Learning-based Spotlight Position Optimization for Non-Line-of-Sight Human Localization and Posture Classification

Sreenithy Chandran
Arizona State University
schand56@asu.edu

Tatsuya Yatagawa
Hitotsubashi University
tatsuya.yatagawa@r.hit-u.ac.jp

Hiroyuki Kubo
Chiba University
hkubo@chiba-u.jp

Suren Jayasuriya
Arizona State University
sjayasur@asu.edu

Abstract

Non-line-of-sight imaging (NLOS) is the process of estimating information about a scene that is hidden from the direct line of sight of the camera. NLOS imaging typically requires time-resolved detectors and a laser source for illumination, which are both expensive and computationally intensive to handle. In this paper, we propose an NLOS-based localization and posture classification technique that uses an off-the-shelf projector and camera system. We leverage a message-passing neural network to learn a visible scene geometry and predict the best position to be spotlighted by the projector that can maximize the NLOS signal. The neural network is trained end-to-end and the network parameters are optimized to maximize the NLOS performance. Unlike prior deep-learning-based NLOS techniques that assume planar relay walls, our system allows us to handle line-of-sight scenes where scene geometries are more arbitrary. Our method demonstrates state-of-the-art performance in object localization and position classification using both synthetic and real scenes.

1. Introduction

Non-line-of-sight (NLOS) imaging refers to the technique of imaging hidden parts of a scene that are not within the field of view of a camera. This involves interpreting the illumination reflected/scattered from the NLOS object onto visible surfaces. NLOS imaging has been employed for the identification, tracking and 3D shape reconstruction of hidden objects. NLOS imaging techniques are rapidly developing [11] and currently have numerous applications, such as search and rescue [43], endoscopy [26], and hidden pedestrian detection for autonomous driving [2].

NLOS imaging was first demonstrated by Velten et al. [41] using an ultra-fast laser and a streak camera. Subsequent research in transient imaging leveraged a pulsed laser with high-resolution temporal detectors such as single-photon avalanche diodes (SPADs) [4, 30, 31, 43]. Active transient imaging pulses a fast laser into the scene and measures the time that the photon takes to arrive back at the temporal detector. However, high temporal resolution with SPADs requires precise calibration and long acquisition times. Furthermore, the time efficiency of processing SPAD data processing is insufficient for large scenes and high-resolution images [25, 43]. Another alternative is to use continuous wave Time-of-Flight(ToF) cameras with modulated light sources [15, 17, 27]. ToF cameras are cheaper than streak cameras and SPADs and are popular in real-time NLOS applications when high resolution is not needed [27].

Cameras are by far the cheapest detectors, albeit lacking

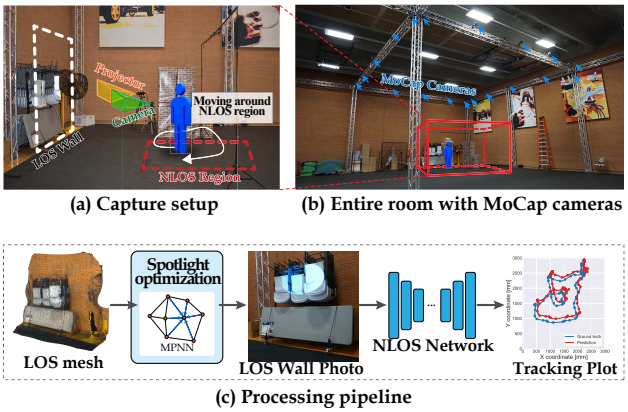

Figure 1. Given the polygonal mesh of a target scene, our method predicts which area of the scene to illuminate with a spotlight and maximize light scatter information from a hidden person. Then, we capture RGB images of the wall visible from the camera under optimal illumination. Finally, our neural network predicts the 2D position and posture of the hidden person.

information on light transport, such as ToF of light. Therefore, researchers have also explored NLOS techniques using conventional cameras and lasers [8, 16, 20, 23] and ambient illumination [1, 3, 33–35, 39]. Recently, the use of scene priors and deep learning has become popular to overcome the ill-posedness of NLOS imaging problem [17, 29, 33].

In this work, we present an active data-driven NLOS posture classification and tracking pipeline that works with a standard RGB camera and single spotlight illumination. Our approach does not require optical alignment or system calibration. It combines a graph neural network with a physics-based differentiable renderer to optimally determine a spotlight position to maximize NLOS performance. The goal of illumination estimation is to learn the best illumination direction that maximizes the NLOS radiance that reaches the camera, since we have knowledge of the LOS geometry. We leverage this to improve downstream NLOS imaging tasks. A major focus of our method is to move beyond small-scale imaging setups with line-of-sight(LOS) walls/ visible surfaces that are mostly planar to work across scenes that are practically present in the real world. Chandran et al. [7] proposed an approach to handle LOS scenes with occlusions. However, their imaging model assumed diffuse reflectance for the LOS wall and handled scenes with limited complexity and very small NLOS volumes ($30\text{cm} \times 30\text{cm} \times 30\text{cm}$). We build a large dataset of realistic looking synthetic scenes with complex geometry, textures, occlusions, etc. for this purpose. We also captured highly accurate real data with human NLOS subjects and validated our method using this dataset. Our specific contributions include the following:

- An end-to-end neural computational imaging method to learn the best illumination for a LOS scene mesh to maximize NLOS performance. Our pipeline consists of a novel message-passing neural network for estimating spotlight position, a physics-based renderer, and a neural network for NLOS localization/posture classification.
- Owing to the use of differentiable rendering in our pipeline, the proposed method works significantly well for realistic-scale scenes with non-diffuse surfaces and self-illuminating objects.
- We used synthetic and real data to demonstrate superior performance compared to several baselines.

Our method achieves a highly accurate localization of unknown human subjects. We surpass the best competing methods by more than 45 cm in terms of root mean square error. Compared to methods that use only a single-intensity LOS wall image, our method based on optimizing the spotlight has clear advantages, as shown by experimental results and ablative studies. Check our project page for more details.

2. Related Work

Active illumination in NLOS: Active illumination methods employ controlled illumination sources (e.g., lasers and projectors) and detectors to explore the hidden parts of scenes. Kirmani et al. [18] proposed the first framework for transient imaging to "look around the corner." Velten et al. [41] introduced a backpropagation technique for NLOS scene reconstruction, this was later used in gated systems [22] and SPADs [4]. Furthermore, the non-impulse illumination was also shown worthy for NLOS tasks [19].

Passive illumination in NLOS: Passive illumination methods [1, 3, 21, 28, 36] employ ambient light for NLOS imaging tasks. For instance, some considered the objects in the scene as pinspecks or pinholes [33, 34, 39], while others utilized occluders [3, 45], such as doorways [21], to reconstruct the hidden scene. Moreover, Sharma et al. [36] leveraged raw signals from a LOS wall to perform NLOS tasks, while Medin et al. [28] leveraged cast shadows of objects on LOS diffuse walls and inferred biometric information of humans in an NLOS region.

Deep learning for NLOS: For NLOS tasks, deep learning techniques have been used with both ToF and conventional RGB data. Carmazzo et al. [6] introduced a neural network, which was trained with the data captured using a SPAD setup, to perform localization and identification tasks. Chen et al. [9] proposed a deep-learning-based method that uses scene priors. They trained a neural network using a differentiable transient renderer to perform the NLOS imaging tasks. Xu et al. [44] performed human pose recognition for a transient NLOS dataset characterized by the confocal NLOS model. Chen et al. [8] utilized a U-net-like architecture to reconstruct the scene geometry from steady-state NLOS data. Cao et al. [5] introduced the CNN-Based NLOS Localization Under Changing Ambient Illumination (NLOS-LUCAI). He et al. [13] introduced a deep learning framework for simultaneous real-time imaging and tracking of dynamic targets using an RGB camera.

The work closest to ours is by Chandran et al. [7]. They proposed an adaptive lighting framework using physics-based optimization, estimating where on a LOS wall the projector should illuminate to maximize NLOS information. They also proposed a deep learning-based approach to predict the locations of NLOS objects from intensity images. They, however, worked with only approximately planar diffuse walls with small NLOS region dimensions. In contrast, our work goes beyond this to handle walls with complexities, occlusions, and varying materials.

Differentiable rendering for NLOS: The utilization of differentiable rendering has been increasing in recent times, especially for the purpose of analysis-by-synthesis (AbS), also known as inverse rendering. Klein et al. [20] used AbS to track NLOS objects, formulating the problem as a non-

linear optimization based on data from light transport simulation and real measurements. Tsai et al. [40] employed a SPAD setup to simultaneously acquire NLOST objects' shape geometries and reflectance properties in the AbS manner. We propose an end-to-end approach that utilizes a differentiable path tracer to transmit information from the image domain to the polygon mesh that represents the scene domain.

3. Method

This section outlines our proposed method. Section 3.1 describes our problem statement. Then, we describe the proposed processing pipeline consisting of three components. Section 3.2 introduces the first component, the Illumination Estimation Network (IEN), a graph neural network that estimates the optimal lighting position to maximize the quantity of NLOS information. Section 3.3 discusses the second component, a differentiable rendering engine that uses the illumination information given by the first component. Section 3.4 describes the last component, a neural network that involves estimating the position and posture of the human subject from the RGB picture calculated by the second component.

3.1. Problem Statement

Our imaging system consists of a projector P as our illumination source and a camera C as our detector. We use the projector only to illuminate a single spot, as opposed to projecting spatially varying illumination. The imaging system is positioned without direct field of view over the NLOS object as shown in Figure 2. We represent the visible surface as a polygonal mesh. The light from the projector P hits the visible surface at, triangle $t = (v_0, v_1, v_2)$, then reaches the NLOS object O before returning to the LOS surface at another triangle t', and finally captured by the camera C. The hidden NLOS object has a location $l = (x, y)$ and a posture associated with it. We restrict our attention to light effects from three-bounce paths of the form, $P \rightarrow t \rightarrow O \rightarrow t' \rightarrow C$, which represents a path connecting the source P and camera C interacts with the NLOS surface only once, as shown in Fig. 2. This simplification is motivated by previous observations that photons following higher-order paths are difficult to detect using existing sensors. The image of the LOS surface I is related to the location of an NLOS object $l = (x, y)$ by a function F, i.e.,

$$I = F(l, \alpha, \phi), \tag{1}$$

where α refers to the position of the illumination on the LOS surface and ϕ refers to the other parameters that affect the captured image, such as the material of the LOS surface, NLOS subject posture, and noise. The forward function F models the light transport matrix of the setup. The goal of our study is to invert this function F and optimize α to more accurately recover the object location l.

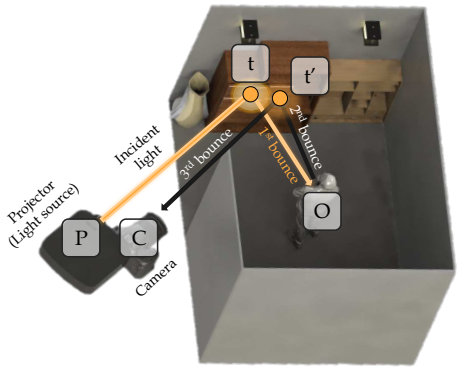

Figure 2. Problem Setup: The light source P and the camera C are focused on the LOS surface. The human subject O moves around in the NLOS region. Light from projector P hits a visible surface at triangle t, travels to the NLOS subject O, bounces off the LOS surface at another triangle t', and then, comes back into the camera C.

3.2. Illumination Estimation Network

The active light source used for the NLOS problem plays an important role in improving the signal-to-noise ratio (SNR) of the NLOS information, as demonstrated by Chandran et al. [7]. The primary question that our study aims to address is finding out where to shine the spotlight on a visible surface. To address this, we introduce an illumination estimation network (IEN). The IEN takes a mesh of a scene and outputs the nodes of the triangle that have to be illuminated, to maximize NLOS information. Here, the size of the mesh and the relative camera position against the visible surface can be specified arbitrarily by a unit distance that does not necessarily need to correspond to physical units (e.g., centimeters and millimeters).

The IEN is based on a message-passing neural network (MPNN) of Gilmer et al. [12] to handle the LOS meshes of arbitrary sizes. We represent the input LOS mesh (acquired through 3D scanning in practice) as a triangle mesh $M = (V, F)$, where V and F correspond to sets of vertices and faces, respectively. A 3D mesh is transformed into a graph $G = (X, A)$ where X has dimension $(|V|, 3)$ and defines the spatial xyz-features for each node, and the adjacency matrix A with dimension $(|V|, |V|)$ defines the connected neighborhood of each node. The vertex attributes of the graph are passed to a multilayer perceptron (MLP), i.e., the vertex-wise feed-forward network, to obtain the vertex-level features. Then, the output from this encoder is passed to the graph convolutional network of Verma et al. [42]. The feature update in each graph convolution layer is given as

$$\mathbf{h}_v^{(l)} = \mathbf{b} + \sum_{m=1}^{M} \frac{1}{|\mathcal{N}_v|} \sum_{u \in \mathcal{N}_v} \alpha_m^{(l)}(v, u) \mathbf{W}_m^{(l)} \mathbf{h}_u^{(l-1)}, \tag{2}$$

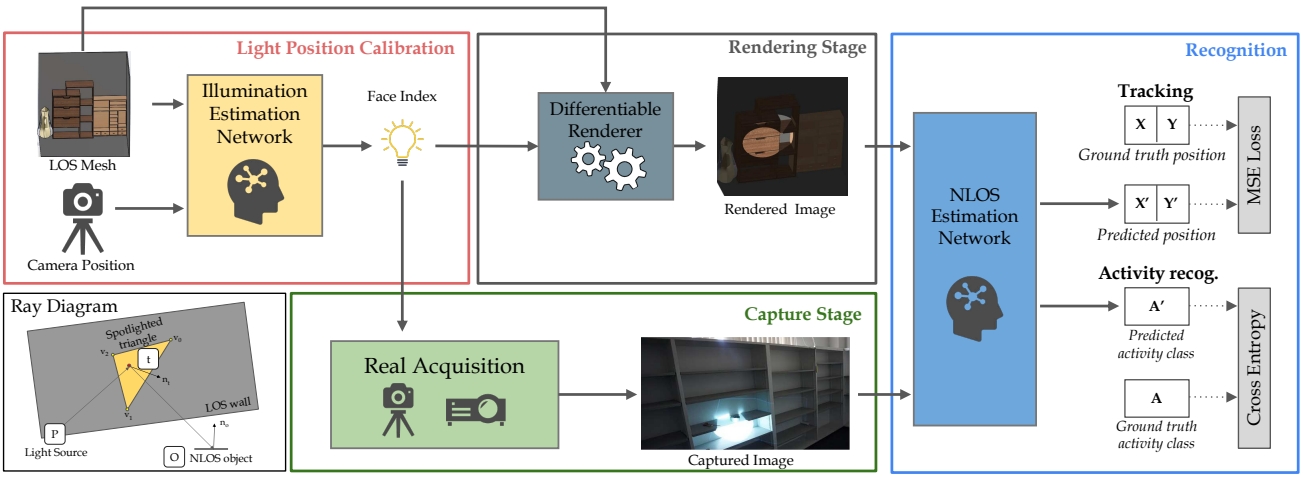

Figure 3. Full training and inference pipeline of our system. The light position calibration, rendering stage, and recognition blocks are used in training, while the light position calibration, capture stage and recognition blocks are used in real data inference.

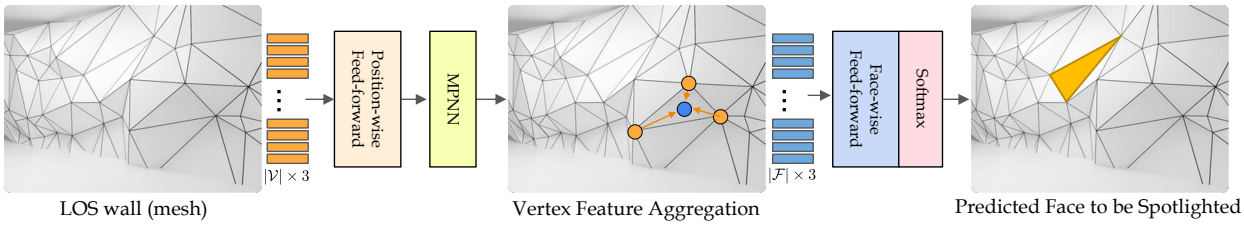

Figure 4. An overview of our Illumination Estimation Network. The network consists of a graph convolutional network, it is preceded by a position-wise feed-forward network and followed by a face-wise feed-forward network.

where $\mathbf{b}$ is a bias vector, $\alpha_m^{(l)}(v,u)$ is the attention weight obtained by the m-th attention block, $\mathbf{W}_m^{(l)}$ is the linear transformation matrix associated with a graph convolution layer, $\mathcal{N}_v$ is the set of adjacent vertices of a target vertex v (including v itself), and $|\mathcal{N}_v|$ is the set's cardinal. The attention weights $\alpha_m^{(l)}(v,u)$ are calculated as

$$\alpha_m^{(l)}(v,u) = \frac{\exp\left(\mathbf{u}_m^{(l)} \cdot (\mathbf{h}_u^{(l-1)} - \mathbf{h}_v^{(l-1)}) + c_m^{(l)}\right)}{\sum_{m=1}^{M} \exp\left(\mathbf{u}_m^{(l)} \cdot (\mathbf{h}_u^{(l-1)} - \mathbf{h}_v^{(l-1)}) + c_m^{(l)}\right)}, \quad (3)$$

where $\mathbf{u}_m^{(l)}$ and $c_m^{(l)}$ are learnable parameters, specific to each layer l. The attention coefficients are normalized so that they sum to 1, i.e., $\sum_{m=1}^{M} \alpha_m^{(l)}(i,j) = 1$. The encoded node-level features are then sequentially passed through a stack of three feature-steered convolutional layers. Each of these layers aggregates messages from two attention heads. The two labels correspond to either light on or light off. Finally, the refined node-level features are passed the prediction block built with an MLP, which outputs the probability of how likely each triangle should be spotlighted.

3.3. Physics-Based Differentiable Rendering

We exploit a differentiable renderer in our proposed pipeline. Our rendering engine is built upon "redner" [24], a differentiable renderer based on edge sampling. With this engine, we can obtain an RGB picture of the LOS surface visible from the camera through physically-based rendering in a differentiable manner. Since our pipeline is trained end-to-end, the differentiable path tracer is essential to back-propagate the image-domain features to the mesh domain. In our case, the goal of the renderer is to compute the gradients of an illuminated LOS surface with respect to the position of the light used in the illumination. This offers the core of our contribution, identifying the best position at which a spotlight should shine on the LOS surface.

3.4. NLOS Network

The goal here is to perform NLOS localization and posture classification, that is, we identify the posture performed by the human and also obtain the 2D location of the person. We assume that the position of the light is largely based on the location of the human and not the posture performed by the hidden human. Thus, to train our pipeline, we use the mean square error (MSE) between the predicted loca-

tion l' and the NLOS ground truth location l. But, as shown in Fig. 3, we also have an NLOS subnetwork that predicts posture based on input RGB images, for which we use the standard cross-entropy loss between the predicted posture label and the ground-truth posture label. We use a ResNet-18 [14] as our feature extractor, this is then fed to two subnetworks. For both the tracking and posture classification tasks, we use an MLP decoder, which consists of three fully connected layers with a ReLU activation and is followed by the last fully connected layer that outputs the (x, y) coordinates. On the contrary, the last layer is activated with the softmax function for posture classification.

3.5. Training and Inference

All of our training is done on synthetic data, and the inference performance is evaluated with both synthetic and real data. Refer to Sec. 4.1 for details of the simulated data used for the training. During training time, the entire pipeline including the IEN, differentiable renderer, and NLOS network is trained from end to end. During inference on real data, we used the trained weights of the IEN to estimate where the spotlight should be placed. Refer to Sec. 4.2 for specific details of real data capture. The captured LOS mesh is decimated and then passed into the IEN which gives the estimate of the spotlight position. After that, we proceed to capture an RGB image of the visible surface with the given illumination. Lastly, the RGB image is passed to the NLOS network to obtain localization or posture classification results. During inference, our method takes about 7ms per estimation on average to process the RGB input and output the posture classification and tracking predictions. More details are available in the supplemental material.

4. Dataset

4.1. Simulated Data

Our goal for training was to generate a dataset that is close to real-world scenarios both in terms of realism (textures, compositions, objects, occlusions, etc.) and also in terms of scale. We generated 30 LOS scenes for this purpose. Most learning-based active or passive NLOS methods make use of very small-scale NLOS setups and NLOS objects. For example, NLOS objects, such as a 3D-printed bunny, dragon, etc., have been conventionally used in the imaging community. They are not as realistic as the variety of objects that can be found in real-world settings. Thus, to enhance the realism of synthetic scenes, we collect publicly available meshes and arrange them in the scenes using SolidWorks.

Since our goal is posture classification and localization of human subjects, we use human models to generate our data. We perform similar activities to the ones in Sharma et al. [36]. This includes standing, sitting, crouching, hands at 90° with respect to the floor, and hands at 45° (to mimic waving). In addition to this, we also have random gestures that are generated for classification as *unrecognized activity*.Generating a great deal of scene data is a lengthy process. To increase the number of synthetic scenes, we have implemented a data augmentation step in Blender. We have created a plugin for Blender to do this, the specifics of which are in the supplemental material.

4.2. Real Data

We collected a real-world dataset consisting of 8 indoor scenes with 5 human subjects of varying heights between 5.0–6.2 ft. This includes LOS surfaces in classrooms, conference rooms, storage rooms, and bedrooms. Some sample scenes are shown in the supplemental document. The subject was at a distance of approximately 2.0–8.0 ft from the LOS surface. We used an InFocus IN3138HDa projector to create a spotlight illumination and a Sony α6000 mirrorless camera to capture the illuminated LOS scene. We also considered the presence of ambient light while adjusting the exposure parameters.

LOS mesh: We use the Polycam LIDAR capture feature app on the iPhone 13 Pro to get a mesh of the visible surface. The LIDAR sensor on the iPhone has a maximum range of 5.00 m. The captured LOS meshes originally consisted of 3000–10,000 vertices, depending on the complexity of the wall. These meshes were decimated to consist of 500–1000 vertices to reduce computational complexity.

Ground truth acquisition: We used the OptiTrack motion capture system to get high-quality localization as ground truth values with 0.50 mm precision. The human subjects wore a suit with IR markers for motion capturing. To increase the diversity of data, we also captured several indoor scenes without a motion-capture rig. This was performed with a USB camera on the ceiling and an ArUco marker put on the subject's head. Given the marker size in the image captured by a camera with calibrated intrinsic and extrinsic parameters, we obtained the 2D position of the subject in an NLOS region using off-the-shelf pose estimation software.

5. Experiments

In this section, we will cover the training specifics, the metrics used to assess the performance of our approach, and the competing methods we compared it to. We will then present the results of our proposed method and provide a more in-depth analysis of it. Here are several assumptions that we made in our experiments. When we shine a light on the spot proposed by the IEN, we manually focus the projector on that spot, although there could be some illumination on adjacent triangles too. For all of our experiments, we consider that there is only one human subject acting around the NLOS region at a time.

5.1. Training Details

The pipeline is implemented using PyTorch, where the graph convolutional network is constructed using the `MessagePassing` module provided by the PyG library [10]. We train the network using Adam optimizer with a learning rate of 10^{-2} with a weight decay of 10^{-5}. On a computer with two NVIDIA GTX 1080 Ti graphics cards, the training takes approximately two days. Note that the test data set consisted of LOS surfaces that were not present in the training data set.

5.2. Comparisons

Since our method works on RGB image inputs to the NLOS network, it fundamentally distinguishes it from methods based on SPADs. Given this difference in the input data, a direct comparison with SPAD-based methods would not provide meaningful insights. Instead, we compare our method against the following state-of-the-art methods, chosen specifically for their similarity in acquisition setup.

RGB Images: We directly used the RGB images without active illumination (only ambient light) in the scene to train the proposed NLOS localization/posture classification networks. The goal of this baseline is to reveal the importance of the IEN of our method.

Adaptive Lighting: This setup is the one presented by Chandran et al. [7], which proposes an adaptive lighting method to determine which one or more spots of light should be focused on the scene. This approach uses an optimization technique rather than our learning-based approach to determine where to shine the light. We leverage the code shared by the authors for our implementation. For all the NLOS scenes in our training dataset, we used only the LOS geometry and obtained the best illumination patch to shine light on according to their method. Then we render the scenes in the dataset using the given illumination and train on their CNN architecture.

Flash Photography: This setup is the one presented by Tanick et al. [38], they use a regression network and generative network for NLOS-based scene reconstruction. This setup is similar to ours in terms of involving a flashlight and a normal RGB camera. We re-implement the network described in [38] and train it in our data set. Their regression network performs both localization and classification, and we adapt the architecture of the classification network so that the last layer accounts for our 6 posture classes.

DL-NLOS: This setup is the one presented by He et al. [13], which introduces deep learning for NLOS localization solely on RGB images captured under ambient illumination. Their localization consists of five convolutional layers followed by three fully connected layers. We reimplement their proposed network architecture based on the details in the paper. We update the last layers with softmax to perform posture classification as well. We have made adjustments to the baselines to the best of our ability to match and adopt to our problem statement.

Table 1. Result of posture classification against synthetic and real scenes, where the numbers in this table show correct recognition ratio in units of %.

Scene type	NLOS Posture	RGB Images	Adaptive Lighting	Flash Photo.	DL-NLOS	Ours
Synthetic	Standing	56.6	75.5	69.2	75.2	**97.1**
	Sit	52.6	77.3	66.3	74.7	**96.8**
	Crouch	52.1	74.3	66.1	73.2	**96.2**
	Hands (90)	53.8	76.2	68.4	74.8	**94.5**
	Hands (45)	54.2	75.8	67.9	74.3	**94.3**
	Unknown	50.8	70.1	65.3	71.3	**97.7**
Real	Standing	50.9	72.9	63.5	72.6	**94.1**
	Sit	48.1	72.1	61.2	71.5	**88.8**
	Crouch	49.7	70.1	62.1	69.6	**87.2**
	Hands (90)	47.2	70.6	61.2	71.1	**86.2**
	Hands (45)	44.3	71.7	62.8	70.9	**85.0**
	Unknown	44.9	65.2	59.9	66.4	**82.9**

5.3. Posture Classification Performance

The full results for both synthetic and real data are shown in Table 1. Our posture classification network identified human posture with 96.1% accuracy for 10 unknown LOS synthetic scenes. The average performance by RGB-only training is 53.2%, flash photography method is 67.2%, this was bettered by He et al. [13] with 73.9% and Chandran et al. [7] by about 74.8%. For real scenes, our method has a classification accuracy of 87.4%, and the closest best-performing methods were [7, 13] with 70.4% accuracy. Refer to supplemental material for further analysis.

5.4. Localization Performance

We evaluate the accuracy in localization using the average distance (i.e., localization error) between the ground truth positions in the moving trajectory and those predicted by our method. For both synthetic and real data, Tab. 2 shows comparisons of our method with competing methods, where the average distances are denoted in units of centimeters. The average localization error for synthetic scenes for our method is 6.33 cm for subjects performing known activities, while 9.86 cm for subjects performing unknown activities that the network did not see during training. For real scenes, the errors for known and unknown activities are 31.45 cm and 45.14 cm, respectively. Compared to baseline methods, errors are 124.76 cm for RGB-only training, 87.09 cm for the adaptive lighting method [7], 100.83 cm for flash photography [38], and 85.90 cm for DL-NLOS [13]. According to these results, the performance improvement over the network trained only on RGB images validates the importance of the IEN. Moreover, our

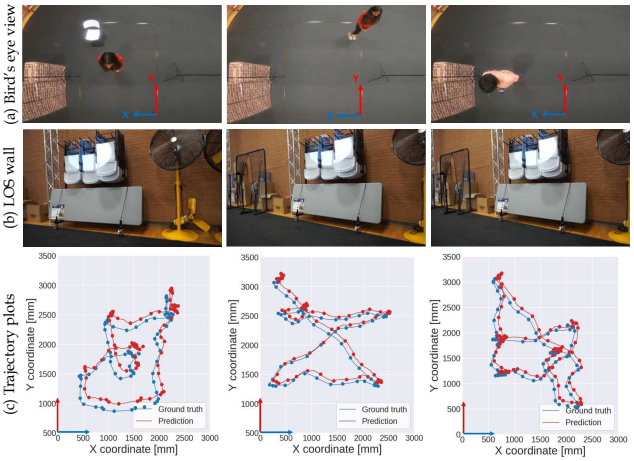

Figure 5. Results of real-world tracking. (a) shows a bird's eye view of the human walking in the NLOS space, (b) shows photos of the LOS wall, and (c) shows the trajectory plot of the ground truth and our prediction for test video sequences. Refer to our supplementary video for more information.

Table 2. Results of localization against synthetic and real scenes, where the average distance between ground truth and predicted positions are shown in units of centimeters.

Scene type	NLOS Posture	RGB Images	Adaptive Lighting	Flash Photo.	DL-NLOS	Ours
Synthetic	Standing	19.89	10.34	13.02	9.56	**5.43**
	Sit	18.16	12.84	14.78	11.41	**7.56**
	Crouch	20.02	12.91	17.45	13.16	**7.81**
	Hands (90)	20.31	14.72	19.87	12.02	**5.12**
	Hands (45)	21.90	15.06	19.73	12.89	**5.71**
	Unknown	29.14	17.89	20.72	15.21	**9.86**
Real	Standing	107.12	82.80	92.98	82.11	**28.43**
	Sit	130.89	84.86	97.41	85.31	**30.71**
	Crouch	125.67	86.77	100.43	86.01	**32.51**
	Hands (90)	110.73	88.12	102.50	85.76	**28.67**
	Hands (45)	121.56	88.64	102.98	86.52	**33.89**
	Unknown	156.90	90.53	110.32	90.62	**45.14**

method outperforms all competing methods and, furthermore, its accuracy surpasses that of the best of the competing methods by more than 50.00 cm for both known and unknown activities. Fig. 5 shows tracking trajectories obtained by our method for real data. We have included real video test results in our supplemental video. It should be noted that the trajectories of our method in the figure have been smoothed by the Savitzky–Golay filter [32] to improve the estimation of the trajectory by refining the noisy raw output from the network. This smoothing operation is a practical step, which can be seamlessly integrated into our system, making it a justified part of the evaluation process.

5.5. Importance of Spotlight Position Optimization

We assess the contribution of the IEN to the NLOS task by conducting an additional experiment. We compare our method with the following four alternatives.

Table 3. Results of ablative studies to validate the effectiveness of the illumination predictions. We group the results based on the average trajectory error and average posture classification across all the test data. The localization metrics are presented in units of centimeters, and the correct recognition ratios are in units of %. The results shown here are for simulated data.

Task	IEN +CNN	AL +ResNet	Random +ResNet	Center +ResNet	Ours
Localization (↓) (Average Error [cm])	10.71	14.23	19.16	22.63	**8.10**
Posture Classification (↑) (Accuracy [%])	91.28	79.84	67.36	64.09	**96.13**

Figure 6. (a) shows the sample scene with ambient lighting present, (b) is the scene with the LOS surface illuminated by our IEN prediction direction, (c) is the scene with spotlight direction selected by [7], (d) is illumination in the center of the scene, (e) is a randomly selected spotlight direction that is chosen.

IEN+CNN: We construct a model that comprises the IEN followed by the CNN for localization and classification proposed by Chandran et al [7].

AL+ResNet: We use the spotlight position selected by adaptive lighting [7] and use that as input to the NLOS network consisting of ResNet+MLP used in our proposed method. It is assumed that the walls of the line-of-sight (LOS) are diffuse, as is the case with the adaptive lighting technique.

Random+ResNet: We also compare with alternatives in which the location of the spotlight is selected randomly somewhere on the LOS surface. For NLOS tasks, the same network consisting of our ResNet+MLP is utilized.

Center+ResNet: As with the above, we also compare an alternative in which the spotlight always illuminates the center of the field of view. Again, the same network consisting of ResNet + MLP as ours is utilized for NLOS tasks.

The visual comparison of our method and the following alternatives is shown in Fig. 6. Table 3 shows the comparison between the proposed method and these alternatives. This table demonstrates that our approach is superior to the other options, indicating that our method was successful in identifying the most suitable area to illuminate, resulting in maxmimizing NLOS signal to the detector. Obviously, the proposed method outperforms Random+ResNet and Center+ResNet which are based on simple heuristics.

Our method also outperforms AL+ResNet, the adaptive

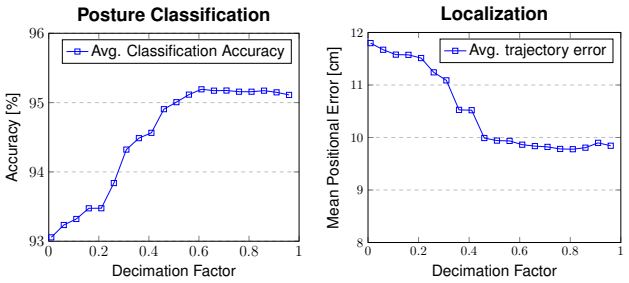

Figure 7. Plot shows the effect of decimation on posture classification accuracy and average trajectory error. The decimation factor varies between 0.1 to 1, where 1 refers to the highest resolution of capture and 0.1 refers to decimating the total number of vertices in the LOS mesh is reduced by a factor of 10.

lighting method [7] extended by our NLOS network. The lower performance of AL + ResNet suggests that the diffuse assumption by [7] does not work well when the scene includes specular surfaces (e.g., mirrors, glasses), metallic surfaces, translucent materials (e.g., wax, plastics), and strongly textured surfaces. The adaptive lighting method is indeed prone to shining a light on a position on the diffuse surface. For example, the bottom row of Fig. 6 shows that the adaptive lighting [7] overlooks the refrigerator on the right, which may reflect more light. In contrast, our method appropriately shines the light on the refrigerator, which reflects the light from the NLOS object the most. Clearly, when IEN+CNN and our method are compared, it is evident that the ResNet backbone does improve the NLOS performance of our method. We did not conduct any experiments to evaluate the effects of different network backbones, feature extractors, etc. on the NLOS task, as our aim is to demonstrate the significance of spotlight optimization. Also, it must be noted that, the size of spotlight is directly related to the area of the decimated patch that has to be illuminated.

5.6. Effect of Mesh Decimation

To understand how the mesh resolution of the LOS area affects the NLOS performance, we alter the resolution of the scene mesh at different ratios by decimating it. The LOS meshes we captured have a diverse number of vertices, as described in Sec. 4.2. To ensure a fair evaluation across the test set, we select LOS meshes with approximately the same number of vertices (i.e., 9000–11000 vertices), and reduce them up to about one-tenth of their original size (approximately 1000 vertices). The meshes with different resolutions are then input to the pipeline. The experimental results in Fig. 7 show that the performance of the NLOS task does not increase significantly only by using a high-resolution mesh. It is attributed to the increasing difficulty of obtaining an adequate feature from a higher-resolution mesh. This observation suggests that the original high-resolution meshes contain much more geometric details than what is required to interpret the scene geometry. Therefore, we may decrease the mesh resolution to approximately 50% of the original, where the geometric details are visually retained. This also indicates the robustness of our technique to the accuracy of the LOS scan.

6. Conclusion

In this work, we demonstrate the importance of choosing the optimal position to be illuminated in an active LOS imaging system using a projector and standard RGB camera. We verified our method with synthetic and real data from real-world scenes with a human in the NLOS region. We showed the proposed method's state-of-the-art tracking and posture classification performance in challenging scenarios where the LOS region may be partly occluded and consist of components with non-diffuse materials. The proposed method was successful in posture classification for unknown real-world scenes, achieving an accuracy of approximately 87%. It also achieved a highly accurate localization of unknown human subjects moving around the NLOS region, with a root mean square error of approximately 45 cm.The localization error of our method is approximately one-half of those obtained by the best of the state-of-the-art methods that we compared. These results highlight the importance of optimizing the position of the spotlight, the primary focus of this study.

For future work, we plan to explore the use of spatially varying illumination that could be more optimal than a single spotlight. The NLOS region size that can be handled by our method is currently limited by the low SNR signals from the NLOS objects. Hence, our method was tested only on a single human subject in the NLOS region. To overcome the limitation of subject type and number of subjects, we would like to investigate incorporating computational imaging hardware into the end-to-end optimization loop [27, 37].

Acknowledgements

This work was supported by JSPS KAKENHI (JP19H04138, JP22K17907) and JST FOREST (JPMJFR206I), as well as NSF grant IIS-1909192. The authors would like to thank Shenbagaraj Kannapiran for helping to conduct tracking experiments at the ASU Drone Studio. The authors acknowledge Research Computing at Arizona State University for providing GPU resources for this research.

References

[1] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, and Fredo Durand. Computational mirrors: Blind inverse light transport by deep

matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[2] Sven Bauer, Robin Streiter, and Gerd Wanielik. Non-line-of-sight mitigation for reliable urban gnss vehicle localization using a particle filter. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 1664–1671. IEEE, 2015. 1

[3] 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. 2

[4] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics Express*, 23(16):20997–21011, 2015. 1, 2

[5] Yanpeng Cao, Rui Liang, Jiangxin Yang, Yanlong Cao, Zewei He, Jian Chen, and Xin Li. Computational framework for steady-state nlos localization under changing ambient illumination conditions. *Optics Express*, 30(2):2438–2452, 2022. 2

[6] Piergiorgio Caramazza, Alessandro Boccolini, Daniel Buschek, Matthias Hullin, Catherine F Higham, Robert Henderson, Roderick Murray-Smith, and Daniele Faccio. Neural network identification of people hidden from view with a single-pixel, single-photon detector. *Scientific Reports*, 8(1):11945, 2018. 2

[7] Sreenithy Chandran and Suren Jayasuriya. Adaptive lighting for data-driven non-line-of-sight 3d localization and object identification. *British Machine Vision Conference (BMVC)*, 2019. 2, 3, 6, 7, 8

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

[9] 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. 2

[10] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 6

[11] Ruixu Geng, Yang Hu, Yan Chen, et al. Recent advances on non-line-of-sight imaging: Conventional physical models, deep learning, and new scenes. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2021. 1

[12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017. 3

[13] JinHui He, ShuKong Wu, Ran Wei, and YuNing Zhang. Non-line-of-sight imaging and tracking of moving objects based on deep learning. *Optics Express*, 30(10):16758–16772, 2022. 2, 6

[14] 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. 5

[15] 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

[16] Connor Henley, Tomohiro Maeda, Tristan Swedish, and Ramesh Raskar. Imaging behind occluders using two-bounce light. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 573–588. Springer, 2020. 2

[17] Achuta Kadambi, Hang Zhao, Boxin Shi, and Ramesh Raskar. Occluded imaging with time-of-flight sensors. *ACM Transactions on Graphics (ToG)*, 35(2):1–12, 2016. 1, 2

[18] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using transient imaging. In *2009 IEEE 12th International Conference on Computer Vision*, pages 159–166. IEEE, 2009. 2

[19] Ahmed Kirmani, Haris Jeelani, Vahid Montazerhodjat, and Vivek K Goyal. Diffuse imaging: Creating optical images with unfocused time-resolved illumination and sensing. *IEEE Signal Processing Letters*, 19(1):31–34, 2011. 2

[20] 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):32491, 2016. 2

[21] William Krska, Sheila W Seidel, Charles Saunders, Robinson Czajkowski, Christopher Yu, John Murray-Bruce, and Vivek Goyal. Double your corners, double your fun: the doorway camera. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022. 2

[22] Martin Laurenzis and Andreas Velten. Nonline-of-sight laser gated viewing of scattered photons. *Optical Engineering*, 53(2):023102–023102, 2014. 2

[23] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2019. 2

[24] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph.*, 37(6):222:1–222:11, 2018. 4

[25] Xiaochun Liu, Sebastian Bauer, and Andreas Velten. Analysis of feature visibility in non-line-of-sight measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10140–10148, 2019. 1

[26] Tomohiro Maeda, Guy Satat, Tristan Swedish, Lagnojita Sinha, and Ramesh Raskar. Recent advances in imaging around corners. *arXiv preprint arXiv:1910.05613*, 2019. 1

[27] Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6):1–12, 2017. 1, 8

[28] Safa C Medin, Amir Weiss, Frédo Durand, William T Freeman, and Gregory W Wornell. Can shadows reveal biometric information? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 869–879, 2023. 2

[29] 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

[30] 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. 1

[31] 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

[32] William H Press and Saul A Teukolsky. Savitzky-golay smoothing filters. *Computers in Physics*, 4(6):669–672, 1990. 7

[33] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019. 2

[34] Sheila W Seidel, Yanting Ma, John Murray-Bruce, Charles Saunders, William T Freeman, C Yu Christopher, and Vivek K Goyal. Corner occluder computational periscopy: Estimating a hidden scene from a single photograph. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2019. 2

[35] Sheila W Seidel, John Murray-Bruce, Yanting Ma, Christopher Yu, William T Freeman, and Vivek K Goyal. Two-dimensional non-line-of-sight scene estimation from a single edge occluder. *IEEE Transactions on Computational Imaging*, 7:58–72, 2020. 2

[36] Prafull Sharma, Miika Aittala, Yoav Y Schechner, Antonio Torralba, Gregory W Wornell, William T Freeman, and Frédo Durand. What you can learn by staring at a blank wall. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2330–2339, 2021. 2, 5

[37] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018. 8

[38] Matthew Tancik, Guy Satat, and Ramesh Raskar. Flash photography for data-driven hidden scene recovery. *arXiv preprint arXiv:1810.11710*, 2018. 6

[39] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. *International Journal of Computer Vision*, 110:92–112, 2014. 2

[40] Chia-Yin Tsai, Aswin C Sankaranarayanan, and Ioannis Gkioulekas. Beyond volumetric albedo–a surface optimization framework for non-line-of-sight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1545–1555, 2019. 3

[41] 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):745, 2012. 1, 2

[42] Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2598–2606, 2018. 3

[43] Cheng Wu, Jianjiang Liu, Xin Huang, Zheng-Ping Li, Chao Yu, Jun-Tian Ye, Jun Zhang, Qiang Zhang, Xiankang Dou, Vivek K Goyal, et al. Non–line-of-sight imaging over 1.43 km. *Proceedings of the National Academy of Sciences*, 118(10):e2024468118, 2021. 1

[44] Qianqian Xu, Liquan Dong, Lingqin Kong, Yuejin Zhao, and Ming Liu. Active non-line-of-sight human pose estimation based on deep learning. In *2021 International Conference on Optical Instruments and Technology: Optical Systems, Optoelectronic Instruments, Novel Display, and Imaging Technology*, volume 12277, pages 123–131. SPIE, 2022. 2

[45] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12231–12239, 2019. 2