Fast Non-line-of-sight Imaging with Two-step Deep Remapping

Dayu Zhu
Georgia Institute of Technology
dzhu71@gatech.edu

Wenshan Cai
Georgia Institute of Technology
wcai@gatech.edu

Abstract

Conventional imaging only records photons directly sent from the object to the detector, while non-line-of-sight (NLOS) imaging takes the indirect light into account. Most NLOS solutions employ a transient scanning process, followed by a physical based algorithm to reconstruct the NLOS scenes. However, the transient detection requires sophisticated apparatus, with long scanning time and low robustness to ambient environment, and the reconstruction algorithms are typically time-consuming and computationally expensive. Here we propose a new NLOS solution to address the above defects, with innovations on both equipment and algorithm. We apply inexpensive commercial Lidar for detection, with much higher scanning speed and better compatibility to real-world imaging. Our reconstruction framework is deep learning based, with a generative two-step remapping strategy to guarantee high reconstruction fidelity. The overall detection and reconstruction process allows for millisecond responses, with reconstruction precision of millimeter level. We have experimentally tested the proposed solution on both synthetic and real objects, and further demonstrated our method to be applicable to full-color NLOS imaging.

1. Introduction

Imaging is a seemingly mature technique that is employed to record the real-world scene, which has been developed over centuries. One central technique of imaging is to record the distribution of photons which are emitted from or scattered by the target and further received by the camera. Apart from the planar retrievals, i.e. pictures, people also strive to extract the depth or phase information of objects and reconstruct three-dimensional scenes, which is a thriving field of research in computer vision [45, 17]. Most imaging and reconstruction methods can only recover the information in the line-of-sight (LOS) field of the camera, which implies that there should be no obstacle between the direct light path connecting the object and the camera. Otherwise, the light from the object will be reflected or de-
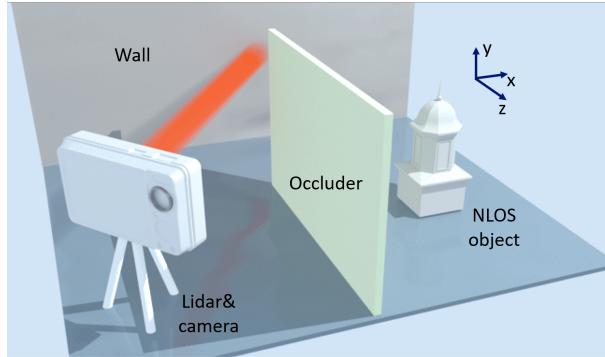

Figure 1. Illustrator of the non-line-of-sight (NLOS) imaging. The light path between the object and the camera is blocked by an occluder, thus the camera cannot directly capture the information of the object. With the aid of a wall or a secondary object which scatters the photons from the NLOS object to the camera, NLOS imaging may be achieved. Lidar is to actively probe the NLOS scene, and information of the reflected light is used to reconstruct the NLOS object.

flected, which alters the intensity and directionality of light and further inevitably impedes and encrypts the original photon distribution. As a result, the object is 'hidden' from the plain sight and becomes non-line-of-sight (NLOS) for observation, and the encrypted information is usually too weak and has to be treated as noise in imaging.

Although the NLOS information is greatly distorted, it is still possible to recover the original photon distribution and achieve NLOS imaging and reconstruction [18]. Numerous attempts have been made to analytically decode the NLOS information from the captured pictures [9, 2, 5, 4, 7, 8]. Despite effective in some cases, the LOS and NLOS data are greatly entangled on the image plane and there is always no guarantee of reasonable separation of them. Thus, instead of passively decoding the images, an alternative approach is to actively probe the NLOS scene [50, 36, 52, 20, 39, 49, 24, 1, 48, 26]. This active method provides much more information than a passive image and much higher degree of freedom for NLOS imaging and reconstruction. This genre usually applies a laser to scan the scene spa-

tially, and a photon detector is used to detect the transient events of back-scattered photons. The NLOS reconstruction is then achieved by analyzing the back-projection conditions of the transients or through other geometric optical transformations [3, 50]. The algorithms may proceed iteratively until the reconstruction converges. The active scheme can achieve high fidelity of NLOS reconstruction, meanwhile, some major shortcomings hinder the universal application of this NLOS imaging approach. On the physical level, laser scanning and transient detection processes are necessary. The laser scanning process may take minutes to hours with a lab-built scanning system, which is not applicable to real-time scenarios, and the ultrafast lasers are expensive and may not satisfy the safe standard for out-of-lab applications. As for the detection equipment, the transient events require lab-built optical systems with sophisticated apparatus, such as single-photon avalanche photodetectors (SPAD), ultrafast photon counting modules, streak cameras, etc [21, 6, 27]. The overall equipment settings are highly complicated and vulnerable to ambient perturbations, thus they are more suitable for ultrafast optical experiments than real-world imaging scenes. At the algorithm level, the transient detection will collect a huge amount of data, and the reconstruction process always deals with matrices or tensors of gigantic sizes. These properties lead to high consumption of memory, storage and computational resources, and the reconstruction time will hardly fulfill the expectation of real-time NLOS imaging and reconstruction.

To address these pressing challenges, in this paper we introduce a novel methodology for real time NLOS imaging. Instead of using sophisticated devices for transient experiment, we employ a low-end Lidar as both the probe and detection devices [51]. Compared to in-lab scanning systems with ultrafast laser and SPAD, Lidar is commercially available at a much lower cost (only a few hundred dollars), with substantially faster operation speed (scanning time at the millisecond level) and strong robustness to environmental factors. Not relying on transient time-of-flight (ToF) information of photons, we utilize the depth map and intensity map collected by the Lidar to achieve real-time, full-color NLOS imaging. Our algorithm is deep learning based and consists of two neural networks: a compressor and a generator [33]. There are some pioneering works that have introduced deep learning into NLOS imaging and revealed the promising future of deep learning assisted NLOS imaging [13, 15, 35, 37, 2]. The reconstruction speeds of deep learning approaches are inherently much faster than geometric physical-based methods, with better versatility and generalization abilities. The reported methods are mostly built on supervised learning and try to establish the direct mapping between the detected data and target NLOS scenes. However, in sharp contrast to most computer vision tasks, in NLOS reconstruction the

input data and output scenes share almost no common geometric features (such as edges or shapes), thus the supervised mapping of input and the label (i.e. the ground truth NLOS scene) may not be well-learnt by a conventional convolutional neural network (CNN) [34, 32]. Besides, generating high dimensional data with explicit semantics, such as images or 3D scenes, is never a closed-form problem, and a generative model such as generative adversarial network (GAN) or variational autoencoder (VAE) is necessary [22, 30]. Therefore, the reconstruction performance of these reported methods is limited by the direct mapping with supervised learning, such as blurriness of reconstructed results and false retrieval of complicated objects. In this work we propose a deep learning framework that consists of two components and achieves NLOS reconstruction in two steps. The generator is the decoder of a trained VAE responsible for generating various scenes, and the CNN compressor converts the detected depth and intensity information into a low-dimensional latent space for the generator. When the Lidar collects new depth and intensity maps, the compressor will compress them into a latent vector, then the generator will recover the NLOS scene by decoding the vector. This methodology overcomes the flaws of previous deep learning approaches, which were based on a single network and trained in a supervised fashion. Our framework is capable of achieving state-of-the-art reconstruction performance, with strong robustness to ambient surroundings and the position, gesture and viewpoint of the NLOS objects. We have demonstrated the functionality of our approach on a synthetic dataset, then applied transfer learning and verified the performance through real-world experiments. To illustrate that our approach is not limited to the reconstruction of geometric features of NLOS scenes, we step further to experimentally prove the effectiveness of our method for the recovery of full-color information. The methodology presented here potentially offers a paradigm shift for applications of real-time NLOS imaging, such as driving assistant, geological prospecting, biomedical imaging, animal behavioral experiments, and many more [31, 47, 46, 20, 11, 29, 53, 40, 38].

2. Methods

2.1. Lidar-based detection

The typical NLOS imaging scenario is depicted in Fig. 1. When an obstacle blocks the direct light path between the object and the camera, the conventional LOS imaging would fail to function as the detector receives no direct information from the target. However, another object or a wall between the camera and the target may serve as a 'relay station' and provide an indirect light path for NLOS imaging: a photon sent from the target object can be reflected by the relay object and further collected by the cam-

era. After that, information from both the target and the relay objects is embedded in the photon distribution received by the camera. To disentangle the two sources of data, we need to actively probe the surroundings to extract data under different perturbations. The probe is usually conducted through sending a laser in the field of view (FOV) to scan the environment and then collecting the perturbed photon distributions. In this work we use a low-cost Lidar (Intel RealSense, the functioning details of the Lidar are provided in the Supplementary Material) as both the laser source and the camera. The laser beam on the Lidar scans the FOV of 70 by 55 degrees, then the detector on it sequentially collects the photons reflected from the surrounding objects. Based on the ToF information of the received photons, the on-chip processor of the Lidar will infer a depth map and a light intensity map. In general cases (Fig. 2a), most of the photons received by the Lidar are the ones directly reflected back and travel along the reverse path. There are also photons deflected by the wall to different directions, further scattered by surrounding objects and finally collected by the Lidar. However, the light intensities along these multi-reflection pathways will be much lower than that of the direct-reflection light. In this case, the Lidar will count the travel time of the direct-reflection photons and the calculated depth will represent the orthogonal distance between the Lidar plane and the incidence point on the wall, which is typical for the regular Lidar operation.

In another case (Fig. 2b), if the incidence angle is very oblique (>45 degree) and the point of illumination on the wall is close to the target object (<0.3 m in our cases), the multi-reflection paths may overshadow the direct-reflection path. Considering the three-bounce multi-reflection cases (Lidar-wall-object-wall-Lidar), there will be numerous possible three-bounce light paths with the same optical path length, and their total light intensity can be stronger than the direct-reflection one. As a result, the Lidar will denote the distance as half of the multi-reflection optical path length, i.e., the NLOS distance, rather than the direct distance. Although this so-called 'multipath artifacts' is undesired in conventional Lidar applications, it is favorable in our NLOS imaging [19]. In the common case as shown in Fig. 2a, the Lidar will provide the correct depth and intensity, and the information represents the geometric and reflection features of the relay wall, which also implies the position of the Lidar relative to the wall. While in the NLOS case (Fig. 2b), even though the received photons by the Lidar contain information of actual light intensities and optical paths, the Lidar will incorrectly map them to 'fake' locations. In this way, the ground truth photon distribution that reveals the NLOS information will be distorted and overshadowed by the Lidar mechanism. Our methodology focuses on redistributing the intensity and depth maps to extract the original NLOS information and thus achieving imaging and reconstruction
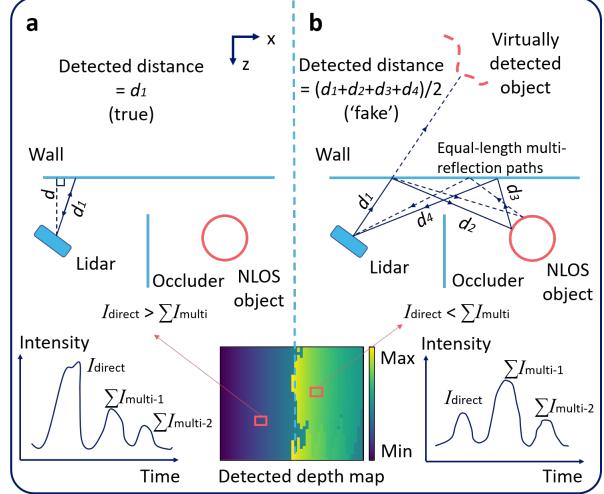

Figure 2. Two cases of Lidar detection (top view). (a) Generally, when the light directly reflected from the wall is far stronger than the light experiencing multi-reflection, the Lidar will denote the correct depth as the actual distance between itself and the wall. (b) If the total intensity of the beams undergoing equal-length multi-reflection is more intense than the direct-reflected light, the Lidar will store the intensity of multi-reflected light and treat the multi-reflection optical path length as the direct-reflection one, which leads to a 'fake detected distance'.

in a highly efficient and faithful manner.

2.2. Two-step deep remapping framework

Since the redistributing or remapping function varies with different NLOS scenes, here we apply deep learning to learn the connection between the ground truth NLOS objects and detected maps of depth and intensity. There are some pioneering works that have introduced supervised learning and applied deep neural networks (NN, such as U-Net [42]) into NLOS imaging. However, the remapping task has two intrinsic requirements: (1) transform the collected information into NLOS representations, and (2) generate NLOS scenes with clean semantics [14]. As a result, a regression-oriented NN trained in the supervised fashion cannot simultaneously meet these two requirements with decent performance. In contrast, here we propose a reconstruction framework consisting of two networks, each fulfilling one requirement: a compressor transforms the detected depth and intensity maps into a latent vector, and a generator decodes the latent vector to the predicted NLOS scene. With the sequential functions of the two components, the algorithm can reconstruct NLOS scenes with high fidelity, comparable to the performance of the state-of-the-art transient-based methods, with a processing time of no more than milliseconds.

The pipeline of the framework is depicted in Fig. 3a. The generator is the decoder of a VAE. Similar to GAN, VAE is
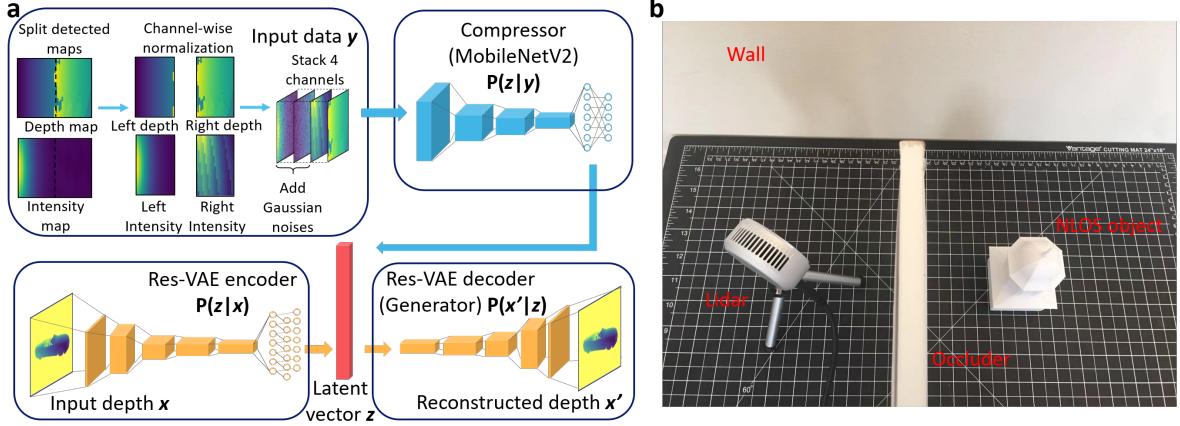
Figure 3. (a) Algorithm workflow and (b) experimental setup for NLOS imaging. (a) The NLOS reconstruction framework consists of two neural networks, a generator and a compressor. The reconstruction process is two-step: the compressor compresses the detected data as a latent vector, and the generator decodes the latent vector into a depth map of the NLOS object. The compressor is adapted from MobileNetV2, and the generator is the decoder of a variational autoencoder (VAE) with ResNet blocks. The input contains the depth and intensity maps detected by the Lidar, stacked as a 4-channel tensor with channel-wise normalization and Gaussian noises. (b) The experimental setup for NLOS imaging used in this work.

a powerful tool for scene and image synthesis, and offers better controllability on the sampling of the latent space. A typical CNN-based VAE consists of an encoder and a decoder with mirror structures to each other. The VAE applied here is composed of ResNet blocks to improve the resolving power and training effectiveness, and we denote it as Res-VAE [23]. Specifically, the encoder down-samples an input image x (64 by 64 in our experiments) to a 256-dimension latent vector z, which means the encoder learns the posterior $p(z|x)$, and z is reparameterized to force its distribution $p(z)$ to be a standard Gaussian distribution. Further, the decoder ($p(x'|z)$) expands z to a reconstructed image x'. The training process is to lower the difference between x and x' (reconstruction loss), as well as the distance between $p(z)$ and a standard Gaussian distribution (Kullback-Leibler divergence). Once the training is finished, the decoder will be able to generate images similar to the ones in the training dataset, and the latent vectors can be randomly sampled from the standard Gaussian distribution. In our case, the input x is the depth map of the NLOS scene relative to the wall, which corresponds to the perpendicular distances between the wall and the elements on the point clouds of NLOS objects. Since the mapping between x and z is mostly one-to-one, the reconstruction is no longer necessary to transform the detected information y to a high-dimension x. Instead, it only needs to compress y to a low dimensional vector z, which can be facilitated by a CNN-based compressor. The role of the compressor is rather standard in computer vision, which is fed with images and generates the predicted labels. Here the compressor is adapted from a lightweight network, MobileNetV2, to achieve good

performance and avoid overfitting [43]. The input of the compressor is the depth-and-intensity map detected by the Lidar, and the output predicts the latent vectors of the target NLOS depth map encoded by the pretrained Res-VAE.

3. Experiments

To validate the efficacy of the Lidar-based approach and the deep learning reconstruction algorithm, we have conducted 4 categories of experiments. The first one is based on a synthetic dataset, and the second experiment is on real objects with transfer learning applied. We further demonstrate the power of our method for full-color NLOS imaging, with experiments on images of everyday objects and complex scenes, respectively.

3.1. Experiment on a synthetic dataset

NLOS renderer. We first train and test the functionality of our framework on a synthetic dataset, and the weights can be further optimized when applied to real-world experiments through transfer learning. The synthetic dataset is composed of the NLOS depth maps and the corresponding simulated depth and intensity maps detected by the Lidar. As a result, a NLOS renderer is necessary to simulate the behavior of the Lidar [28]. Our rendering framework is based on three-bounce simulation. For each incidence, if there is one multi-reflection optical path length whose total intensity (total intensity is the intensity summation of all the three-bounce light paths with the same optical path length) is greater than the direct-reflection intensity, the NLOS depth and intensity will be denoted. Otherwise, it

| | Bowl | Soundbox | Table | Car | Cup | Sofa | Tea table | Motorbike | Chair | Monitor |

Ground truth depth map
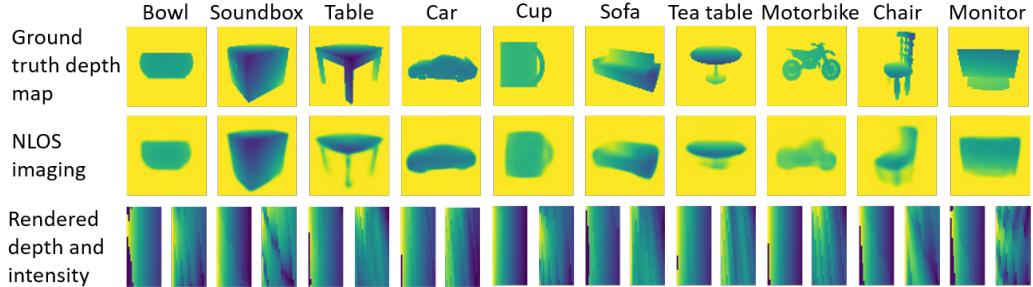
NLOS imaging

Rendered depth and intensity

Figure 4. The NLOS reconstruction experiment on a synthetic dataset. The reconstruction results of randomly selected everyday items in the test set are presented. The first row is the ground truth depth maps of the items, and the second row shows the reconstructed depth maps. The last row presents the inputs for the NLOS reconstruction, which are the simulated depth and intensity maps by our NLOS renderer. Since we assume the NLOS objects are situated on the right of the Lidar, the left halves of the depth and intensity maps are almost identical for different items. Therefore, here we only present the right halves of these maps. For each case, the left image represents the depth map, and the one on the right is for the intensity.

will be a common LOS case. The implementation details of the renderer are described in the Supplementary Material. The final depth intensity maps consist of 80 by 64 pixels. Since the renderer conducts matrix operations and multi-processing calculation, the rendering for one object takes only seconds, which is much more efficient than traditional ray-tracing renderers. The targets for rendering are around 30,800 depth maps of everyday objects (cars, tables, lamps, boats, chairs, etc.) from ShapeNet [12, 54]. To improve the variance of the dataset, the objects are situated with various altitudes, angles and distances to the wall.

Networks training. Once the dataset is formed, it is time to train the neural networks in our framework. The Res-VAE needs to be trained first since we need to push the corresponding latent vectors of the depth maps into a dictionary. During the training of Res-VAE, the Kullback-Leibler distance is not necessary to be suppressed to a very low level. Since the compressor will infer the prior distribution $p(z)$ through $p(z|y)$, it is insignificant if $p(z)$ is not strictly Gaussian. With the acquisition of the latent vectors, the compressor is trained to map the rendered depth and intensity maps to the latent vectors. As illustrated in Figs. 1 and 3b, we assume the Lidar is located on the left of the object, so the left halves of the depth and intensity maps will mostly denote the LOS information of the wall, while the right parts are encoded with the information of both the wall and the NLOS objects. The right halves of the depth maps correspond to larger distances, thereby leading to weaker intensities. To prevent the subtle signals on the right from being flooded by the overwhelming information on the left, we split the maps to the left and right halves, and vertically stack them into a 4-channel input image (ordered as right-depth, right-intensity, left-depth, left-intensity), with each channel normalized independently as illustrated in Fig. 3(a). Besides, the left-depth and left-

intensity are almost the same for different data points since they mostly document the LOS information. Hence, we introduce Gaussian noises to these two channels in order to add randomness and prevent overfitting. The 30,800 data points are randomly split into a training dataset of 24,600 and a test dataset of 6,200. The details of the training process and the structures of the generator and the compressor are provided in the Supplementary Material.

Reconstruction performance. With the training of both the generator and the compressor, we will be able to predict the NLOS scenes based on the captured depth and intensity maps. The NLOS reconstruction results of different categories of ShapeNet-items from the test dataset are plotted in Fig. 4. The first two rows are the ground truth depth maps and the reconstructed results, respectively. The third row of the figure presents the rendered depth and intensity maps which are used as the input for the reconstruction framework. Since their left halves are almost identical and indistinguishable to the eye, we only show their right halves for each case. Although the examples are randomly chosen and with a wide range of shapes, the reconstruction algorithm is able to recover significant features of them. This experiment demonstrates the high quality of the NLOS imaging through the developed method in this work.

3.2. Real-world NLOS reconstruction

Having confirmed the performance of our NLOS imaging method on the synthetic dataset, we further apply it to real-world NLOS scenes. In real-world applications we usually cannot conduct an enormous amount of detections to generate a big training dataset, therefore we utilize the networks with weights pretrained on the synthetic dataset, further continue to train the networks on a relatively small real-world dataset through transfer learning. We have 54 3D-printed objects as the target objects, which are mod-
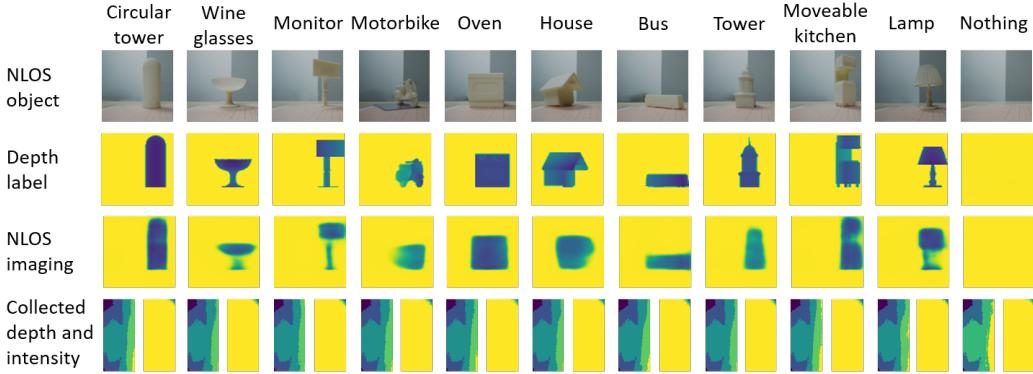
Figure 5. The NLOS reconstruction experiment on real objects. The first row displays the pictures of the actual objects used in this experiment, and the second row corresponds to the depth labels for training. The reconstructed depth maps are shown in the third row. The last row presents the inputs for the NLOS reconstruction, which are the depth and intensity maps collected by the Lidar (right halves only as explained before).

els of common objects like buses, tables, pianos, etc. The experimental environment is presented in Fig. 3b, with the objects positioned at 9 different locations with 5 rotational angles (rotation only for non-circular items), generating a dataset of around 1,900 NLOS cases. Although the generator does not need to retrain, the compressor trained on the synthetic dataset is not directly applicable to the real-world environment. Besides, such a tiny dataset (randomly split into 1,600 as the training dataset and 300 as the test set) is not compatible to a large-scale network due to inevitable overfitting, and a simple network is unable to handle the task. In this case, we leverage transfer learning to mitigate the training dilemma and transfer the knowledge that the compressor has learnt from the synthetic dataset to the real-world dataset. The target labels for the compressor are the latent vectors of the virtual depth maps corresponding to the 1,900 scenes. Since the compressor has been previously trained on the synthetic dataset, we continue to train it on the real-world training dataset for several epochs until the training loss and test loss are about to diverge. Next, we freeze all the network weights except for the last linear layer, and train it until the loss becomes steady. The entire process of the transfer learning takes only a couple of minutes, and the reconstruction performance on the test dataset is displayed in Fig. 5. Each column presents one example, and the four rows are the pictures of the actual objects, the virtual depth labels, the reconstructed depth maps, and the right halves of detected depth and intensity maps, respectively. It is worth noting that our methodology is able to reconstruct sufficient geometric and positional information of the NLOS objects. And if there is nothing behind the occluder, the framework will not provide a false-positive NLOS result. Certain details are not fully recovered (such as the motorbike), which is expected since some of the reflected light cannot be captured by the Lidar, thus the

loss of information is inevitable. Nevertheless, the reconstruction quality is comparable or superior to those of the state-of-the-art methods, with much lower requirements on equipment and ambient environment along with much faster imaging speed. Since the imaging scale and data format are all different from other methods, it is impossible to compare them quantitatively, while a qualitative comparison and error analysis are provided in Supplementary Material. The average accuracy of the reconstructed depth is 98.27% (root mean square error of 4.68 mm, see Supplementary Material for metric definition). Such a performance indicates our methodology is applicable to small-scale NLOS scenarios that require good precision and fast speed, such as biomedical imaging, animal behavioral surveillance, etc. If we further use open-source Lidar that allows access to detailed ToF information, the NLOS signal will be captured even though it is no stronger than the LOS signal. In this case, the imaging will not be limited to the cases where the multipath artifacts happen. Besides, the maps acquired by the Lidar will indicate the position of Lidar and bidirectional reflectance distribution function (BRDF) of the wall, which functions for calibration and makes the model readily transferable to other imaging setups (such as positional displacement of Lidar, material change of wall or objects) by meta-learning or phase compensation. As all deep learning algorithms are data-driven, we will enlarge our dataset in further study, and setup-independent NLOS imaging will also be formalized in future. Still, since the model has learnt the inherent physics of photon remapping, it will be able to recover some NLOS features even if the imaging condition is altered or the target is far from the distribution of the dataset.

3.3. Full-color NLOS imaging

NLOS reconstruction mostly refers to the reconstruction of the geometric features of NLOS objects, while regular LOS imaging also captures and records the color information. Here we step forward to demonstrate the capacity of our methodology for the recovery of NLOS scenes with full color information.

Most commercial Lidars utilize infrared laser to perceive the surrounding environment. However, it is impossible to detect the color information in the visible wavelength range through an infrared laser at a single wavelength. One promising solution is to build an optical system with red, green and blue (RGB) lasers as the scanning seed [13]. While stick to commercially available devices, we demonstrate a Lidar and a camera are able to cooperatively recover the color images displayed in the NLOS scenes. Since the Lidar is equipped with a camera, we do not need to add additional detectors or components to the system. The experimental setup for full-color NLOS imaging is illustrated in Fig. 6. A screen displays images but is invisible to the camera, while we aim to recover the colored images based on the light projected to the wall. In this experiment, the reconstruction objective will not be single-channel depth maps, but three-channel RGB images instead [44]. The algorithm is similar to previous, as depicted in Fig. 6a. The training dataset is composed of RGB images of everyday objects. Each time the screen displays a new image, the camera will capture an image of the light spot on the wall, and the Lidar also denotes the depth map. We have collected 48,800 data points and 43,000 of them are grouped into a training set. During the training phase, the Res-VAE is trained to generate RGB images of the objects, and the latent vectors of the images in the training set are documented. As for the compressor, the input is a 4-channel tensor, which is stacked with the RGB image collected by the camera along with the depth map detected by the Lidar. Here we add random Gaussian noises to the depth maps to introduce variance. The depth maps will function more significantly if the angle and position of the Lidar are dynamic, which is left to future research. To illustrate that the functionality of our NLOS framework is not limited to specific neural networks, here the compressor is adapted from ResNet50, which is also more compatible to the large-scale dataset with RGB information. The reconstructed RGB images of different categories of everyday items are displayed in Fig. 7, with original images, reconstructed images, and RGB light intensity maps arranged from the top row to the bottom. It is evidenced that the geometric features and color information are retained to a great extent, which validates the capacity of the proposed approach. Providing the RGB scanning light source is introduced, we will be able to recover both the shape and color of nonluminous objects.
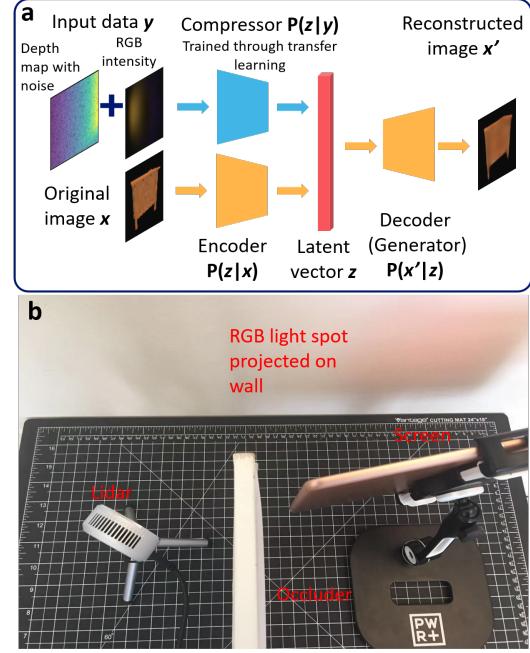

Figure 6. (a) Algorithm workflow and (b) experimental setup for full-color NLOS imaging. (a) The algorithm is essentially the same as the one presented in Fig. 3, while the input here is the concatenated tensor of captured RGB light intensity map and the depth map with Gaussian noise. The reconstruction objective is to obtain the full-color image. (b) When the screen displays an image, the device is able to detect the depth map along with the RGB light intensity map projected on the wall, which will be used for full-color NLOS imaging.

3.4. Full-color NLOS imaging for complex scenes

A mature imaging technique should not have constraints on the number of objects to be captured, and the imaging quality should not degrade with the increased complexity of the scene. To explore the full capacity of our solution, we test it on a much more complicated RGB dataset, STL10 [16]. Instead of images of isolated objects in pure black background, STL10 contains real pictures taken with different devices, angles, positions, and surrounding environments. The pictures may have more than one objects, and the semantic meanings are intricate and entangled with each other. To accommodate the increased amount of information in the image, we extend the dimension of the latent vector from 256 to 512, while other parameters and the training pipeline remain the same as those in the previous experiments. Some examples from the test set (50,000 data points in the training set, 10,000 in the test set) are presented in Fig. 8. Since the images in STL10 are not well-categorized, the examples are randomly picked from the whole test set. It is noticeable that major profiles and color information are correctly unfolded, while some fine

	Sofa	Single Sofa	Table	Cabinet	Desk	Tea table	Cellphone	Chair	Soundbox	Monitor

Original image

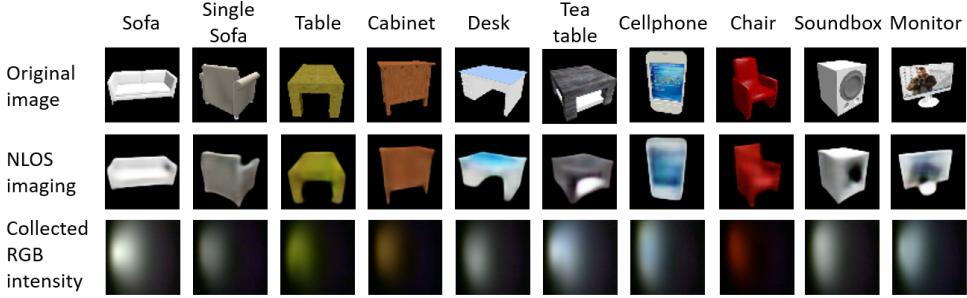NLOS imaging

Collected RGB intensity

Figure 7. Full-color NLOS reconstruction experiment on displayed images of everyday items. The results of different categories of everyday items are presented, with the first row as the original images, and the second row as the reconstructed full-color images. The third row presents the RGB light intensity maps captured by the camera on the Lidar.

Original image

NLOS imaging

Collected RGB intensity

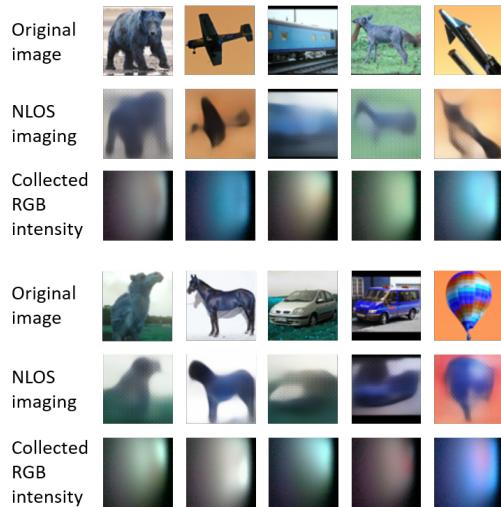Original image

NLOS imaging

Collected RGB intensity

Figure 8. Full-color NLOS reconstruction experiment on displayed images of complex scenes. Randomly selected examples from the whole test set are presented, with the original images, the reconstructed full-color images, and the RGB light intensity maps shown sequentially.

details are sacrificed. The performance is mainly limited by the generative capability of the generator. Introducing advanced VAE models, such as β-VAE, IntroVAE, VQ-VAE-2, is expected to mitigate this issue [10, 25, 41].

4. Conclusion and discussion

We have developed a general methodology for real-time NLOS imaging, with innovation in both imaging equipment and reconstruction algorithm. Physically, we choose low-cost Lidar to replace the complicated devices commonly used for ultrafast experiment, and our scheme features much faster detection speed and better robustness to ambient environment. Algorithmically, we have proposed a deep learning framework consisting of two networks, a compressor and a generator. The framework is not in the typical manner of supervised learning, and the reconstruction process is two-step. The efficacy of the methodology has been verified experimentally, on both real-object NLOS reconstruction and full-color NLOS imaging, with the state-of-the-art performance. Our approach is directly pertinent to some real-world applications, such as corner-detection for unmanned vehicles, NLOS remote sensing, NLOS medical imaging, to name a few.

We realize additional efforts are needed to bring the proposed method to perfection. The Lidar used in this work has a very low laser power, which is designed for small range detection. We plan to transfer our technique to high-performance Lidar with open-source access to ToF information, such as those equipped on self-driving cars, to fulfill large-scale NLOS tasks, including but not limited to NLOS imaging of street scenes, buildings, and vehicles. Other follow-up efforts include the introduction of RGB light sources for full-color detection of nonluminous objects, development of techniques to disentangle the effects of the position and angle of Lidar, etc. As for the algorithm, the projected bottleneck for complex NLOS tasks will be the generative capability of the generator. Compared to GANs, the generated images of most VAEs are blurrier. Introducing most advanced VAE architectures is expected to resolve this problem. Since the compressor is responsible for extracting NLOS information from the depth and intensity maps, it would be favorable to update attention blocks to energize the compressor with better efficiency on locating the NLOS features. In addition, instead of training the generator and the compressor sequentially as performed in this work, we expect improved performance if they are trained concurrently, similar to the mechanism of GAN.

5. Acknowledgements

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 International Conference on Computer Vision*, pages 7889–7899, 2019. 1

[2] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, and Frédo Durand. Computational mirrors: Blind inverse light transport by deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 14311–14321, 2019. 1, 2

[3] Victor Arellano, Diego Gutierrez, and Adrian Jarabo. Fast back-projection for non-line of sight reconstruction. *Optics express*, 25(10):11574–11583, 2017. 2

[4] Manel Baradad, Vickie Ye, Adam B. Yedidia, Frédo Durand, William T. Freeman, Gregory W. Wornell, and Antonio Torralba. Inferring light fields from shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

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

[6] Wolfgang Becker. *Advanced time-correlated single photon counting techniques*, volume 81. Springer Science & Business Media, 2005. 2

[7] Jacopo Bertolotti, Elbert G Van Putten, Christian Blum, Ad Lagendijk, Willem L Vos, and Allard P Mosk. Non-invasive imaging through opaque scattering layers. *Nature*, 491(7423):232–234, 2012. 1

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

[9] Katherine L Bouman, Vickie Ye, Adam B Yedidia, Frédo Durand, Gregory W Wornell, Antonio Torralba, and William T Freeman. Turning corners into cameras: Principles and methods. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2270–2278, 2017. 1

[10] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018. 8

[11] Susan Chan, Ryan E Warburton, Genevieve Gariepy, Jonathan Leach, and Daniele Faccio. Non-line-of-sight tracking of people at long range. *Optics express*, 25(9):10109–10117, 2017. 2

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

[13] Wenzheng Chen, Simon Daneau, Fahim Mannan, and Felix Heide. Steady-state non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6799, 2019. 2, 7

[14] 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. 3

[15] 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. 2

[16] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, 11–13 Apr 2011. 7

[17] Boguslaw Cyganek and J Paul Siebert. *An introduction to 3D computer vision techniques and algorithms*. John Wiley & Sons, 2011. 1

[18] Daniele Faccio, Andreas Velten, and Gordon Wetzstein. Non-line-of-sight imaging. *Nature Reviews Physics*, pages 1–10, 2020. 1

[19] Stefan Fuchs. *Calibration and multipath mitigation for increased accuracy of time-of-flight camera measurements in robotic applications*. PhD thesis, Universitätsbibliothek der Technischen Universität Berlin, 2012. 3

[20] Genevieve Gariepy, Francesco Tonolini, Robert Henderson, Jonathan Leach, and Daniele Faccio. Detection and tracking of moving objects hidden from view. *Nature Photonics*, 10(1):23–26, 2016. 1, 2

[21] Massimo Ghioni, Angelo Gulinatti, Ivan Rech, Franco Zappa, and Sergio Cova. Progress in silicon single-photon avalanche diodes. *IEEE Journal of selected topics in quantum electronics*, 13(4):852–862, 2007. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

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

[25] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in neural information processing systems*, pages 52–63, 2018. 8

[26] Mariko Isogawa, Dorian Chan, Ye Yuan, Kris Kitani, and Matthew O'Toole. Efficient non-line-of-sight imaging from transient sinograms. In *European Conference on Computer Vision*, pages 193–208. Springer, 2020. 1

[27] J Itatani, F Quéré, Gennady L Yudin, M Yu Ivanov, Ferenc Krausz, and Paul B Corkum. Attosecond streak camera. *Physical review letters*, 88(17):173903, 2002. 2

[28] Adrian Jarabo, Julio Marco, Adolfo Muñoz, Raul Buisan, Wojciech Jarosz, and Diego Gutierrez. A framework for transient rendering. *ACM Transactions on Graphics (ToG)*, 33(6):1–10, 2014. 4

[29] Ori Katz, Pierre Heidmann, Mathias Fink, and Sylvain Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics*, 8(10):784–790, 2014. 2

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[31] Jonathan Klein, Christoph Peters, Jaime Martín, Martin Laurenzis, and Matthias B Hullin. Tracking objects outside the line of sight using 2d intensity images. *Scientific reports*, 6(1):1–9, 2016. 2

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 2

[34] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[35] Xin Lei, Liangyu He, Yixuan Tan, Ken Xingze Wang, Xinggang Wang, Yihan Du, Shanhui Fan, and Zongfu Yu. Direct object recognition without line-of-sight using optical coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2019. 2

[36] Xiaochun Liu, Ibón Guillén, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature*, 572(7771):620–623, 2019. 1

[37] Christopher A Metzler, Felix Heide, Prasana Rangarajan, Muralidhar Madabhushi Balaji, Aparna Viswanath, Ashok Veeraraghavan, and Richard G Baraniuk. Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging. *Optica*, 7(1):63–71, 2020. 2

[38] Ji Hyun Nam, Eric Brandt, Sebastian Bauer, Xiaochun Liu, Eftychios Sifakis, and Andreas Velten. Real-time non-line-of-sight imaging of dynamic scenes. *arXiv preprint arXiv:2010.12737*, 2020. 2

[39] Matthew O'Toole, David B Lindell, and Gordon Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338–341, 2018. 1

[40] Adithya Pediredla, Akshat Dave, and Ashok Veeraraghavan. Snlos: Non-line-of-sight scanning through temporal focusing. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–13. IEEE, 2019. 2

[41] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pages 14866–14876, 2019. 8

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[44] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019. 7

[45] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. *Advances in neural information processing systems*, 18:1161–1168, 2005. 1

[46] Nicolas Scheiner, Florian Kraus, Fangyin Wei, Buu Phan, Fahim Mannan, Nils Appenrodt, Werner Ritter, Jurgen Dickmann, Klaus Dietmayer, Bernhard Sick, and Felix Heide. Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2068–2077, 2020. 2

[47] Brandon M. Smith, Matthew O'Toole, and Mohit Gupta. Tracking multiple objects outside the line of sight using speckle imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[48] Christos Thrampoulidis, Gal Shulkind, Feihu Xu, William T Freeman, Jeffrey H Shapiro, Antonio Torralba, Franco NC Wong, and Gregory W Wornell. Exploiting occlusion in non-line-of-sight active imaging. *IEEE Transactions on Computational Imaging*, 4(3):419–431, 2018. 1

[49] Chia-Yin Tsai, Kiriakos N Kutulakos, Srinivasa G Narasimhan, and Aswin C Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7216–7224, 2017. 1

[50] Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications*, 3(1):1–8, 2012. 1, 2

[51] Claus Weitkamp. *Lidar: range-resolved optical remote sensing of the atmosphere*, volume 102. Springer Science & Business, 2006. 2

[52] Shumian Xin, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis Gkioulekas. A theory of fermat paths for non-line-of-sight shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6800–6809, 2019. 1

[53] Feihu Xu, Gal Shulkind, Christos Thrampoulidis, Jeffrey H Shapiro, Antonio Torralba, Franco NC Wong, and Gregory W Wornell. Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *Optics express*, 26(8):9945–9962, 2018. 2

[54] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems*, volume 32, pages 492–502, 2019. 5