Compressive Non-Line-of-Sight Imaging with Deep Learning

Shenyu Zhu,[1,2] Yong Meng Sua,[1,2,*] Ting Bu,[1,2] and Yu-Ping Huang[1,2,†]

[1] *Department of Physics, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, New Jersey 07030, USA*

[2] *Center for Quantum Science and Engineering, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, New Jersey 07030, USA*

(Received 19 July 2022; revised 19 January 2023; accepted 23 February 2023; published 28 March 2023)

In non-line-of-sight (NLOS) imaging, the spatial information of hidden targets is reconstructed from the time-of-light (TOF) of the multiple bounced signal photons. The need for NLOS imagers to perform extensive scanning in the transverse spatial dimensions constrains the imaging speed and reconstruction quality while limiting their applications on static scenes. Utilizing a photon TOF histogram with picosecond temporal resolution, we develop compressive non-line-of-sight imaging enabled by deep learning. Two-dimensional images (32 × 32 pixels) of the NLOS targets can be reconstructed with superior reconstruction quality via a convolutional neural network (CNN), using significantly downscaled data (8 × 8 scanning points) at a downsampling ratio of 6.25% compared to the traditional methods. The CNN is end-to-end trained purely using simulated data but robust for image reconstruction with experiment data. Our results suggest that deep learning is effective for reducing the scanning points and total capture time towards scanningless NLOS imaging and videography.

DOI: 10.1103/PhysRevApplied.19.034090

I. INTRODUCTION

Imaging and sensing a hidden target outside of the direct line-of-sight has been an emerging research topic in various fields, such as autonomous driving [1], remote sensing [2], and biomedical imaging [3]. To image a hidden object around the corner, the geometric information of the observable "wall" as the diffuser is first measured. The time-of-flight (TOF) information of the returning signal photons is measured at a series of scanning points on the wall and then processed regarding the geometric information [4–7]. The temporal profile and statistics of the returning TOF histograms of the signal photons are related to the travel path length in the scene of detection, from which the computational imaging methods can image and sense the target in various environments. Thus, the three-dimensional position information of the hidden target can be retrieved. At each scanning point, the scattered probe light reaches a certain area on the target (defined by the scattering angle of the wall); thus, the measured temporal histogram of the returning photons contains the spatial information of that area of the target [5]. This scanning area coverage opens up the possibility for reconstructing a higher pixel number NLOS image of the hidden target using much fewer scanning points, which can be achieved using compressed sensing [8]. Providing adequate temporal resolution, the reconstruction using fewer scanning points can also be done using deep-learning-based methods, which has recently played a role in extracting effective information and reconstructing the target image regarding various fields [9], such as TOF imaging [10,11], compressed sensing [12], polarimetric imaging [13], and optoacoustic tomography [14].

Various algorithms have been used to reconstruct the NLOS image. One branch of methods first models the imaging scenario and then computes the three-dimensional position of the hidden target based on the model. Back-projection-based methods estimate the most probable position and shape for the target [4,15]. Deconvolution-based methods compute the least-error estimation of the target position according to the Poisson statistics and light propagation model [16,17]. Fast Fourier transform increases the processing speed for solving the deconvolution problem [5,18–20]. These methods can reconstruct a three-dimensional point cloud of the hidden target. However, these methods usually require the same dimension for the input data and the result, which increases the data acquisition time if higher pixel numbers are needed. Another branch of methods utilizes deep learning to train a mathematical model projection from the three dimensional TOF information to the desired feature of the hidden target [21]. The feature can be the three-dimensional positions of the target, the two-dimensional intensity image, or the depth map of the hidden target. Recent studies have

[*] ysua@stevens.edu

[†] yuping.huang@stevens.edu

demonstrated that the depth map and image of the hidden scene can be reconstructed using various kinds of deep learning methods, such as the U-Net [22], the nonlocal neural network [23], the neural transient field [24], and so on [25]. These methods provide an alternative method for image reconstruction and may extract the desired information from the input data more effectively [11,26]. In both cases, higher temporal resolution is required for higher reconstruction quality.

Conventional single-photon detection systems are limited by the timing jitter and the "pile-up" effect. The timing resolution is several tens of picoseconds for the state-of-art single-photon avalanche diode (SPAD) imaging systems [17,27]. Recently, up-conversion single-photon detection has achieved picosecond-resolution NLOS imaging and sensing by performing nonlinear optical gating for the picosecond pulses of the returning NLOS signal photons [28,29]. Here, we utilize a single-pixel nonlinear gated single-photon imager for NLOS data acquisition [30,31], which has a 10-ps timing resolution and is independent from the timing jitter of the associated electrical device. It also isolates the three-time bounced signal photons from the environmental noise, especially the one-time scattered photons from the "wall," which is usually several magnitudes stronger. As a result, the imager gets rid of the "pile-up" effect. Thus the higher temporal resolution provides more precise temporal information of the hidden target, and is able to yield a more detailed spatial information of the target.

Utilizing picosecond temporal resolution photon TOF histogram, we demonstrate a NLOS image reconstruction method using a convolutional neural network (CNN), which projects the three-dimensional input data (photon counting histograms on different pixels) to the two-dimensional NLOS image. The CNN can reconstruct the image of the hidden target using spatially downscaled data [8,12] at a downsampling ratio of 6.25 %. Thanks to the results that the simulated photon counting temporal histograms are very similar to the experiment ones, the CNN can be trained by only using the simulated data with end-to-end training, yet it is very robust in image reconstruction with experiment data. This shows that the high timing resolution shrinks the difference between the simulation and experiment data, enabling CNN training without real-world data [22]. Additionally, the simulated data are generated using a set of simple geometric shapes [32], which are different from the shapes of experimental targets. This shows that fed by the high temporal resolution input data, the reconstruction model from three-dimensional data to two-dimensional NLOS image can be built using the CNN [23,33]. Using the CNN, the input data are downscaled in the temporal domain and upscaled in the spatial domain for reconstructing the NLOS image, showing the capability of retrieving the spatial information from the temporal data [26].

II. IMAGING SYSTEM SETUP

The NLOS imaging system utilizes a nonlinear gated single-photon detection (NGSPD) system [29] to capture the three-bounced signal photons. The NLOS imaging system is shown in Fig. 1. The imaging system uses a 50-MHz femtosecond mode-locked laser as the light source. The pump and probe laser pulses are generated by filtering the mode-locked laser using a pair of 200-GHz dense wavelength-division multiplexers (DWDMs). The center wavelength of the pump pulse is 1565.5 nm, and that of the probe is 1554.1 nm. The pulse widths of the pump and the probe are 6.8 and 6.3 ps, respectively, shown in the upper left and middle in Fig. 1. The pump laser passes through an optical delay line (ODL). The probe laser pulse is sent out from an optical transceiver at a beam size of 2.2 mm FWHM, from which the probe beam is first steered by a MEMS mirror, and then illuminates different scanning points on the diffuser to probe the hidden target. The transceiver receives the returning signal photons on the same scanning point for illumination, which forms a confocal configuration [5]. The optical transceiver is made of a fiber collimator, which is composed of an angle-polished single mode fiber and an aspheric lens ($f = 11.0$ mm, NA $= 0.25$). The returning signal pulses are first isolated from the outgoing probe using an optical circulator (55 dB isolation ratio), then mixed with an optical pump pulse train using another DWDM. Then the pump and signal are fiber coupled into a commercial quasi-phase-matching nonlinear optical waveguide, where the pump up-converts the returning signal into sum-frequency photons.

The nonlinear optical waveguide has the center quasi-phase-matching wavelength at 1559.8 nm, and the internal conversion efficiency at 137 %/(W cm^2). When the delay time of the pump is scanned by the ODL, the pump pulses temporally sweep across the signal photons. At each optical delay, the photons at the band of the designed sum-frequency wavelength are detected using a silicon-based SPAD (approximately 70 % efficiency at 780 nm), then counted by a field-programmable gate array (FPGA). At each scanning point, the photon count versus the delay time of the pump forms a temporal histogram. The up-conversion process can happen effectively only when the signal is at certain temporal-frequency mode [30]. The impulse response of the system is defined by the FWHM of the sum-frequency temporal histogram, as shown in the upper right inset figure of Fig. 1. This FWHM is 10 ps, which defines the temporal resolution [31,34] of the system. The background count rate is about 5.5 kHz, including the intrinsic dark count rate of the SPAD (200 Hz) and the Raman noise (5.3 kHz) in the nonlinear optical waveguide.

The detected scene is composed of a 2-inch diameter metallic diffuser (120 grit, reflectivity > 96 %) as the "wall," and the hidden target, which is 12 cm away from the diffuser. The hidden targets, made of retroreflective tape cut into various shapes, are attached on an ordinary BK-7 glass plate. The transmittance of the glass plates is about 92 %, so that most of the light can transmit through the glass plate.

III. NLOS DATA SIMULATION AND EXPERIMENT ACQUISITION

The training dataset for the CNN is generated from the confocal light-cone model [5] using simulated simple geometric shapes as the targets. First, eight kinds of simple geometric shapes (circle, arc, square, triangle, semicircle, rectangle, ring, and L shape) are randomly generated in different geometric parameters (i.e., position, size, and rotation angle, etc.) [32], which are later used as the training labels (output) for the CNN. These shapes are used since they contain the basic geometric elements of more complex targets, such as letters, while not containing the exact experiment target themselves. The total number of the shapes are 20 000, and the size of each shape is 32 × 32 pixels. The simulated data, i.e., the temporal histograms of each shape are generated from the confocal model as

$$c(u,v,t) = h(t) * \left(\frac{1}{r^b}\iiint_{x,y,z} s(\alpha)\delta\left((u-x)^2+(v-y)^2 + z^2 - \left(\frac{ct}{2}\right)^2\right) o(x,y,z)\right), \tag{1}$$

where x,y,z are the three-dimensional coordinates in the object space, u,v are the scanning points on the diffuser space, t is the measured photon arrival time. In such a way, we define $o(x,y,z)$ as the reflectivity of the target in the object space, and $c(u,v,t)$ is the retrieved photon counting temporal histograms at different scanning points (u,v). The function $\delta((u-x)^2+(v-y)^2+z^2-(ct/2)^2)$ describes the TOF (t) for the probe beam to travel from point (u,v) on the diffuser to the object (x,y,z). We assume the object surface is even, and the reflectivity is normalized to 1. $s(\alpha)$ is the scattering angle profile of the diffuser (about 60° FWHM), where α is the angle between the incident laser and scattering direction toward point (x,y,z) at the scanning point (u,v). The optical power falloff caused by the scattering is described by the term $1/r^b$, where $r=\sqrt{(u-x)^2+(v-y)^2+z^2}$, and b is the falloff factor. The falloff factor is evaluated by comparing the temporal

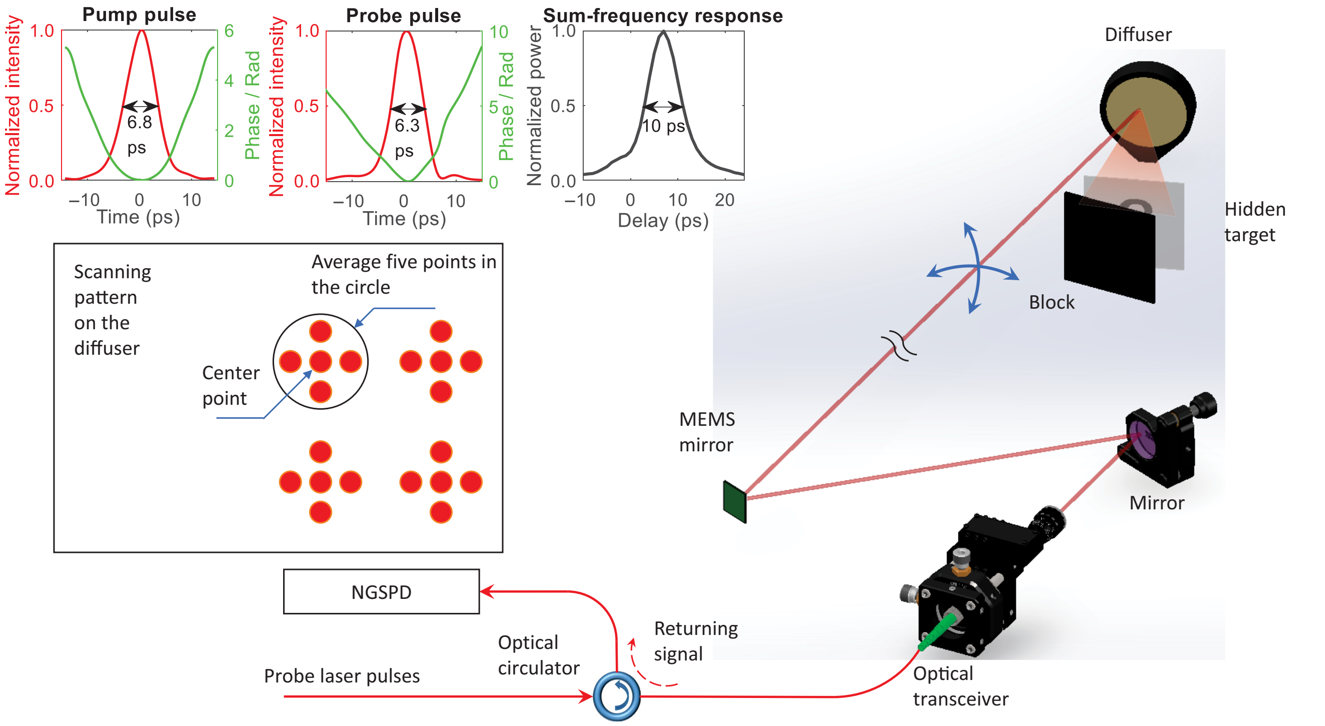

FIG. 1. Setup of the NLOS imaging system. MEMS, microelectromechanical system; NGSPD, nonlinear optical gated single photon detector. The upper three figures show the pulse profile of the pump and signal as well as the sum-frequency impulse response of the system. The pump and the probe pulse profiles are measured using frequency-resolved optical gating (FROG), while the sum-frequency power is measured using a power meter at different optical delays. The middle left subplot is the scanning pattern on the diffuser for NLOS imaging.

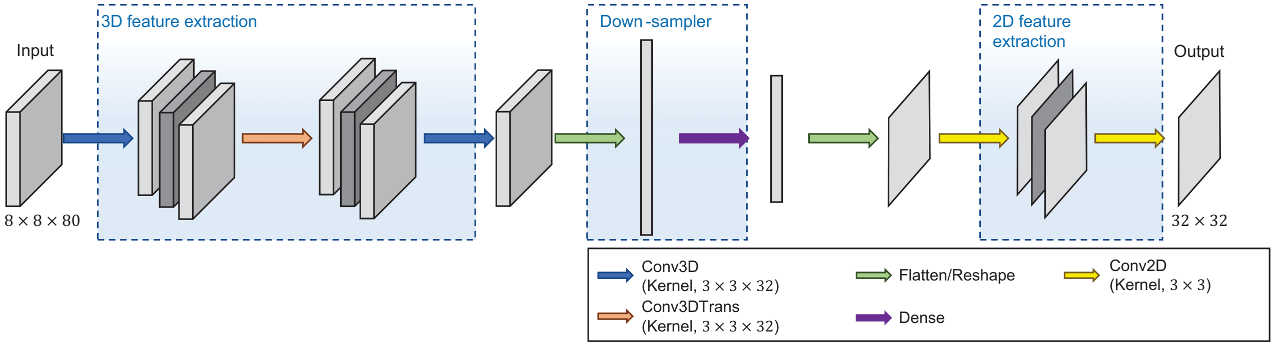

FIG. 2. Structure of the convolutional neural network. The size of the input and the output layer is labeled at the bottom of the block. The network structure between layers is labeled with arrows of different colors, labeled in the right-bottom part.

histograms of the experiment results and simulated results, and we set b to be approximately 2.3 for the retroreflective target. $h(t)$ is the impulse response of the system. The simulated dataset is then computed using Eq. (1). We simulate 8 × 8 scanning points for each shape as the input of the training set, while the time bin width is set to 1 ps. Using the NGSPD, the returning photons from the diffuser and the background are gated out from those from the target. Thus, the temporal histogram has no effective signals at arrival time earlier than the TOF position of the target. Hence, we discard most of the time bins beyond the position of the target, leaving only 80 time bins that contain the signal from the target as the input of CNN. The temporal histograms are normalized before being put into the CNN.

The CNN is built using Keras [35]. We first use three-dimensional convolutional layers for feature extraction, and use one dense layer to downsample the features and then reshape them into two-dimensional images, and finally use a series of two-dimensional convolutional layers to optimize the result. The structure of the neural network is shown in Fig. 2. All the layers use ReLU function as activation function, except for the dense layer, which uses sigmoid function. All the simulated data are used as the training dataset. During the training process, we first randomize the sequence of the dataset and divide the dataset into five parts, with each part having 4000 shapes. Then, we use the k-fold cross validation to address the possible overfitting problem caused by the randomness of training and validation datasets. Finally the CNN will be fit for all the training data, and mean-squared error function is used as the optimization function. Consequently, we build a three-dimensional to two-dimensional NLOS image reconstruction model using the CNN, with the input size of 8 × 8 scanning positions ×80 time bins, and output size of 32 × 32 pixels. We also work on a more compressed 4 × 4 scanning positions case and a lower temporal resolution case, where the structure of the CNNs are the same but trained using respective simulated data sets. The training is conducted on a computer with an Intel Core I9-10900 CPU, 64 GB RAM, and an NVIDIA GeForce RTX 3080 GPU. In the training process, the learning rate is set at 1×10^{-5}, and the CNN converges in 200 epochs. Even though a simple CNN model is being used in this work, note that other deep learning models with more sophisticated architectures incorporating physical priori [36] may provide better reconstruction quality; see the Supplemental Material [37] where a simple U-Net is used as a comparison.

To capture the experiment data for each target, the MEMS mirror is steered to raster scan the probe laser beam on the diffuser. To train the CNN by using purely simulated data, the experiment data has to closely resemble the simulated data. This places stringent requirements on noise rejection and temporal resolution of the photon detection. The former will help to mitigate anomaly spike in the photon arrival-time histogram, and the latter is crucial to fully capture the geometrical information of the target. To minimize the effect of speckle noise (arising from the rough surface of diffusive wall) that typically causes discrepancy between the simulated and experiment temporal histograms (see Appendix I) [38], at each scanning position, we measure the temporal histogram of five neighboring scanning points and calculate their average as the final histogram for each pixel. The scanning pattern is shown in the middle-left inset of Fig. 1. In this way, the random photon-number spike in the temporal histogram caused by the speckle field can be mitigated, reducing the dissimilarity between experiment data to the simulated data. At each scanning point, the dwell time per delay is 2 ms, so that the total dwell time per delay for each pixel is summed up to 10 ms [29]. The temporal scanning interval is 1 ps. We scan 16 × 16 pixels for each target in order to reconstruct the image using the LCT algorithm [5] shown in the third row in Fig. 3, and then downsample the

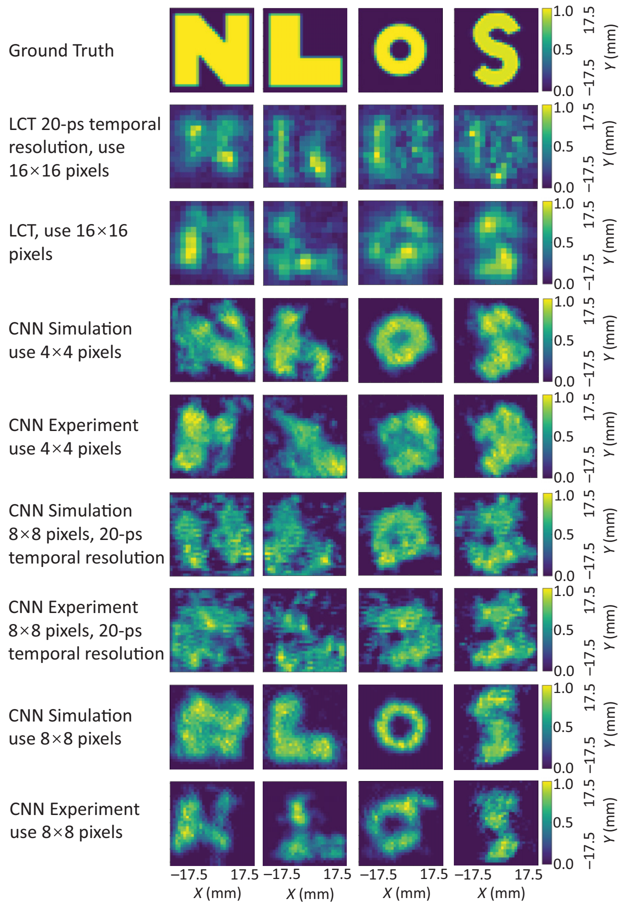

FIG. 3. NLOS imaging results for different letters of “NLOS.” NLOS imaging results for different letters of “NLOS.” Each row lists the ground truth and reconstruction results of one letter. The ground truth of the letters are listed in the first row. The next two rows list the reconstructed image using the LCT method, the second for the simulated 20-ps temporal resolution reconstructed images, and the third for the experiment 10-ps resolution ones. The other six rows are results reconstructed using the CNN. The fourth and fifth are the reconstruction results using 4 × 4 pixels using simulated data and the experiment data, respectively. The sixth and seventh rows are the reconstruction results of the 20-ps temporal resolution data using 8 × 8 pixels. The eighth and ninth rows are the reconstruction results using 8 × 8 pixels. The color bar indicates the normalized reflectivity of the surface.

experiment data to 8 × 8 pixels as the input of the neural network (the ninth row in Fig. 3).

IV. COMPRESSIVE NLOS IMAGE RECONSTRUCTION

Four shapes of English letters “NLOS” are prepared as the target in the experiment. The results are shown in Fig. 3. The CNN retrieves the 32 × 32 pixel spatial image of the target from 8 × 8 temporal histograms, showing excellent agreement with simulation result. The eighth row in Fig. 3 shows the predicted results from the simulated data, and the ninth row shows the results of the experiment data. As a comparison, although using 16 × 16 scanning points, the LCT reconstructed results is not as clear as the CNN reconstructed results, shown in the third row in Fig. 3. A three-dimensional Gaussian filter has been added to the LCT results to remove the background noises. The quality of the reconstructed results are evaluated by cross-correlating the result figures with the ground truth. The similarity evaluation criterion is defined as the maximum value of the cross-correlation result divided by the maximum of the autocorrelation of the ground truth, listed in Table I. The CNN reconstructed results (per image size = 32 × 32) are directly compared with the training labels, while the LCT reconstructed results (per image size = 16 × 16) are compared with the resized training labels. The evaluation indicates that the CNN reconstructed images have higher similarity to the ground truth than the LCT ones; see Supplemental Material [37] for the raw experimental data and the three-dimensional point clouds of the LCT results.

Several results using other reconstruction conditions are also shown in Fig. 3. A more compressed case spatially downsamples the input data further more and uses only 4 × 4 pixels as input, as shown in the fourth and fifth row of Fig. 3. The other decreases the temporal resolution to about 20 ps in both the simulation and experiment. For this case, we simulate the temporal histograms with 20-ps temporal resolution as the simulated data. For the experiment data, we convolute the experiment temporal histograms on each pixel by a Gaussian function with 17-ps FWHM, so that the result impulse response would be 20 ps. The time-bin width of the histograms remains 1 ps. For each shape, the CNN predicted experiment results using 8 × 8 pixels have the highest similarity of the ground truth, as shown in the seventh and the eighth row of Table I. The 4 × 4 pixel input case also provides less similar reconstruction results. The reconstructed shapes of the letters are distorted but still have a coarse profile indicating the letter, which is evaluated in the third and fourth row of Table I. The reduced temporal resolution results, although using 8 × 8 scanning points as the input, have even lower reconstruction quality comparing with the results using 4 × 4 scanning points. The lower temporal resolution experiment results using the LCT method with 16 × 16 pixels are shown in the second row of Fig. 3, and the CNN reconstructed results are shown in the sixth and seventh row of Fig. 3. The result image becomes coarser, and the letters cannot be distinguished, indicating the loss of detailed spatial information, as the evaluation results in the second, fifth, and sixth row in Table I are among the lowest. This indicates the relevance of adequate temporal information for NLOS image reconstruction. Regarding the spatial resolution of the system $\Delta x = (c\sqrt{(x/2)^2 + z^2})/(x)\Delta t \approx$ 1.1 cm, the reconstruction results show a clear shape beyond the spatial resolution [32]. The reconstructed

TABLE I. Evaluation of the reconstruction results using cross-correlation.

Target	N shape	L shape	O shape	S shape
LCT reconstruction using 16 × 16 scanning points (experiment)	0.588	0.526	0.707	0.702
LCT reconstruction of 20-ps resolution using 16 × 16 scanning points (experiment)	0.450	0.456	0.520	0.540
CNN reconstruction using 4 × 4 scanning points (simulation)	0.502	0.549	0.808	0.689
CNN reconstruction using 4 × 4 scanning points (experiment)	0.458	0.490	0.713	0.692
CNN reconstruction of 20-ps resolution using 8 × 8 scanning points (simulation)	0.453	0.488	0.740	0.587
CNN reconstruction of 20-ps resolution using 8 × 8 scanning points (experiment)	0.418	0.459	0.642	0.600
CNN reconstruction using 8 × 8 scanning points (simulation)	0.644	0.642	0.850	0.786
CNN reconstruction using 8 × 8 scanning points (experiment)	0.596	0.507	0.898	0.798

letters do not exist in the training dataset, while the CNN still provides results that match with the ground truths, which indicates that the trained CNN builds a model projecting from the temporal data to the spatial image.

The capability of capturing the NLOS signal in high temporal resolution guarantees the CNN reconstruction results. First, the differences in the temporal histograms of different targets lead to different reconstructed image. Several samples of the normalized temporal histograms of both simulation and experiment are shown in Fig. 4. The temporal histograms in the center scanning point do not have much difference between different targets as shown in the third row of Fig. 4. The reason is that the TOF of the back-scattered photons from the edge of the target to the center scanning point do not differ much, thus the "tail" of the histograms is not significant. At corner scanning points, however, the temporal histograms differ from each other because of the target shape difference. On these scanning points, the TOF difference from different locations of the target contributes to a longer tail in the histogram, which provides the information for reconstructing the shape. Second, considering that the CNN is trained using simulated data, it is required that the experiment result histograms match those of the simulation. In this case, high temporal resolution is required for catching the slight difference between the histograms of the targets. As a comparison, reconstruction results of lower temporal resolution have even lower quality than the spatially downsampled case using 4 × 4 pixels. Additionally, the fluctuation in the temporal histogram, caused by the scattering randomness of the diffuser and the target, is mitigated by averaging the histograms from nearby

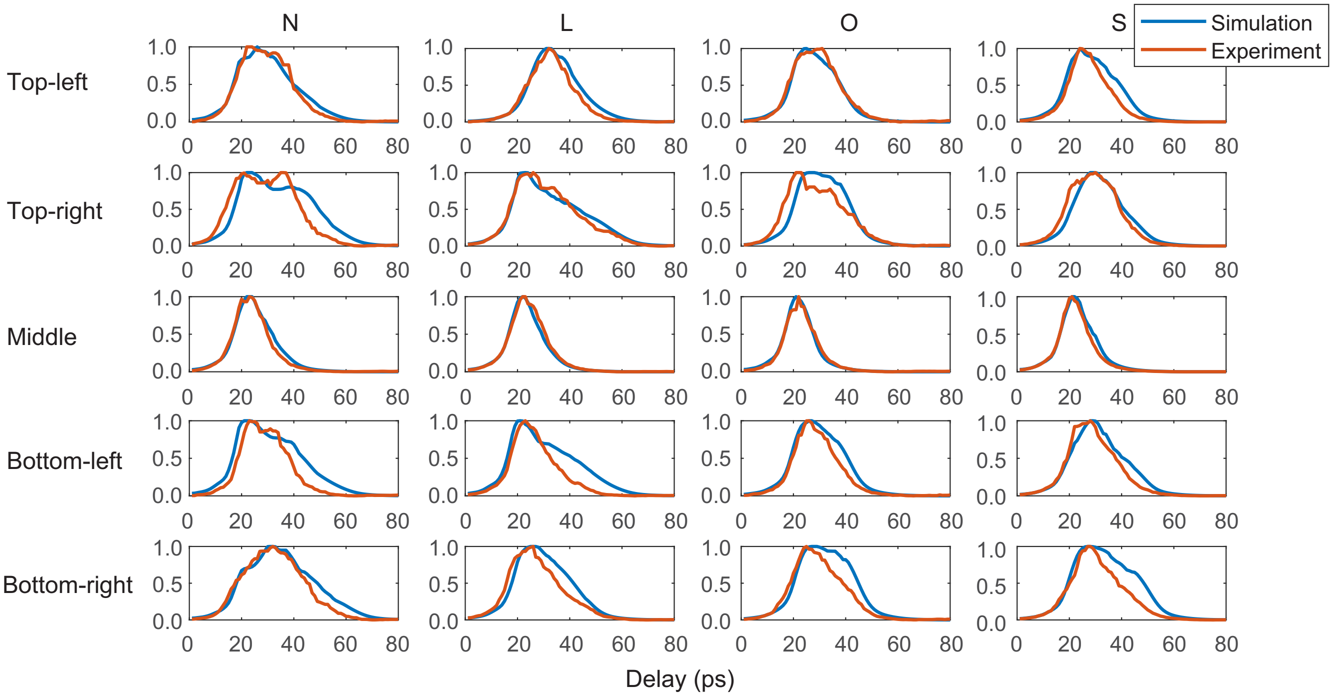

FIG. 4. Comparison of the temporal histograms between the simulated data and the experiment captured data. Each column in the figure indicates one target letter, while each row lists the temporal histogram from the same scanning point. The picked scanning points are at the corner of the scanning area, as labeled. The histograms are normalized on each scanning point.

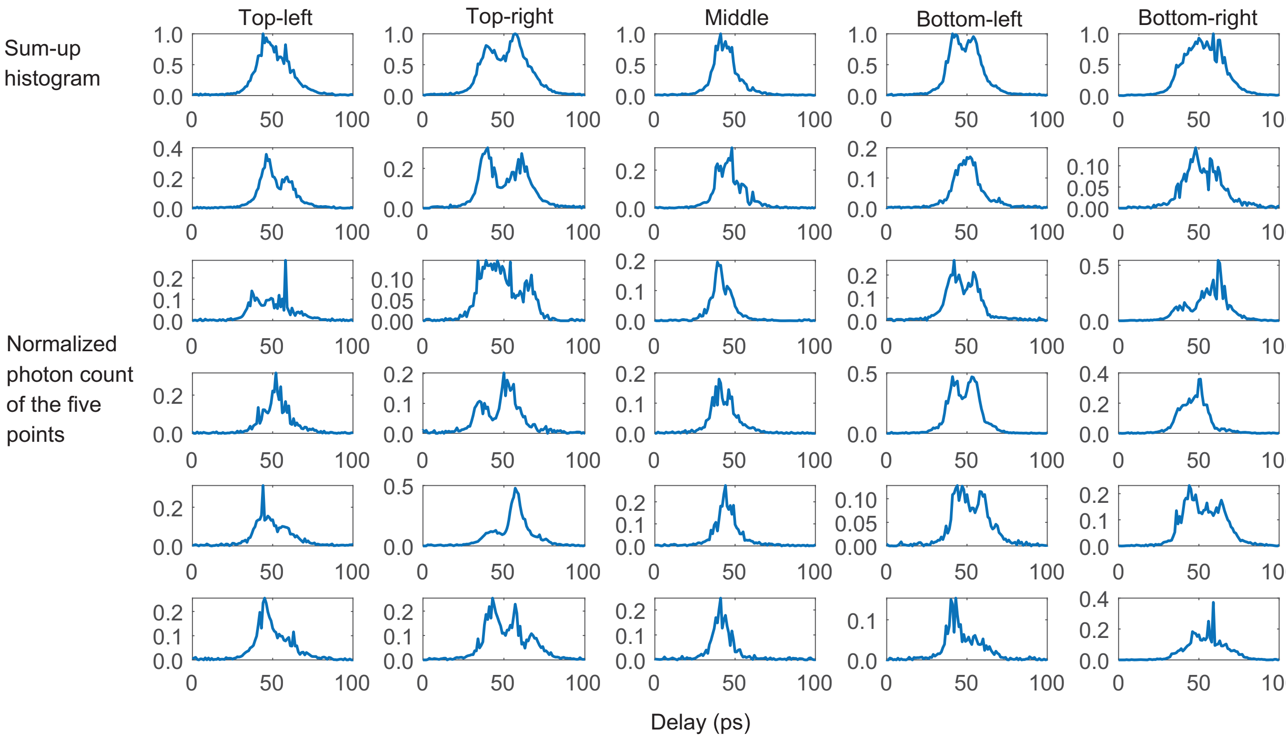

FIG. 5. The temporal histograms of the target “N” letter at different points. The first row are the summed-up histograms on the top-left, top-right, middle, bottom-left, and bottom-right positions. The second to the sixth row list the temporal histograms at the five scanning points of the combined temporal histogram. Each column indicates the histogram at one combined temporal histogram on the diffuser. Each row of temporal histograms are normalized to the maximum value of the sum-up histogram to better show the amplitude and shape difference.

scanning points. The remaining discrepancy between the simulation and experiment data causes the mismatch between the reconstructed images. From the experiment results, we show that the nonlinear gated single-photon detection system is able to capture the NLOS signal in high temporal resolution such that the temporal histogram from the experiment can match those from the simulation.

V. DISCUSSION

Image reconstruction of NLOS hidden targets from the temporal signal has been a long-pursued task, where higher temporal resolution gives better reconstruction results. We demonstrate that high temporal resolution enables compressive NLOS image reconstruction through deep learning. Using a simulated-data-trained CNN, a 32×32 pixel image can be reconstructed using 8×8 temporal histograms. The CNN is trained using the simulated data generated from the LCT model. Given that the experiment histograms match the simulation histograms, the CNN can interpret the spatially downsampled experiment captured histograms into the NLOS image. Our results show a potential possibility of enhancing the NLOS data-acquisition speed, which can be helpful for NLOS videography [6,27]. One potential advantage of this deep-learning method is that the CNN, once trained, does not need the iteration-based optimization calculation for compressive sensing [8], therefore, requiring less computational resources and reconstruction speed. Moreover, the spatial resolution of the system can be enhanced by further improving temporal resolution using narrower optical pulses as the optical gating [28]. In conclusion, we demonstrate that a NLOS imaging modality from three-dimensional data to two-dimensional images is built by combining the nonlinear gating single-photon imaging system and deep-learning methods. Such a method may extend some imaging and sensing applications such as pose estimation [39] and item recognition [40] into the NLOS scenario.

ACKNOWLEDGMENTS

This material is based upon work supported by the ACC-New Jersey under Contract No. W15QKN-18-D-0040. We thank Yayuan Li for his support and suggestion.

APPENDIX: MINIMIZING DISCREPANCY BETWEEN SIMULATION AND EXPERIMENT DATA

In the data-acquisition process of the NLOS imaging, we scan the probe beam on five scanning points at each pixel. This is to mitigate the photon-counting randomness caused by the scattering effect from the diffuser and the target. An example of the temporal histogram differences between the scanning points are shown in Fig. 5. The second to the sixth rows show that even at very near scanning points, the temporal histograms differ in both the amplitude and the shape, and are also quite different from the simulation results. This difference is mainly brought in by the scattering randomness of the diffuser and the target. We sum up these five histograms and normalize the result histogram as shown in the first row in Fig. 5. This average partially compensates the random fluctuation from the scattering, and makes the histogram look more similar to the simulated case.

[1] C. Rablau, in *Fifteenth Conference on Education and Training in Optics and Photonics: ETOP 2019*, Vol. 11143, International Society for Optics and Photonics (SPIE, Quebec City, Quebec, Canada, 2019), p. 84.

[2] K. Zhang, B. Li, X. Zhu, H. Chen, and G. Sun, NLOS signal detection based on single orthogonal dual-polarized GNSS antenna, Int. J. Antennas Propag. **2017**, 8548427 (2017).

[3] P. Bruza, A. Petusseau, A. Ulku, J. Gunn, S. Streeter, K. Samkoe, C. Bruschini, E. Charbon, and B. Pogue, Single-photon avalanche diode imaging sensor for subsurface fluorescence LIDAR, Optica **8**, 1126 (2021).

[4] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging, Nat. Commun. **3**, 745 (2012).

[5] M. O'Toole, D. B. Lindell, and G. Wetzstein, Confocal non-line-of-sight imaging based on the light-cone transform, Nature **555**, 338 (2018).

[6] X. Feng and L. Gao, Ultrafast light field tomography for snapshot transient and non-line-of-sight imaging, Nat. Commun. **12**, 2179 (2021).

[7] G. Musarra, A. Lyons, E. Conca, Y. Altmann, F. Villa, F. Zappa, M. Padgett, and D. Faccio, Non-Line-of-Sight Three-Dimensional Imaging with a Single-Pixel Camera, Phys. Rev. Appl. **12**, 011002 (2019).

[8] J.-T. Ye, X. Huang, Z.-P. Li, and F. Xu, Compressed sensing for active non-line-of-sight imaging, Opt. Express **29**, 1749 (2021).

[9] G. Barbastathis, A. Ozcan, and G. Situ, On the use of deep learning for computational imaging, Optica **6**, 921 (2019).

[10] J. Peng, Z. Xiong, X. Huang, Z.-P. Li, D. Liu, and F. Xu, in *European Conference on Computer Vision* (Springer, Glasgow, UK, 2020), p. 225.

[11] A. Turpin, G. Musarra, V. Kapitany, F. Tonolini, A. Lyons, I. Starshynov, F. Villa, E. Conca, F. Fioranelli, R. Murray-Smith, and D. Faccio, Spatial images from temporal data, Optica **7**, 900 (2020).

[12] F. Wang, C. Wang, C. Deng, S. Han, and G. Situ, Single-pixel imaging using physics enhanced deep learning, Photon. Res. **10**, 104 (2022).

[13] L. Si, T. Huang, X. Wang, Y. Yao, Y. Dong, R. Liao, and H. Ma, Deep learning Mueller matrix feature retrieval from a snapshot Stokes image, Opt. Express **30**, 8676 (2022).

[14] J. Li, C. Wang, T. Chen, T. Lu, S. Li, B. Sun, F. Gao, and V. Ntziachristos, Deep learning-based quantitative optoacoustic tomography of deep tissues in the absence of labeled experimental data, Optica **9**, 32 (2022).

[15] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten, Non-line-of-sight imaging using a time-gated single photon avalanche diode, Opt. Express **23**, 20997 (2015).

[16] F. Xu, G. Shulkind, C. Thrampoulidis, J. H. Shapiro, A. Torralba, F. N. Wong, and G. W. Wornell, Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging, Opt. Express **26**, 9945 (2018).

[17] C. Wu, J. Liu, X. Huang, Z.-P. Li, C. Yu, J.-T. Ye, J. Zhang, Q. Zhang, X. Dou, V. K. Goyal, F. Xu, and J.-W. Pan, Non–line-of-sight imaging over 1.43 km, Proc. Natl. Acad. Sci. **118**, 10 (2021).

[18] D. B. Lindell, G. Wetzstein, and M. O'Toole, Wave-based non-line-of-sight imaging using fast f-k migration, ACM Trans. Graph. **38**, 116 (2019).

[19] X. Liu, I. Guillén, M. La Manna, J. H. Nam, S. A. Reza, T. H. Le, A. Jarabo, D. Gutierrez, and A. Velten, Non-line-of-sight imaging using phasor-field virtual wave optics, Nature **572**, 620 (2019).

[20] X. Liu, S. Bauer, and A. Velten, Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems, Nat. Commun. **11**, 1 (2020).

[21] R. Geng, Y. Hu, and Y. Chen, Recent advances on non-line-of-sight imaging: Conventional physical models, deep learning, and new scenes, arXiv preprint arXiv:2104.13807 (2021).

[22] J. G. Chopite, M. B. Hullin, M. Wand, and J. Iseringhausen, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, USA, 2020), p. 957.

[23] J. Peng, F. Mu, J. H. Nam, S. Raghavan, Y. Li, A. Velten, and Z. Xiong, Towards non-line-of-sight photography, arXiv preprint arXiv:2109.07783 (2021).

[24] S. Shen, Z. Wang, P. Liu, Z. Pan, R. Li, T. Gao, S. Li, and J. Yu, Non-line-of-sight imaging via neural transient fields, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2257 (2021).

[25] W. Chen, F. Wei, K. N. Kutulakos, S. Rusinkiewicz, and F. Heide, Learned feature embeddings for non-line-of-sight imaging and recognition, ACM Trans. Graphics (Proc. SIGGRAPH Asia) **39**, 230 (2020).

[26] A. Turpin, V. Kapitany, J. Radford, D. Rovelli, K. Mitchell, A. Lyons, I. Starshynov, and D. Faccio, 3D Imaging from Multipath Temporal Echoes, Phys. Rev. Lett. **126**, 174301 (2021).

[27] J. H. Nam, E. Brandt, S. Bauer, X. Liu, M. Renna, A. Tosi, E. Sifakis, and A. Velten, Low-latency time-of-flight non-line-of-sight imaging at 5 frames per second, Nat. Commun. **12**, 6526 (2021).

[28] B. Wang, M.-Y. Zheng, J.-J. Han, X. Huang, X.-P. Xie, F. Xu, Q. Zhang, and J.-W. Pan, Non-Line-of-Sight Imaging with Picosecond Temporal Resolution, Phys. Rev. Lett. **127**, 053602 (2021).

[29] S. Zhu, Y. M. Sua, P. Rehain, and Y.-P. Huang, Single photon imaging and sensing of highly obscured objects around the corner, Opt. Express **29**, 40865 (2021).

[30] A. Shahverdi, Y. M. Sua, I. Dickson, M. Garikapati, and Y.-P. Huang, Mode selective up-conversion detection for LIDAR applications, Opt. Express **26**, 15914 (2018).

[31] P. Rehain, Y. M. Sua, S. Zhu, I. Dickson, B. Muthuswamy, J. Ramanathan, A. Shahverdi, and Y.-P. Huang, Noise-tolerant single photon sensitive three-dimensional imager, Nat. Commun. **11**, 921 (2020).

[32] A. A. Pushkina, G. Maltese, J. I. Costa-Filho, P. Patel, and A. I. Lvovsky, Superresolution Linear Optical Imaging in the Far Field, Phys. Rev. Lett. **127**, 253602 (2021).

[33] C. Pei, A. Zhang, Y. Deng, F. Xu, J. Wu, D. U.-L. Li, H. Qiao, L. Fang, and Q. Dai, Dynamic non-line-of-sight imaging system based on the optimization of point spread functions, Opt. Express **29**, 32349 (2021).

[34] S. Maruca, P. Rehain, Y. M. Sua, S. Zhu, and Y. Huang, Non-invasive single photon imaging through strongly scattering media, Opt. Express **29**, 9981 (2021).

[35] F. Chollet *et al.*, Keras, https://keras.io (2015).

[36] F. Mu, S. Mo, J. Peng, X. Liu, J. H. Nam, S. Raghavan, A. Velten, and Y. Li, Physics to the rescue: Deep non-line-of-sight reconstruction for high-speed imaging, arXiv preprint arXiv:2205.01679 (2022).

[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.19.034090 for (i) the simulated and the experimental results using a U-Net; and (ii) additional figures for the raw data and the three-dimensional light-cone transformation reconstructed point clouds.

[38] I. Starshynov, O. Ghafur, J. Fitches, and D. Faccio, Coherent Control of Light for Non-Line-of-Sight Imaging, Phys. Rev. Appl. **12**, 064045 (2019).

[39] P. Kirkland, V. Kapitany, A. Lyons, J. Soraghan, A. Turpin, D. Faccio, and G. D. Caterina, in *Emerging Imaging and Sensing Technologies for Security and Defence V; and Advanced Manufacturing Technologies for Micro- and Nanosystems in Security and Defence III*, Vol. 11540, International Society for Optics and Photonics (SPIE, 2020), p. 66.

[40] G. Mora-Martín, A. Turpin, A. Ruget, A. Halimi, R. Henderson, J. Leach, and I. Gyongy, High-speed object detection with a single-photon time-of-flight image sensor, Opt. Express **29**, 33184 (2021).