NLOST: Non-Line-of-Sight Imaging with Transformer

Yue Li* Jiayong Peng* Juntian Ye Yueyi Zhang Feihu Xu Zhiwei Xiong†
University of Science and Technology of China
{yueli65,jiayong,jt141884}@mail.ustc.edu.cn {zhyuey,feihuxu,zwxiong}@ustc.edu.cn

Abstract

Time-resolved non-line-of-sight (NLOS) imaging is based on the multi-bounce indirect reflections from the hidden objects for 3D sensing. Reconstruction from NLOS measurements remains challenging especially for complicated scenes. To boost the performance, we present NLOST, the first transformer-based neural network for NLOS reconstruction. Specifically, after extracting the shallow features with the assistance of physics-based priors, we design two spatial-temporal self attention encoders to explore both local and global correlations within 3D NLOS data by splitting or downsampling the features into different scales, respectively. Then, we design a spatial-temporal cross attention decoder to integrate local and global features in the token space of transformer, resulting in deep features with high representation capabilities. Finally, deep and shallow features are fused to reconstruct the 3D volume of hidden scenes. Extensive experimental results demonstrate the superior performance of the proposed method over existing solutions on both synthetic data and real-world data captured by different NLOS imaging systems.

1. Introduction

Traditional imaging methods mainly focus on recovering information in line-of-sight scenarios, where there are no obstacles on the direct light path between the target and the camera. In contrast, non-line-of-sight (NLOS) imaging targets recovering the hidden scene beyond the direct line of the cameras' sight, where a diffuse relay surface scatters the light from the scene with dramatic loss. Recently, NLOS imaging has brought tremendous revolutions to autonomous driving [2, 20, 35], disaster rescue [18, 43], and medical diagnosis [27].

The time-of-flight (ToF) based imaging scheme is a common configuration in NLOS [9, 14, 21, 29, 36], where a laser source projects a short-pulse light to the relay wall. The light propagates from the relay wall to the hidden object, then reflects back to the relay wall and is finally captured with a time-resolved single-photon avalanche diode (SPAD) detector. The hidden volume could be reconstructed by modeling the three bounces of the traveling light, achieving "seeing around corners".

Existing NLOS reconstruction algorithms have achieved decent results, but are still confronted with great challenges. Methods based on filter back projection [17, 41] or light path transport [13, 14, 29] often impose restrictive conditions, such as an ideal diffuse surface and no occlusions behind the wall, resulting in detailed texture loss and heavy noise. Methods based on wave propagation [21, 24] are sensitive to the depth range of hidden objects, making distant regions indistinct. Recently, deep-learning-based methods [9, 10, 28, 36] have been introduced to NLOS reconstruction with improved detailed textures and geometries. However, there still remains a large room for boosting their performance on complicated scenes and generalization capabilities toward different real-world systems.

Inspired by the success of transformer [25, 40] in a variety of vision tasks including 3D reconstruction [7, 19, 39, 42, 45], we propose the first transformer-based method for NLOS reconstruction, termed as NLOST. Our method leverages the powerful representation capability of the transformer for capturing local and global spatial-temporal correlations in 3D NLOS measurements. Specifically, to exploit these correlations, we design an end-to-end neural network with two elaborate attention mechanisms tailored for NLOS reconstruction. The network first extracts the shallow features from the NLOS measurements with a feature extractor incorporating physics-based priors. Then, two spatial-temporal self attention encoders built on transformers are proposed to extract local and global information from the shallow features, respectively. For the local encoder, the input features are split into patches, and the local information is exploited in each patch along the spatial and temporal dimensions, successively. For the global encoder, the input features are downsampled to a smaller scale, and the global information is exploited along spatial and temporal dimensions in the whole feature space. The complementary local and global information is further integrated with

*Equal contribution. † Corresponding author.

each other into the token space of transformers by the proposed spatial-temporal cross attention decoder, generating deep local and global features with high representation capabilities. Finally, the above-obtained shallow, deep local, and deep global features are fused together to reconstruct the 3D volume of hidden scenes.

Extensive experiments are performed on both synthetic and real-world datasets. In addition to the publicly available data, we also capture a set of real-world measurements with a self-built NLOS imaging system. Compared with existing traditional and deep-learning-based solutions, our method achieves superior reconstruction performance as well as improved generalization capability to real-world scenarios. Contributions are summarized as follows:

- We propose the first transformer-based neural network for NLOS reconstruction.
- We exploit the complementary local and global correlations in 3D NLOS measurements with two elaborate spatial-temporal attention mechanisms.
- Our method achieves superior performance on both synthetic data and real-world data from different imaging systems.
- We capture a set of NLOS transient measurements with a self-built confocal system and release them for future researches in this field (https://github.com/Depth2World/NLOST).

2. Related Work

NLOS Imaging System. Existing NLOS imaging systems can be divided into two categories: passive and active. Passive systems [1, 3, 4, 34, 44] seek to perform the reconstruction solely with the light emitted from the ambient environment or the hidden object, which capture the NLOS measurements with a conventional camera and remain very challenging for general scenes. Active systems [14, 15, 21, 24, 29, 41] illuminate the scene with a controlled light source, usually a laser, and reconstruct the hidden scenes from the active transient measurements. Generally, the most effective and robust setup uses a pulsed illumination source and a fast SPAD detector to measure the ToF measurements through the scenes.

Traditional NLOS Reconstruction. Many algorithms have been developed for time-resolved NLOS reconstruction since Kirmani *et al*. [15] propose to recover the hidden object out of the visible line of sight. As a precursory work in the field of NLOS, Velten *et al*. [41] propose a filtered back-projection (FBP) method to recover the hidden objects from NLOS measurements. O'Toole *et al*. [29] facilitate the light-cone transform (LCT) for NLOS reconstruction under the following assumptions: light scatters isotropically and only once behind the wall, and the scene contains no occlusions. They simplify the transient formation in a linear 3D convolution form, and the reconstruction can be expressed as a deconvolution process and solved efficiently. Following [29], Heide *et al*. [13] further model the partial occlusions and surface normals in NLOS imaging and develop a factorization approach for nonlinear inverse time-resolved light transport. Recent researches have transitioned from geometrical optics models to wave propagation models [21, 23]. Lindell *et al*. [21] introduce a wave-based image formation model for NLOS imaging and adopt frequency-wavenumber migration (FK). Liu *et al*. [23] start from the phasor field formalism and present a Rayleigh Sommerfeld Diffraction (RSD) algorithm for general transient data. However, the traditional algorithms are either restricted by the ideal assumptions or fragile for the distant targets in real-world scenarios.

Deep NLOS Reconstruction. Recently, deep learning has demonstrated success in computational imaging [8, 16, 33, 46], which sparks interest into NLOS reconstruction. Chopite *et al*. [10] first employ a convolutional neural network for NLOS depth estimation, with a 3D encoder and a 2D decoder in U-Net [11] architecture. Due to the lack of special network design for the transient measurement, this model behaves no better than physics-based solutions [21, 23, 29] on synthetic data and fails to generalize to real-world scenarios. Chen *et al*. [9] propose a learned feature embedded network (LFE) to reduce the domain gap between synthetic and real-world datasets, which incorporates the physical-based method [21] at the feature level and then projects the features from 3D spatial domains to 2D planes directly to reconstruct final intensity and depth maps. While promising results are achieved, the 3D to 2D projection may lead to information loss, and LFE requires multi-view supervision during training which burdens the training data generation. Inspired by the recently proposed Neural Radiance Field (NeRF), Shen *et al*. [36] introduce Neural Transient Field (NeTF) to recover the 3D volume from the transient measurements, which uses the multi-layer perception to represent a 3D density volume. Nevertheless, NeTF suffers from severe noise on smooth surfaces when recovering the geometry. In addition, the transient field has to be rendered for each measurement.

As the first transformer-based method, NLOST enjoys a performance boost over previous solutions, thanks to our specially designed attention mechanisms tailored for 3D NLOS measurements. Instead of mapping the features to the 2D images as in [9], our model works on 3D domains all the way, which avoids information loss. In addition, NLOST can be trained with single-view supervision and also avoids test time rendering as required in [36].

3. Preliminary

Forward Model. In this paper, we focus on ToF-based NLOS imaging, which mainly contains a laser source, a

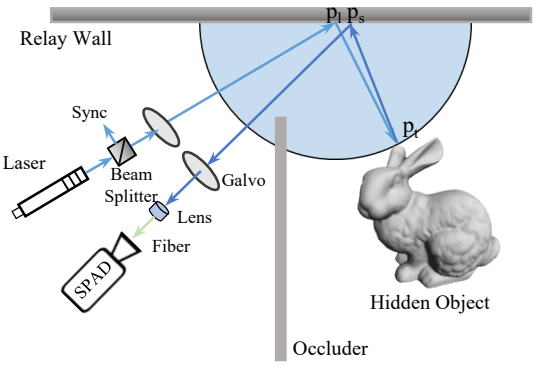

Figure 1. A schematic diagram of the NLOS imaging system.

time-resolved SPAD detector, a relay wall, and a hidden object, shown in Fig. 1. The system works in a confocal manner, where the laser projects short periodic light pulses $\delta(t)$ toward the relay wall at illumination point p_l, from where the light is diffusely scattered at time $t = 0$ and targets the hidden object. After integrated with the object at a certain target point p_t, a fraction of the light is reflected back to the relay wall after time interval t and finally captured by the SPAD at the sampling point p_s, resulting in a 3D spatial-temporal volume $\tau(p_s, p_t, t)$, known as transient measurement. The transient measurement, containing both geometric and photometric information of the hidden object, is a function of illuminated point p_l, sampling point p_s, and target point p_t and can be modeled as

$$\tau(p_s,t) = \iiint_{\Omega} \rho(p_t)\cdot f(n_{p_l\to p_t}, n_{p_t\to p_s})\cdot\varphi \cdot \delta(r_l + r_s - t\cdot c)\mathrm{d}\Omega, \quad (1)$$

where Ω denotes the spherical surface of scattered pulse light from the relay wall. $\rho(\cdot)$ denotes the albedo of the hidden object. $n_{a\to b}$ means the normalized direction from point a to point b. $f(\cdot)$ represents the bidirectional reflectance distribution function, containing diffuse, specular, and retroreflective components. c is the speed of light. r_l is the distance between the illumination point and the target point, while r_s is the distance between the sampling point and the target point. φ is the geometry radiometric term modeled as

$$\varphi = \frac{(n_{p_l\to p_t}\cdot n_{p_t})(n_{p_t\to p_s}\cdot n_{p_t})\cdot \upsilon_{p_l\to p_t}\cdot \upsilon_{p_t\to p_s}}{r_l^2\cdot r_s^2}, \quad (2)$$

where $\upsilon_{a\to b}$ represents the visibility of point a to point b and n_{p_t} is the surface orientation of the target point p_t.

This forward model is general and only assumes no interreflections in the hidden scene. Similar to [9, 10, 29], we model the photon detection of SPAD with an inhomogeneous Poisson process [38] and store the transient measurement in the form of a histogram matrix $H[n]$ with discrete temporal bins as

$$H[n] \sim \text{Poisson}\,(\tau(p_s,t) + B), \quad (3)$$

where n is the index of the temporal bins, and B is the noise photon detections, including both background photons and dark counts [5] of SPAD sensors.

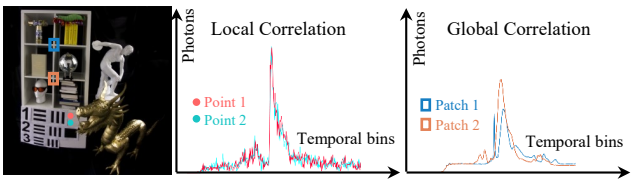

Figure 2. An example of the local and global correlations in the NLOS measurement. The patches and points in the scene indicate the sampling locations on the corresponding relay wall.

Local Correlation. For natural scenes, a certain location usually has similar intensity and depth values to its neighborhoods. This short-range correlation (denoted as local correlation) generally holds in a small range of regions and has been exploited in many vision tasks [33, 37]. As shown in Fig. 2, the histograms of two adjacent points in the transient data are close to each other, which demonstrates that the NLOS measurements also contain local correlation. This kind of correlation is exploited by our spatial-temporal self-attention encoder under the constraint of local continuity for the reconstructed scene.

Global Correlation. For natural scenes, distant patches with similar geometry may have similar intensity and depth values. This long-range correlation (denoted as a global or nonlocal correlation) generally exists in different regions of the scene, which has also been exploited in many vision tasks [12, 22, 31, 32]. As shown in Fig. 2, if we average the histograms of two patches with similar geometry in the transient data, the resulting curves are quite similar. It suggests that the global correlation also holds in the NLOS measurements. This kind of correlation is further exploited by our spatial-temporal self-attention encoder under the constraint of global consistency for the reconstructed scene.

4. Proposed Method

We propose the first transformer-based neural network for NLOS reconstruction by fully exploiting the local and global correlations in the transient measurements, as shown in Fig. 3. Firstly, the input transient data is fed to a feature extractor to extract the shallow features F_S and F_S^* with physics-based priors. Then, two spatial-temporal self attention encoders exploit the local and global correlations along spatial and temporal dimensions from the shallow features F_S, respectively, and generate the local features F_L and the global features F_G. After that, two spatial-temporal cross attention decoders integrate the complementary local and global features, respectively, and generate the deep local features F_L^* and the deep global features F_G^* with improved representation capabilities. Finally, the shallow features F_S^* and the deep features F_L^* and F_G^*, are fused together to reconstruct the 3D volume of hidden objects, generating the intensity image and the depth map.

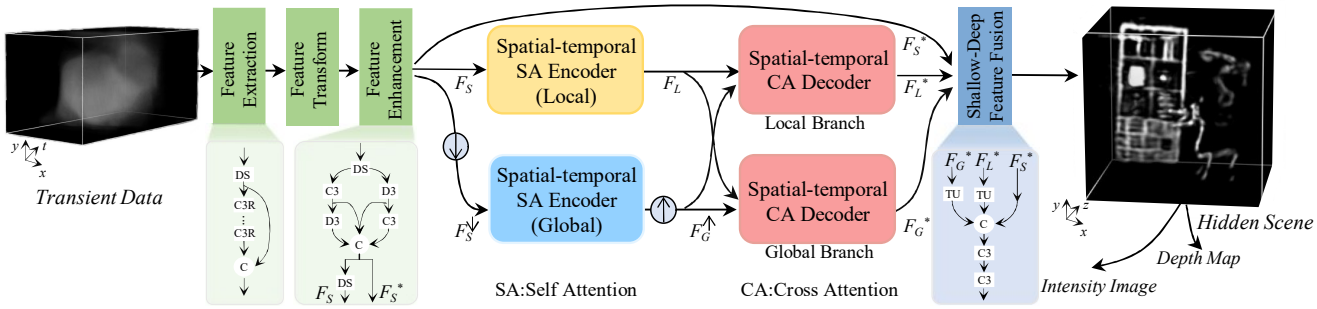

Figure 3. The flowchart of our proposed transformer-based neural network for NLOS reconstruction from the input NLOS transient measurement. "C" and "D" with rectangular blocks denote the 3D convolution and 3D dilated convolution, respectively, with their kernel sizes behind. "C" with circular blocks denotes the concatenation. "DS" with a rectangular block denotes the downsampling operators along the temporal dimension. "TU" with a rectangular block denotes the upsampling operator along the temporal dimension. "↓" and "↑" with a circular block denote the downsampling and upsampling operators along the spatial dimension.

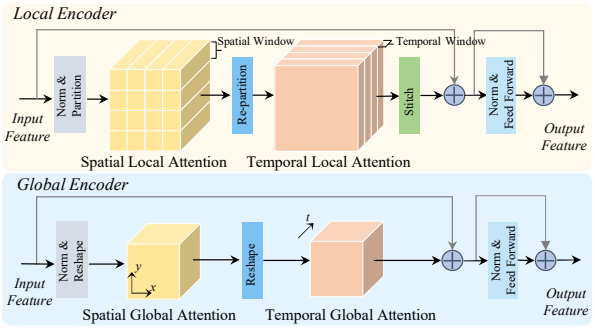

Figure 4. An overview of spatial-temporal self attention encoder.

4.1. Shallow Feature Extraction

The shallow feature extractor consists of a feature extraction layer, a feature transform layer, and a feature enhancement layer, as shown in Fig. 3. Given a transient measurement, we first downsample the input in the temporal dimension and extend the channel dimension by several residual convolutions with the feature extraction layer. Inspired by the existing learning-based methods [9, 28], we transform the spatial-temporal features to the 3D spatial domain with a physics-based prior FK [21] in the feature transform layer. We further enhance the output features by several interlaced 3D convolutions and 3D dilated convolutions to enlarge the receptive field with the feature enhancement layer, producing the shallow features F_S^* and F_S, where $F_S \in \mathbb{R}^{H\times W\times T\times C}$. H, W, T, and C denote the height, width, time, and channel dimension of the feature volume.

4.2. Spatial-Temporal Self Attention Encoder

Local Correlation Exploration. To exploit the local correlation in NLOS measurement, we design a local spatial-temporal encoder based on the multi-head self attention (MSA) mechanism. As shown in Fig. 4, the local encoder consists of a spatial window-based MSA layer (W_s-MSA), a temporal window-based MSA layer (W_t-MSA), and a feed-forward network (FFN). Given the shallow features F_S previously extracted, the local encoder first partitions the features into patches (with a size of $P_s^2 \cdot T$, a number of $N_s = HW/P_s^2$) along spatial dimensions and processes these patches with W_s-MSA, individually. Then, the output features are reshaped and partitioned into patches (with a size of $P_t \cdot HW$, a number of $N_t = T/P_t$) along the temporal dimensions again and processed by W_t-MSA, individually. Finally, the output features are stitched and fed into the FFN, generating the features with local information. This process can be modeled as

$$F_L = FFN\{W_t\text{-}MSA\{W_s\text{-}MSA\{F_S\}\}\}, \quad (4)$$

where $F_L \in \mathbb{R}^{H\times W\times T\times C}$ denotes the output local features. By partitioning the input features into patches and extracting information within patches along spatial and temporal dimensions successively, the local encoder maintains the continuity of depth and intensity in a local region of the 3D transient measurement, which helps to provide more details for the reconstructions of hidden scenes.

Global Correlation Exploration. To exploit the global correlations in NLOS measurements, we design a global spatial-temporal encoder based on the MSA mechanism. As shown in Fig. 4, the global encoder consists of a full spatial MSA layer (F_s-MSA), a full temporal MSA layer (F_t-MSA), and an FFN. Given the shallow features F_S, the global encoder first downsamples the features along the spatial dimensions and processes the features with F_s-MSA. Then, the output features are reshaped and processed with F_t-MSA along the temporal dimension. Finally, the output features are fed into the FFN, generating the features with global information. This process can be modeled as

$$F_G = FFN\{F_t\text{-}MSA\{F_s\text{-}MSA\{F_S^{\downarrow}\}\}\}, \quad (5)$$

where $F_G \in \mathbb{R}^{\frac{H}{k}\times\frac{W}{k}\times T\times C}$ denotes the output global features and k denotes the downsampling factor. By downsampling the input features to a smaller scale and extracting information within the whole feature space along spatial and temporal dimensions successively, the global encoder maintains the consistency of depth and intensity in the whole 3D transient measurements, which helps to recover hidden

scenes with large depth ranges and complicated geometries.

As demonstrated in the ablation study in Sec. 5.4, our elaborate spatial-temporal self attention encoder effectively captures the local and global correlations in NLOS measurements, which improves the reconstruction performance for challenging real-world scenes.

4.3. Spatial-Temporal Cross Attention Decoder

To integrate both local and global information, we further design a spatial-temporal cross attention decoder to improve the feature representation capability. The decoder consists of a local branch and a global branch based on our devised spatial-temporal cross attention (STCA) mechanism, as shown in Fig. 5. Both local and global branches contain an STCA layer, and an FFN interleaved with normalization. For the local branch, the local features F_L and the upsampled global features $F_G^{\uparrow}$ (with the same scale as F_L) are fed into the STCA and FFN in sequence, generating the deep local features $F_L^* \in \mathbb{R}^{H\times W\times T\times C}$ as

$$F_L^* = FFN\left\{STCA\left[Q,K,V\right]\right\},$$
$$Q = F_G^{\uparrow}, K = V = F_L. \tag{6}$$

For the global branch, the upsampled global features $F_G^{\uparrow}$ and the local features F_L are fed into STCA and FFN in sequence generating the deep global features $F_G^* \in \mathbb{R}^{H\times W\times T\times C}$ as

$$F_G^* = FFN\left\{STCA\left[Q,K,V\right]\right\},$$
$$Q = F_L, K = V = F_G^{\uparrow}. \tag{7}$$

As shown in Fig. 5, the STCA integrates the local and global features in a 3D token space, where local and global features are adopted as the query in turn. Given the two input features, a $1\times1\times1$ convolution is conducted to produce the query (Q), key (K), and value(V), respectively. The space of Q, K, and V is reshaped to $\mathbb{R}^{HW\times D\times C}$ to calculate the spatial cross attention by matrix multiplication. After that, the output features are fed into a $1\times1\times1$ convolution resulting in a new K, and a new V. The space of the initial Q, the new K, and the new V is reshaped to $\mathbb{R}^{D\times HW\times C}$ for calculating the temporal cross attention by matrix multiplication as well. As such, the two input features are integrated, and the local and global information complement each other simultaneously. The integration greatly improves the representation capability of the output features and promotes the reconstruction performance, as demonstrated in Sec. 5.4.

4.4. Shallow-Deep Feature Fusion

Finally, the deep local features F_L^* and deep global features F_G^* are fed to 3D deconvolutions to upsample the temporal dimension (with the same scale as F_S^*). Then the upsampled deep local and global features, and the shallow features F_S^* are concatenated along the channel dimension and then fused with 3D convolutional layers to reconstruct the 3D volume V of the hidden scene. The intensity image $\hat{I}$ is

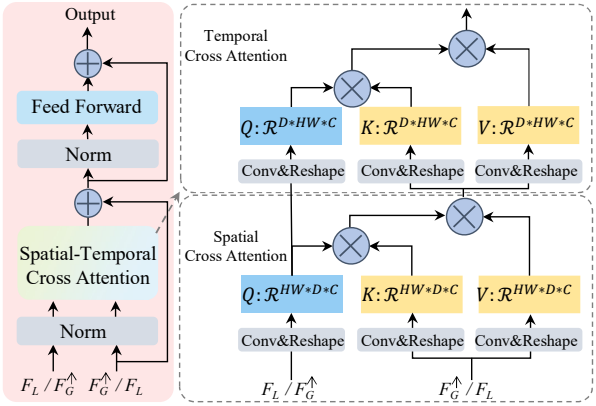

Figure 5. An overview of spatial-temporal cross attention decoder. "×" with a circle denotes the matrix multiplication.

obtained with a max operator along the z axis, while depth map $\hat{D}$ is obtained with an argmax operator along the z axis, which can be modeled as

$$V = \text{fusion}[\text{cat}(F_S^*, F_L^{*\uparrow}, F_G^{*\uparrow})],$$
$$\hat{I} = \max_z(V), \hat{D} = \text{argmax}_z(V), \tag{8}$$

where cat denotes the concatenation along the channel dimension, while fusion contains several 3D convolutions layers. The effectiveness of the shallow-deep feature fusion is demonstrated in Sec. 5.4.

4.5. Loss Function

The loss function is twofold: the intensity loss $\mathcal{L}_I$ and the depth loss $\mathcal{L}_D$. The former is defined as the Manhattan distance between the reconstructed intensity image $\hat{I}$ and the ground-truth I, while the latter is the Manhattan distance between the reconstructed depth map $\hat{D}$ and the ground-truth D, which can be denoted as

$$\mathcal{L}_I(I,\hat{I}) = \|I-\hat{I}\|_1, \mathcal{L}_D(D,\hat{D}) = \|D-\hat{D}\|_1. \tag{9}$$

The final loss function to train the network is

$$\mathcal{L} = \mathcal{L}_I(I,\hat{I}) + \alpha\mathcal{L}_D(D,\hat{D}), \tag{10}$$

where α is a weighting factor.

5. Experiments on Simulated Data

5.1. Data Simulation and Evaluation Metrics

Following [9, 10, 28], we simulate the training and testing data using the transient rasterizer [9] with default settings. A total of 6925 transient measurements with corresponding RGB images are rendered from the motorcycles in the ShapeNet Core dataset [6]. Each measurement has a spatial resolution of 128×128×512 with a bin width of 33 ps. A total of 6250 samples are adopted for training while the remaining 675 samples are used for testing, denoted as **Seen** testing data. To validate the generalization capability of our network, we also render 500 transient measurements from other objects (*e.g.* baskets, helmets, cars, and so on) , denoted as **Unseen** testing data.

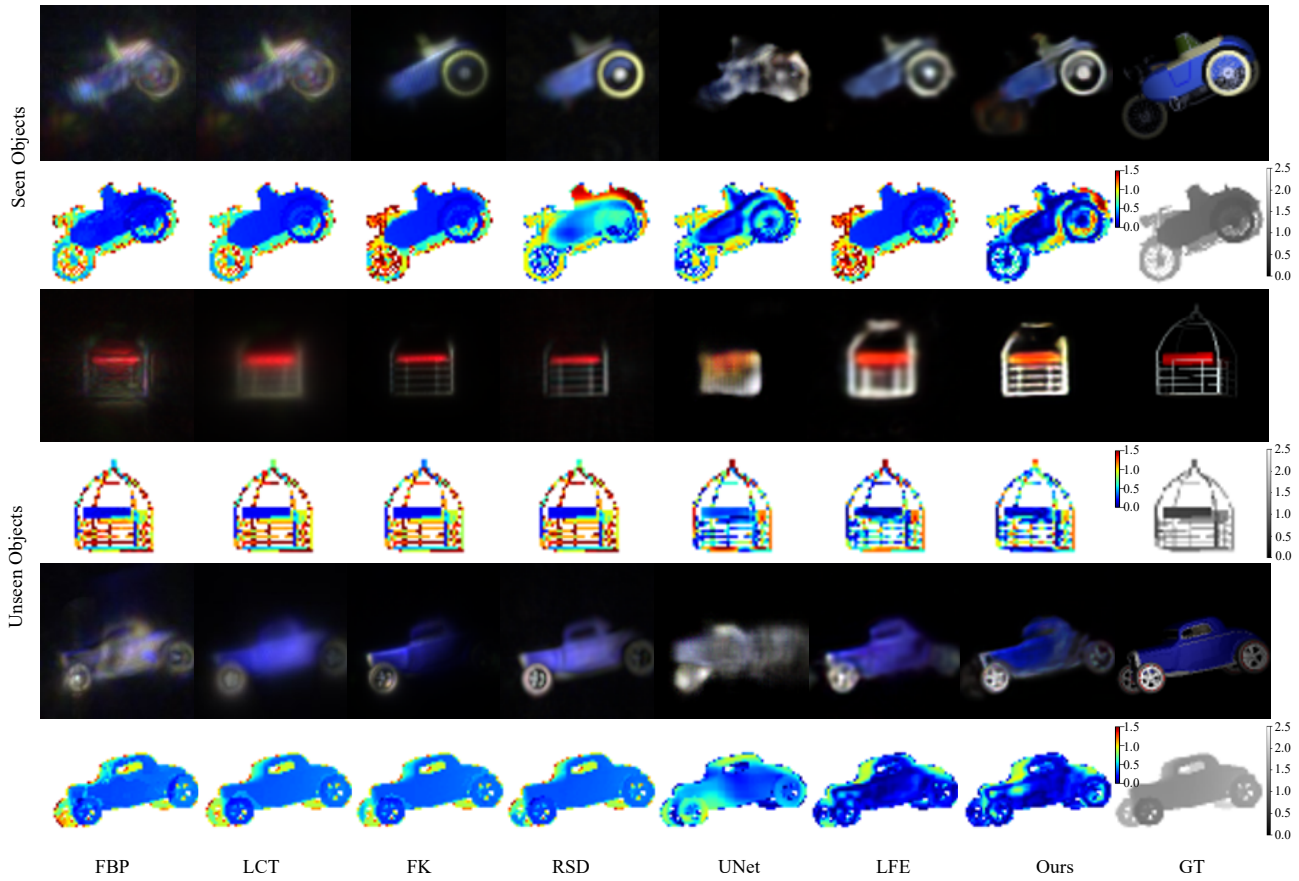

Figure 6. Reconstructed results from **Seen** and **Unseen** test sets on the simulated datasets. The odd and even rows are the intensity images and the depth error maps, respectively. GT denotes the ground truth. The color bars show the value of depth and the error map.

The quantitative evaluation metrics are twofold. For the intensity image, we compute the peak signal-to-noise ratio (PSNR), and structural similarity metrics (SSIM) averaged on the corresponding test samples. For the depth map, we compute the root mean square error (RMSE) and mean absolute distance (MAD) averaged on the test samples.

5.2. Implementation Details

We implement our method using PyTorch [30] and train the network on the simulated data with a batch size of 4. We initialize the network randomly and adopt the AdamW [26] solver with a learning rate of 10^{-4} and an exponential decay of 0.95. The hyper-parameter α is set to 1. We make comparisons with the existing baselines, including physics-based methods: FBP [41], LCT [29], FK [21], and RSD [23]; and deep-learning-based methods: UNet [10], LFE [9] and NeTF [36]. The implementations of the baseline methods follow their publicly available codes. UNet and LFE are trained on the same simulated data as ours, while NeTF [36] is directly trained on the test measurement. We only include NeTF for real-world experiments due to its computational burden for hundreds of simulated scenes. See more details of the deep models in the supplement.

Table 1. Quantitative comparisons of different methods in terms of reconstructing intensity images and depth maps on the **Seen** and **Unseen** test sets.

Data	Methods	Intensity		Depth	
		PSNR↑	SSIM↑	RMSE↓	MAD↓
Seen	FBP [41]	19.96	0.1846	0.7053	0.6694
	LCT [29]	19.78	0.4477	0.6694	0.6321
	RSD [24]	22.17	0.4257	0.7156	0.6846
	FK [21]	23.11	0.7996	0.5558	0.5332
	UNet [10]	24.38	0.7792	0.0820	0.0317
	LFE [9]	26.90	0.8661	0.0769	0.0455
	Ours	**28.17**	**0.9018**	**0.0666**	**0.0221**
Unseen	FBP [41]	17.81	0.2114	0.6986	0.6479
	LCT [29]	18.54	0.4962	0.6604	0.6152
	RSD [24]	19.58	0.4151	0.7335	0.6938
	FK [21]	19.92	0.7729	0.5896	0.5526
	UNet [10]	17.87	0.6932	0.1326	0.0555
	LFE [9]	23.40	0.8100	0.1220	0.0561
	Ours	**23.99**	**0.8286**	**0.1107**	**0.0444**

5.3. Simulated Results

We first evaluate our method on the **Seen** test set. The quantitative results of different methods are listed in Table 1. As can be seen, our method achieves the best performance in both intensity and depth reconstruction and sig-

Table 2. Ablation results on spatial-temporal attention. Spa and Tem denote that the attention mechanism only operates on the spatial or temporal dimension.

SA		CA		Intensity		Depth	
Spa	Tem	Spa	Tem	PSNR	SSIM	RMSE	MAD
✓	✗	✓	✓	24.80	0.8738	0.0688	0.0247
✗	✓	✓	✓	26.09	0.9065	0.0674	0.0236
✓	✓	✓	✗	24.80	0.8739	0.0688	0.0247
✓	✓	✗	✓	25.60	0.9010	0.0676	0.0236
✓	✓	✓	✓	**26.41**	**0.9158**	**0.0640**	**0.0205**

Table 3. Ablation results on local-global feature integration.

Integration	Intensity		Depth	
	PSNR	SSIM	RMSE	MAD
NoInt	25.82	0.8867	0.0718	0.0274
GloInt	25.94	0.9013	0.0668	0.0244
LocInt	26.06	0.9045	0.0652	0.0218
LGInt(Ours)	**26.41**	**0.9158**	**0.0640**	**0.0205**

nificantly surpasses existing baselines. For the intensity image, our method improves the reconstruction performance by a large margin over the physics-based methods, *i.e.*, 5.06 dB over FK and 6.00 dB over RSD in terms of PSNR, which demonstrates the superiority of modeling the NLOS reconstruction with transformer. Meanwhile, compared with the deep-learning-based methods, our method achieves 1.27 dB and 3.79 dB improvements over UNet and LFE, respectively, which demonstrates the effectiveness of exploiting local and global correlations in transient measurements. For the depth map, our method decreases RMSE by 19% and 13% over UNet and LFE, respectively. In addition, we also provide the quantitative results on the **Unseen** test set in Table 1, which includes more complicated scenes. As can be seen, our method also performs the best in both intensity and depth reconstruction, demonstrating the superior generalization capability of our network to unseen objects.

In addition to the quantitative comparisons, we also provide qualitative results for reconstructed intensity images and depth maps on exemplar **Seen** and **Unseen** objects, as shown in Fig. 6. For the intensity image, FBP, LCT, and UNet generate blurry results. FK and RSD recover main structures but without details. LFE behaves better than physics-based methods but still misses details. In contrast, our network recovers both structures and fine details. For the depth map, FBP, LCT, FK, and LFE can hardly reconstruct the details, *e.g.* the wheel of the motorcycle. RSD and UNet have difficulties in recovering even the main structures of the hidden objects. In contrast, our network reconstructs as many as details, especially for the wheel of the motorcycle. Moreover, our network generalizes well to unseen objects. See more qualitative results in the supplement.

5.4. Ablation Study

We conduct ablation experiments to further validate the efficiency of the spatial-temporal attention mechanisms, local-global feature integration, and shallow-deep feature fusion. The models are trained in grayscale for quantitative results on **Seen** test set and qualitative results on real-world transient data.

Table 4. Ablation results on shallow-deep feature fusion. Glo, Loc, and Sha indicate the deep global features, deep local features, and shallow features, respectively.

Feature			Intensity		Depth	
Glo	Loc	Sha	PSNR	SSIM	RMSE	MAD
✓	✗	✗	25.58	0.8736	0.0700	0.0256
✓	✓	✗	26.33	0.9146	0.0646	0.0206
✓	✓	✓	**26.41**	**0.9158**	**0.0640**	**0.0205**

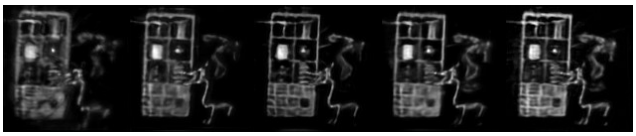

Figure 7. Reconstructed intensity images from a real-world scene under different attention mechanisms. "-S" and "-T" indicate that the attention only operates on the spatial or temporal dimensions.

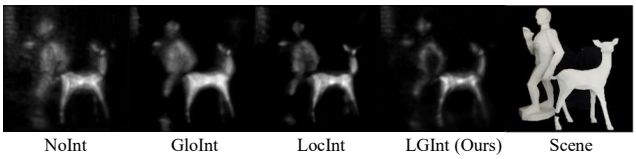

Figure 8. Reconstructed intensity images from a real-world scene under different integration operations.

Spatial-temporal Attention. The spatial-temporal self and cross attention mechanisms are designed to exploit the local and global correlations in 3D transient measurements in both spatial and temporal dimensions. We thus investigate the efficiency of individual spatial and temporal attention in the self attention encoder and the cross attention decoder, with results listed in Table 2 and Fig. 7. As can be seen, spatial and temporal attentions contribute differently to the reconstruction performance, and fusing both of them boosts the performance.

Local-global Feature Integration. In the cross attention decoder, the local and global features are integrated by querying each other in the token space. We further investigate the effectiveness of this integration (LGInt) by comparing it with other alternatives: NoInt (no integration between two branches), LocInt (only integration on local branch), and GloInt (only integration on global branch). The results are listed in Table 3 and Fig. 8. NoInt performs the worst in both intensity and depth, while the performance improves with one kind of information being integrated. When both local and global information is integrated, the performance is further promoted.

Shallow-deep Feature Fusion. We fuse shallow and deep features to recover the 3D volume of hidden objects. We thus study their contributions to the reconstruction perfor-

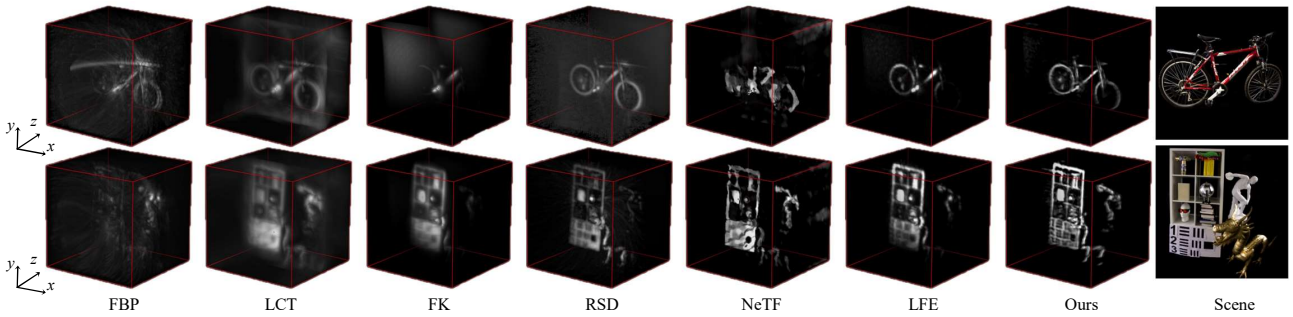

Figure 9. Reconstructed hidden scenes from the public real-world dataset in [21]. Zoom in for details.

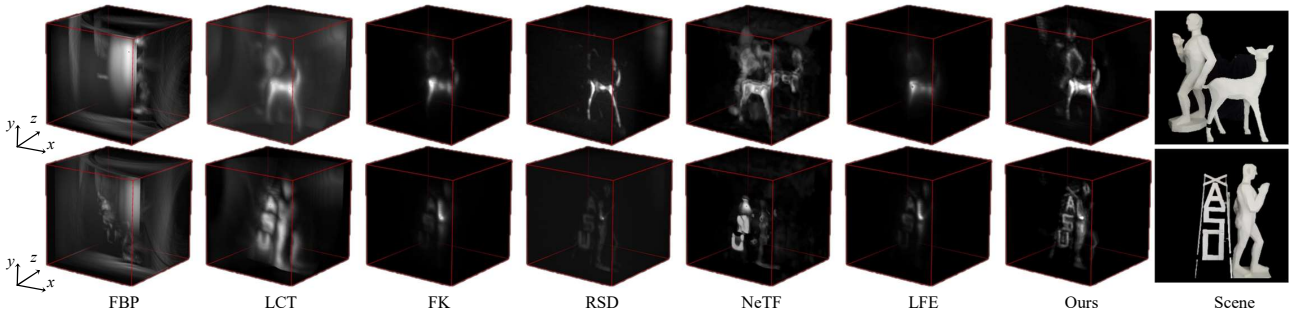

Figure 10. Reconstructed hidden scenes from real-world measurements captured by our NLOS imaging system. Zoom in for details.

mance in Tab. 4. As can be seen, the shallow and deep features contribute differently to the intensity and depth reconstructions, and fusing both of them further improves the performance, which demonstrates the efficiency of our shallow-deep fusion. Qualitative results tested on real-world data can be seen in the supplementary material.

6. Experiments on Real-world Data

Data Preparation. In addition to the simulated data, we also provide qualitative results on real-world data. Due to the heavy workload to build an NLOS system, the real-world data in this field is relatively scarce compared with other fields. We first use the public real-world dataset from Lindell *et al*. [21], which has 7 different scenes. To demonstrate the generalization capability of our method, we then capture 6 different scenes with a self-built confocal imaging system. We release our data for future research in this field.

Real-world Results. The qualitative results on real-world data are shown in Fig. 9 and Fig. 10, with more in the supplement. As can be seen, our method generates promising results with fine details and sharp boundaries of the hidden scenes, especially the girder of the bike, the bookshelf, and the pedestrian. FBP and LCT generate blurry results. FK and RSD can reconstruct main structures but suffer from heavy noise. NeTF can only recover cursory shapes. LFE behaves better than the above methods but still misses some details. The encouraging results produced by our method demonstrate its superiority over existing solutions.

Further Analysis. Beyond improvement of the reconstruction performance, it is worth mentioning that: (a) Our network is trained on the simulated data and directly adopted to process the real-world measurements from two different imaging systems, which demonstrates its superior generalization capability. (b) Our network can recover results with fine details and large depth ranges far beyond the existing physics-based methods, which demonstrates its high representation capability. (c) Our network is trained without multi-view supervision and avoids test time rendering, which alleviates the burdens in data simulation and inference computation encountered in previous deep networks.

7. Conclusions

In this paper, we present the first transformer-based neural network for NLOS reconstruction. By designing spatial-temporal attention mechanisms tailored for 3D NLOS transient measurements to exploit the local and global correlations, transformer is successfully adapted to the challenging NLOS reconstruction task for the first time. The proposed method outperforms existing state-of-the-art approaches on both simulated data and real-world data from different imaging systems. Our future work includes extending NLOST to the non-confocal imaging system and the downstream tasks.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 61901435, and Innovation Program for Quantum Science and Technology under Grant 2021ZD0300300.

References

[1] Mufeed Batarseh, Sergey Sukhov, Zhiqin Shen, Heath Gemar, Reza Rezvani, and Aristide Dogariu. Passive sensing around the corner using spatial coherence. *Nature communications*, 9(1):1–6, 2018. 2

[2] Sven Bauer, Robin Streiter, and Gerd Wanielik. Non-line-of-sight mitigation for reliable urban gnss vehicle localization using a particle filter. In *International Conference on Information Fusion*, pages 1664–1671, 2015. 1

[3] Jeremy Boger-Lombard and Ori Katz. Passive optical time-of-flight for non line-of-sight localization. *Nature communications*, 10(1):1–9, 2019. 2

[4] Katherine L Bouman, Vickie Ye, Adam B Yedidia, Frédo Durand, Gregory W Wornell, Antonio Torralba, and William T Freeman. Turning corners into cameras: Principles and methods. In *International Conference on Computer Vision*, pages 2270–2278, 2017. 2

[5] Danilo Bronzi, Federica Villa, Simone Tisa, Alberto Tosi, and Franco Zappa. Spad figures of merit for photon-counting, photon-timing, and imaging applications: a review. *IEEE Sensors Journal*, 16(1):3–12, 2015. 3

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[7] Wenjie Chang, Yueyi Zhang, and Zhiwei Xiong. Transformer-based monocular depth estimation with attention supervision. In *British Machine Vision Conference*, page 136. BMVA Press, 2021. 1

[8] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Real-world image denoising with deep boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):3071–3087, 2019. 2

[9] 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. 1, 2, 3, 4, 5, 6

[10] Javier Grau Chopite, Matthias B Hullin, Michael Wand, and Julian Iseringhausen. Deep non-line-of-sight reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages 960–969, 2020. 1, 2, 3, 5, 6

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 424–432, 2016. 2

[12] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 3

[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. 1, 2

[14] Felix Heide, Lei Xiao, Wolfgang Heidrich, and Matthias B Hullin. Diffuse mirrors: 3d reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Conference on Computer Vision and Pattern Recognition*, pages 3222–3229, 2014. 1, 2

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

[16] Dilip Krishnan and Richard Szeliski. Multigrid and multilevel preconditioners for computational photography. *ACM Transactions on Graphics*, 30(6):1–10, 2011. 2

[17] Martin Laurenzis and Andreas Velten. Non-line-of-sight active imaging of scattered photons. In *Electro-Optical Remote Sensing, Photonic Technologies, and Applications VII; and Military Applications in Hyperspectral Imaging and High Spatial Resolution Sensing*, volume 8897, page 889706, 2013. 1

[18] Martin Laurenzis, Andreas Velten, and Jonathan Klein. Dual-mode optical sensing: three-dimensional imaging and seeing around a corner. *Optical Engineering*, 56(3):031202, 2016. 1

[19] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *International Conference on Computer Vision*, pages 6197–6206, 2021. 1

[20] David B Lindell, Gordon Wetzstein, and Vladlen Koltun. Acoustic non-line-of-sight imaging. In *Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2019. 1

[21] David B Lindell, Gordon Wetzstein, and Matthew O'Toole. Wave-based non-line-of-sight imaging using fast fk migration. *ACM Transactions on Graphics*, 38(4):1–13, 2019. 1, 2, 4, 6, 8

[22] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[23] Xiaochun Liu, Sebastian Bauer, and Andreas Velten. Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems. *Nature Communications*, 11(1):1–13, 2020. 2, 6

[24] 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. 1, 2, 6

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[27] Tomohiro Maeda, Guy Satat, Tristan Swedish, Lagnojita Sinha, and Ramesh Raskar. Recent advances in imaging around corners. *arXiv preprint arXiv:1910.05613*, 2019. 1

[28] 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*, pages 1–12, 2022. 1, 4, 5

[29] 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, 6

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 6

[31] Jiayong Peng, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. Photon-efficient 3d imaging with a non-local neural network. In *European Conference on Computer Vision*, pages 225–241, 2020. 3

[32] Jiayong Peng, Zhiwei Xiong, Hao Tan, Xin Huang, Zheng-Ping Li, and Feihu Xu. Boosting photon-efficient image reconstruction with a unified deep neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[33] Jiayong Peng, Zhiwei Xiong, Yicheng Wang, Yueyi Zhang, and Dong Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020. 2, 3

[34] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019. 2

[35] Nicolas Scheiner, Florian Kraus, Fangyin Wei, Buu Phan, Fahim Mannan, Nils Appenrodt, Werner Ritter, Jurgen Dickmann, Klaus Dietmayer, Bernhard Sick, et al. Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In *Conference on Computer Vision and Pattern Recognition*, pages 2068–2077, 2020. 1

[36] 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. 1, 2, 6

[37] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 3

[38] Donald L Snyder and Michael I Miller. *Random point processes in time and space*. Springer Science & Business Media, 2012. 3

[39] Qing Su and Shihao Ji. Chitransformer: Towards reliable stereo from cues. In *Conference on Computer Vision and Pattern Recognition*, pages 1939–1949, 2022. 1

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[41] 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):1–8, 2012. 1, 2, 6

[42] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *International Conference on Computer Vision*, pages 2563–2572, 2021. 1

[43] Cheng Wu, Jianjiang Liu, Xin Huang, Zheng-Ping Li, Chao Yu, Jun-Tian Ye, Jun Zhang, Qiang Zhang, Xiankang Dou, Vivek K Goyal, et al. Non–line-of-sight imaging over 1.43 km. *Proceedings of the National Academy of Sciences*, 118(10), 2021. 1

[44] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 12231–12239, 2019. 2

[45] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Conference on Computer Vision and Pattern Recognition*, pages 6677–6686, 2022. 1

[46] Hongyan Zhu, Bingquan Chu, Chu Zhang, Fei Liu, Linjun Jiang, and Yong He. Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers. *Scientific Reports*, 7(1):1–12, 2017. 2