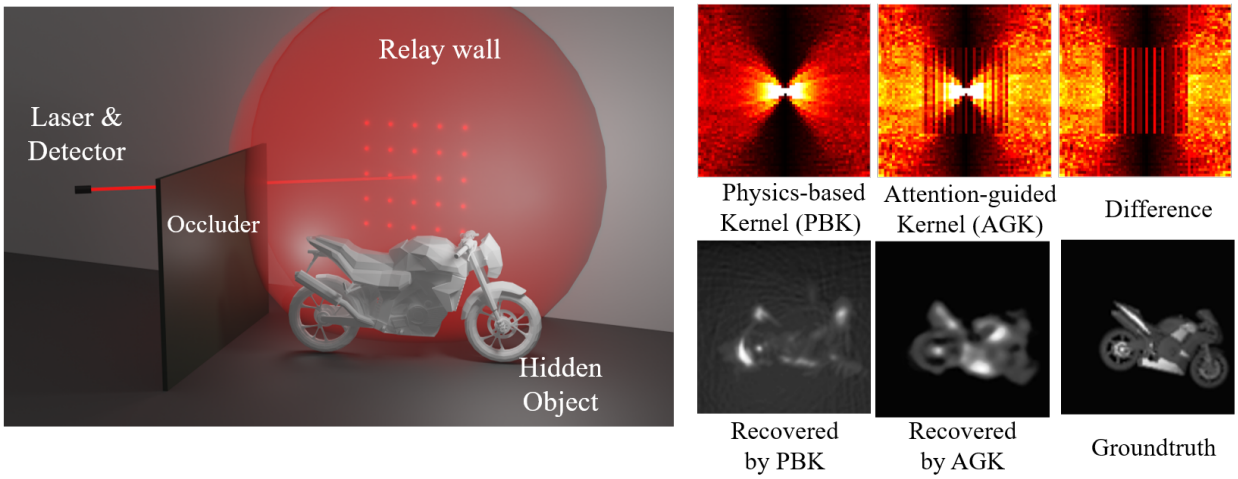
Enhancing Non-line-of-sight Imaging via Learnable Inverse Kernel and Attention Mechanisms

Yanhua Yu, Siyuan Shen, Zi Wang, Binbin Huang, Yuehan Wang, Xingyue Peng, Suan Xia,
Ping Liu, Ruiqian Li, and Shiying Li
Shanghai Engineering Research Center of Intelligent Vision and Imaging,
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{yuyh, shensy, wangzi, huangbb, wangyh8, pengxy, xiasa, lirq1, lishy1}@shanghaitech.edu.cn



(a) A NLOS Setting

(b) Fourier Spectrum of Different Kernels and Depth Recovery

Figure 1: (a) A typical NLOS imaging setup includes a pulsed laser that illuminates a part of the relay wall, and a time-resolved detector that captures the returning photons after multiple bounces between the wall and the hidden object. (b) Top: Fourier spectrum of different kernels. Bottom: Recovered depth maps using different kernels. Our proposed attention-guided kernel has similar low-frequency components (centered in the image) to the physics-based kernel, but contains more high-frequency components (side parts in the image), resulting in a finer reconstruction.

Abstract

Recovering information from non-line-of-sight (NLOS) imaging is a computationally-intensive inverse problem. Most physics-based NLOS imaging methods address the complexity of this problem by assuming three-bounce reflections and no self-occlusion. However, these assumptions may break down for objects with large depth variations, preventing physics-based algorithms from accurately reconstructing the details and high-frequency information. On the other hand, while learning-based methods can avoid these assumptions, they may struggle to reconstruct details without specific designs due to the spectral bias of neural networks. To overcome these issues, we propose a novel approach that enhances physics-based NLOS imaging methods by introducing a learnable inverse kernel in the Fourier domain and using an attention mechanism to improve the neural network to learn high-frequency information. Our method is evaluated on publicly available and new synthetic datasets, demonstrating its commendable performance compared to prior physics-based and learning-based methods, especially for objects with large depth variations. Moreover, our approach generalizes well to real data and can be applied to tasks such as classification and depth reconstruction. We will make our code and dataset publicly available:https://sci2020.github.io.

1. Introduction

Current non-line-of-sight (NLOS) imaging techniques typically adopt pulse lasers and time-resolved detectors to image hidden objects behind obstacles or around corners, as illustrated in Fig.1(a). To reconstruct the hidden objects from the detector's measurements, known as NLOS transients, most physics-based NLOS imaging methods, such as Light-cone transform (LCT) [25], assume that the NLOS scene only involves three-bounce reflections and has no self-occlusion, which simplifies the problem into a linear one in the Fourier domain. However, these assumptions may not hold for more complex objects with large depth variations, which are common in practical NLOS tasks, such as tilted vehicles in autonomous driving. In these scenarios, the invalid assumptions of most physics-based methods make it challenging to recover the high-frequency details of the hidden objects. Moreover, studies have found that objects with large depth variations in NLOS problems typically have complex geometries and normal distributions, which may result in a loss of high-frequency information in the Fourier domain [17] due to the limited NLOS aperture [18, 22], rendering the NLOS imaging problem highly ill-posed.

Learning-based methods avoid the assumptions, such as the three-bounce reflections and no self-occlusion, in physics-based NLOS reconstruction methods and leverage learned scene priors to compensate for the loss of high-frequency information in NLOS imaging process [7, 6, 23]. These methods achieve better results in scenes with large depth variations compared to physics-based methods. However, existing learning-based methods lack dedicated designs to specifically address the problem of high-frequency information loss in scenes with significant depth variations. Furthermore, due to the spectral bias of neural networks [28, 4], the networks tend to learn low-frequency components, which makes them less sensitive to high-frequency details during NLOS reconstruction. Therefore, the current deep learning methods still face challenges in reconstructing high-frequency details in NLOS imaging, resulting in unsatisfactory reconstruction for complex objects with large depth variations.

To address the challenge of reconstructing high-frequency information and details in NLOS imaging, we propose an end-to-end deep learning framework that primarily operates in the Fourier domain, which is consistent with physics-based methods. We utilize a 3D convolutional neural network (CNN) and the fast Fourier transform (FFT) to extract frequency features (F-features) from raw NLOS transients. The inverse kernel of physics-based methods does not accurately capture high-frequency information from non-linear effects like self-occlusions, especially for objects with large depth variations. We thus make the inverse kernel learnable in the Fourier domain and use self-attention and cross-attention to guide the network to embed learned scene priors into the low-frequency and high-frequency components of the kernel, respectively. As shown in Fig.1(b), the kernel guided by attention mechanisms contains richer high-frequency information than the physics-based kernel, resulting in finer reconstruction. With the extracted F-features and the learned inverse kernel, we multiply them and use iFFT to obtain spatial features (S-features), which resemble physics-based methods. The network-obtained S-features are suitable for end-to-end training for various tasks, such as 2D imaging, depth reconstruction, and object classification [3].

We evaluate our proposed method on three syntheic datasets, as well as additional experimental data. First, we validate our method on our new synthetic datasets for different tasks like 2D imaging, depth estimation and classifications. Additionally, we test our method on a public dataset of motorbikes [6]. We demonstrate that our method can recover high-frequency details and get higher accuracy. Finally, we show the effectiveness of our method on a public dataset [23].

We summarize our contributions as follows:

- We propose an end-to-end deep learning framework that learns an inverse kernel in the Fourier domain to reconstruct high-frequency information and details in NLOS imaging, particularly for objects with large depth variations.
- We evaluate our method on three different synthetic datasets and additional experimental data on various tasks, such as 2D imaging, depth reconstruction, and object classification.

2. Related Work

We briefly review most relevant prior works and refer readers to recent surveys [8, 9] for an overview.

Physics-based reconstruction. Kirmani et al. [14] first propose non-line-of-sight imaging for recovering hidden scenes around corners, using time-resolved equipment to capture the time-of-flight of rebounced light, which is known as transients. However, the intractable multipath light transport incurs prohibitive computation cost to recover information of the hidden object, which is an ill-posed inverse problem. In order to approximate a linear inverse operator, a variety of approaches makes efforts to simplify the light transport by making certain assumptions [14, 25, 1, 29, 11, 26, 37, 20]. Backprojection (BP-) based methods build a 3D probability map of the hidden object geometry instead of solving the accurate light transport and use a Laplacian filter to sharpen the volumetric result as post-processing [13, 14, 10, 2, 35]. LCT [25] makes isotropic assumption of the light reflection and introduces the confocal NLOS setting. Ahn et al. [1] report that the inverse kernel, e.g., computed using Wiener filtering [25, 38],

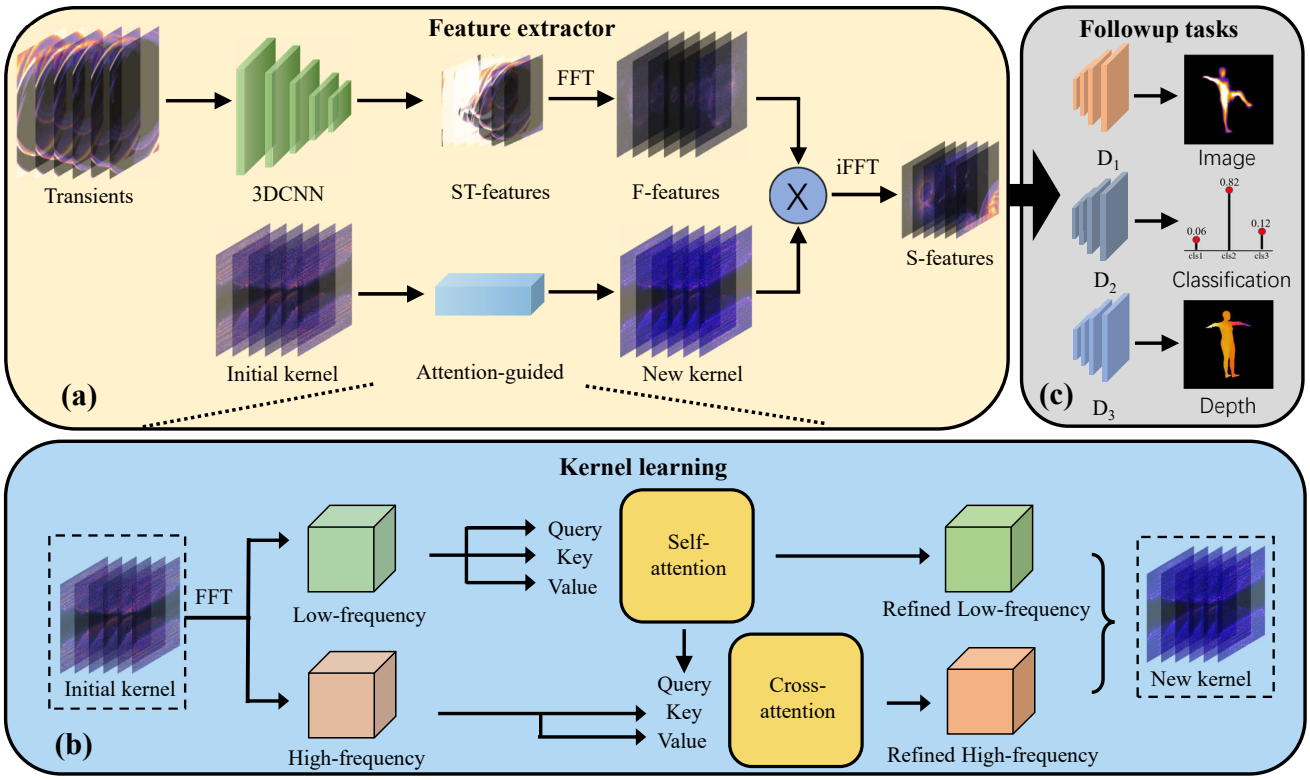


Figure 2: Overview of our framework. In the top row we illustrate our framework. Our framework first takes as input a raw transient and output a downsample feature volume (S-T features) with a 3D CNN. Then we map the feature volume into a Fourier space so as to convolve with our learned kernel $\mathcal{K}$, yielding spatial features (S-feature) that suits followup tasks. In the bottom row, we show how we learn our kernel. Our kernel is implemented as a learnable Fourier volume that is initialized with LCT [25]. We then conduct our proposed attention guidance learning to enhance the $\mathcal{K}$ for light cone transform in the deep Fourier space.

resembles a Laplacian filter and propose a convolutional Gram operator to refine the reconstruction. Wave-based methods transform light transport as wave propagation and can handle complex NLOS scenes [15, 19, 24, 34]. Physical model-based methods, however, suffer from a critical limitation in which they are unable to account for nonlinear light propagation, including self-occlusions. This issue is particularly acute when the depth of the hidden object exhibits significant variations. We utilize neural networks to induce the precise inverse light transport operator using a substantial amount of data, thus overcoming the limitations of traditional physical models.

Learning-based reconstruction. Data-driven solvers [6, 31, 16, 23, 39, 26, 30] learn 2D or 3D features from input transients of hidden objects without strong assumptions. Both traditional and learning-based approaches consider the inverse operator to be spatial-invariant and optimize the NLOS reconstruction. Chen et al. [6] propose an end-to-end learning framework capable of extracting features from the transients for different downstream tasks. They introduce a feature propagation module based on physical model to ensure stable training. Mu et al. [23] incorporate the physical priors of wave propagation and volume rendering into neural networks for robust NLOS reconstruction. Similar to Chen et al. [6], Liu et al. [16] introduce additional priors to enhance output features. However, the utilization of an approximated inverse kernel in these frameworks presents a constraint that hinders the learning of features for scenes with highly variable depth. Additionally, previous works have predominantly focused on the extraction of spatial-temporal features, overlooking frequency domain learning, which limits the ability to obtain a more precise and memory-efficient feature extractor. Our method utilizes the widely used attention mechanism [32, 21, 27, 36]to guide the kernel estimation and learn accurate features in the frequency domain.

3. Methods

Our end-to-end learnable framework is capable of learning attention-guided inverse kernel and versatile feature embeddings from only NLOS transients, which turns out to be useful for downstream tasks like 2D imaging, depth reconstruction and object classification. It first learns an attention-guided inverse kernel that effectively captures high-frequency information during reconstruction. The network then embeds NLOS transient data into a high-dimensional latent space, which helps extracting meaningful cues and priors. This allows mapping learned transient features into the spatial domain so as to conduct downstream tasks. Our pipeline is illustrated in Fig. 2.

3.1. NLOS imaging model

Transients. The input transients, denoted by $\tau \in \mathbb{R}^{T\times H\times W}$, conform to the widely-used confocal setting, where the illumination point and the scanning point overlap. The transients record the photon counts with their arrival time and are formulated as follows:

$$\tau(t, x', y') = \int_{\Omega} \frac{1}{r^4}\rho(x, y, z)\delta(t - \frac{r}{c})dV \tag{1}$$

where (x', y') represents the location of the illumination and scanning points, while Ω denotes the entire hidden scene. The function δ represents a hemisphere centered at (x', y') with radius $r = ct$, representing the length of the light path. The parameter ρ denotes the albedo of the hidden object and encodes the geometric information. As indicated by Eq. 1, the transients represent an integral signal of scattered light. Recovering the hidden objects from these transients is a non-trivial inverse problem. Most physics-based methods make some assumptions to introduce the inverse kernel $\mathcal{K}_0$ (the physics-based kernel (PBK) in Fig. 1), and recover the NLOS image by

$$\rho = \mathcal{R}'\{\text{IFFT3D}\{\text{FFT3D}\{\mathcal{R}\{\tau\}\} \odot \mathcal{K}_0\}\} \tag{2}$$

where $\mathcal{R}$ and $\mathcal{R}'$ are resampling operators at a temporal or depth axis [25]. $\odot$ is an element-wise multiplication. However, the linear inverse kernel $\mathcal{K}_0$ is inaccurate especially in high-frequency domain (see Fig.1(b)). Thus we leverage the attention mechanism to guide to learn the inverse kernel in low-frequency and high-frequency domains, respectively.

3.2. Attention-guided kernel learning

We estimate an inverse kernel $\mathcal{K} \in \mathbb{C}^{\hat{t}\times\hat{h}\times\hat{w}}$ that is jointly learned from the data distribution. We aim to learn the inverse kernel that translates from transients to spatial domain in the deep feature space. Transient data are usually sparse and have difficulties in encoding high-frequency details. We design our framework to learn high-frequency components that traditional methods cannot capture. To learn such a kernel, we propose to model the low frequency part and a high-frequency part separately. We then employ attention mechanisms to enhance both components during the training.

More specifically, we obtain a learnable kernel $\mathcal{K}^o$. We initialize this kernel with LCT [25] and find that other initializations are also plausible. We then separate it into two parts according to the frequency coordinates, as follows

$$\begin{aligned} S_{\text{low}} &= \{\mathcal{K}^o_{ijk} : i < t/2 \;\&\; j < h/2 \;\&\; k < w/2\} \in \mathbb{C}^L \\ S_{\text{high}} &= \{\mathcal{K}^o_{ijk} : i \geq t/2 \;\&\; j \geq h/2 \;\&\; k \geq w/2\} \in \mathbb{C}^L. \end{aligned} \tag{3}$$

where $L = (\hat{t} \times \hat{h} \times \hat{w})/2$. This process simply divides the kernel $\mathcal{K}^o$ into two equal-size components that contain different frequency 'tokens' for the following attention-based learning. We adopt this 50-50 separation in our work because, unlike previous methods, our kernel is learnable. Due to the self- and cross-attention mechanisms, our framework enables weight adjustment by integrating information from different frequencies, even when separation is not optimal. We provide detailed experimental results in Sec. 4.4 to test the effectiveness of different separations, and find that the differences in performance metrics are not significant.

Given two components of low and high frequency extracted from the kernel, we conduct attention guidance learning. Following the processing in [32], we apply position encoding on the two components into $\bar{S}_{low}$ and $\bar{S}_{high}$ so as to learn kernel embedding that correlates to the frequency coordinate. We employ the low-frequency components S_{low} to refine the low-frequency tokens, as low frequency signals represent structural information, such as the coarse geometry and layout of the hidden object. We first project the components S_{low} into two subspaces as query Q_{low} and K_{low} using two tiny linear layers. We then update the low frequency components as follows:

$$\begin{aligned} H_{low} &= \text{softmax}(Q^T_{low}K_{low}) \cdot S_{low} \\ \hat{S}_{low} &= \sigma(H_{low}) \odot S_{low} \end{aligned} \tag{4}$$

where $\odot$ is an element-wise multiplication.

High-frequency details are highly correlated to the low-frequency structure. To refine the high-frequency components, we exploit cross attention from the low frequency representation obtained above to guide the learning of high frequencies. More specifically, we use the updated low-frequency components as the query and high-frequency component as the key and the value. The high-frequency components are then processed with cross attention, as follows:

$$\begin{aligned} H_{high} &= \text{softmax}(\hat{S}^T_{low}S_{high}) \cdot S_{high} \\ \hat{S}_{high} &= \sigma(H_{high}) \odot S_{high} \end{aligned} \tag{5}$$

After acquiring both $\hat{S}_{low}$ and $\hat{S}_{high}$, we recombine the two components into a convolutional kernel $\mathcal{K}$ to conduct inverse projection from the transient latent into spatial features, as depicted in Fig 2.

3.3. Feature extractor

NLOS feature encoding. Given a 3D transient volume $\tau \in \mathbb{R}^{t\times h\times w}$, we first encode it into a 4D feature volume $C \in \mathbb{R}^{d\times t/f\times h/f\times w/f}$ with a 3D convolutional network $\mathcal{F}_{pre}$ modified from [6]. This network has two components: a deep 3D CNN C_d and a short-cut convolutional layer C_s. The deep CNN consists of one 3D convolutional block and two residual blocks to downsample the input transient volume into the latent space by a factor of $f = 4$. The short-cut convolutional layer is implemented with kernel size 1, stride 4 and is initialized with ones, so that at the training beginning this serves as a pooling layer. We find that such design enables our training stable, especially benefiting our optimization of learnable kernel. We then concatenate the outputs to obtain the transient latent:

$$C = C_d(\tau) \oplus C_s(\tau) \in \mathbb{R}^{d\times t/4\times h/4\times w/4} \tag{6}$$

Feature propagation. Given the transient latent from the 3D CNN, we aim to mapping it to spatial domain so as to conduct downstream tasks in a pixel-align manner. Several linear and approximated kernels exist, such as BP, LCT and PF. However, none of these methods produce desirable results because they only model simplified light propagation and are not designed for transients in the latent space. To model accurate light transport for complex scenes, we instead adopt a learned kernel $\mathcal{K}$ that makes good use of geometric cues and priors from the training data distribution, as described in Sec. 3.2.

To map the transient latent into the spatial space, we first pad the output transient latent to $t/2\times h/2\times w/2$, which are then mapped to Fourier representation with FFT. The zero padding improves frequency resolutions as zero padding in the spatial domain results in increase of frequencies in the Fourier domain. We find that this process benefits our learning of a high resolution $\hat{\mathcal{K}}$ and consequently boosts the final results. Specifically, we obtain the spatial features S as follows

$$\bar{C} = \text{PAD}(C) \in \mathbb{R}^{d\times t/2\times h/2\times w/2} \tag{7}$$

$$S = \text{IFFT3D}\{\text{FFT3D}\{C\} \odot \mathcal{K}\} \tag{8}$$

where PAD is a zero padding operation and $\odot$ is a element-wise multiplication.

3.4. Task-specific decoding

The learned feature representation allows us to address different NLOS tasks. Based on a specific downstream task, we design the decoder called $\mathcal{D}_{post}$ accordingly and connect it to the extractor to process our 3D features. For 2D image reconstruction and depth estimation tasks, we collapse the 3D features into a 2D feature map along the depth axis, and then use a convolutional network with upsampling layers to obtain an image with the same resolution as the target image. We then adopt the mean square error (MSE) loss as the objective for these regression tasks. For the digit recognition task, we directly use the 3D features as input and utilize a 3D convolutional network as the backbone to extract distinguishable features for classification. We then map the features to the probability distribution of different categories using the cross entropy loss. In a nutshell, we decode the spatial feature according to tasks,

$$T = \mathcal{D}_{post}(\text{S}) \tag{9}$$

where $\mathcal{D}_{post}$ is a task-specific head that can be implemented as an image regressor or a classifier.

4. Implementation

4.1. Datasets

We evaluate our work using three synthetic datasets, one of which is a publicly available dataset in [6]. The other two datasets are generated at small time bins of 4 ps and are convolved with Poisson noise by mimicking the NLOS imaging system with a SPAD [12].

Public Motorbikes Dataset. The motorbikes dataset comprises a total of 1385 samples, including 277 motorbikes rendered from the ShapeNet [5] dataset in 5 different random positions. The size of the transients is $256 \times 256 \times 512$. We divide the dataset into three subsets: 1000 samples for training, 100 samples for validation, and 100 samples for testing.

New Synthetic Datasets. The new synthetic datasets comprise two sub-datasets. one is **Digits Dataset** including digits from 0 to 9. We randomly sample 144 poses for each digit to synthesize a total of 1440 samples. We then split the dataset into 1200 for training and 240 for testing. The other is **Poses Dataset**, serving as a complementary addition to [16]. This dataset encompasses 2642 transient samples, each of which incorporates a pose with depth variations of at least 1 meter. These samples comprise transients of $256 \times 256 \times 512$ along with corresponding ground truth images of 256×256. The dataset has been partitioned into training, validation, and testing sets, distributed in an 8:1:1 ratio.

Real data. To further evaluate the generalization ability of our approach to real-world data, we acquire a set of real data using a confocal NLOS setup. Our hardware system consists of an ultrafast pulsed laser (SuperK 299 EXTREME FIU-15, 47 mW average power, 670 nm, 20 ps pulse width, 39 MHz repetition rate), a fast-gated SPAD (12%detection efficiency, 200 ps time jitter) using a delayer

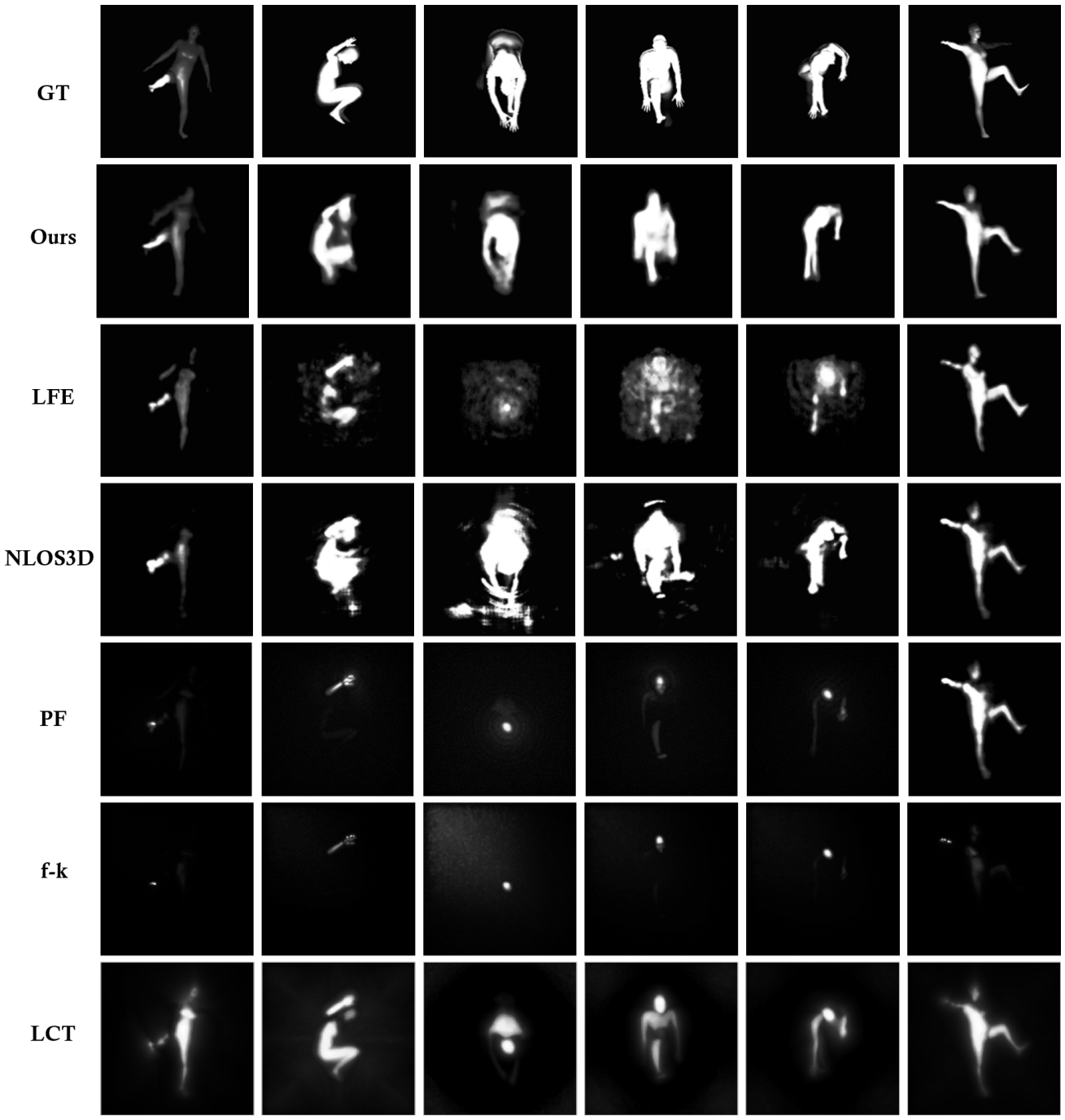


Figure 3: Qualitative comparison of our method and LFE [6], NLOS3D [23], PF [18], LCT [25] and f-k [15]. Our method is capable of reconstructing images of various poses whereas SOTA methods are inaccurate or suffer from different levels of artifacts.

(PicoQuant PSD-065-A-MOD) with gate width 12ns and a galvo scanning system. The laser scans the relay wall while the SPAD focuses on the same position as the laser. For each measurement, we scan a 64×64 grid within $1\text{m}\times1\text{m}$ square, with a 2-second exposure time for each scanning point. We accumulate photon counts within the exposure time at each bin by a photon counter (PicoHarp 300). The captured transient has 6411 bins with a temporal resolution

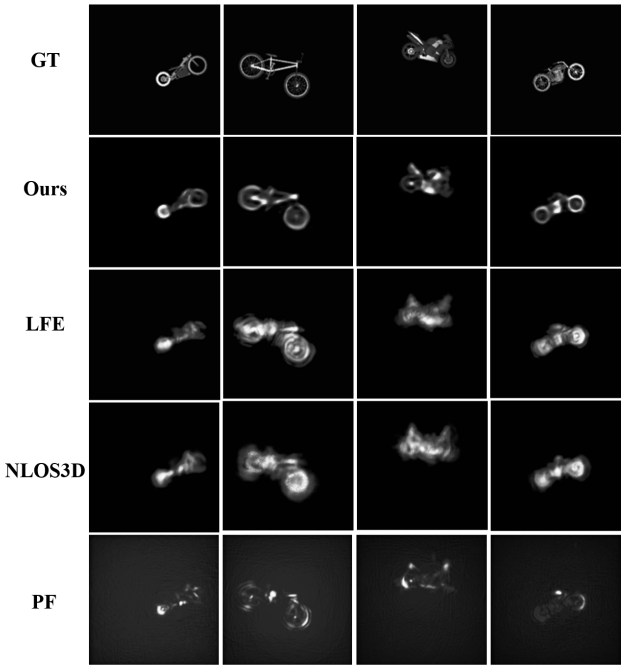


Figure 4: **Comparison of reconstruction results on motorbikes dataset.** The results of our method are sharper and more detailed than SOTA methods, especially in the wheels.

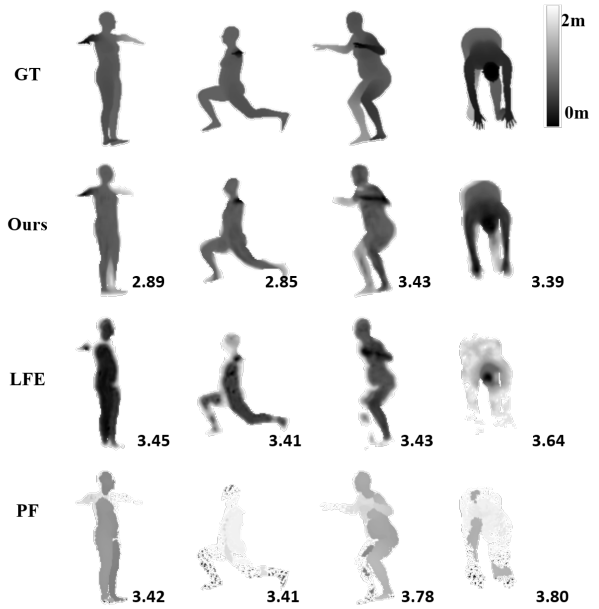


Figure 5: Results of depth estimation on poses dataset. Our method successfully recovers the hidden objects in a larger depth range.

of 4 ps. We reduce the 6411 bins to 4096 bins by discarding the last bins, which contains little information of the hidden object.

4.2. Implementation details

All experiments are conducted on the PyTorch platform using an NVIDIA RTX 3090 GPU card. For input transients, we encode the transients into a feature volume of $32 \times 32 \times 64 \times 32$ and batch size 4. For motorbikes and digits datasets, we encode the transients into a feature volume of $128 \times 128 \times 256 \times 4$ and batch size 1. We adopt AdamW as our optimizer, with $lr = 1e-5$, $\beta_1 = 0.5$ and $\beta_2 = 0.99$. All experiments are trained with 90k iterations except motorbikes, which we only train 50k iterations as we find that it has already converged.

4.3. Comparisons with previous methods

Baselines. We compare our method with state-of-the-art (SOTA) methods, including LFE [6], NLOS3D [23], PF [18], LCT [25], and f-k [15]. Two baselines are LFE [6] and NLOS3D [23]. LFE [6] is an end-to-end learning system that also learns feature embedding and conduct feature propagation with a physics-based kernel. NLOS3D [23] uses an encoder which equips Rayleigh-Sommerfeld diffraction (RSD) as a physical feature propagation, combined with a conditional radiance field to predict both density and color and render images via volume rendering. We implement these baselines and adapt them to our tasks for fair comparisons. For LFE [6], we directly adopt the public code and keep the hyper-parameters the same as described in the paper. For NLOS3D [23], we slightly modify their framework since their method is implemented for non-confocal setting. Specifically, we replace their RSD function with Phasor Fields [18] and only use single-view images as supervision. We train them with the same batch size and iterations as ours, but keep other hyper-parameters as the same as their original implementation.

2D Image reconstruction. We report the quantitative 2D imaging results on poses dataset in Tab. 1. Our method achieves better PNSR and SSIM, surpassing both LFE [6] and NLOS3D [23]. As shown in Fig 3, our method is capable of reconstructing images of various poses and large depth variations because we learn a disentangle representation of the inverse transport kernel that preserves both coarse shape and highly-detail edges. By contrast, LFE [6] and NLOS3D [23] are prone to produce noise and artifacts, harming the final results, especially in the cases that the approximated kernel does not match the actual light transport. While other baselines typically perform on a specific scenes (as is shown in the last column), our method is capable of recovering both shape and details across different poses. We further validate this in the motorbikes experiment, where our technique consistently outperforms these baselines in both quantitative measurements and qualitative visualizations (as Fig. 4 shows). Our method is the only one that clearly reconstructs the high frequency details of the motorbikes, e.g., the shape and contour of the wheels.

Table 1: Quantitative comparison of our technique and the baselines. Our algorithm demonstrates superior performance on imaging task compared to the LFE [6] and NLOS3D [23] algorithms on two distinct datasets.

Dataset	Methods	MSE↓	PSNR↑	SSIM↑
Poses	LFE [6]	0.09	21.43	0.84
	NLOS3D [23]	0.08	22.65	0.91
	Ours	**0.06**	**24.59**	**0.94**
Motorbikes	LFE [6]	0.07	25.79	0.92
	NLOS3D [23]	0.08	25.59	0.91
	Ours	**0.05**	**26.61**	**0.94**

Depth estimation. We conduct depth estimation experiments on the poses dataset with significant depth variations. Figure 5 illustrates the results for four different postures. In the first three poses, since the hands are up, the depth of the arm exhibits significant variation and the self-occlusion is severe. PF [18], which assumes a physical based kernel, only recovers the basic shape and fails to estimate the depth. LFE [6] uses a convolutional neural network to extract dense spatio-temporal domain features, allowing it to obtain a more complete shape. However, the lack of corrections for the approximate physical model prevents LFE from recovering clear details, such as the depth of the arm. In contrast, our method can accurately predict the depth information of hidden objects using a learnable kernel. For the most challenging posture of the bent pose with a most severe self-occlusion, the PF and LFE can only recover the depth of a limited region (e.g., the head). In comparison, our method can achieve accurate depth estimation, including the obscured parts such as the legs, by sharing information at high and low frequencies in the kernel.

Digits classification. We utilize ResNet3D-50 [33] as the downstream decoder for digit classification. We report standard metrics for classification tasks, including precision, recall, and accuracy on average. The results of our experiments are presented in Tab. 2.

Results on real data. Although Our method is trained with synthetic data, we demonstrate its ability to generalize to real-world scenarios. We capture NLOS transients of a mannequin with various poses using our system. Due to limited GPU memory, the time resolution of our method is constrained. As a result, we need to resample the measured data from its original bin resolution of 4e-12 with 4096 bins to a bin resolution of 3.2e-11 with 512 bins. We then feed these raw transients into our network trained on the poses dataset to predict their imaging, as shown in Fig 6. In addition, we compare our results to two methods: a learning-based counter-part LFE [6] and a SOTA backprojection-based PF [18]. We show that our method clearly recovers the overall shape and significantly reduces artifacts.

Table 2: Digits classification results. Our method consistently outperforms LFE [6].

Methods	precision	recall	accuracy
LFE [6]	0.91	0.89	0.89
Ours	**0.95**	**0.94**	**0.94**

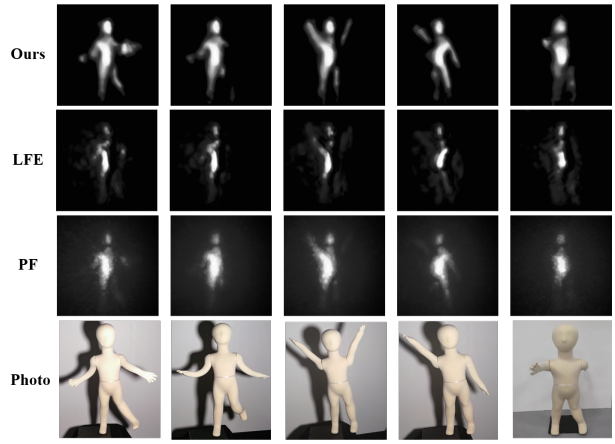


Figure 6: Results on real data. Our method can generalize to the captured real data, matching and even outperforming the learning-based counter-part LFE [6] and the backprojection-based PF [18].

4.4. Ablation studies and analysis

We carry out three ablation analyses: Low- and High-frequency attention mechanisms,Different kernel separation, and Generalization of learned kernels.

Low- and high- frequency attention. We aim to validate the effectiveness of our frequency attention design to leverage the importance between $\mathcal{K}_{low}$ and $\mathcal{K}_{high}$. To ensure the clarity of the experiment, we conduct our experiments using the poses dataset with significant depth variations. We compare three different attention structures. **LS**: we only add self-attention to the low-frequency kernel and do not perform any operations on the high-frequency part. **LSHS**: we perform self-attention on both the low-frequency and high-frequency kernels. **LSHC** is the scheme we use in the paper, which add the refined low-frequency kernel information to guide the learning of the high-frequency kernel.

Figure 7 showcases the 2D image reconstruction on a bending pose. The **LS** scheme enhances the low-frequency information and preserves the fundamental shape at the front. However, the absence of high-frequency information causes the back of the body to be blurred. On the other hand, the **LSHS** scheme utilizes high-frequency attention to obtain sharper and cleaner details, including half of the arm. Instead of allowing high-frequency kernel learning on its own, the **LSHC** technique utilizes low-frequency information to guide the learning of high-frequency details. This enables the utilization of existing priors in low-frequency to

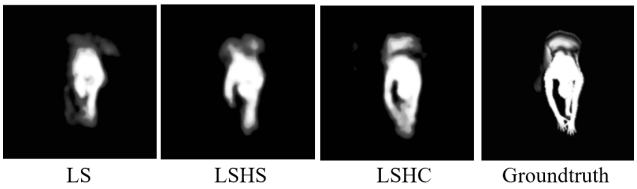


Figure 7: Pose reconstruction using different attention mechanisms: LS, LSHS, LSHC (ours).

infer challenging details, such as the recovery of the entire arm and a clear back body. These results effectively demonstrate the effectiveness of employing cross-attention to learn the frequency domain kernel.

Different kernel separation. The aperture size of different NLOS systems affects the optimal separation of the low- and high-frequency kernels. We include additional results in Tab. 3 to test the effectiveness of different kernel separation, and to show that the differences in performance metrics are not significant because, unlike previous methods, our kernel is learnable. Due to the self- and cross-attention mechanisms, our framework enables weight adjustment by integrating information from different frequencies, even when the separation is not optimal. We also acknowledge that better incorporating prior of frequency distributions is an avenue for future research.

Generalization of learned kernels. Generalization is an inevitable challenge that deep learning methods need to address. We visualize and compare the kernels learned from different datasets. Fig. 8 (a) illustrates the similarity of the kernels. Additionally, we conduct tests across datasets. Fig. 8 (b) presents the recovering results with a model trained on the motorbikes dataset and on the poses dataset. These results showcase good generalization capability of our approach. We observe that our method can learn how to integrate and enhance low- and high-frequency information, rather than relying solely on scene priors.

5. Conclusion

In this paper, we propose an end-to-end deep learning framework that improves the ability of neural networks to learn high-frequency information for NLOS imaging and reconstruction. Our method introduces a learnable inverse kernel in the Fourier domain and using an attention mechanism. Our method avoids the inaccurate inverse kernel that can occur in physics-based methods due to invalid assumptions, and also addresses the limited high-frequency representation problem of neural networks. We evaluate our method on different datasets and demonstrates superior performance compared to previous physics-based and learning-based methods, especially for objects with large depth variations. In addition, our method generalizes well on experimental NLOS data and can be applied to tasks such as NLOS imaging, depth reconstruction and classification.

Table 3: Quantitative results using different kernel separation.

Low - High	MSE ↓	PSNR ↑	SSIM ↑
50% - 50%	**0.05**	**26.61**	**0.94**
25% - 75%	0.06	26.59	0.93
75% - 25%	0.05	**26.61**	0.93

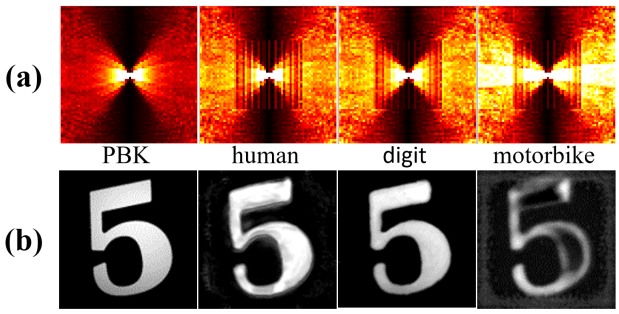


Figure 8: (a) The inverse kernels learned on three datasets. (b) From left to right: Ground truth, images reconstructed by training a model on the respective datasets.

Acknowledgments

The authors appreciate the anonymous reviewers and area chairs for their valuable comments. We also thank Dr. Wenzheng Chen for his helpful discussions. This work is supported in part by Natural Science Foundation of China under contracts Nos. 61977047 and 61976138, and by Science and Technology Commission of Shanghai Municipality under contract No. 21010502400.

References

[1] Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C Sankaranarayanan. Convolutional approximations to the general non-line-of-sight imaging operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7889–7899, 2019.

[2] Victor Arellano, Diego Gutierrez, and Adrian Jarabo. Fast back-projection for non-line of sight reconstruction. In *ACM SIGGRAPH 2017 Posters*, pages 1–2. 2017.

[3] Clara Callenberg, Zheng Shi, Felix Heide, and Matthias B Hullin. Low-cost spad sensing for non-line-of-sight tracking, material classification and depth imaging. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021.

[4] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pages 2205–2211. International Joint Conferences on Artificial Intelligence, 2021.

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

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

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

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

[9] Ruixu Geng, Yang Hu, and Yan Chen. Recent advances on non-line-of-sight imaging: Conventional physical models, deep learning, and new scenes. *APSIPA Transactions on Signal and Information Processing*, 2022.

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

[11] JinHui He, ShuKong Wu, Ran Wei, and YuNing Zhang. Non-line-of-sight imaging and tracking of moving objects based on deep learning. *Optics Express*, 30(10):16758–16772, 2022.

[12] Quercus Hernandez, Diego Gutierrez, and Adrian Jarabo. A computational model of a single-photon avalanche diode sensor for transient imaging, 2017.

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

[14] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using ultrafast transient imaging. *International journal of computer vision*, 95(1):13–28, 2011.

[15] 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.

[16] Ping Liu, Yanhua Yu, Zhengqing Pan, Xingyue Peng, Ruiqian Li, Yuehan Wang, Jingyi Yu, and Shiying Li. Hiddenpose: Non-line-of-sight 3d human pose estimation. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2022.

[17] Xiaochun Liu, Sebastian Bauer, and Andreas Velten. Analysis of feature visibility in non-line-of-sight measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10140–10148, 2019.

[18] Xiaochun Liu, Sebastian Bauer, and Andreas Velten. Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems. *Nature communications*, 11(1):1645, 2020.

[19] 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.

[20] Xintong Liu, Jianyu Wang, Zhupeng Li, Zuoqiang Shi, Xing Fu, and Lingyun Qiu. Non-line-of-sight reconstruction with signal–object collaborative regularization. *Light: Science & Applications*, 10(1):198, 2021.

[21] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Julio Marco, Adrian Jarabo, Ji Hyun Nam, Xiaochun Liu, Miguel Ángel Cosculluela, Andreas Velten, and Diego Gutierrez. Virtual light transport matrices for non-line-of-sight imaging. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2440–2449, 2021.

[23] 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.

[24] Ji Hyun Nam, Eric Brandt, Sebastian Bauer, Xiaochun Liu, Marco Renna, Alberto Tosi, Eftychios Sifakis, and Andreas Velten. Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second. *Nature communications*, 12(1):1–10, 2021.

[25] 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.

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

[27] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11908–11915, 2020.

[28] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[29] Sheila W Seidel, John Murray-Bruce, Yanting Ma, Christopher Yu, William T Freeman, and Vivek K Goyal. Two-dimensional non-line-of-sight scene estimation from a single edge occluder. *IEEE Transactions on Computational Imaging*, 7:58–72, 2020.

[30] Prafull Sharma, Miika Aittala, Yoav Y Schechner, Antonio Torralba, Gregory W Wornell, William T Freeman, and Frédo Durand. What you can learn by staring at a blank wall. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2330–2339, 2021.

[31] 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*, 2021.

[32] 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.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[34] Florian Willomitzer, Prasanna V Rangarajan, Fengqiang Li, Muralidhar M Balaji, Marc P Christensen, and Oliver Cossairt. Fast non-line-of-sight imaging with high-resolution and wide field of view using synthetic wavelength holography. *Nature communications*, 12(1):6647, 2021.

[35] 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):e2024468118, 2021.

[36] Zhang XiaoYu, Shi HaiChao, Li ChangSheng, and LiXin Duan. Twinnet: Twin structured knowledge transfer network for weakly supervised action localization, 2022.

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

[38] Sean I Young, David B Lindell, Bernd Girod, David Taubman, and Gordon Wetzstein. Non-line-of-sight surface reconstruction using the directional light-cone transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1407–1416, 2020.

[39] Shanshan Zheng, Meihua Liao, Fei Wang, Wenqi He, Xiang Peng, and Guohai Situ. Non-line-of-sight imaging under white-light illumination: a two-step deep learning approach. *Optics Express*, 29(24):40091–40105, 2021.