Generalizable Non-Line-of-Sight Imaging with Learnable Physical Priors -Supplementary Material-

Shida Sun Yue Li Yueyi Zhang Zhiwei Xiong
University of Science and Technology of China
{sdsun, yueli65}@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

1. Additional Results

1.1. More Real-world Results

We provide more real-world results under the public data [3, 4] and self-captured data. The reconstructed results from different approaches are shown in Fig. 3 and Fig. 4. Compared with other approaches, our approach successfully recovers the clear boundary and full details for each scene, which demonstrates the effectiveness of our proposed method.

1.2. More Ablation Results

In this section, we provide additional ablation study results on both the synthetic dataset and public real-world data to demonstrate the contributions of each module.

Quantitative experiments. As can be seen in the Tab. 2, integrating two modules achieves the best performance. To clarify the improvement from our design compared to different fixed parameter settings, we ablate LPC in Tab. 3 by replacing it with varying fixed compensation weights, and ablate APF in Tab. 4 by replacing it with varying σ. The above results show that LPC and APF achieve the best performance, which confirms the effectiveness of each module.

Qualitative ablation under real-world data. As shown in Fig. 5, the LPC contributes more details of the object, while the APF suppresses background artifacts. The integration of both modules yields the best results.

1.3. Generalization Evaluation

We provide additional quantitative results to validate the generalization of the approach. In Section 4.3, we present results under different photon acquisition efficiencies. Unlike those settings, in this section, we add different Poisson noise to affect the SNR of the transient measurements. As can be seen in Table 1, in most cases, our approach achieves the best results compared to other approaches. This demonstrates that our network effectively compensates for photon acquisition and reduces noise, resulting in more robust and clearer reconstructed images.

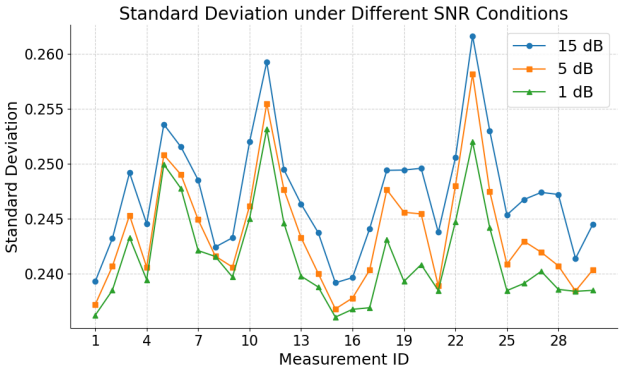

Figure 1. Visualization of the standard deviation predicted by APF. Different colored lines in the figure illustrate how the predicted standard deviation varies under different SNR conditions.

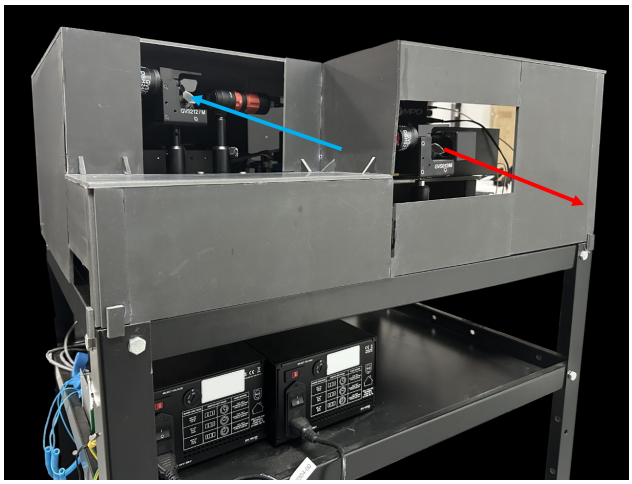

Figure 2. Our own NLOS imaging system.

1.4. Visualization of Standard Deviations

In this section, we provide visualization results of the standard deviation from APF. We randomly select 30 transient measurements from the Seen test set. A lower standard deviation of the Gaussian window in the frequency domain typically indicates a narrower frequency response, leading

Table 1. Quantitative results on the **Unseen** test set under different SNRs by adding varying Poisson noise. The best in **bold**, the second in underline.

Method	Intensity (PSNR↑ / SSIM↑)			Depth (RMSE↓ / MAD↓)		
	15 dB	5 dB	1 dB	15 dB	5 dB	1 dB
LCT	17.86 / 0.1750	17.76 / 0.1301	17.56 / 0.1050	0.6545 / 0.6120	0.7648 / 0.7370	0.7906 / 0.7694
FK	20.57 / 0.5900	20.56 / 0.5867	20.56 / 0.5815	0.8059 / 0.8038	0.7684 / 0.7658	0.7466 / 0.7429
RSD	20.06 / 0.1655	19.60 / 0.1111	18.83 / 0.0833	0.5474 / 0.5134	0.6908 / 0.6610	0.7185 / 0.6919
LFE	21.86 / 0.6489	21.87 / 0.6353	21.69 / 0.5877	0.2881 / 0.1588	0.2882 / 0.1589	0.2892 / 0.1613
I-K	21.74 / 0.7676	21.67 / 0.7315	21.56 / 0.6740	0.2788 / 0.1484	0.2797 / 0.1557	0.2818 / 0.1670
NLOST	21.93 / 0.7716	21.62 / 0.7337	21.81 / 0.6533	0.2659 / 0.1296	0.2715 / **0.1292**	0.2634 / 0.1401
Ours	**22.73 / 0.7950**	**22.52 / 0.7741**	**22.29 / 0.7187**	**0.2645 / 0.1291**	**0.2625** / 0.1308	**0.2623 / 0.1361**

Table 2. Ablation results of each module under the Seen test set. The best in **bold**.

LPC / APF	✗/ ✗	✓/ ✗	✗/ ✓	✓/ ✓
PSNR / SSIM	23.31 / 0.8431	23.69 / 0.8606	23.72 / 0.8603	**24.08 / 0.8704**
RMSE / MAD	0.0987 / 0.0423	0.0957 / 0.0397	0.0914 / 0.0348	**0.0867 / 0.0307**

Table 3. Ablation results between fixed coefficient weights and proposed LPC module. The best in **bold**.

	coeff = 1	coeff = 2	coeff = 4	LPC
PSNR / SSIM	23.25 / 0.8397	23.31 / 0.8431	22.98 / 0.7940	**23.69 / 0.8606**
RMSE / MAD	0.0979 / 0.0438	0.0987 / 0.0423	0.1132 / 0.0626	**0.0957 / 0.0397**

Table 4. Ablation results between fixed σ and proposed APF module. The best in **bold**.

	$\sigma = 0.25$	$\sigma = 0.3$	$\sigma = 0.35$	APF
PSNR / SSIM	23.05 / 0.8208	23.31 / 0.8431	23.30 / 0.8174	**23.72 / 0.8603**
RMSE / MAD	0.1028 / 0.0457	0.0987 / 0.0423	0.1014 / 0.0487	**0.0914 / 0.0348**

Table 5. Quantitative results under varying resolution settings. The approach is trained under the resolution of 256x256.

Resolutions	64×64	128×128	256×256	512×512
Mem / Time	1.5GB / 0.02s	5.2GB / 0.04s	16GB / 0.11s	66GB / 0.57s

to stronger high-frequency noise suppression. As shown in Fig. 1, the standard deviation exhibits a decreasing trend as the SNR decreases, which means APF predicts a tighter Gaussian window for features to filter out more noise.

Furthermore, although numerous values exhibit only minor fluctuations, the superior performance demonstrated in Tab. 4 confirms the necessity of the proposed learning strategy.

1.5. Scaling to Other Resolutions

Our approach is trained with a resolution of 256x256, but it is capable of scaling to other resolutions. The results for scaling to varying resolutions are shown as Tab. 5, demonstrating that our approach can handle both high-resolution data and real-time settings with reasonable efficiency.

2. Imaging Setup

System details. Details of our system are shown in Fig. 2.
Capturing details. The materials in the self-captured scenes can be divided into two types: retro-reflective type (*e.g.*, a bookshelf, and the letters 'NYLT' covered with reflective tape) and diffuse type (e.g., the number '2' covered with white printer paper, a cardboard printed with '123XYZ', and the plaster statues).

3. Details of the Network Structure

Feature extraction module. Given transient measurements as input, the feature extraction module is responsible for downsampling and extracting feature vectors, as well as enhancing the transient measurements. As described in LFE [1], the module consists of two branches. The first branch contains a convolutional layer (kernel size = 3, stride = 1), while the second branch contains a ResNet block [2]. Each ResNet block includes two convolutional layers and one LeakyReLU layer. The output feature dimension of each branch is 1, and the spatial size of the output is 4× smaller than the input data. The two branches are concatenated along the feature dimension and then output.
Spectrum convolution. Firstly, the features of the transient measurements are transformed into the frequency domain using a Fourier transform. Subsequently, a series of 3D convolutional layers with a stride of (1, 2, 2) is applied to

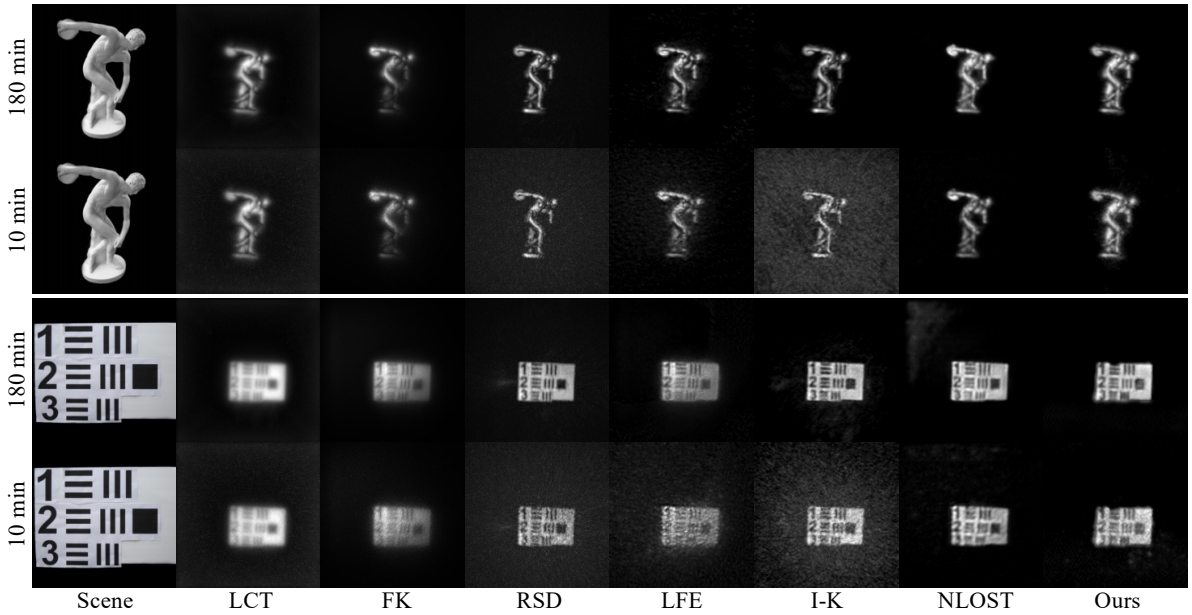

Figure 3. Visualization comparison on the public real-world data [4]. The left annotation indicates the total acquisition time.

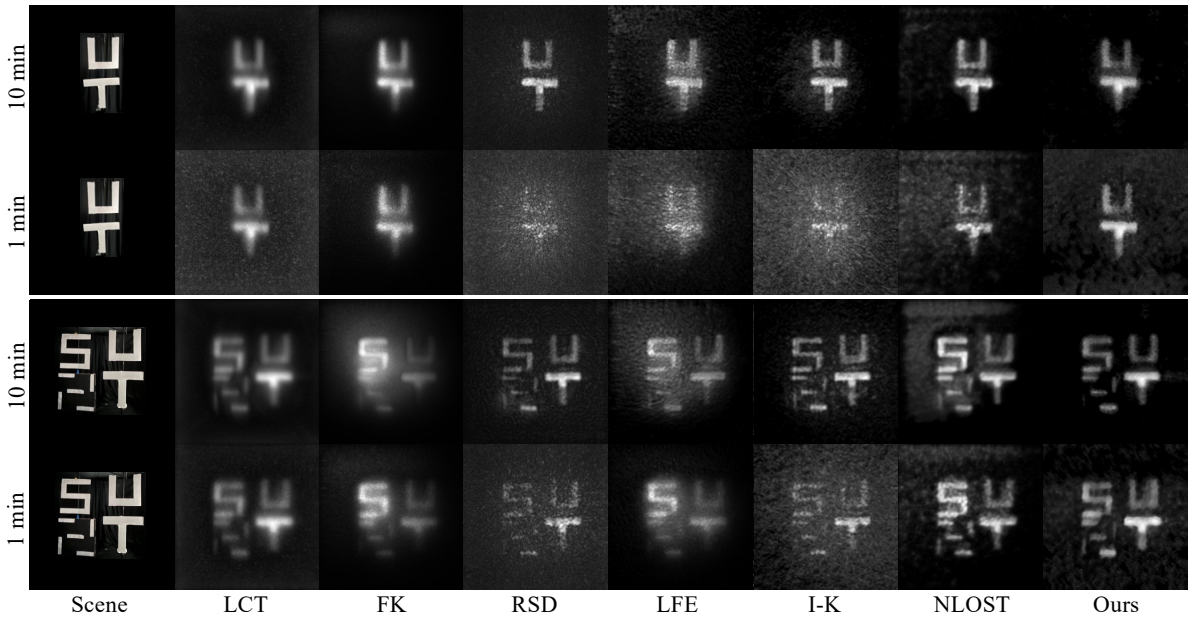

Figure 4. Visualization comparison on the self-captured real-world data. The left annotation indicates the total acquisition time.

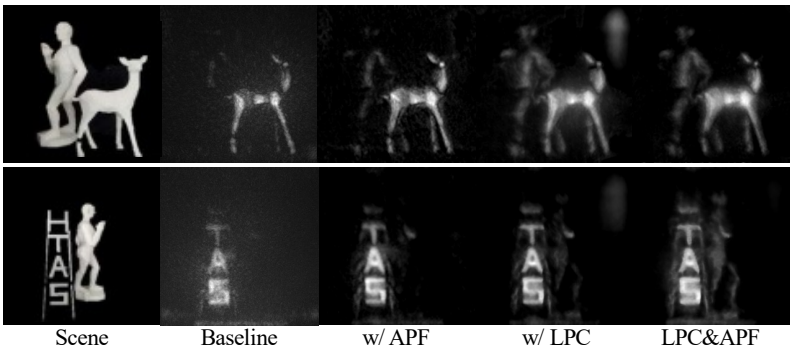

Figure 5. Ablation results on public real-world data. Baseline denotes w/o LPC and APF modules.

reduce the spatial dimensions of the feature vectors, with each convolutional layer followed by a ReLU activation layer. The extracted features are then processed using a 1D convolutional layer, normalized with LayerNorm, and passed through a fully connected layer to fuse the spectrum dimension into a scalar value. Finally, a sigmoid activation computes the desired standard deviation.

Wave propagation module. We utilize the physics-based approach RSD [5, 6] as a wave propagation module. The module transforms the features from the spatial-temporal domain into the linear spatial domain.

Rendering module. The rendering network consists of one convolutional layer and two custom convolutional blocks [1]. Each convolutional block includes one convolutional layer and two ResNet blocks. The structure of the ResNet blocks is the same as in the feature extraction module. The network takes the output from the wave propagation module as input and applies the first convolutional block to obtain enhanced features. Next, the enhanced features are concatenated with the input features along the spatial dimension and processed by the second convolutional block. To ensure more stable training, we employ a residual structure, where the output of each convolutional block is added to the input features, producing the final output of the rendering module (*i.e.*, intensity and depth images).

4. Discussion for the non-confocal system

Eq. (17) in [6] shows σ of the Gaussian-shaped illumination function is determined solely by the virtual source and remains unchanged across systems. Thus, APF also applies unchanged to non-confocal setups. Moreover, since LPC depends on distances between illumination and detection points, retraining on non-confocal transients suffices for extension. In summary, the proposed method is theoretically applicable under the non-confocal setting.

References

[1] Wenzheng Chen, Fangyin Wei, Kiriakos N Kutulakos, Szymon Rusinkiewicz, and Felix Heide. Learned feature embeddings for non-line-of-sight imaging and recognition. *ACM Transactions on Graphics*, 39(6):1–18, 2020. 2, 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] Yue Li, Jiayong Peng, Juntian Ye, Yueyi Zhang, Feihu Xu, and Zhiwei Xiong. Nlost: Non-line-of-sight imaging with transformer. In *CVPR*, 2023. 1

[4] 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, 3

[5] 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. 4

[6] 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. 4