-gradient depth branch. Directly backpropagating depth errors through the volumetric arg max operation is unstable and tends to flatten reconstructed intensity. To avoid this, gradients from the depth head are restricted to its 2D rendering network render_D. Specifically, we block gradients into the volumetric selection operators:

$$\frac{\partial \hat{D}_i}{\partial(\arg\max_z V^{(i)}_{\text{out}})} = 0 \qquad \frac{\partial \hat{D}_i}{\partial(\max_z V^{(i)}_{\text{out}})} = 0. \tag{9}$$

This no-gradient design stabilizes training and prevents depth supervision from washing out volumetric intensity contrast.

4 Experiments and Results

We train all models using AdamW [32] with a weight decay of 1×10^{-4} (excluding biases and normalization parameters). The network is optimized for 35 epochs with an initial learning rate of 1×10^{-4}, which is decayed to 1×10^{-6} using a cosine annealing scheduler, and a batch size of 4.

Hardware. All experiments are conducted on a workstation equipped with an Intel Core Ultra 9 285K CPU and a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB VRAM).

4.1 Dataset

We train on synthetic confocal transients rendered with the public LFE renderer [6]. Each transient has a spatial resolution of 128×128 and $\tau = 512$ time

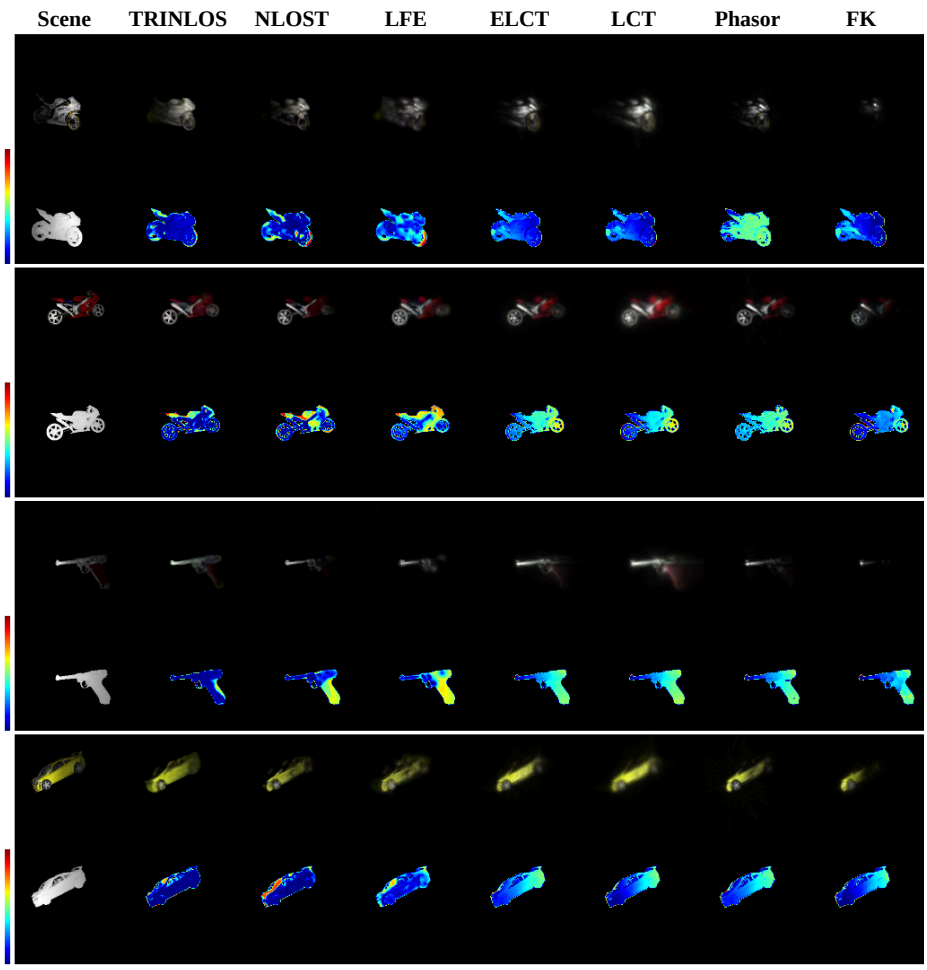


Fig. 4: Qualitative comparisons on synthetic data rendered with [6]. Rows 1–2 depict seen motorcycles, while Rows 3–4 show unseen objects. For each object, the first row presents the reconstructed RGB image and the second row shows the depth error map. Depth errors are normalized to [0,1] and color-coded from blue (low) to red (high).

bins ($\Delta\tau = 32, \text{ps}$), with temporal jitter $\sigma_j = 0.3$ bins to model timing uncertainty. The training set includes 6,740 motorcycle transients from ShapeNet, augmented by planar rotations and flips for viewpoint diversity. For evaluation, we consider two settings: *Seen*, with 400 unseen motorcycle samples, and *Unseen*, with 700 objects from new categories (gun, warcraft, car, etc.).

Noise model. Gaussian and Poisson noise are injected on-the-fly during training to emulate realistic measurement conditions, improving robustness to photon noise and sensor non-idealities.

Table 1: Quantitative comparison on synthetic transients and computational cost analysis. Physics-based and learning-based methods are reported separately. First best values are in **bold**, second best values are underlined.

Method	Reconstruction								Runtime	
	PSNR↑		SSIM↑		$RMSE_{Dep}$ ↓		MAD_{Dep} ↓		Time (ms)	Mem (GB)
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen		
Physics-based Methods										
ELCT	**26.82**	23.63	0.618	0.496	0.680	0.593	0.700	0.614	14.80	8.6
LCT [35]	25.34	22.72	0.644	0.437	0.664	0.576	0.691	0.606	**14.35**	**7.5**
Phasor [26]	26.67	**25.53**	0.507	0.435	0.753	0.721	0.776	0.747	29.91	10.4
FK [23]	26.24	24.96	**0.874**	**0.816**	**0.555**	**0.539**	**0.571**	**0.558**	22.63	9.3
Learning-based Methods										
LFE [6]	27.42	25.43	0.894	0.827	0.066	0.086	0.019	0.034	**48.80**	12.2
NLOST [20]	28.77	25.65	0.946	0.885	0.063	0.106	0.013	0.032	165.29	12.2
TRINLOS	**30.26**	**26.13**	**0.958**	**0.911**	**0.054**	**0.087**	**0.011**	**0.025**	103.22	**8.0**

4.2 Quantitative Comparison and Computational Analysis

Table 1 provides a comprehensive comparison of physics-based and learning-based NLOS reconstruction methods on synthetic transient datasets. Results are reported for both *Seen* and *Unseen* scenes, including reconstruction metrics (PSNR, SSIM, RMSE, MAD) and computational cost (runtime and memory). Additional comparisons are provided in the supplementary material.

Physics-based methods. Among the physics-based approaches, FK achieves the highest reconstruction quality in most metrics, with PSNR of 26.24/24.96 (Seen/Unseen) and the lowest RMSE and MAD values. ELCT, our proposed physics-guided initialization, provides competitive performance, particularly for unseen scenes where it surpasses LCT in PSNR and SSIM, and demonstrates reduced memory usage compared to FK and Phasor. LCT and Phasor exhibit complementary strengths: LCT is fastest and most memory-efficient, whereas Phasor excels in PSNR on unseen data.

Learning-based methods. Among the learning-based methods, TRINLOS consistently achieves the highest reconstruction accuracy across all metrics, with a PSNR of 30.26/26.13 (Seen/Unseen) and the lowest RMSE/MAD values (0.054/0.087 and 0.011/0.025, respectively). NLOST and LFE remain strong baselines, with NLOST performing second-best in most reconstruction metrics (28.77/25.65 PSNR) and LFE achieving the fastest runtime (**48.80 ms**). Due to sequential processing of the three triplanes, TRINLOS requires 103.22 ms, exceeding the runtime of LFE by more than a factor of two. Nevertheless, it remains considerably faster than NLOST, requiring only about two-thirds of its runtime, while also achieving the lowest memory consumption (8.0 GB) among the compared learning-based methods. Notably, even though TRINLOS was trained with depth-gradient propagation blocked in one branch, it still outperforms LFE and NLOST—which allow full depth-gradient propagation—demonstrating superior

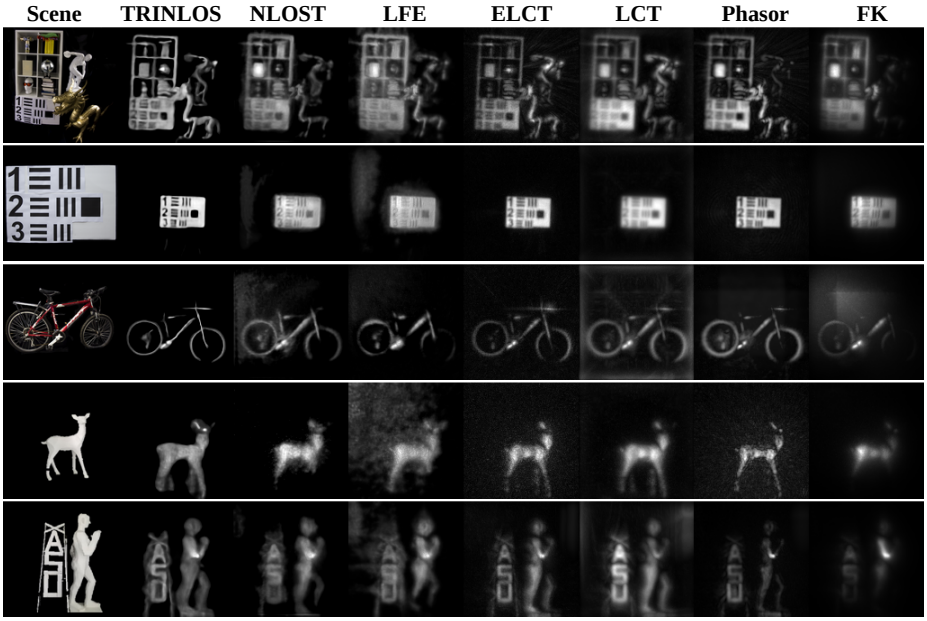


Fig. 5: Real-world reconstruction results comparison across methods

depth reconstruction quality. Figure 4 presents qualitative comparisons of reconstructed RGB intensity images and the corresponding depth error maps with respect to the ground truth. The results demonstrate strong visual reconstruction performance, particularly for foreground regions that are distant or exhibit weak intensity responses. Notably, fine structures in low-signal areas are recovered with high fidelity.

4.3 Real-World Data Results

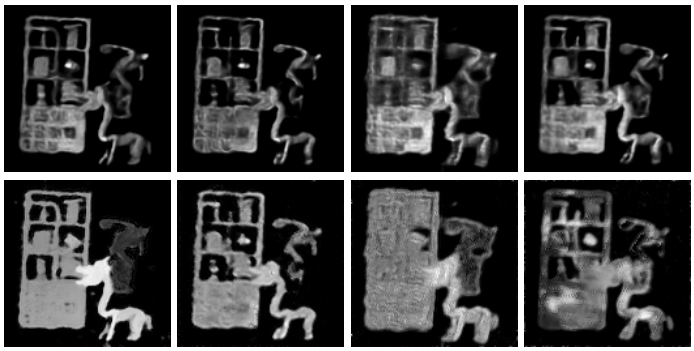
To evaluate cross-dataset generalization, we test our method on public real-world NLOS datasets [20, 23]. As ground truth is unavailable for real-world data, we report qualitative comparisons. Figure 5 contrasts TRINLOS with physics-based (ELCT, LCT, FK, Phasor) and learning-based (LFE, NLOST) approaches.

Among learning-based methods, NLOST reconstructs fine structures but degrades in peripheral regions (e.g., dragon's tail, human/deer legs) and introduces spreading artifacts around foreground objects. LFE exhibits similar behavior with stronger diffusion, resulting in blurrier shapes.

For physics-based baselines, LCT preserves global structure but blurs fine details and can generate X-shaped artifacts. Phasor produces sharp foregrounds yet is noise-sensitive and under-recovers background content. FK is noise-robust but weak in peripheral reconstruction. ELCT balances sharpness and structural recovery, though it remains moderately sensitive to noise.

Table 2: Ablation study on architectural components. We conduct ablation experiments by progressively removing each architectural module to evaluate its individual contribution to the overall performance.

Method	$\mathbf{PSNR_{Int}} \uparrow$	$\mathbf{SSIM_{Int}} \uparrow$	$\mathbf{RMSE_{Dep}} \downarrow$	$\mathbf{MAD_{Dep}} \downarrow$
Proposed	**30.20**	**0.958**	**0.054**	**0.011**
w/o ResConvNeXt	29.31	0.952	0.062	0.014
w/o Restormer attention	28.44	0.944	0.063	0.014
w/o Axis cross-attention	29.21	0.951	0.061	0.014



Full Model w/o RC w/o RA w/o AC

Fig. 6: Qualitative comparison for component ablations on real data (10-min capture). Top: reconstructed intensity, Bottom: reconstructed depth. RC: ResConvNeXt, RA: Restormer attention, AC: Axis cross-attention.

Built upon ELCT, TRINLOS preserves its structural fidelity and sharp boundaries while introducing learnable internal parameters. This allows adaptive modulation of the physics-based initialization, and subsequent neural refinement mitigates ELCT's noise sensitivity. As a result, TRINLOS reconstructs distant objects, retains peripheral structures, and preserves fine details with clear, identifiable quality.

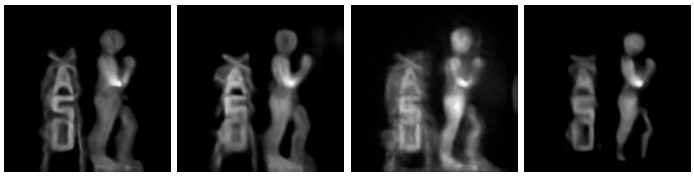
4.4 Ablation Study

We conduct five ablation studies to analyze the effect of: (i) architectural components, (ii) physics-based initialization, (iii) depth-gradient propagation, (iv) triplane projection components and (v) evaluation under a nonconfocal setup. For all synthetic experiments, evaluation is conducted on the seen test dataset.

Architectural components. Table 2 shows that the full model performs best across all metrics, with removal of any component degrading results. Figure 6 shows ablation results for the same object as the first in Figure 5, captured with shorter time and noisier measurements. Eliminating the Restormer causes the largest drop, confirming its key role in refining blurred features and preserv-

Table 3: Quantitative comparison with physics-based reconstruction methods.

Method	$\mathbf{PSNR_{Int}} \uparrow$	$\mathbf{SSIM_{Int}} \uparrow$	$\mathbf{RMSE_{Dep}} \downarrow$	$\mathbf{MAD_{Dep}} \downarrow$
Proposed	**30.20**	**0.958**	**0.054**	**0.011**
LCT [35]	29.20	0.950	0.059	0.015
Phasor [26]	28.95	0.947	0.062	0.013
FK [23]	29.25	0.950	0.063	0.013



Proposed with LCT with Phasor with FK

Fig. 7: Visual comparison of different physics-based initializations. Our proposed method(with ELCT) produces cleaner and more structured reconstructions with fewer artifacts.

ing structural clarity (Figure 6). Removing axis cross-attention leads to the next largest decline, reflecting its importance in compensating for axis-wise information loss in the triplane representation, especially in multi-object scenes where depth cues are partially lost and object separation suffers. Omitting ResConvNeXt further reduces performance: quantitative metrics remain competitive, but qualitative results show fragmented structures. This indicates that ResConvNeXt stabilizes reconstruction by distinguishing foreground from measurement noise in real-world data.

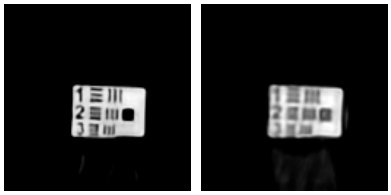
Physics-based initialization. Table 3 compares different physics-based modules for initialization. The proposed ELCT-based model achieves the best quantitative performance, highlighting the advantage of its physical prior.

As shown in Figure 7, the reconstruction results reflect the characteristics of each initialization strategy. With FK initialization, low-intensity foreground components are attenuated, leading to foreground disappearance. With Phasor initialization, significant noise spreads around the foreground regions. With LCT initialization, the global structure is preserved, but the reconstruction remains blurred, especially in regions where multiple objects overlap. In contrast, our method with ELCT initialization achieves superior quantitative performance and provides structurally consistent, sharper reconstructions.

Depth-gradient propagation. Table 4 shows that blocking depth gradients from the shared backbone improves both intensity and depth metrics. Allowing depth gradients to flow into the backbone instead degrades performance, indicating training instability. As shown in Figure 8, the depth-coupled variant produces overall blurrier results. Therefore, we restrict depth gradients to the depth rendering head to ensure stable and sharper reconstructions.

Table 4: Ablation on depth-gradient propagation into the backbone.

Method	D-grad	PSNR↑	SSIM↑	RMSE↓	MAD↓
Proposed	No	**30.20**	**0.958**	**0.054**	**0.011**
w/ depth grad.	Yes	28.47	0.946	0.058	0.012

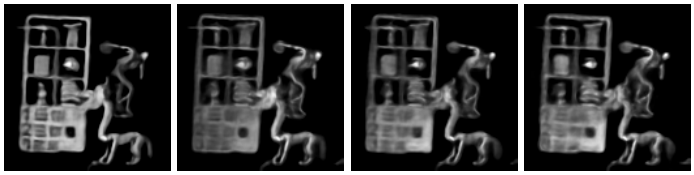


Intensity (Ours) Intensity (D-grad)

Fig. 8: Impact of Depth Gradients on Reconstruction

Table 5: Ablation on triplane projection components.

Method	$\text{PSNR}_{\text{Int}} \uparrow$	$\text{SSIM}_{\text{Int}} \uparrow$	$\text{RMSE}_{\text{Int}} \downarrow$	$\text{RMSE}_{\text{Dep}} \downarrow$	$\text{MAD}_{\text{Dep}} \downarrow$
Proposed	**30.26**	**0.958**	**0.032**	**0.054**	**0.011**
w/o xy	29.44	0.953	0.035	0.058	0.013
w/o xz	29.30	0.952	0.035	0.059	0.012
w/o yz	28.98	0.951	0.036	0.059	0.013



Proposed w/o xy w/o xz w/o yz

Fig. 9: Reconstructed intensity images for each triplane projection ablation.

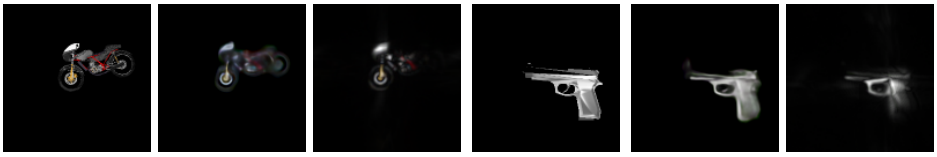
Triplane projection components. Our method fuses volumetric features via three orthogonal 2D projections along the xy, xz, and yz planes. Table 5 reports the effect of ablating each projection individually. Removing the yz projection incurs the largest performance drop, while removing the xy projection has the least impact. As shown in Figure 9, all three components contribute to structural coherence.

The relatively smaller degradation from removing xy is attributable to the architecture's rendering pipeline: xy-plane features are reintegrated at a later stage when rendering the intensity map and depth map, allowing scene information lost at the fusion step to be partially recovered downstream. In contrast, xz and yz features do not benefit from this additional pathway, making their contribution at the fusion stage more critical.

Non-confocal dataset evaluation. TRINLOS originally uses an LCT-based ELCT backbone, which is limited to confocal data. For non-confocal measurements, we replace it with RSD [25] and evaluate both seen and unseen scenes. As Table 6 shows, our model outperforms RSD across all metrics,

Table 6: Non-confocal dataset: quantitative comparison.

Method	PSNR↑		SSIM↑		RMSE_{Dep} ↓		MAD_{Dep} ↓	
	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
Ours	**27.92**	**28.12**	**0.940**	**0.952**	**0.061**	**0.052**	**0.012**	**0.009**
RSD	24.97	25.33	0.670	0.707	0.602	0.605	0.582	0.588



Seen Target Seen Ours Seen RSD Unseen Target Unseen Ours Unseen RSD

Fig. 10: Qualitative comparison on seen and unseen non-confocal scenes. Columns 1-3: seen (Target, Ours, RSD), columns 4-6: unseen (Target, Ours, RSD).

including PSNR, SSIM, depth RMSE, and MAD. Figure 10 illustrates that low-intensity regions, challenging for RSD, are accurately recovered while preserving structure. These results demonstrate that our approach generalizes beyond confocal setups and effectively restores volumetric details.

5 Conclusions

We present TRINLOS, a triplane-backbone framework that combines an ELCT initializer with compact 2D/3D processing for memory-efficient NLOS reconstruction. On synthetic and real data, it matches or surpasses prior physics- and learning-based methods while remaining robust to heavy photon noise. Ablations verify that ELCT, triplane projection, and depth-gradient blocking jointly balance accuracy and efficiency. Future work will pursue higher-resolution reconstruction and lightweight real-time variants with minimal accuracy trade-offs.

Acknowledgements

This work was supported by an INHA University research grant. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00411853). This research was further supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) through the Defense Innovation Vanguard Enterprise Project, funded by the Defense Acquisition Program Administration (DAPA) (No. R230201).

References

1. Ahn, B., Dave, A., Veeraraghavan, A., Gkioulekas, I., Sankaranarayanan, A.C.: Convolutional approximations to the general non-line-of-sight imaging operator. In: ICCV. pp. 7889–7899 (2019)
2. Arellano, V., Gutierrez, D., Jarabo, A.: Fast back-projection for non-line of sight reconstruction. Optics Express **25**(10), 11574 (2017)
3. Buttafava, M., Zeman, J., Tosi, A., Eliceiri, K., Velten, A.: Non-line-of-sight imaging using a time-gated single photon avalanche diode. Optics Express **23**(16), 20997 (2015)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., de Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient Geometry-aware 3D Generative Adversarial Networks. In: CVPR. pp. 16102–16112 (2022)
5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: TensoRF: Tensorial Radiance Fields. In: ECCV. pp. 333–350 (2022)
6. Chen, W., Wei, F., Kutulakos, K.N., Rusinkiewicz, S., Heide, F.: Learned feature embeddings for non-line-of-sight imaging and recognition. ACM TOG **39**(6), 1–18 (2020)
7. Cho, I., Shim, H., Kim, S.J.: Learning to enhance aperture phasor field for non-line-of-sight imaging. In: European Conference on Computer Vision. pp. 72–89. Springer (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
9. Faccio, D., Velten, A., Wetzstein, G.: Non-line-of-sight imaging. Nature Reviews Physics **2**(6), 318–327 (2020)
10. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In: CVPR. pp. 12479–12488 (2023)
11. Fujimura, Y., Kushida, T., Funatomi, T., Mukaigawa, Y.: NLOS-NeuS: Non-line-of-sight Neural Implicit Surface. In: ICCV. pp. 10498–10507 (2023)
12. Geng, R., Hu, Y., Chen, Y.: 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), 1–48 (2021)
13. Grau Chopite, J., Hullin, M.B., Wand, M., Iseringhausen, J.: Deep Non-Line-of-Sight Reconstruction. In: CVPR. pp. 960–969 (2020)
14. Gupta, O., Willwacher, T., Velten, A., Veeraraghavan, A., Raskar, R.: Reconstruction of hidden 3D shapes using diffuse reflections. Optics Express **20**(17), 19096 (2012)
15. He, J., Li, B., Yang, G., Liu, Z.: Blaze3dm: Integrating triplane representation with diffusion for solving 3d inverse problems in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–66 (2025)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778 (2016)
17. Heide, F., O'Toole, M., Zang, K., Lindell, D.B., Diamond, S., Wetzstein, G.: Non-line-of-sight Imaging with Partial Occluders and Surface Normals. ACM TOG **38**(3), 1–10 (2019)

18. Isogawa, M., Chan, D., Yuan, Y., Kitani, K., O'Toole, M.: Efficient Non-Line-of-Sight Imaging from Transient Sinograms. In: ECCV. pp. 193–208 (2020)
19. Khatib, R., Giryes, R.: TriNeRFLet: A Wavelet Based Triplane NeRF Representation. In: ECCV. pp. 358–374 (2024)
20. Li, Y., Peng, J., Ye, J., Zhang, Y., Xu, F., Xiong, Z.: NLOST: Non-Line-of-Sight Imaging with Transformer. In: CVPR. pp. 13313–13322 (2023)
21. Li, Y., Sun, Y., Sun, S., Ye, J., Zhang, Y., Xu, F., Xiong, Z.: Toward dynamic non-line-of-sight imaging with mamba enforced temporal consistency. Advances in Neural Information Processing Systems **37**, 126452–126473 (2024)
22. Li, Y., Zhang, Y., Ye, J., Xu, F., Xiong, Z.: Deep non-line-of-sight imaging from under-scanning measurements. Advances in Neural Information Processing Systems **36**, 59095–59106 (2023)
23. Lindell, D.B., Wetzstein, G., O'Toole, M.: Wave-based non-line-of-sight imaging using fast f-k migration. ACM TOG **38**(4), 1–13 (2019)
24. Liu, J., Zhou, Y., Huang, X., Li, Z.P., Xu, F.: Photon-Efficient Non-Line-of-Sight Imaging. IEEE Transactions on Computational Imaging **8**, 639–650 (2022)
25. Liu, X., Bauer, S., Velten, A.: Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems. Nature Communications **11**(1) (2020)
26. Liu, X., Guillén, I., La Manna, M., Nam, J.H., Reza, S.A., Huu Le, T., Jarabo, A., Gutierrez, D., Velten, A.: Non-line-of-sight imaging using phasor-field virtual wave optics. Nature **572**(7771), 620–623 (2019)
27. Liu, X., Velten, A.: The role of Wigner Distribution Function in Non-Line-of-Sight Imaging. In: IEEE International Conference on Computational Photography (ICCP). pp. 1–12 (2020)
28. Liu, X., Wang, J., Li, Z., Shi, Z., Fu, X., Qiu, L.: Non-line-of-sight reconstruction with signal–object collaborative regularization. Light: Science & Applications **10**(1) (2021)
29. Liu, X., Wang, J., Xiao, L., Shi, Z., Fu, X., Qiu, L.: Non-line-of-sight imaging with arbitrary illumination and detection pattern. Nature Communications **14**(1) (2023)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: ICCV. pp. 9992–10002 (2021)
31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR. pp. 11966–11976 (2022)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
33. Mu, F., Mo, S., Peng, J., Liu, X., Nam, J.H., Raghavan, S., Velten, A., Li, Y.: Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
34. Nam, J.H., Brandt, E., Bauer, S., Liu, X., Renna, M., Tosi, A., Sifakis, E., Velten, A.: Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. Nature Communications **12**(1) (2021)
35. O'Toole, M., Lindell, D.B., Wetzstein, G.: Confocal non-line-of-sight imaging based on the light-cone transform. Nature **555**(7696), 338–341 (2018)
36. Riccardo, S., Conca, E., Sesta, V., Velten, A., Tosi, A.: Fast-Gated 16 × 16 SPAD Array With 16 on-Chip 6 ps Time-to-Digital Converters for Non-Line-of-Sight Imaging. IEEE Sensors Journal **22**(17), 16874–16885 (2022)
37. Shen, S., Wang, Z., Liu, P., Pan, Z., Li, R., Gao, T., Li, S., Yu, J.: Non-line-of-Sight Imaging via Neural Transient Fields. IEEE TPAMI **43**(7), 2257–2268 (2021)

38. Shim, H., Cho, I., Kwon, D., Kim, S.J.: Domain reduction strategy for non-line-of-sight imaging. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)
39. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3D Neural Field Generation Using Triplane Diffusion. In: CVPR. pp. 20875–20886 (2023)
40. Su, X., Zhu, T., Liu, L., Chen, Z., Zhang, Y., Li, S., Ye, J., Xu, F., Yuan, X.: Dual-branch graph feature learning for nlos imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7051–7059 (2025)
41. Sun, S., Li, Y., Zhang, Y., Xiong, Z.: Generalizable non-line-of-sight imaging with learnable physical priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25040–25049 (2025)
42. Tsai, C.Y., Sankaranarayanan, A.C., Gkioulekas, I.: Beyond Volumetric Albedo — A Surface Optimization Framework for Non-Line-Of-Sight Imaging. In: CVPR. pp. 1545–1555 (2019)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
44. Velten, A., Willwacher, T., Gupta, O., Veeraraghavan, A., Bawendi, M.G., Raskar, R.: Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. Nature Communications **3**(1) (2012)
45. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., Guo, B.: RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In: CVPR. pp. 4563–4573 (2023)
46. Willomitzer, F., Rangarajan, P.V., Li, F., Balaji, M.M., Christensen, M.P., Cossairt, O.: Fast non-line-of-sight imaging with high-resolution and wide field of view using synthetic wavelength holography. Nature Communications **12**(1) (2021)
47. Wu, B.S., Chen, H.E., Huang, S.Y., Wang, Y.C.F.: TPA3D: Triplane Attention for Fast Text-to-3D Generation. In: ECCV. pp. 438–455 (2024)
48. Ye, J.T., Huang, X., Li, Z.P., Xu, F.: Compressed sensing for active non-line-of-sight imaging. Optics Express **29**(2), 1749 (2021)
49. Young, S.I., Lindell, D.B., Girod, B., Taubman, D., Wetzstein, G.: Non-Line-of-Sight Surface Reconstruction Using the Directional Light-Cone Transform. In: CVPR. pp. 1404–1413 (2020)
50. Yu, J., Xie, G., Yang, J., Tian, X., Shi, X., Tang, M., Zhang, S., Jin, C.: High-resolution non-confocal non-line-of-sight imaging based on spherical-slice transform from spatial and temporal frequency to space and time. Optics Letters **49**(13), 3806 (2024)
51. Yu, Y., Shen, S., Wang, Z., Huang, B., Wang, Y., Peng, X., Xia, S., Liu, P., Li, R., Li, S.: Enhancing non-line-of-sight imaging via learnable inverse kernel and attention mechanisms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10563–10573 (2023)
52. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient Transformer for High-Resolution Image Restoration. In: CVPR. pp. 5718–5729 (2022)
53. Zhang, S., Jin, S., Liu, H., Li, Y., Jiang, X., Xu, M.: Cmformer: Non-line-of-sight imaging with a memory-efficient metaformer network. Optics and Lasers in Engineering **187**, 108875 (2025)
54. Zhang, Z., Ehinger, K.A., Drummond, T.: Tcam-diff: Triplane-aware cross-attention medical diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22732–22740 (2025)

55. Zou, Z., Yu, Z., Guo, Y., Li, Y., Liang, D., Cao, Y., Zhang, S.: Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers. In: CVPR. pp. 10324–10335 (2024)