

Steady-state Non-Line-of-Sight Imaging

Wenzheng Chen^{1,2*} Simon Daneau^{1,3} Fahim Mannan¹ Felix Heide^{1,4}

¹Algolux ²University of Toronto ³Université de Montréal ⁴Princeton University

Abstract

Conventional intensity cameras recover objects in the direct line-of-sight of the camera, whereas occluded scene parts are considered lost in this process. Non-line-of-sight imaging (NLOS) aims at recovering these occluded objects by analyzing their indirect reflections on visible scene surfaces. Existing NLOS methods temporally probe the indirect light transport to unmix light paths based on their travel time, which mandates specialized instrumentation that suffers from low photon efficiency, high cost, and mechanical scanning. We depart from temporal probing and demonstrate steady-state NLOS imaging using conventional intensity sensors and continuous illumination. Instead of assuming perfectly isotropic scattering, the proposed method exploits directionality in the hidden surface reflectance, resulting in (small) spatial variation of their indirect reflections for varying illumination. To tackle the shape-dependence of these variations, we propose a trainable architecture which learns to map diffuse indirect reflections to scene reflectance using only synthetic training data. Relying on consumer color image sensors, with high fill factor, high quantum efficiency and low read-out noise, we demonstrate high-fidelity color NLOS imaging for scene configurations tackled before with picosecond time resolution.

1. Introduction

Recovering objects from conventional monocular imagery represents a central challenge in computer vision, with a large body of work on sensing techniques using controlled illumination with spatial [50, 42] or temporal coding [32, 24, 19, 39], multi-view reconstruction methods [18], sensing via coded optics [47], and recently learned reconstruction methods using single-view monocular images [49, 11, 16]. While these sensing methods drive applications across domains, including autonomous vehicles, robotics, augmented reality, and dataset acquisition for scene understanding [52], they only recover objects in

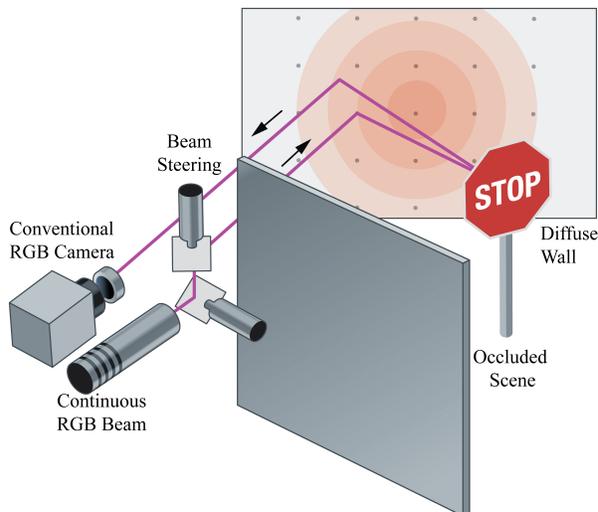


Figure 1: We demonstrate that it is possible to image occluded objects outside the direct line-of-sight using continuous illumination and conventional cameras, without temporal sampling. We sparsely scan a diffuse wall with a beam of white light and reconstruct “hidden” objects only from spatial variations in steady-state indirect reflections.

the direct line-of-sight of the camera. This is because objects outside the line-of-sight only contribute to a measurement through indirect reflections via visible diffuse object surfaces. These reflections are extremely weak due to the multiple scattering, and they lose (most) angular information on the diffuse scene surface (as opposed to a mirror surface in the scene). NLOS imaging aims at recovering objects outside a camera’s line-of-sight from these indirect light transport components.

To tackle the lack of angular resolution, a number of NLOS approaches have been described that temporally probe the light-transport in the scene, thereby unmixing light path contributions by their optical path length [1, 30, 36, 43] and effectively trading angular with temporal resolution. To acquire temporally resolved images of light transport, existing methods either directly sample the temporal impulse response of the scene by recording the temporal echoes of laser pulses [54, 43, 17, 7, 53, 3, 40], or they use amplitude-coded illumination and time-of-flight sensors [21, 26, 25].

*The majority of this work was done while interning at Algolux.

While amplitude coding approaches suffer from low temporal resolution due to sensor demodulation bandwidth limitations [32] and the corresponding ill-posed inverse problem [19], direct probing methods achieve high temporal resolution already in the acquisition phase, but in turn require ultra-short pulsed laser illumination and detectors with < 10 ps temporal resolution for macroscopic scenes. This mandates instrumentation with high temporal resolution, that suffers from severe practical limitations including low photon efficiency, large measurement volumes, high-resolution timing electronics, excessive cost and monochromatic acquisition. Early streak-camera setups [54] hence require hours of acquisition time, and, while emerging single photon avalanche diode (SPAD) detectors [7, 40] are sensitive to individual photons, they are in fact photon-inefficient (diffuse experiments in [40]) due to very low fill factors and pileup distortions at higher pulse power. To overcome this issue without excessive integration times, recent approaches [40, 20] restrict the scene to retro-reflective material surfaces, which eliminates quadratic falloff from these surfaces, but effectively also constrains practical use to a single object class.

In this work, we demonstrate that it is possible to image objects outside of the direct line-of-sight using conventional intensity sensors and continuous illumination, without temporal coding. In contrast to previous methods, that assume perfectly isotropic reflectance, the proposed method exploits directionality of the hidden object’s reflectance, resulting in spatial variation of the indirect reflections for varying illumination. To handle the shape-dependence of these variations, we learn a deep model trained using a training corpus of simulated indirect renderings. By relying on consumer color image sensors, with high fill factor, high quantum efficiency and low read-out noise, we demonstrate full-color NLOS imaging at fast imaging rates and in setup scenarios identical to those tackled by recent pulsed systems with picosecond resolution.

Specifically, we make the following contributions:

- We formulate an image formation model for steady-state NLOS imaging and an efficient implementation without ray-tracing. Based on this model, we derive an optimization method for the special case of planar scenes with known reflectance.
- We propose a learnable architecture for steady-state NLOS imaging for representative object classes.
- We validate the proposed method in simulation, and experimentally using setup and scene specifications identical to the ones used in previous time-resolved methods. We demonstrate that the method generalizes across objects with different reflectance and shapes.
- We introduce a synthetic training set for steady-state

NLOS imaging. The dataset and models will be published for full reproducibility.

2. Related Work

Transient Imaging Kirmani et. al [30] first proposed the concept of recovering “hidden” objects outside a camera’s direct line-of-sight using temporally resolved light transport measurements in which short pulses of light are captured “in flight” before the global transport reaches a steady state. These transient measurements are the temporal impulse response of light transport in the scene. Abramson [1] first demonstrated a holographic capture system for transient imaging, and Velten et al. [55] showed the first experimental NLOS imaging results using a femto-second laser and streak camera system. Since these seminal works, a growing body of work has been exploring transient imaging with a focus on enabling improved NLOS imaging [43, 36, 56, 17, 21, 19, 7, 38].

Impulse Non-Line-of-Sight-Imaging A major line of research [43, 54, 17, 40, 53, 3, 45, 41, 58] proposes to acquire transient images directly, by sending pulses of light into the scene and capturing the response with detectors capable of high temporal sampling. While the streak camera setup from Velten et al. [55] allows for temporal precision of < 10 ps, corresponding to a path length of 3 mm, the high instrumentation cost and sensitivity has sparked work on single photon avalanche diodes (SPADs) as a detector alternative [7, 41]. Recently, O’Toole et al. [41] propose scanned SPAD capture setup that allows for computational efficiency by modeling transport as a shift-invariant convolution. Although SPAD detectors can offer comparable resolution < 10 ps [37], they typically suffer from low fill factors typically around a few percent [44] and low spatial resolution in the kilo-pixel range [35]. Compared to ubiquitous intensity image sensors with > 10 megapixel resolution, current SPAD sensors are still five orders of magnitude more costly, and two orders of magnitude less photon-efficient.

Modulated and Coherent Non-Line-of-Sight-Imaging

As an alternative to impulse-based acquisition, correlation time-of-flight setups have been proposed [19, 25, 21, 26] which encode travel-time indirectly in a sequence of phase measurements. While correlation time-of-flight cameras are readily available, e.g. Microsoft’s Kinect One, their application to transient imaging is limited due to amplitude modulation bandwidths around 100 MHz, and hence temporal resolution in the nanosecond range. A further line of work [29, 28] explores using correlations in the carrier wave itself, instead of amplitude modulation. While this approach allows for single-shot NLOS captures, it is limited to scenes at microscopic scales [28].

Tracking and Classification Most similar to the proposed

method are recent approaches that use conventional intensity measurements for NLOS vision tasks [31, 8, 9, 5]. Although not requiring temporal resolution, these existing approaches are restricted to coarse localization and classification to a limited extent, in contrast to full imaging and geometry reconstruction applications.

3. Image Formation Model

Non-line-of-sight imaging methods recover object properties outside the direct line-of-sight from third-order bounces. Typically, a diffuse wall patch in the direct line-of-sight is illuminated, where the light then scatters and partially reaches a hidden object outside the direct line-of-sight. At the object surface, the scattered light is reflected back to the visible wall where it may be measured. In contrast to existing methods that rely on temporally resolved transport, the proposed method uses stationary third-bounce transport, i.e. without time information, to recover reflectance and geometry of the hidden scene objects.

3.1. Stationary Light Transport

Specializing the Rendering Equation [27] to non-line-of-sight imaging, we model the radiance L at a position \mathbf{w} on the wall as

$$L(\mathbf{w}) = \int_{\Omega} \rho(\mathbf{x} - \mathbf{l}, \mathbf{w} - \mathbf{x}) (\mathbf{n}_{\mathbf{x}} \cdot (\mathbf{x} - \mathbf{l})) \frac{1}{r_{\mathbf{x}\mathbf{w}}^2} \frac{1}{r_{\mathbf{x}\mathbf{l}}^2} L(\mathbf{l}) d\mathbf{x} + \delta(\|\mathbf{l} - \mathbf{w}\|) L(\mathbf{l}), \quad (1)$$

with \mathbf{x} , $\mathbf{n}_{\mathbf{x}}$ the position and corresponding normal on the object surface Ω , \mathbf{l} being a given beam position on the wall, and ρ denoting the bi-directional reflectance distribution function (BRDF). This image formation model assumes three indirect bounces, with the distance function r modeling intensity falloff between input positions, and one direct bounce, when \mathbf{l} and \mathbf{w} are identical in the Dirac delta function $\delta(\cdot)$, and it ignores occlusions in the scene outside the line-of-sight. We model the BRDF with a diffuse and specular term as

$$\rho(\omega_i, \omega_o) = \alpha_d \rho_d(\omega_i, \omega_o) + \alpha_s \rho_s(\omega_i, \omega_o) \quad (2)$$

The diffuse component ρ_d models light scattering, resulting in almost orientation-independent low-pass reflections without temporally coded illumination. In contrast, the specular reflectance component ρ_s contributes high-frequency specular highlights, i.e. mirror-reflections blurred by a specular lobe. These two components are mixed with a diffuse albedo α_d and specular albedo α_s . While the spatial and color distributions of these two albedo components can vary, they are often correlated for objects composed of different materials, changing only at the boundaries of materials on the same surface. Although the

proposed method is not restricted to a specific BRDF model, we adopt a Phong model [46] in the following.

3.2. Sensor Model

We use a conventional color camera in this work. We model the raw sensor readings with the Poisson-Gaussian noise model from Foi et al. [15, 14] as samples

$$b \sim \frac{1}{\kappa} \mathcal{P} \left(\int_T \int_W \int_{\Omega_A} L(\mathbf{w}) d\omega d\mathbf{w} dt \frac{\kappa}{E} \right) + \mathcal{N}(0, \sigma^2), \quad (3)$$

where we integrate Eq. (1) over the solid angle Ω_A of the camera's aperture, over spatial position W that the given pixel maps to, and exposure time T , resulting in the incident photons when divided by the photon energy E . The sensor measurement b at the given pixel is then modeled with the parameters $\kappa > 0$ and $\sigma > 0$ in a Poisson and Gaussian distribution, respectively, accurately reflecting the effects of analog gain, quantum efficiency and readout noise. For notational brevity, we have not included sub-sampling on the color filter array of the sensor.

4. Inverse Indirect Transport for Planar Scenes

In this section, we address the special case of planar objects. Assuming planar scenes in the hidden volume allows us to recover reflectance and 3D geometry from indirect reflections. Moreover, in this case, we can formulate the corresponding inverse problem using efficient optimization methods with analytic gradients. In the remainder of this paper, we assume that the shape and reflectance of the directly visible scene parts are known, i.e. the visible wall area. The proposed hardware setup allows for high-frequency spatially coded illumination, and hence the wall geometry can be estimated using established structured-light methods [50]. Illuminating a patch \mathbf{l} on the visible wall, a hidden planar scene surface produces a diffuse low-frequency reflection component, encoding the projected position independently of the orientation [31], and higher-frequency specular reflection components of the blurred specular albedo mapped to orientation-dependent positions on the wall. Assuming a single point light source at \mathbf{l} on the wall, see Fig. 2, the specular direction at a plane point \mathbf{p} is the mirror direction $\mathbf{r} = (\mathbf{p} - \mathbf{l}) - 2((\mathbf{p} - \mathbf{l}) \cdot \mathbf{n})\mathbf{n}$ with the plane normal being \mathbf{n} . The center of the specular lobe \mathbf{c} on the wall is the mirror point of \mathbf{l} , i.e. the intersection of the reflected ray in direction \mathbf{r} with the wall. Conversely, if we detect a specular lobe around \mathbf{c} in a measurement, we can solve for the corresponding plane point as

$$\mathbf{p}(\mathbf{v}, \mathbf{n}) = \mathbf{c} + ((\mathbf{v} - \mathbf{c}) \cdot \mathbf{n}) \left(\mathbf{n} - \frac{\mathbf{v} - \mathbf{l} - ((\mathbf{c} - \mathbf{l}) \cdot \mathbf{n})\mathbf{n}}{\mathbf{n} \cdot (2\mathbf{v} - \mathbf{c} - \mathbf{l})} \right), \quad (4)$$

that is a function of the planar surface represented by its normal \mathbf{n} and a point \mathbf{v} on the plane. Eq. (4) follows imme-

diately from the constraint that the orthogonal projections of the points l and c onto the plane result in equal triangles with p and the respective point, see Supplemental Material for a detailed derivation. The plane has three degrees of freedom (DOF), which we parametrize

$$\mathbf{n}(\theta, \phi) = \begin{bmatrix} \cos(\theta) \sin(\phi) \\ \sin(\theta) \sin(\phi) \\ \cos(\theta) \end{bmatrix}, \quad \mathbf{v}(\nu) = \vec{o} + \nu \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (5)$$

Expressing the normals in spherical coordinates ensures unit-length normals without explicit constraints. The plane position is here modeled as z-axis offset from the volume origin \vec{o} . With this parametrization we can estimate the specular albedo α_s and plane geometry θ, ϕ, ν . Specifically, we sequentially illuminate the wall in N spots uniformly sampled on the visible wall area and acquire a capture of the wall for each of the illumination points. Next, we extract sparse features and perform feature matching between the capture \mathbf{b}_t with the most detected features and all other captures $\mathbf{b}_{\{1, \dots, N\} \setminus t}$. We use SIFT features [34] and RANSAC [13] matching. For every feature $f \in \{1, \dots, F\}$, this results in a set of matched positions \mathbf{c}_i^f with $i \in \Psi_f$, and Ψ_f being here the set of images with matches for feature f . We select the top \tilde{F} features, with descending number of matches, and solve for the plane geometry by minimizing the re-projection distance on the plane as

$$\theta^*, \phi^*, \nu^* = \underset{\theta, \phi, \nu}{\operatorname{argmin}} \sum_{f=1}^{\tilde{F}} \sum_{i \in \Psi_f} \left\| \mathbf{p}_i^f(\mathbf{v}, \mathbf{n}) - \bar{\mathbf{p}}^f \right\|_2^2 \quad (6)$$

with $\bar{\mathbf{p}}^f = \frac{1}{|\Psi_f|} \sum_{i \in \Psi_f} \mathbf{p}_i^f(\mathbf{v}, \mathbf{n})$,

where we use the notational shortcut $\mathbf{p}_i^f(\mathbf{v}, \mathbf{n})$ for the plane point from Eq. 4 with reflected point \mathbf{c}_i^f and laser point l_i . With this objective, we solve for consensus between the re-projected points on the plane for all features. The variable $\bar{\mathbf{p}}^f$ represents here the mean position of all reprojected points for a given feature f . We solve the optimization problem from Eq. (6) using limited-memory BFGS [33] which is a highly efficient quasi-Newton method. The analytic gradient of the least-squares objective from Eq. (6), i.e. the partials w.r.t. θ, ϕ, ν , are derived in the Supplemental Material. Given the planar geometry estimate θ^*, ϕ^*, ν^* , we estimate homography mappings \mathbf{H}_i between every image plane i and the reprojected plane bounds. Specifically, we re-project the four image plane corners for measurement \mathbf{b}_i on the plane using Eq. 4 with laser position l_i and reflection points as the four image plane corners. Given the homographies, we estimate the plane reflectance as the following

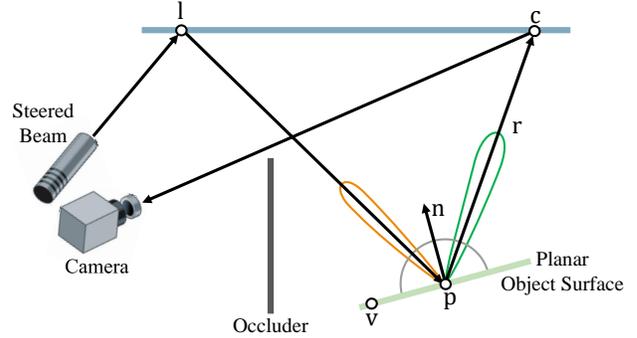


Figure 2: Indirect reflection on planar surface. A virtual source l on the diffuse wall indirectly illuminates p on a planar scene surface with \mathbf{n} and position \mathbf{v} . Depending on the surface BRDF, some light will be scattered back diffusely (uniformly in gray), retro-reflected back around l (red lobe), and specularly reflected in direction \mathbf{r} to a visible point c .

maximum likelihood estimation problem

$$\underset{\mathbf{x}}{\operatorname{argmin}} - \sum_{i=1}^N \log \left(p(\mathbf{b}_i | \mathbf{W}_{\mathbf{H}_i} \mathbf{K}_{\mathbf{H}_i}^\beta \mathbf{x}) \right) + \Gamma_{\text{TV}}(\mathbf{x}) \quad (7)$$

subject to $0 \leq \mathbf{x}$

which is a linear inverse problem with $p(\mathbf{b}_i | \mathbf{W}_{\mathbf{H}_i} \mathbf{K}_{\mathbf{H}_i}^\beta \mathbf{x})$ as the likelihood of observing a measurements \mathbf{b}_i given a specular reflectance \mathbf{x} on the plane, and $\Gamma_{\text{TV}}(\mathbf{x})$ as total variation (TV) [51] prior on the specular reflectance itself. We follow [12, 38] and assume a Poissonian-Gaussian likelihood term p and adopt their variant of the Alternating Direction Method of Multipliers (ADMM) [6] to solve the resulting linear inverse problem. While the solver method is established and described in detail in the Supplemental Material, our linear forward operator $\mathbf{W}_{\mathbf{H}_i} \mathbf{K}_{\mathbf{H}_i}^\beta$ is the main difference in the proposed approach. This forward operator consists of the spatially varying convolution matrix \mathbf{K} which blurs the specular reflectance based on the angular falloff β and distance to the wall (encoded by the homography), and a subsequent warping matrix \mathbf{W} which warps plane coordinates to the image plane and resamples the blurred specular reflectance coefficients using bi-linear interpolation. Although the proposed method is general, we assume a Gaussian spectral falloff with known standard deviation β . Note that the warping and falloff operators are only linear once the plane geometry is known.

5. Learning Inverse Indirect Illumination

In this section, we describe a trainable network and loss functions that allow us to recover occluded objects with arbitrary shape from intensity images. In contrast to planar geometries with only three DOF, arbitrary objects can have complex shapes with orders of magnitudes more parameters, i.e. for non-parametric surfaces three DOF per surface patch. To tackle this reconstruction challenge, we rely on

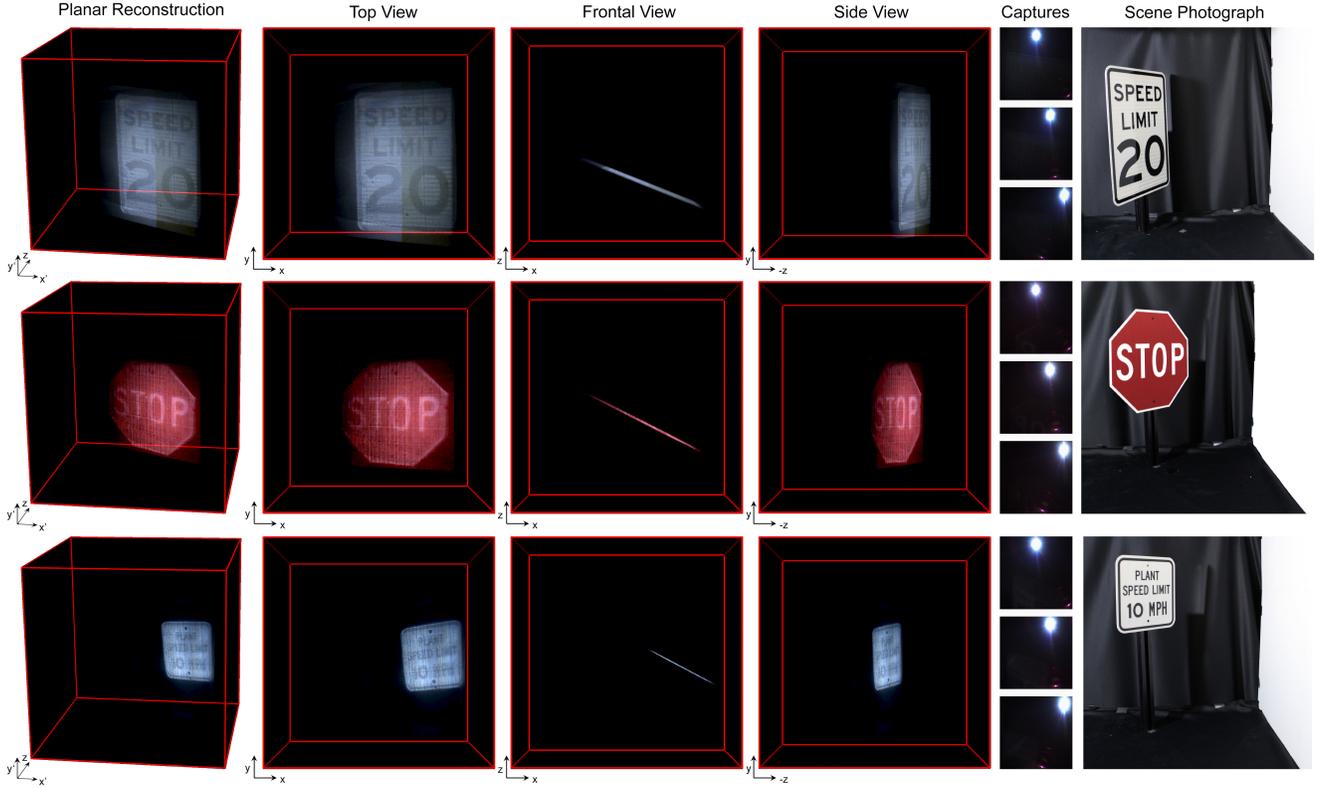


Figure 3: Experimental geometry and albedo reconstructions for the special case of planar objects, captured with the prototype from Sec. 7.2 and in setup geometry from [41]. We demonstrate reconstructions for three different surface materials. The first row shows an object with diamond grade retroreflective surface coating as they are found on number plates and high-quality street signs, identical to the objects in [41], which surprisingly contain faint specular components visible in the measurements (please zoom into electronic version of this document). The second and third row show a conventionally painted road sign and an engineering-grade street sign. The proposed method runs at around two seconds including capture and reconstruction and achieves high resolution results without temporal sampling.

strong priors on scene semantics which recent deep neural networks have been shown to encode efficiently [4, 48].

Input and Latent Parametrization In the proposed capture setup, we project individual beams of light on the diffuse wall and capture the intensity image that is formed by steady-state global illumination in the scene. In general, projecting light beams to different positions on the wall results in different observations which we dub indirect reflection maps, i.e. indirect component of the image on the wall without the direct reflection. Each map contains information about the object shape and normal information in specific direction if the BRDF is angle-dependent. Note that this is not only the case for highly specular BRDFs, but also for lambertian BRDFs due to foreshortening and varying albedo. Hence, by changing the beam position we acquire variational information about shape and reflectance.

Assuming locally smooth object surfaces, we sample the available wall area uniformly in a 5×5 grid and acquire multiple indirect reflection maps. We stack all the captured images, forming a $h \times w \times (5 \cdot 5 \cdot 3)$ dimension tensor as network input. The virtual source position is a further important information that may be provided to the network. However, since we use uniform deterministic sampling, we

found that the model learns this structured information, in contrast to random source sampling.

We use the orthogonal view of the scene as our ground truth latent variable, as if the camera had been placed in the center of the visible wall in wall normal direction and with ambient illumination present. Given the stack of indirect reflection maps, the proposed network is trained to estimate the corresponding orthogonal view into the hidden scene.

Network Architecture We propose a variant of the U-Net architecture [48] as our network backbone structure, shown in Fig. 4. It contains a 8 layers encoder and decoder. Each encoder layer reduces the image size by a factor of two in each dimension and doubles the feature channel. This scaling is repeated until we retrieve a 1024 dimension latent vector. In corresponding convolution and deconvolution layer pairs with the same size, we concatenate them to learn residual information.

Loss functions We use a multi-scale ℓ_2 loss function

$$V_{multi-scale} = \sum_k \gamma_k \|\mathbf{i}^k - \mathbf{o}^k\|^2, \quad (8)$$

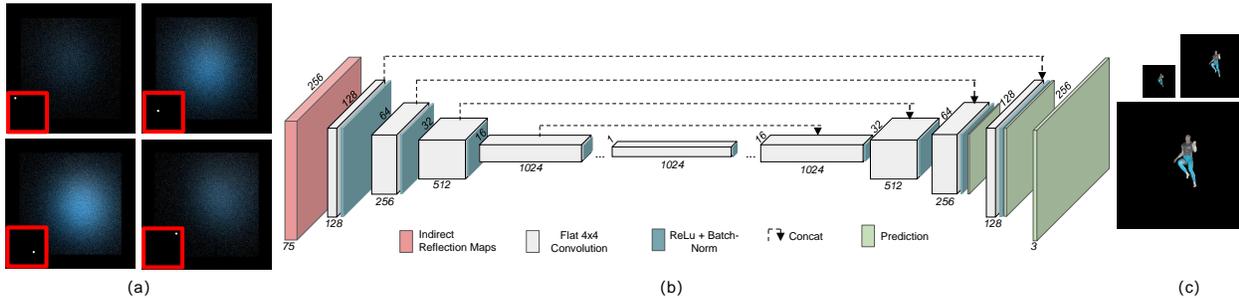


Figure 4: Learning NLOS imaging for arbitrary scene geometries and reflectance. We propose an encoder-decoder architecture that takes as input a stack of 5×5 synthetic indirect reflection measurements uniformly sampled on the diffuse wall. Here we show 4 examples of these 25 indirect reflection maps. The inset indicates the projected light beam position. The network outputs an orthogonal projection of the unknown scene as latent parametrization (c). We use a variant of the U-Net [48] architecture (b) which predicts these projections at three scales.

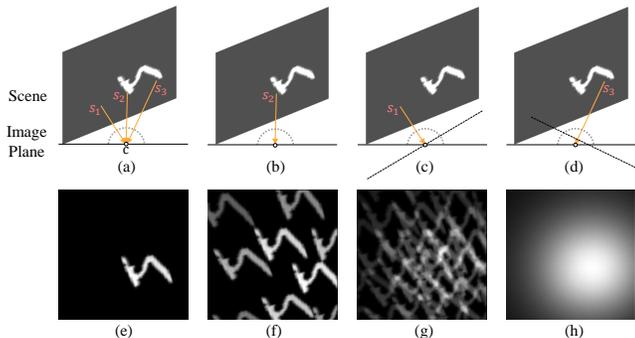


Figure 5: Training set rendering. Top row: (a) The radiance at a pixel c is the integral over the hemisphere of incoming indirect radiance, see Eq. (1). We uniformly sample the unit hemisphere (b-d), render each sample direction s_i using direct rendering with orthographical projection, and accumulate the sample views. Bottom row: (e) Diffuse character with albedo sampled from MNIST. (f), (g) and (h) 5×5 , 10×10 , and 100×100 directional samples.

where \mathbf{i} is the predicted network output and \mathbf{o} is the ground-truth orthogonal image. Here, k represents different scales and γ_k is the corresponding weight of that layer. With feature map at k -the layer, we adopt an extra one deconvolution layer to convert the feature to an estimate at the target resolution. We predict 64×64 , 128×128 and 256×256 ground truth images and set the weights γ_k as 0.6, 0.8 and 1.0. See the Supplemental Material for training details.

6. Training Datasets

The proposed deep model requires a large training set to represent objects with arbitrary shapes, orientations, locations and reflectance. While localization tasks for known objects [8, 9] may only require small datasets, that could be acquired experimentally, handling unknowns scenes requires sampling a large combinatorial space, which we tackle by synthetic data generation. Although more practical than experimental acquisition, ray-traced rendering of indirect reflections still requires minutes per scene [23, 22], and, for 5 min per measurement [22], a training set of 100,000 images would require one year of render time.

Instead, we propose a novel pipeline for indirect third-bounce simulations using direct rendering. As shown in Fig. 5, a given wall pixel c integrates light over the hemisphere of incoming light directions, see Eq. (1). We sample the unit hemisphere to estimate this integral, rendering each sample direction using direct rendering with orthographical projection, and finally accumulating the sampled views. Hardware-accelerated OpenGL allows for microsecond render times for a single view direction. We synthesize a full third-bounce measurement in 0.1 seconds for 10000 hemisphere samples, which is more than $600\times$ faster than [22], see Supplement Material. We render the training data adopting the setup geometry from [41]. Fig. 6 shows examples for hidden objects from the following classes.

MNIST Digits A number of recent works capture character-shaped validation objects [54, 19, 7, 41]. We represent this object class by placing MNIST digits on a plane, with randomly sampled rotations and shifts. We also sample specular coefficients in $[0, 512]$ with a Phong [46] BRDF to represent different materials. We generate 20000 examples with albedo randomly sampled from MNIST.

ShapeNet Models We synthesis measurements from ShapeNet [57] to represent more complex objects. We select the representative classes 'car', 'pillow', 'laptop' and 'chair', and train models for each class. Each class contains hundreds of models, and we render 20000 examples with random location, orientation, and reflectance as above.

Human Models Finally, we synthesize data for human models with non-rigid, varying body shapes and poses. We sample these models from the SCAPE [2] dataset and implement realistic appearance variation using the clothing samples from [10]. We generate 18000 examples with location-orientation sampling as above.

7. Evaluation

In this section, we validate the proposed methods in simulation and using experimental measurements.

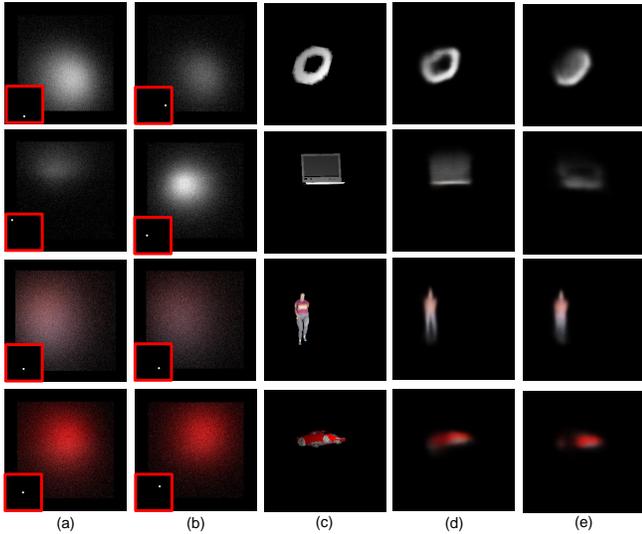


Figure 6: Qualitative results on synthetic data (including sensor noise simulation). (a)(b) We show two examples of 25 rendered indirect reflection maps. (c) Unknown scene (orthogonal projection). (d)(e) Reconstruction by 5×5 and 1×1 indirect reflection maps, respectively.

Beam Samples	MNIST	Car	Pillow	Laptop	Chair	Human
5×5	24.0	26.3	26.7	25.9	22.5	25.5
3×3	22.5	25.9	25.9	24.8	22.3	25.3
1×1	21.4	25.9	24.8	24.6	22.1	25.0

Table 1: Reconstruction performance PSNR [dB] for decreasing virtual source sampling and varying object class. While the reconstruction performance drops significantly when reducing the source sampling, it does not completely fail even for a single source position. In this case, the method does not provide accurate shape, but only rough location information, see Fig. 6.

7.1. Analysis

We first assess the proposed network model, which runs at 32 FPS reconstruction rates, on unseen synthetic data from MNIST digits, ShapeNet, and the Human dataset. Fig. 6 shows two examples of indirect reflection maps and their corresponding light positions. We simulate Poissian-Gaussian noise $\sigma = 0.05$ and $\kappa = 1/0.03$ according to Eq. (3), calibrated for our experimental setup. The qualitative results in Fig. 6 show that the proposed model can precisely localize objects and recover accurate reflectance for large objects. Although recovery for smaller objects with diffuse reflectance becomes an extremely challenging task, our method still recovers coarse object geometry and reflectance. Table 1 shows a quantitative evaluation of the reconstruction performance for different sampling patterns. While denser beam sampling on the wall results in higher recovery performance, even a single sampling position does provide enough information to perform recovery. However, Fig. 6 shows that fine geometry is lost in this case and only

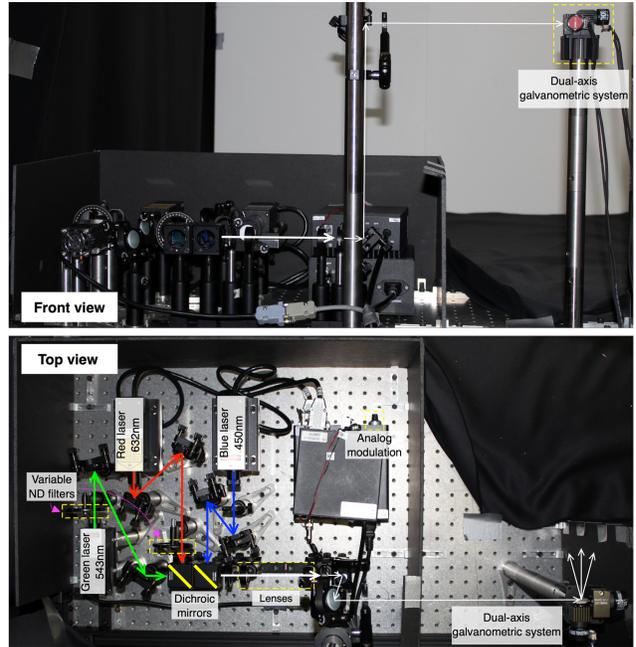


Figure 7: Prototype RGB laser source showing the path of the light, that ends in a dual-axis galvo. Top view: variable neutral density (ND) filters and one analog modulation is used to control the individual power of lasers. Long focal lenses combination and several irises reduce the beam diameter. Front view: only the white light path is shown. The camera is placed right next to the galvo.

rough location and shape can be recovered.

7.2. Experimental Setup

To illuminate a hidden scene, we require a high-power, steerable and narrow white light source shown in Fig. 7. Unfortunately, these goals are challenging to achieve with wide sources like diodes (following from Gaussian beam considerations). We built a high-intensity collimated RGB light beam with three continuous 200 mW lasers of different wavelengths collinearized with two dichroic mirrors. Inexpensive modules with similar specifications are also available in a very small footprint as so-called RGB lasers. The wavelengths are chosen to correspond roughly to the maximums of the three-color sensor of our high-quantum-efficiency Allied Vision Prosilica GT 1930C camera (outside to the right of the Fig.). White balance is achieved by adjusting the power of each laser. Through the combination of several irises and lenses, the diameter of the beam is reduced to less than 5 mm, which is fed into a Thorlabs GVS102 dual-axis galvo. For each laser spot we acquire a single 50 ms exposure, leading to a full capture time of around 1.25 s for 5×5 spots. Please see the Supplemental Material for additional details.

7.3. Experimental Validation.

Fig. 3 shows three planar reconstruction examples acquired using the described experimental setup. Surprising-

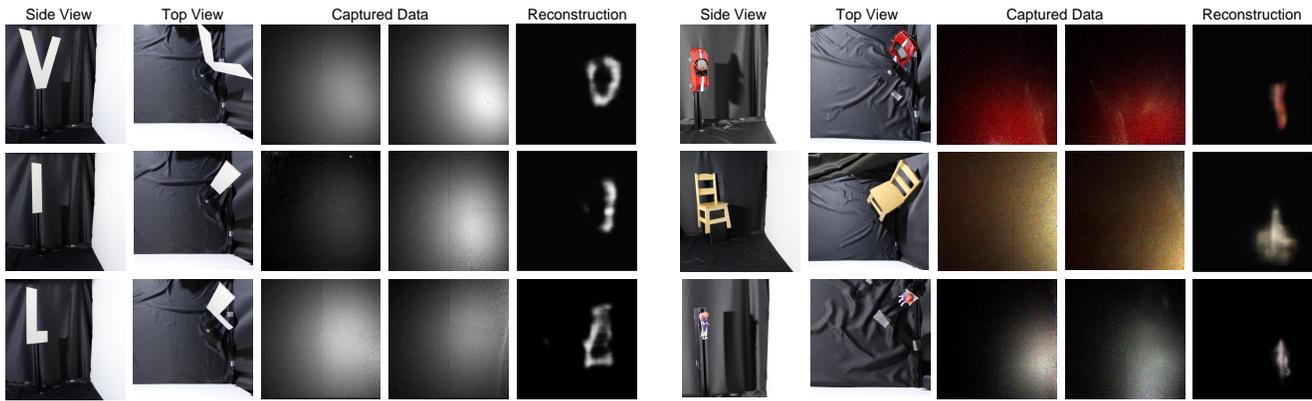


Figure 8: Non-line-of-sight reconstructions (orthogonal projections) from the proposed learned model for the setup geometry from [41]. We demonstrate that the model can recover projected location on the wall and reflectance of fully diffuse scene objects such as the three diffuse letters on the left side. The scene objects are cutouts from diffuse foam-board material. The measurements have been contrast-normalized and the light position has been cropped to visualize the very low-signal diffuse component. The reconstructions on the right show captured scenes of more significantly smaller, more complex objects. For these challenging cases, the proposed method recovers rough shape, reflectance and projected position.

ly, even high-grade retroreflective surface coatings as they are found on number plates and high-quality street signs (also identical to the objects in [41]), contain faint specular components visible in the measurements. The proposed optimization method achieves high quality reflectance and geometry recovery at almost interactive reconstruction times of about one second. While the dominating retroreflective component is returned to the virtual source position, a residual diffuse component is still present and appears as a halo-like artefact in the reconstruction. This diffuse halo is more apparent for the painted road sign, but still does not prohibit us to recover high-quality geometry and reflectance for these planar objects without temporal coding.

Fig. 8 shows reconstruction results for diffuse objects without requiring planar scene geometry. We demonstrate that the learned network, trained entirely on synthetic data, generalizes to experimental captures. The character-shaped objects on the left are cutouts of diffuse white foam boards, comparable to the validation scenes used in transient method [19, 7, 41]. The proposed data-driven method accurately recovers shape and location on the wall for these large characters from their diffuse indirect reflections. For the smaller complex objects on the right, the diffuse reflections are substantially dimmer, but the proposed approach still recovers rough shape and position. Note that the mannequin figurine is only 3 cm in width and still recovered by the learned reconstruction method.

Finally, we demonstrate that the proposed method can also be applied to depth recovery of non-planar occluded scenes. Using the same network architecture as before, but with depth maps now as labels, the resulting model can recover reasonable object depth, which we validate in Fig. 9 for both synthetic and experimental measurements.

8. Conclusion

We have demonstrated that it is possible to image objects outside of the direct line-of-sight using conventional RG-

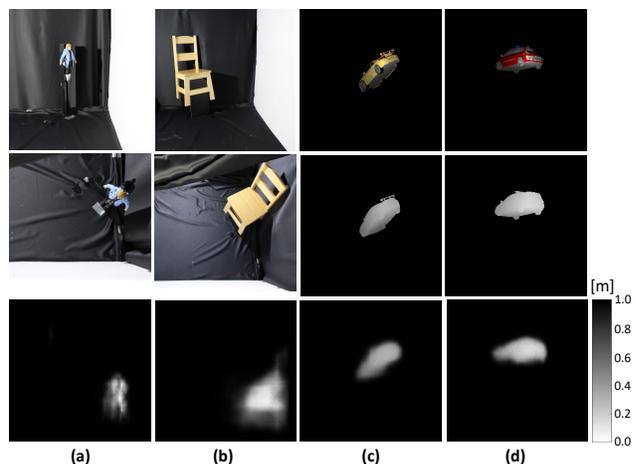


Figure 9: Learning Hidden Geometry Reconstruction. The proposed method allows not only for albedo recovery, but also geometry reconstruction from steady-state indirect reflections. In (a,b), we show depth recovery of hidden scenes. From top to bottom: side and top view photographs of the hidden scene, and reconstructed depth map in meters. In (c,d) we show simulated results. From top to bottom: orthogonal view of hidden scene, ground truth depth of hidden scene, and reconstructed depth.

B cameras and continuous illumination, without temporal coding. Relying on spatial variations in indirect reflections, we show high-quality NLOS geometry and reflectance recovery for planar scenes and a learned architecture which handles the shape-dependence of these variations. We have validated the proposed steady-state NLOS imaging method in simulation and experimentally.

Promising directions for future research include the reduction of the laser spot size using single-model laser systems and achromatic optics, and mechanical occlusion before the lens system to discard the strong direct component along with lens flare. A further exciting opportunity are multiple inter-reflections in the hidden scene which may al-

low to conceptually turn every scene surface into a sensor. Relying on consumer color image sensors in our prototype system, the proposed method makes a first step towards this vision by achieving full-color non-line-of-sight imaging at fast imaging rates and in scenarios identical to those targeted by recent pulsed systems with picosecond resolution.

9. Acknowledgements

The authors thank Colin Brosseau for assisting with the experiments and contributing in many fruitful discussions.

References

- [1] N. Abramson. Light-in-flight recording by holography. *Optics Letters*, 3(4):121–123, 1978. 1, 2
- [2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 6
- [3] V. Arellano, D. Gutierrez, and A. Jarabo. Fast back-projection for non-line of sight reconstruction. *Optics Express*, 25(10):11574–11583, 2017. 1, 2
- [4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 5
- [5] K. L. Bouman, V. Ye, A. B. Yedidia, F. Durand, G. W. Wornell, A. Torralba, and W. T. Freeman. Turning corners into cameras: Principles and methods. In *International Conference on Computer Vision*, volume 1, page 8, 2017. 3
- [6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found.&Trends® in Machine Learning*, 3(1):1–122, 2011. 4
- [7] M. Buttafava, J. Zeman, A. Tosi, K. Eliceiri, and A. Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–21011, 2015. 1, 2, 6, 8
- [8] P. Caramazza, A. Boccolini, D. Buschek, M. Hullin, C. F. Higham, R. Henderson, R. Murray-Smith, and D. Faccio. Neural network identification of people hidden from view with a single-pixel, single-photon detector. *Scientific reports*, 8(1):11945, 2018. 3, 6
- [9] S. Chan, R. E. Warburton, G. Garipey, J. Leach, and D. Faccio. Non-line-of-sight tracking of people at long range. *Optics express*, 25(9):10109–10117, 2017. 3, 6
- [10] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV)*, 2016. 6
- [11] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1
- [12] M. A. Figueiredo and J. M. Bioucas-Dias. Restoration of poissonian images using alternating direction optimization. *IEEE transactions on Image Processing*, 19(12):3133–3145, 2010. 4
- [13] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4
- [14] A. Foi. Clipped noisy images: Heteroskedastic modeling and practical denoising. *Signal Processing*, 89(12):2609–2629, 2009. 3
- [15] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Trans. Image Process.*, 17(10):1737–1754, 2008. 3
- [16] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 1
- [17] O. Gupta, T. Willwacher, A. Velten, A. Veeraraghavan, and R. Raskar. Reconstruction of hidden 3d shapes using diffuse reflections. *Opt. Express*, 20(17):19096–19108, Aug 2012. 1, 2
- [18] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1
- [19] F. Heide, M. B. Hullin, J. Gregson, and W. Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics (ToG)*, 32(4):45, 2013. 1, 2, 6, 8
- [20] F. Heide, M. O’Toole, K. Zhang, D. Lindell, S. Diamond, and G. Wetzstein. Robust non-line-of-sight imaging with single photon detectors. *arXiv preprint arXiv:1711.07134*, 2017. 2
- [21] F. Heide, L. Xiao, W. Heidrich, and M. 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, 2
- [22] W. Jakob. Mitsuba renderer, 2010. <http://www.mitsuba-renderer.org>. 6
- [23] A. Jarabo, J. Marco, A. Munoz, R. Buisan, W. Jarosz, and D. Gutierrez. A framework for transient rendering. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 33(6), nov 2014. 6
- [24] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):167, 2013. 1
- [25] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics (ToG)*, 32(6):167, 2013. 1, 2
- [26] A. Kadambi, H. Zhao, B. Shi, and R. Raskar. Occluded imaging with time-of-flight sensors. *ACM Transactions on Graphics (ToG)*, 35(2):15, 2016. 1, 2
- [27] J. T. Kajiya. The rendering equation. In *Proc. SIGGRAPH*, pages 143–150, 1986. 3
- [28] O. Katz, P. Heidmann, M. Fink, and S. Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics*, 8(10):784, 2014. 2
- [29] O. Katz, E. Small, and Y. Silberberg. Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nature photonics*, 6(8):549–553, 2012. 2
- [30] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar. Looking around the corner using transient imaging. In *Proc. ICCV*, pages 159–166, 2009. 1, 2
- [31] J. Klein, C. Peters, J. Martín, M. Laurenzis, and M. B. Hullin. Tracking objects outside the line of sight using 2d intensity images. *Scientific reports*, 6:32491, 2016. 3
- [32] R. Lange. 3d time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology. 2000. 1, 2

- [33] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. 4
- [34] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 4
- [35] Y. Maruyama and E. Charbon. A time-gated 128x128 cmos spad array for on-chip fluorescence detection. In *Proc. Intl. Image Sensor Workshop (IISW)*, number EPFL-CONF-178145, 2011. 2
- [36] N. Naik, S. Zhao, A. Velten, R. Raskar, and K. Bala. Single view reflectance capture using multiplexed scattering and time-of-flight imaging. *ACM Trans. Graph.*, 30(6):171, 2011. 1, 2
- [37] F. Nolet, S. Parent, N. Roy, M.-O. Mercier, S. Charlebois, R. Fontaine, and J.-F. Pratte. Quenching circuit and spad integrated in cmos 65 nm with 7.8 ps fwhm single photon timing resolution. *Instruments*, 2(4):19, 2018. 2
- [38] M. O’Toole, F. Heide, D. B. Lindell, K. Zang, S. Diamond, and G. Wetzstein. Reconstructing transient images from single-photon sensors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2289–2297, July 2017. 2, 4
- [39] M. O’Toole, F. Heide, L. Xiao, M. B. Hullin, W. Heidrich, and K. N. Kutulakos. Temporal frequency probing for 5d transient analysis of global light transport. *ACM Transactions on Graphics (ToG)*, 33(4):87, 2014. 1
- [40] M. O’Toole, D. B. Lindell, and G. Wetzstein. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature*, 555(7696):338, 2018. 1, 2
- [41] M. O’Toole, D. B. Lindell, and G. Wetzstein. Confocal non-line-of-sight imaging based on the light cone transform. *Nature*, pages 338–341, 2018. 2, 5, 6, 8
- [42] M. O’Toole, R. Raskar, and K. N. Kutulakos. Primal-dual coding to probe light transport. *ACM Trans. Graph.*, 31(4):39–1, 2012. 1
- [43] R. Pandharkar, A. Velten, A. Bardagjy, E. Lawson, M. Bawendi, and R. Raskar. Estimating motion and size of moving non-line-of-sight objects in cluttered environments. In *Proc. CVPR*, pages 265–272, 2011. 1, 2
- [44] L. Parmesan, N. A. Dutton, N. J. Calder, A. J. Holmes, L. A. Grant, and R. K. Henderson. A 9.8 μm sample and hold time to amplitude converter cmos spad pixel. In *Solid State Device Research Conference (ESSDERC), 2014 44th European*, pages 290–293. IEEE, 2014. 2
- [45] A. K. Pediredla, M. Buttafava, A. Tosi, O. Cossairt, and A. Veeraraghavan. Reconstructing rooms using photon echoes: A plane based model and reconstruction algorithm for looking around the corner. In *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2017. 2
- [46] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 3, 6
- [47] R. Raskar and J. Tumblin. *Computational Photography: Mastering New Techniques For Lenses, Lighting, and Sensors*. A K Peters, Limited, 2007. 1
- [48] O. Ronneberger, P. Fischer, and T. 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. 5, 6
- [49] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006. 1
- [50] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003. 1, 3
- [51] D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappra, V. K. Goyal, F. N. Wong, and J. H. Shapiro. Photon-efficient imaging with a single-photon camera. *Nature communications*, 7:12046, 2016. 4
- [52] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *The IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1
- [53] C.-Y. Tsai, K. N. Kutulakos, S. G. Narasimhan, and A. C. Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2
- [54] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. Bawendi, and R. Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications*, 3:745, 2012. 1, 2, 6
- [55] A. Velten, D. Wu, A. Jarabo, B. Masia, C. Barsi, C. Joshi, E. Lawson, M. Bawendi, D. Gutierrez, and R. Raskar. Femto-photography: Capturing and visualizing the propagation of light. *ACM Trans. Graph.*, 32, 2013. 2
- [56] D. Wu, M. O’Toole, A. Velten, A. Agrawal, and R. Raskar. Decomposing global light transport using time of flight imaging. In *Proc. CVPR*, pages 366–373, 2012. 2
- [57] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 6
- [58] F. Xu, G. Shulkind, C. Thrampoulidis, J. H. Shapiro, A. Torralba, F. N. C. Wong, and G. W. Wornell. Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *OSA Opt. Express*, 26(8):9945–9962, 2018. 2