Supplement of Non-line-of-sight Imaging with Signal Superresolution Network

Jianyu Wang[1], Xintong Liu[1], Leping Xiao[1], Zuoqiang Shi[1,2], Lingyun Qiu[1,2], and Xing Fu[1]
[1]Tsinghua University
[2]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications

S1. Introduction

In Sec. S2 of this supplement, we provide more details about training the proposed signal superresolution network (SSN).

In Sec. S3 of this supplement, we show additional experimental results using synthetic data from the test set, measured data from the Stanford dataset [5], and higher scale signal recovery. We also show additional comparisons with other interpolation methods, end-to-end NLOS imaging method and decreasing the exposure duration.

In Sec. S4 of this supplement, we perform the ablation study to show the superiority of the 3-D kernel.

The complete video of the reconstructed dynamic scene is provided in "Dynamic scene.mp4".

S2. Training details

To train the network, we use the L_2 norm as the loss function, which can be written as

$$\mathcal{L} = \sum_{(d_l, d_h)\in\mathbb{S}} ||\Phi(d_l) - d_h||_2^2 \tag{1}$$

in which $||\cdot||_2$ denotes the L_2 norm. $\mathbb{S}$ is the training set. d_l represents the low resolution signal which is the input of the network, Φ is the signal superresolution network, and $\Phi(d_l)$ is the output of the network. d_h represents the high resolution signal.

For both confocal and non-confocal settings, we set the number of 3-dimensional attention-in-attention blocks (3D-A^{2}B) as 16, and we use the Adam [3] optimizer as well as cosine annealing with the initial learning rate as 1×10^{-4}. The batch size is set as 256. It takes about 80 s to run each epoch, and 4.5 days to train the network till convergence on 4 Tesla V100.

S3. Additional results

In this section, we provide additional results to illustrate the effectiveness of the proposed pipeline. The reconstructions are obtained with F-K [5], LCT [7], LOG-BP [4], D-LCT [8] and SOCR [6] methods. For each instance, the parameters are fixed in each algorithm.

S3.1. Results of synthetic data

The proposed pipeline is tested with the instance of the shoe in the test set under the confocal scenario. As shown in Fig. S1b, the SNR of the signal recovered by the proposed network is almost three times higher than the one recovered by nearest neighbor interpolation. Besides, the first arrival time is correctly recovered by the proposed method, while the nearest neighbor method fails at some virtual sources. In addition, state-of-the-art methods can provide high quality reconstructions with the signal recovered by the proposed network. The PSNR and SSIM values of the reconstructed targets obtained with the proposed pipeline are very close to those obtained with the original signal.

S3.2. Results of measured confocal data

In this subsection, we show additional results of the statue from the Stanford dataset. The exposure time for each virtual source is 0.0023 s, 0.0137 s, and 0.0412 s respectively.

For the shortest exposure time, the SNR of the recovered signals of both the nearest neighbor method and SSN are low. However, as shown in Fig. S2b, the signal recovered by SSN is much more similar to the original measurement. This is also illustrated by the reconstruction results shown in Fig. S2c. When the exposure time for each point increases to 0.0137 s (Fig. S3), both methods yield better results. However, the background noise is much higher in the signal recovered by the nearest neighbor method, and the corresponding reconstructions contain more artifacts. For the longest exposure time, the nearest neighbor method still fails to recover the first arrival time at most virtual sources(Fig. S4b).

As illustrated in these three comparisons, the proposed method outperforms the conventional interpolation method in all cases and is more robust to the measurement noise. Furthermore, as shown in the main paper, the proposed pipeline can provide high resolution reconstructions when

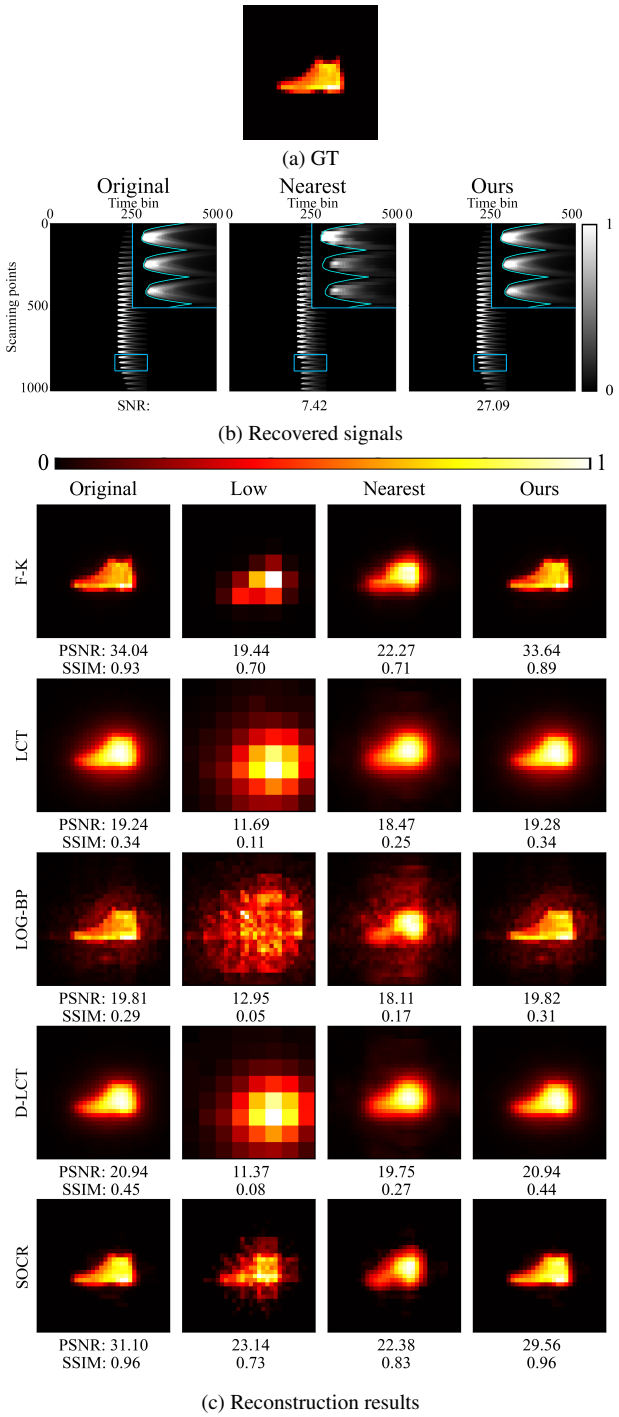


Figure S1. Recovered signals and reconstruction results of the shoe contained in the test set. (a) Ground truth of the hidden object. (b) A comparison of the recovered signals. The first arrival time of the original signal is marked by the blue curve in the zoom-in window of each sub-figure. (c) Reconstruction results of state-of-the-art methods with different signals.

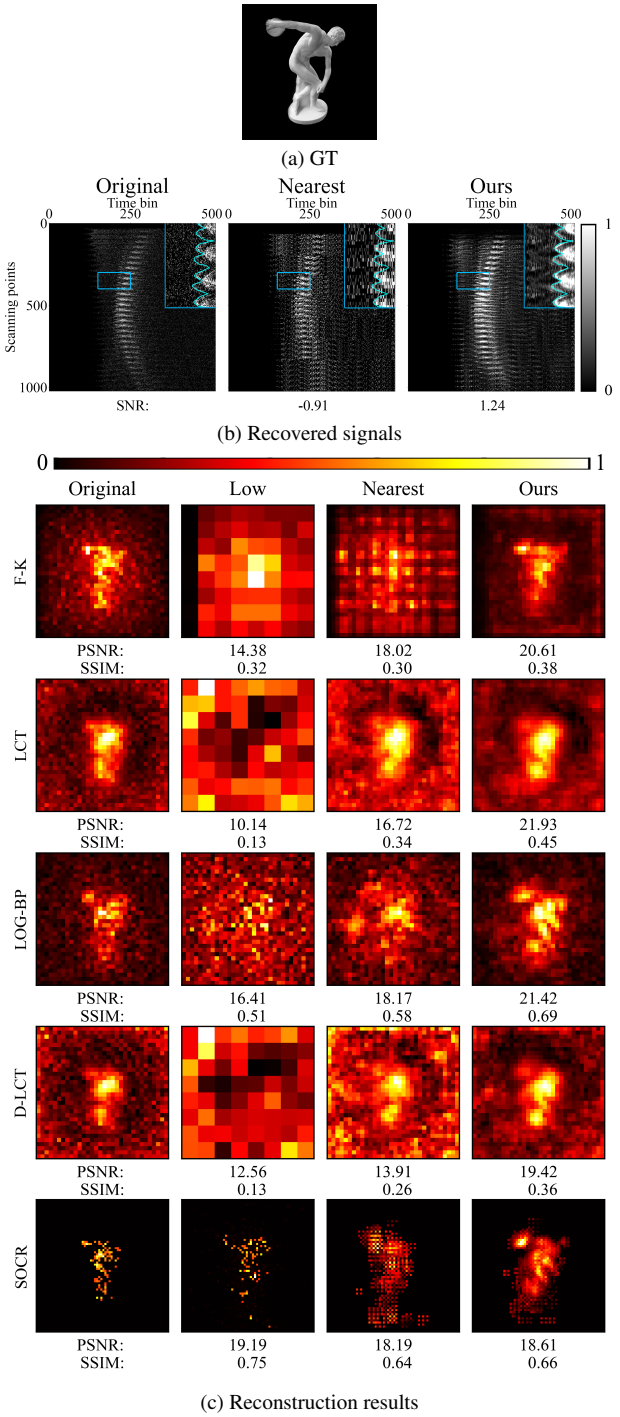


Figure S2. Recovered signals and reconstruction results of the statue. The exposure time of each virtual source is 0.0023 s. (a) Ground truth of the hidden object. (b) A comparison of recovered signals. The first arrival time of the original signal is marked by the blue curve in the zoom-in window of each sub-figure. (c) Reconstruction results of state-of-the-art methods with different signals.

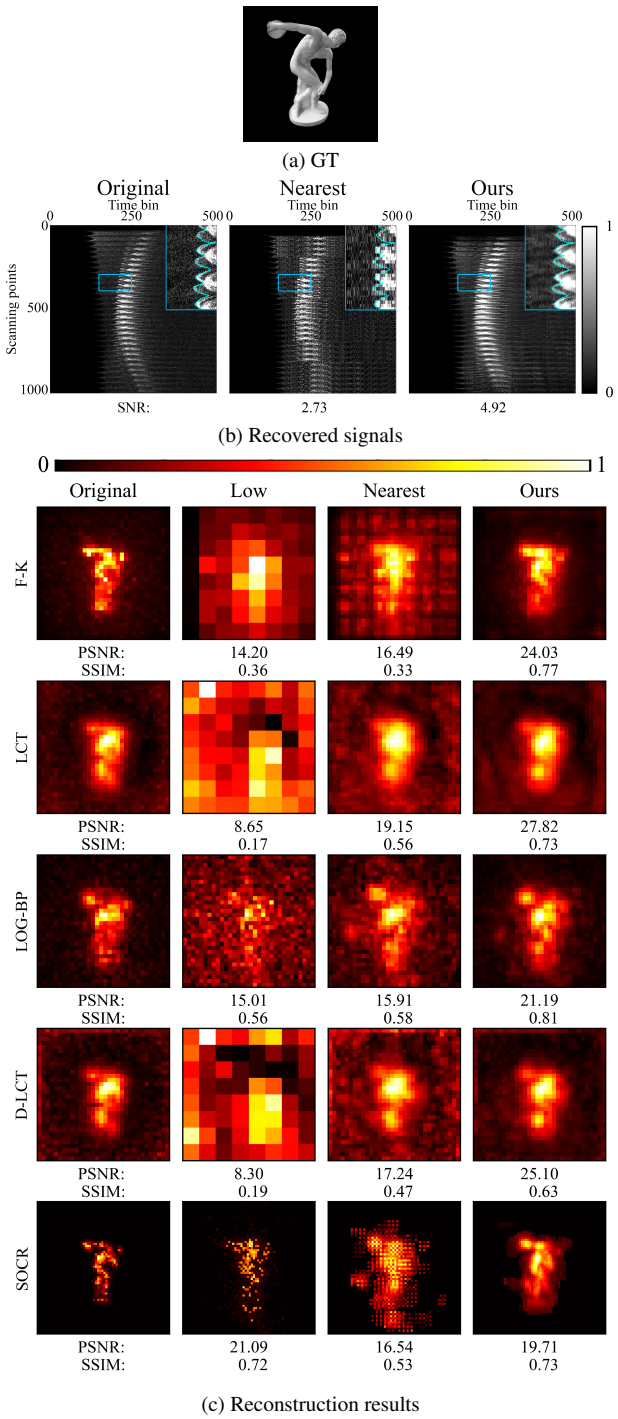


Figure S3. Recovered signals and reconstruction results of the statue. The xposure time of each virtual source is 0.0137 s. (a) Ground truth of the hidden object. (b) A comparison of recovered signals. The first arrival time of the original signal is marked by the blue curve in the zoom-in window of each sub-figure. (c) Reconstruction results of state-of-the-art methods with different signals.

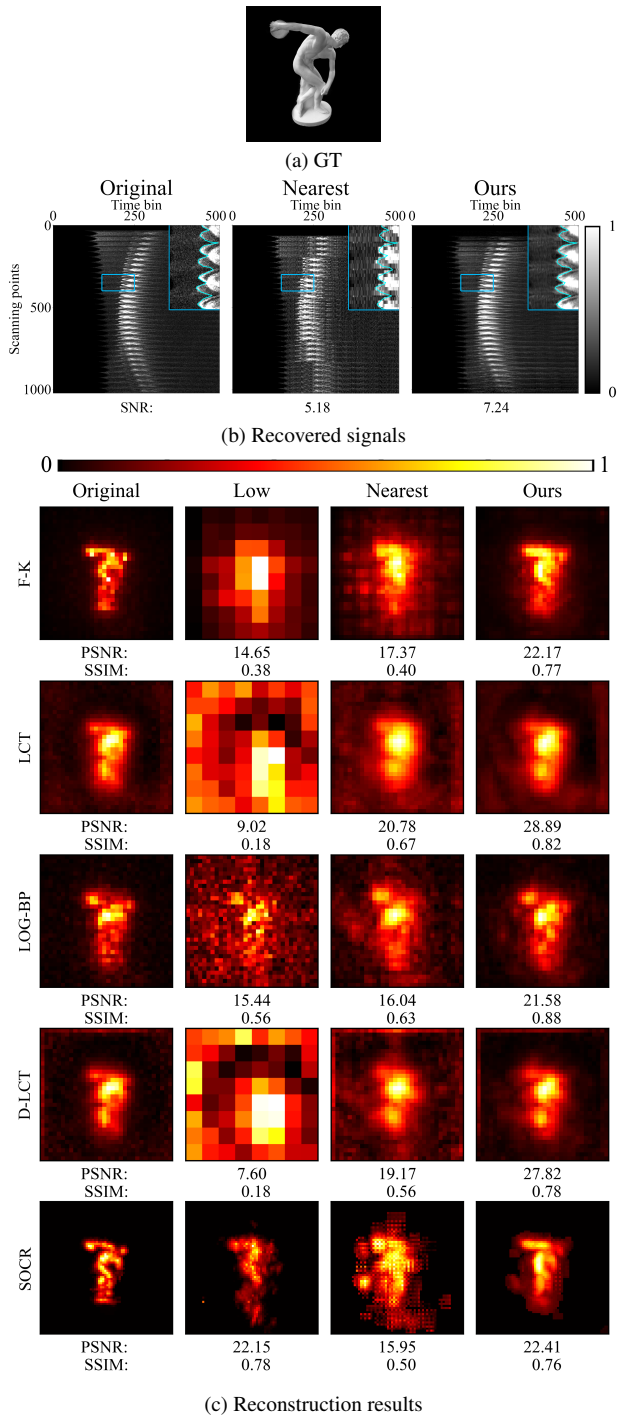


Figure S4. Recovered signals and reconstruction results of the statue. The exposure time of each virtual source is 0.0412 s. (a) Ground truth of the hidden object. (b) A comparison of recovered signals. The first arrival time of the original signal is marked by the blue curve in the zoom-in window of each sub-figure. (c) Reconstruction results of state-of-the-art methods with different signals.

the exposure time is only 0.0069 s at each virtual source.

S3.3. Higher scale singal recovery with the signal superresolution network

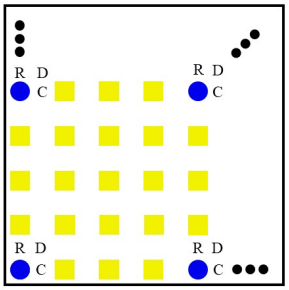
In this subsection, we briefly introduce the $\times 8$ scale recovery, using the SSN trained for $\times 4$ scale recovery. Since the network learns a local recovery operator, we can achieve a higher scale recovery by reusing the SSN several times and combining the results.

The process of $\times 8$ scale recovery involves four steps. The first step is to increase the spatial resolution of the input signal from 8×8 to 32×32. The spacing between adjacent virtual sources of the recovered signal is small enough, which fulfills the assumption introduced in the manuscript. In the second step, three additional 8×8 signals are obtained using nearest neighbor interpolation. As shown in Fig. S5a, the interpolation is done in the adjacent row (R), column (C), and diagonal (D) directions of the blue circle points, respectively. All virtual sources of these signals are not included in the signal from the first step. The third step involves using SSN three times to recover three 32×32 sub-signals. Finally, in the fourth step, the four 32×32 sub-signals are positioned appropriately to create a signal with 64×64 virtual sources.

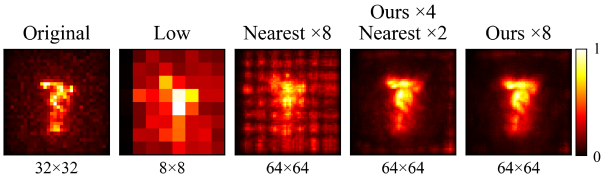
Reconstruction results are shown in Fig. S5b. The first column shows the reconstruction obtained with the original signal. The second column shows the reconstruction obtained from the low resolution signal. The third column shows the reconstruction obtained with the signal interpolated by the nearest neighbor interpolation method, which contains many artifacts. In the fourth column, the signal is first recovered by SSN and then is interpolated by the nearest neighbor interpolation method. The last column shows the reconstruction obtained with the method introduced above. Compared to the fourth column, the background of the result provided by the proposed method contains less noise.

S3.4. Comparison with other interpolation methods

The comparison of the proposed framework with other interpolation methods is shown in Fig. S6. All results are reconstructed by the F-K method. The exposure time of each virtual source is 0.0412 s. As shown in the figure, different interpolation methods lead to different continuity of the results. The nearest neighbor interpolation technique provides discontinuous results, the trilinear and cubic interpolations result in C^0 and C^1 continuity. However, due to the delta function contained in the forward model, the transient images with complete spatial measurements are not continuous. Thus, the nearest neighbor interpolation method is used for comparisons in the manuscript. Furthermore, the proposed framework outperforms all conventional interpolation methods.



(a) Interpolated virtual sources



(b) Reconstruction results

Figure S5. Illustration of interpolated virtual sources and reconstructions of higher scale signal recovery. The exposure time of each virtual source is 0.0069 s. (a) The blue circle points represent the virtual sources where the low resolution signal is measured; the yellow square points represent the virtual sources where the signal is recovered by the proposed pipeline. In the second step, three additional signals are interpolated with adjacent row (R), column (C), and diagonal (D). (b) Reconstruction results of different signals. All reconstructions are obtained with the F-K method. The number of virtual sources is shown below each subfigure.

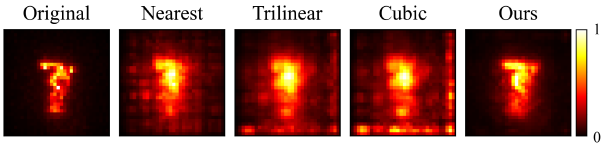


Figure S6. Comparison with other interpolation methods. The exposure time of each virtual source is 0.0412 s. All reconstructions are obtained with the F-K method.

S3.5. Comparison with end-to-end NLOS imaging method

We also compare the proposed method with existing end-to-end NLOS imaging method. Among many methods, we choose the "Learned Feature Embeddings" (LFE) method [2] to compare with the proposed method. The comparisons are shown in Fig. S7.

The first column shows the results reconstructed from the signal with spatial resolution as 32×32. In the second column, the results are obtained using the pretrained weight of LFE which is publicly available. The low resolution signal is first interpolated to 256×256 with the nearest neighbor interpolation and then processed by the LFE method.

The gap in the spatial resolution is so large that the recon-

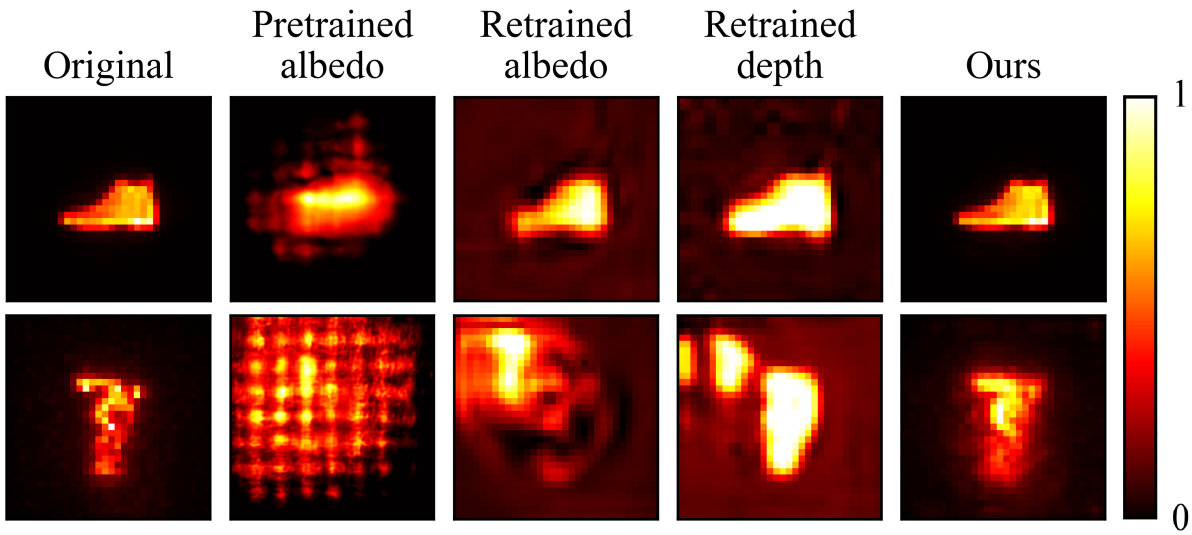


Figure S7. Comparisons with end-to-end NLOS imaging method. The first row shows the results of synthetic data, the second row shows the results of measured data (The exposure time of each virtual source is 0.0412 s).

structions are full of artifacts. Thus, we retrain the network with the same dataset used for training SSN. Through training, the low resolution signal is first interpolated to 32×32 and then processed by the LFE method. The parameters are the same as those provided in the supplement of [2].

Reconstruction results of the retrained network are shown in the third and fourth columns. Although the network performs better after retraining, it can only provide an approximate depth estimation of the statue but fails to reconstruct the albedo. The reconstruction results of the proposed framework are shown in the fifth column, which has much sharper boundaries. According to the comparison, the generalization ability of the proposed method is much better.

S3.6. Comparison with decreasing the exposure duration

There are two direct approaches to reduce the total exposure time in NLOS imaging under the confocal setting: decreasing the exposure duration of each virtual source or decreasing the number of virtual sources. In this subsection, we compare the reconstruction results of these two approaches while keeping the total exposure time the same. The comparisons are shown in Fig. S8. All the results are reconstructed by the F-K method.

We first compare these two methods with synthetic data. To simulate the measurement, we add Poisson noise to the signal. The letter T is placed 1 m away from the visible wall, and the illumination region on the visible wall is 2×2 m^2. A refrence signal with 32×32 virtual sources is generated. For the first method, pulse number of each virtual source is reduced to $\frac{1}{16}$ of the refrence's. For the second method, the signal is subsampled from the reference and processed by the proposed pipeline. As shown in the first row, the proposed framework provides a sharper reconstruction.

These two methods are then tested on measured data. From the Stanford dataset, we choose the same instance with different exposure time to compare these two methods. For the instance of the statue, the shortest exposure time of each virtual source is 0.0023 s, which is $\frac{1}{18}$ of the longest one, 0.0412 s. Thus, to keep the total exposure time close enough, we subsample a signal with 34×34 virtual sources from the shortest one, and subsample a signal with 8×8 virtual sources from the longest one. The total exposure time of these two signals are 2.6588 s and 2.6368 s respectively. As shown in the second row, the contrast of the reconstruction obtained with the proposed pipeline is much higher.

S4. Ablation study

In this section, an ablation study is performed to show the effectiveness of the proposed framework. Since SSN is generalized from A^{2}N [1], we only compare A^{2}N (baseline) with SSN to show the superiority of the employed 3-D kernel. We refer interested readers to A^{2}N for more details.

To make a fair comparison, we train the baseline model with the same training set as SSN, except the inputs are 2-D images. After training, these two networks are tested on the statue instance from the Stanford dataset. The recovered signals are shown in Fig. S9a. Although the baseline model can recover a signal with high SNR, the first arrival time of most virtual sources and signal values are incorrect, which

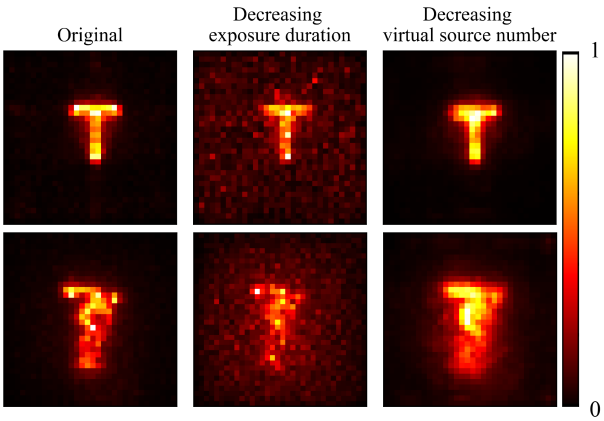
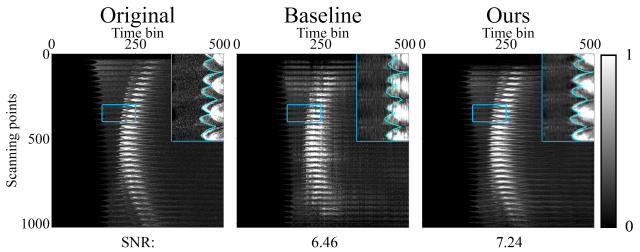
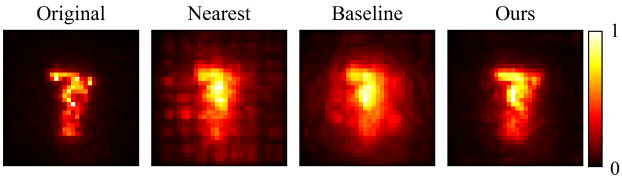


Figure S8. Comparisons with decreasing the exposure duration. The first row shows the reconstruction results of synthetic data, the second row shows the results of measured data from the Stanford dataset. The first column shows the reconstructions obtained from the original measurement. The second column shows the reconstructions obtained from the signal which has a shorter exposure time at each virtual source. The signals used in the third column are measured at 8×8 virtual sources. The reconstructions are obtained with the proposed pipeline. The total exposure times of the second and third columns are the same or close enough.

leads to poor reconstructions. As shown in Fig. S9b, the reconstructed object of the baseline model is much blurrier than the proposed framework.



(a) Recovered signals



(b) Reconstruction results

Figure S9. Recovered signals and reconstruction results for baseline and SSN. The exposure time of each virtual source is 0.0412 s. All results are reconstructed by the F-K method.

References

[1] Haoyu Chen, Jinjin Gu, and Zhi Zhang. Attention in attention network for image super-resolution. *arXiv preprint arXiv:2104.09497*, 2021. 5

[2] 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 (TOG)*, 39(6):1–18, 2020. 4, 5

[3] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 1

[4] Martin Laurenzis and Andreas Velten. Feature selection and back-projection algorithms for nonline-of-sight laser–gated viewing. *Journal of Electronic Imaging*, 23:063003, 11 2014. 1

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

[6] 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*, 08 2021. 1

[7] Matthew O'Toole, David Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555, 2018. 1

[8] 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. 1