

Non-line-of-Sight Imaging via Neural Transient Fields

Siyuan Shen*, Zi Wang*, Ping Liu, Zhengqing Pan, Ruiqian Li, Tian Gao, Shiyong Li, and Jingyi Yu

Abstract—We present a neural modeling framework for Non-Line-of-Sight (NLOS) imaging. Previous solutions have sought to explicitly recover the 3D geometry (e.g., as point clouds) or voxel density (e.g., within a pre-defined volume) of the hidden scene. In contrast, inspired by the recent Neural Radiance Field (NeRF) approach, we use a multi-layer perceptron (MLP) to represent the neural transient field or NeTF. However, NeTF measures the transient over spherical wavefronts rather than the radiance along lines. We therefore formulate a spherical volume NeTF reconstruction pipeline, applicable to both confocal and non-confocal setups. Compared with NeRF, NeTF samples a much sparser set of viewpoints (scanning spots) and the sampling is highly uneven. We thus introduce a Monte Carlo technique to improve the robustness in the reconstruction. Comprehensive experiments on synthetic and real datasets demonstrate NeTF provides higher quality reconstruction and preserves fine details largely missing in the state-of-the-art.

Index Terms—Non-Line-of-Sight (NLOS) Imaging, Neural Radiance Field, Neural Rendering

1 INTRODUCTION

NON-line-of-sight (NLOS) imaging employs time-resolved measurements for recovering hidden scenes beyond the direct line of sight from a sensor [17], [37]. Applications are numerous, ranging from remote sensing to autonomous driving and to rescue missions in hazardous environments. Most existing NLOS setups orient an ultra-fast pulsed laser beam towards a relay wall in the line of sight where the wall diffuses the laser into spherical wavefronts towards the hidden scene. As the wavefront hits the scene and bounces back onto the wall, a time-of-flight (ToF) detector with a picosecond resolution (such as the streak camera [1], [2], [3] or recently the more affordable single-photon avalanche diodes (SPADs) [9], [15], [16], [31], [32], [34]) can be used to record the arrival time and the number of the returning photons [5], [6], [7]. SPAD sensors in a time correlated single photon counting (TCSPC) mode can thus produce transients, of which a single pixel corresponds to a specific pair of illumination and detection spots on the wall and a histogram of the photon counts versus time bins.

The measured transients contain rich geometric informa-

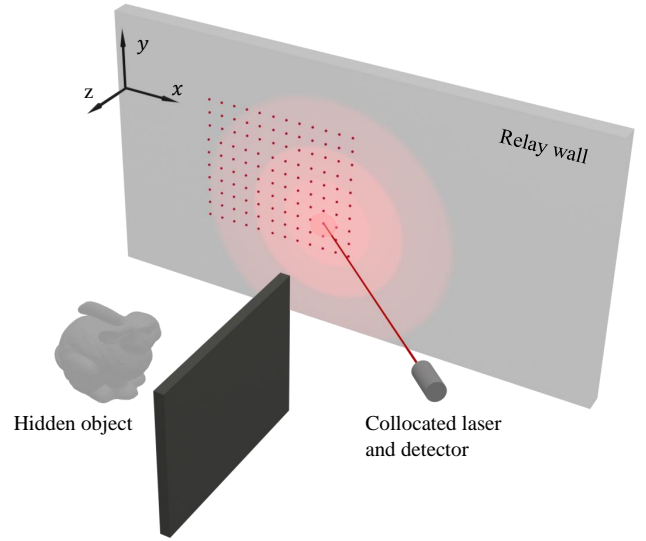


Fig. 1. A typical confocal NLOS imaging system aims a laser towards a diffuse wall that serves as a virtual reflector. The hidden scene is indirectly illuminated as spherical waves intersect with the scene and are reflected back onto the wall. A SPAD sensor measures at different spots on the wall to form transient images.

tion of a hidden scene, potentially usable for scene recovery. In reality, the process corresponds to a typical inverse imaging problem that generally incurs high computational cost, especially because the transients are high dimensional signals. To make the problem tractable, the pioneering back-projection (BP) technique and its variations assume smooth objects so that scene recovery can be modeled as deconvolution [1], [2], [8], [31]. Alternatively, the light-cone transform (LCT) based methods collocate the illumination and sensing spots on the relay wall so that the forward imaging model can be simplified as 3D convolution [4], [15], [34], [35] where

- * indicates equal contribution.
- S. Shen*, P. Liu, R. Li, T. Gao, S. Li and J. Yu are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China. E-mail: shensy, liuping, lirq1, gaotian, lishiyong, yujingyi@shanghaitech.edu.cn.
- Z. Wang* is with the Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China, and the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: wangzi@shanghaitech.edu.cn.
- Z. Pan is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: panzhq@shanghaitech.edu.cn.

advanced signal processing techniques such as Wiener filters [34], [35], [38] can further reduce noise. Assuming the scene is near diffuse, analysis-by-synthesis algorithms can improve reconstruction [18], [36]. The seminal work of the Fermat path based approaches [16], [32] can handle highly specular objects by simultaneously recovering the position and normal of Fermat points on the surface.

We present a novel volumetric NLOS imaging framework by modeling the transient field via deep networks. Our Neural Transient Field (NeTF) technique is inspired by the recent neural radiance field (NeRF) that conducts 3D reconstruction and view synthesis from a set of input images. Different from existing multi-view stereo (MVS) techniques, NeRF assumes a volume rendering model and sets out to use a multi-layer perception (MLP) to recover per-voxel scene density and per-direction color. We observe that NLOS resembles MVS in that each scanning point on the wall resembles a virtual camera and therefore a similar deep learning technique may be potentially used for scene recovery. Different from NeRF though, NeTF measures the transient over spherical wavefronts rather than the radiance along lines. We therefore first formulate volumetric transient fields under the spherical coordinate and devise an MLP that trains on the measurements to predict per-voxel density and view-dependent albedo. Our NeTF formulation is applicable to both confocal and non-confocal setups.

Compared with NeRF, NeTF captures a much sparser set of viewpoints (scanning spots) and the distribution of scene points on the spherical wavefronts can be highly uneven. We therefore develop a Markov chain Monte Carlo (MCMC) technique based on importance sampling for matching the actual scene distribution. We conduct comprehensive experiments on existing synthetic and real datasets. We demonstrate that NeTF provides higher quality reconstruction and preserves fine details largely missing in the state-of-the-art. In particular, the trained MLP provides a continuous 5D representation of the hidden scene without requiring digitizing the NLOS volume or optimizing surface parameters. The 5D representation with directional encoding can handle view-dependent albedo of non-Lambertian surface reflectance and strong self-occlusions, under both confocal or non-confocal setups.

2 RELATED WORK

As an emerging computational imaging technique, NLOS imaging has found broad applications in computer vision and computer graphics, ranging from recovering 3D shape of hidden objects [2], [14], [15], [31], [32], [34], [35] to tracking hidden moving objects [2], [4], [10], [13]. Existing solutions employ time-resolved optical detectors such as streak cameras [1], [2], SPADs [5], [31] and interferometry [12], [16] or non-optical acoustic [11] and thermal [10] sensors to indirectly measure the hidden scene and then apply inverse imaging techniques for recovery. We refer the reviewers to recent surveys [17], [21], [37] for a comprehensive overview.

Confocal vs. Non-Confocal. Kirmani et al. [19], [20] designed and implemented the first prototype non-confocal NLOS system and derived a linear time-invariant model amenable for multi-path light transport analysis. In reality, varying both the laser beam and the measuring spot yield

to a high-dimensional transient field analogous to the light field. Many efforts have since been focused on imposing priors and constraints to accelerate data processing. Velten et al. [2] proposed a back-projection technique with ellipsoidal constraints: the observing point and the laser projection point on the wall correspond to the foci of a set of ellipsoids, each corresponding to a specific transient. The hidden scene can then be reconstructed by intersecting the ellipsoids. To further improve reconstruction quality and speed, subsequent works have applied filtering techniques such as sharpening and thresholding [1], [8], [31]. Alternatively, one can directly model the scene using parametric surfaces and then optimize the parameters over the observations [18], [36], [38]. Ahn et al. [38] models parameter fitting as a linear least-square problem using a convolutional Gram operator. It is also possible to adopt wave optics for NLOS imaging [14], [15], [23], by characterize the problem as specific properties of a temporally evolving wave field in the Fourier domain. To reduce data dimensionality, several recent approaches adopt a confocal setting [34] where the laser and the detector (e.g., a SPAD) collocate, e.g., via a beam splitter. Consequently, the ellipsoidal constraints degenerates to be spherical, simplifying the inverse problem with a 3D deconvolution and system calibration. The seminal work of light-cone transform (LCT) [34] casts the NLOS reconstruction problem as Wiener filtering in the Fourier domain and can achieve a low computational complexity of $O(N^3 \log N)$ for N^3 voxels, compared to $O(N^5)$ in the traditional BP methods. Yong et al. [35] formulate the albedo and normal recovery based on directional LCT (DLCT) as a vector deconvolution problem. The confocal setting results in an overwhelming contribution of direct light first-bounce off the diffuse wall and subsequently produces useful geometric constraints. The seminal work by Xin et al. [16] exploits the Fermat flow induced by the transients for estimating surface normals. Lindell et al. [15] adapt an f-k migration in seismology to convert the surface reconstruction problem to a boundary value problem. The f-k migration method enables faster reconstruction and supports planar or nonplanar diffuse walls.

Volume vs Surface. Existing NLOS methods can also be categorized in terms of the form of the reconstruction results. Two most adopted forms are volume density and points/surfaces. Methods for recovering the former generally discretize the scene into voxels and compute the density, either using intersections of wavefronts under ellipsoidal [1], [2], [5], [8], [29], [31], [38] and spherical [34], [35] constraints, or via modeling the imaging process as convolution and recovering the volume via specially designed deconvolution filters. Methods for recovering the latter have relied on light transport physics [18], [36] for optimizing the shape and reflectance of the hidden scenes. Such methods are generally mathematically tractable but are computationally expensive particularly because higher order geometry such as the surface normal needs to be integrated into the optimization process. Reconstruction results are either sparse as in Fermat [16] where only discontinuities in the transient were used, or rely heavily on the quality of the basis shape as in [18].

Our approach falls into the category of volume based technique. We are inspired by the recent multi-view re-

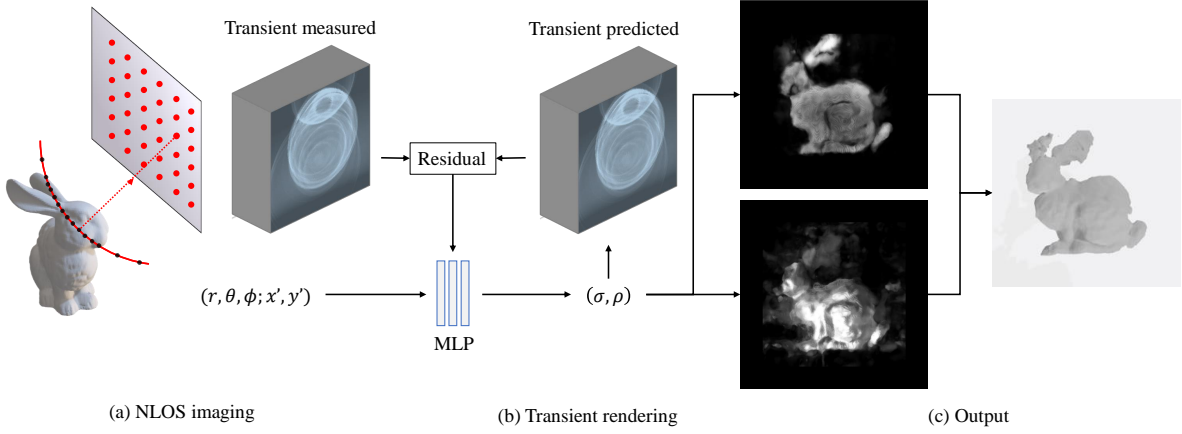


Fig. 2. Our neural transient field (NeTF) reconstruction pipeline. We parameterize every point on a spherical wavefront in terms of the origin on the wall, and the direction and radius of its corresponding spherical coordinates. We set out to recover the transient field under this parameterization via a multi-layer perception under the spherical volume rendering modeling.

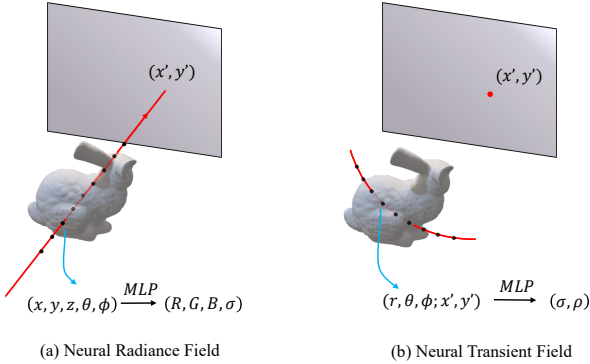


Fig. 3. NeRF vs. NeTF. In NeRF, volume density is accumulated along every line (ray) whereas in NeTF it is accumulated on a spherical wavefront.

construction framework Neural Radiance Field (NeRF) that aims to recover the density and color at every point along every ray, implicitly providing a volumetric reconstruction. NeRF adopts a volume rendering model and sets out to optimize volume density that best matches the observation using a Multi-Layer Perception (MLP). It is also possible to modify NeRF to tackle photometric stereo (PS) problems where the camera is fixed but the lighting conditions vary. We observe the non-confocal NLOS imaging process greatly resembles MVS/PS: fixing the laser beam but measuring transient at different spots on the wall resembles MVS where as fixing the measuring spot but varying the laser beam resembles PS. In fact, the confocal setting is very much similar to the NeRF AA setting [41] where the lighting and the camera move consistently. We therefore call our reconstruction scheme Neural Transient Field or NeTF.

Both NeTF and NeRF use MLP as an optimizer. However, there are several major differences between NeRF and NeTF. First, the volume rendering model used in NeRF is not directly applicable to NeTF. We therefore derive a novel volumetric image formation model under NLOS. Second, NLOS measures the transient rather than the radiance. In fact, the transient is measured from the sum of returning

photons on a wavefront instead of a single ray. We hence formulate a spherical volume reconstruction pipeline. Finally, NeRF generally assumes dense ray samples whereas the NLOS setting is much sparse. We thus introduce a Monte Carlo technique to improve the robustness in the reconstruction.

3 NEURAL TRANSIENT FIELD

We recognize that the NLOS reconstruction problem resembles multi-view reconstruction in the line-of-sight (LOS) and adopt a neural reconstruction framework analogous to NeRF [39]. Each detection spot on the relay wall can be viewed as a virtual *camera*. These cameras capture the transients of the NLOS scene as if viewed from the wall. We adopt the plenoptic radiance field notion of NeRF and represent the NLOS scene as a continuous 5D function of transients, i.e., a plenoptic transient field. We then set out to infer scene density at every point along every spherical wavefront via deep network based optimization. It is important to note that, same as NeRF, our neural transient field (NeTR) representation chooses not to explicitly discretize the scene into volumes. Rather, we use a multi-layer perception (MLP) to virtually represent the volume.

3.1 Scene Representation

The NeRF framework [39] uses the neural radiance field $L(x, y, z, \theta, \phi)$ as scene representation where (x, y, z) corresponds to a point on a ray and (θ, ϕ) the direction of the ray. Their trained network outputs both the density σ at every position (x, y, z) and the (view-dependent) color $c = (r, g, b)$ along direction (θ, ϕ) . The density can be further used for scene reconstruction and the color for image-based rendering. In our case, NeTF, instead of sampling on a single camera ray, samples a hemisphere of rays as light propagates as a spherical wave from the relay wall towards the hidden scene. We hence adopt a continuous 5D function of transients L_{NLOS} under the spherical coordinates as:

$$L_{\text{NLOS}}(x', y', r, \theta, \phi) \rightarrow (\sigma, \rho)$$

where $P(x', y')$ is a detection spot on the wall that serves as the origin of the hemisphere. $Q(r, \theta, \phi)$ is a scene point parameterized using the spherical coordinate (r, θ, ϕ) w.r.t $P(x', y')$. Similar to NeRF though, we set out to design a fully connected neural network, i.e., an MLP, to estimate L_{NLOS} . Different from NeRF, L_{NLOS} in NeTF outputs a volume density σ and a surface reflectance (albedo ρ) rather than color along the direction (θ, ϕ) .

Recall that NLOS needs to scan different spots on the relay wall, resulting in inconsistent spherical coordinates and casting challenges in network training and inference. We thus first transform the spherical coordinates $(x', y', r, \theta, \phi)$ to their corresponding Cartesian coordinates, i.e., (x, y, z, θ, ϕ) as

$$\begin{cases} x = r \sin \theta \cos \phi + x' \\ y = r \sin \theta \sin \phi + y' \\ z = r \cos \theta \end{cases}$$

The transform R ensures that the position of a 3D voxel is consistent when we scan over different detection spots. All subsequent training under MLP should be conducted under the Cartesian coordinate for density and view dependent albedo inferences.

$$L_{\text{NLOS}} : (x', y', r, \theta, \phi) \xrightarrow{R} (x, y, z, \theta, \phi) \xrightarrow{\text{MLP}} (\sigma, \rho)$$

Same as NeRF, a key benefit of NeTF is that we no longer need to discretize the scene into a fixed-resolution volume representation. Instead, the deep network representation can provide scene reconstructions at an arbitrary resolution, recovering fine details largely missing in prior art.

3.2 Forward Model

We first reformulate the NLOS reconstruction problem as a forward model under our NeTF representation. Under the confocal setting [15], [27], [34], the illumination and detection collocate at a same spot $P(x', y')$ on a relay wall, producing a spherical wave anchored at the spot. The transient $\tau_{\text{iso}}(x', y', t)$ recorded at each spot $P(x', y')$ is the summation of photons that are reflected back at a specific time instant t from the NLOS scene in the hemisphere Ω with a radius $ct/2$ as:

$$\begin{aligned} \tau_{\text{iso}}(x', y', t) = & \iiint_{\Omega} \frac{1}{r^4} \rho_{\text{iso}}(x, y, z) g(x', y', x, y, z) \\ & \cdot \delta(2\sqrt{(x' - x)^2 + (y' - y)^2 + z^2} - tc) dx dy dz \end{aligned} \quad (1)$$

where c is the speed of light, r is the distance between the wall and the NLOS scene as $r = \sqrt{(x' - x)^2 + (y' - y)^2 + z^2} = tc/2$, and $\frac{1}{r^4}$ is the light fall-off term. $\rho_{\text{iso}}(x, y, z)$ is the albedo of an NLOS point, where x, y, z are the spatial coordinates of the point. Function g models the time-independent effects, including the surface normal, bidirectional reflectance distribution functions (BRDFs), occlusions patterns, etc. The Dirac delta function relates the time of flight t to the distance r .

Notice that function g makes the imaging process nonlinear. To make the problem tractable, previous linear approximation schemes such as [34] adopts $g = 1$ by assuming that

the NLOS scene scatters isotropically and that no occlusions occur within the NLOS scene. Such assumptions, however, restrict NLOS scenes to being Lambertian and convex. In contrast, NeTF, by adopting a deep network to model the imaging process, can tackle non-linearity without imposing explicit constraints on g .

We intend to specify how much an NLOS point in the hemisphere contributes to the transient through photons propagation. Consider a detection spot $P(x', y')$ to an NLOS point $Q(r, \theta, \phi)$ in a hemisphere centered at P . Recall that the scattering equation [30] serves as the foundation to volume rendering and thereby NeRF. We first derive a photon version of the scattering equation, with more details are provided in the supplementary materials.

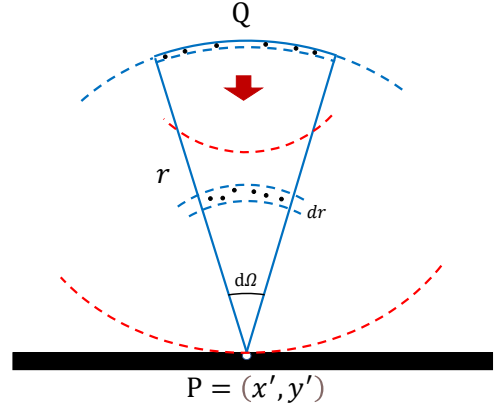


Fig. 4. The volume rendering model derived under the spherical coordinate system, suitable for processing in NeTF.

In our NLOS setting, photons travel in a hemisphere. We assume that the spot P is a patch with radius r_0 , and that the location Q in the hidden scene with its neighbours that contribute to P forms a spherical cross-section with radius r , thickness dr , and a solid angle $d\Omega$. Figure 4 shows that photons travel from P to Q and back from Q to P . When dr is sufficiently small, the inner and outer surface areas of the cross-section are $S = r^2 d\Omega$, and the volume is $S dr$. Recall that σ is the density of particles in the cross-section, therefore the number of particles is $\sigma S dr$. Assume a particle has a radius a , the projected area on the surface can then be computed as $A = \pi a^2$. We assume that light energy E is attenuated due to particles' absorption and scattering and consequently the energy loss dE can be computed as:

$$dE = -\frac{\pi a^2 \sigma r^2 d\Omega dr}{r^2 d\Omega} E = -A \sigma E dr \quad (2)$$

The Beer-Lambert law can be used to compute the attenuation coefficient as $e^{\int_0^r -A\sigma(r', \theta, \phi) dr'}$ along the radius r . Recall that the spot P has a radius r_0 and emits radiant energy as a constant E_P . Take integral of Equation 2, energy received at Q is defined as:

$$E_Q = e^{\int_0^r -A\sigma(r', \theta, \phi) dr'} E_P \frac{r^2 d\Omega}{r^2 2\pi} \quad (3)$$

We now consider the reflection at Q . Assume that the cross-section is sufficiently thin, e.g., $dr = 2a$, the radiant

energy at Q attenuated due to absorption and reflection w.r.t the reflectance ρ can be defined as

$$E'_Q(r, \theta, \phi) = A \cdot \sigma(r, \theta, \phi) \cdot 2a \cdot \rho(r, \theta, \phi) \cdot E_Q \quad (4)$$

On the returning path from Q to P , the wavefronts form hemispheres centered at Q with radius r . The spot P with area πr_0^2 receives the photons back to the relay wall, we have the energy at P w.r.t the solid angle $d\Omega$ as:

$$E'_P(r, \theta, \phi) = e^{\int_r^0 -A\sigma(r', \theta, \phi) dr'} E'_Q \frac{r_0^2 \cdot 2\pi}{r^2 \cdot 2\pi} \quad (5)$$

By taking the integral of Equation 5 w.r.t the solid angle $d\Omega$, we obtain energy received at the detection $P(x', y')$ in the hemisphere at a time instant t , as:

$$\tau(x', y', t) = \iint_{H(x', y'; \frac{ct}{2})} E'_P(r, \theta, \phi) d\Omega \quad (6)$$

Equation 6 serves our forward imaging model. It essentially maps an NLOS point $Q(r, \theta, \phi)$ to a transient τ detected at a spot $P(x', y')$ on a diffuse surface at a time instant t . Substituting Equations 3, 4 and 5, Equation 6 is rewritten as:

$$\begin{aligned} \tau(x', y', t) &= \Gamma_0 \iint_{H(x', y'; \frac{ct}{2})} \frac{1}{r^2} \sigma(r, \theta, \phi; x', y') \\ &\quad \cdot \rho(r, \theta, \phi; x', y') e^{2 \int_{r_0}^r -A\sigma(r', \theta, \phi) dr'} d\Omega \\ &= \Gamma_0 \iint_{H(x', y'; \frac{ct}{2})} \frac{\sin \theta}{r^2} \sigma(r, \theta, \phi; x', y') \\ &\quad \cdot \rho(r, \theta, \phi; x', y') e^{2 \int_{r_0}^r -A\sigma(r', \theta, \phi) dr'} d\theta d\phi \end{aligned} \quad (7)$$

where $\Gamma_0 = Aar_0^2 E_P / \pi$ correspond to the constant terms of particle radius a , initial energy E_P , and patch radius r_0 . The integration domain $H(x', y'; \frac{ct}{2})$ is a hemisphere centered at $P(x', y')$ on a relay wall with a radius of $r = ct/2$. θ and ϕ are the angles of elevation and azimuth in the viewing direction from $P(x', y')$ to an NLOS point, equivalent to those in the direction of reflection from the NLOS scene. $\rho(r, \theta, \phi; x', y')$ models view-varying BRDFs of the NLOS scene. $e^{2 \int_{r_0}^r -A\sigma(r', \theta, \phi) dr'}$ is an exponential attenuation coefficient and reveals the visibility of an NLOS point with respect to varying detection spots $P(x', y')$.

Notice though that the forward plenoptic transient field model in Equation 7 has a high computational complexity unsuitable for direct training using the MLP. We can assume that the NLOS scene is all opaque and has no self-occlusions to simplify the equation as:

$$\tau(x', y', t) = \Gamma_0 \iint_{H(x', y'; \frac{ct}{2})} \frac{\sin \theta}{r^2} \sigma(r, \theta, \phi; x', y') \cdot \rho(r, \theta, \phi; x', y') d\theta d\phi \quad (8)$$

Both NeRF and our NeTF derive the forward model based on volume rendering. The difference though is that traditional volume rendering as used in NeRF approximates a ray propagates between two points as a cylinder whereas

NeTF accounts for the spherical wavefront propagation by using a cone model to characterize attenuation. In addition, the volume rendering model used in NeRF only considered one-way accumulation as light travels through light emitting particles towards the camera sensor. In contrast, our NeTF accounts for both how light illuminates the scene and how scenes illuminates the wall, a two-way propagation process.

3.3 Differentiable Rendering

Our forward model is differentiable, we can therefore numerically compute the continuous integral Equation 8 using quadrature, as:

$$\begin{aligned} \tau(x', y', t) &= \frac{\Delta\theta\Delta\phi}{r^2} \sum_{i,j} \sin \theta_{ij} \\ &\quad \cdot \sigma(r, \theta_{ij}, \phi_{ij}; x', y') \rho(r, \theta_{ij}, \phi_{ij}; x', y') \end{aligned} \quad (9)$$

where $Q(r, \theta_{ij}, \phi_{ij})$ are scene points uniformly sampled along the hemispherical rays. We transform these points into their corresponding Cartesian coordinates as inputs to the MLP. The network outputs the density and reflectance at each point. We then sum all the outputs as neural transient fields from the transients.

We optimize our NeTF by minimizing the following L_2 loss function as the difference between the predicted $\tau(x', y', t)$ and measured $\tau_m(x', y', t)$ transients:

$$L = \sum_{x', y', t} (\tau_m(x', y', t) - \tau(x', y', t))^2 \quad (10)$$

Notice that the use of MLP allows minimizing arbitrary losses as long as they are differentiable with respect to our prediction $\tau(x', y', t)$, although L_2 is most commonly adopted same as in NeRF. Gradient descent can be used to optimize the loss function where the gradients can be calculated as

$$\frac{\partial L}{\partial w} = \sum_{x', y', t} \frac{\partial L}{\partial \tau(x', y', t)} \frac{\partial \tau(x', y', t)}{\partial w} \quad (11)$$

where w represent the parameters in the network, and

$$\frac{\partial L}{\partial \tau(x', y', t)} = 2(\tau_m(x', y', t) - \tau(x', y', t)) \quad (12)$$

$$\begin{aligned} \frac{\partial \tau(x', y', t)}{\partial w} &= \frac{\Delta\theta\Delta\phi}{r^4} \sum_{i,j} \sin \theta_{ij} \\ &\quad \frac{\partial \sigma(r, \theta_{ij}, \phi_{ij}; x', y') \rho(r, \theta_{ij}, \phi_{ij}; x', y')}{\partial w} \end{aligned} \quad (13)$$

where $\frac{\partial \sigma \rho}{\partial w}$ presents the derivatives in the network.

4 NEURAL TRANSIENT FIELDS OPTIMIZATION

Our NeTF forward model allows modeling the plenoptic transient field using an MLP. However, data acquired by NeTF are quite different from the ones in NeRF. In NeRF, a dense set of high resolution images are generally required to produce satisfactory density estimation and view interpolation. Under the dense viewpoint setting, the problem

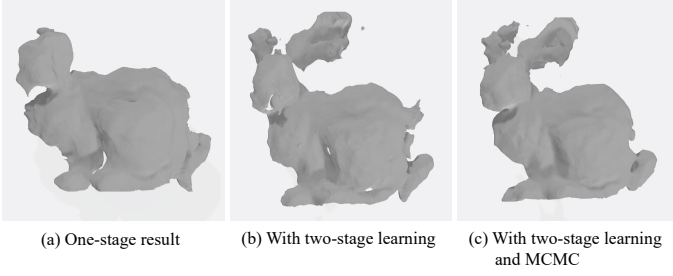


Fig. 5. From left to right: recovered results using one-stage training, two-stage training without and with MCMC. Two-stage significantly improves the reconstruction quality by recovering the ear of the Bunny, largely missing prior art. With MCMC, the back of the right ear can be further recovered whereas the noise on the back is suppressed.

of occlusions are less significant as there will be a sufficient number of views capturing the occluded point to ensure reliable reconstruction. In NeTF, however, our SPAD only captures a sparse set of spots on the wall and an occluded point may be only captured from a very small number of viewpoints (spots). Consequently occlusion can lead to strong reconstruction artifacts if not handled properly. We develop a two-stage learning strategy along with a hierarchical sampling technique to address this sampling bias.

4.1 Two-stage Learning

We observe the sampling bias resembles the long-tailed classification problem in machine learning. A conventional solution is to resample the dataset to achieve a more balanced distribution by over-sampling the minority classes [28]. We therefore adopt a two-stage learning strategy. We first conduct training with all transients to obtain an initial reconstruction. We then calculate the loss function between the predicted and measured transients and use the result as a probability density function (PDF). Next, we resample the detection positions according to the PDF and rebuild the training dataset of transients. We then retrain our network with the new dataset to refine the reconstruction result. We can iterate the two procedures until convergence.

The two-stage learning process provides a viable solution to tackle imbalanced sampling for achieving more accurate reconstruction. Figure 5 demonstrates their corresponding reconstruction results with and without using the second stage. Results with the second stage manages to recover many fine details largely missing from the first stage, e.g., the ear of the bunny and details on the abdomen regions.

4.2 Hierarchical Sampling

Denser samples produce higher quality reconstruction. At the same time, they lead to much higher computation overhead. For example, by uniformly sampling L hemispherical rays at each detection spot and N^2 scene points on each L , the resulting training process requires computational complexity of $O(N^2L)$. We observe that under the confocal setting, spherical wavefronts only intersect with a very small portion of the NLOS scene. These wavefronts tend to converge at specific patches and contribute greatly to

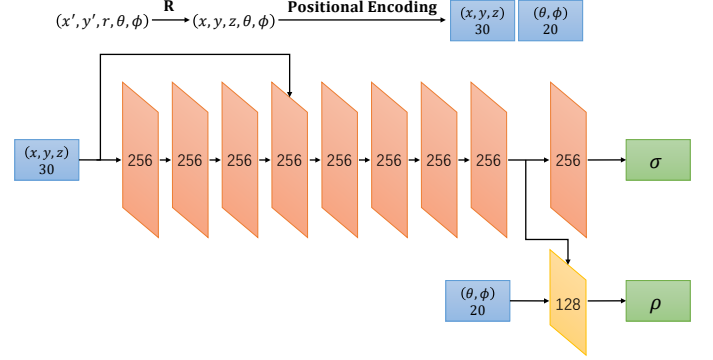


Fig. 6. NeTF network architecture: we adopt an MLP structure analogous to the one used in NeRF. The key differences though are (1) NeTF uses ReLU vs. NeRF uses sigmoid and (2) the last four layers in NeRF is simplified to one layer.

the final integral where the contributions from the rest are negligible.

We hence develop a *coarse-to-fine* sampling scheme. Specifically, we first sample N_c^2 uniform scene points in the hemisphere and evaluate our *coarse* network with the estimated PDF $k(\theta, \phi)$. Given the PDF, we employ a Markov chain Monte Carlo (MCMC) algorithm to produce a *fine* sampling of N_f scene points as $K(\theta_{f,ij}, \phi_{f,ij})$ along the hemispherical rays where samples are relevant to the NLOS volume. We then combine the coarse and fine samples $N_c^2 + N_f$ to re-evaluate the reconstruction quality from our *fine* network.

Specifically,

$$\tau(x', y', t) = \frac{\tau_c(x', y', t) + \tau_f(x', y', t)}{2} \quad (14)$$

where $\tau_c(x', y', t)$ is the integral with the coarse samples N_c^2 , and $\tau_f(x', y', t)$ is estimated with samples N_f from MCMC, as:

$$\tau_f(x', y', t) = \frac{1}{r^4} \sum_{i,j} \frac{\sigma(r, \theta_{f,ij}, \phi_{f,ij}; x', y') \rho(r, \theta_{f,ij}, \phi_{f,ij}; x', y')}{K(\theta_{f,ij}, \phi_{f,ij})} \quad (15)$$

It is important to note that our MCMC sampling is intrinsically differentiable (3.3). Previous volume-based methods, e.g., [2], [15], [34], [35], in theory can also apply such hierarchical sampling technique to refine their reconstruction. In reality, these methods use an explicit volumetric representation with a fixed resolution, making resampling on the hemisphere intractable.

5 EXPERIMENTAL RESULTS

In this section, we discuss our NeTF implementation and experimental validations.

5.1 MLP Settings

We train the NeTF using an MLP. Figure 6 shows the structure of our MLP. Analogy to NeRF [39], we construct a fully connected network with nine 256-channel layers, and with one 128-channel layer. We use ReLU activations for all the

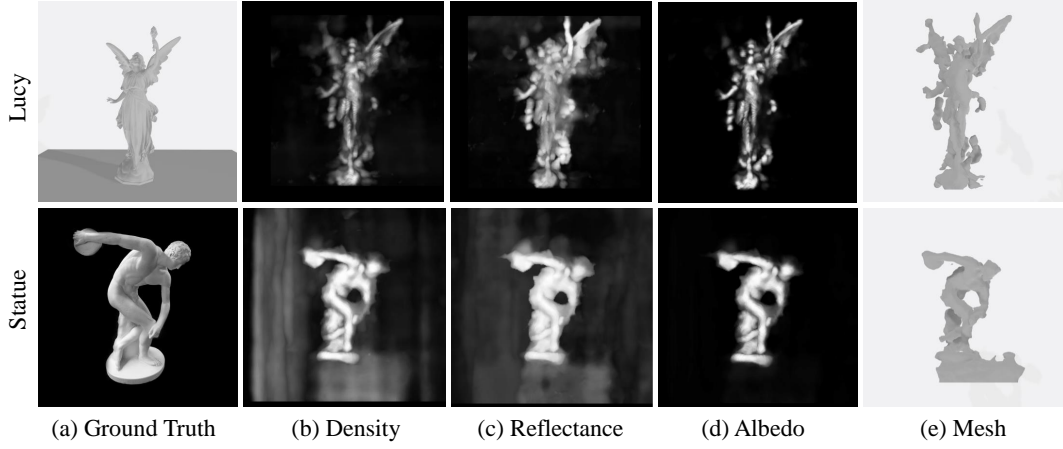


Fig. 7. From left to right: the ground truth, the recovered volume density, reflectance, albedo, and 3D mesh reconstruction using NeTF. Top shows the results on the Lucy model and bottom on the Statue.

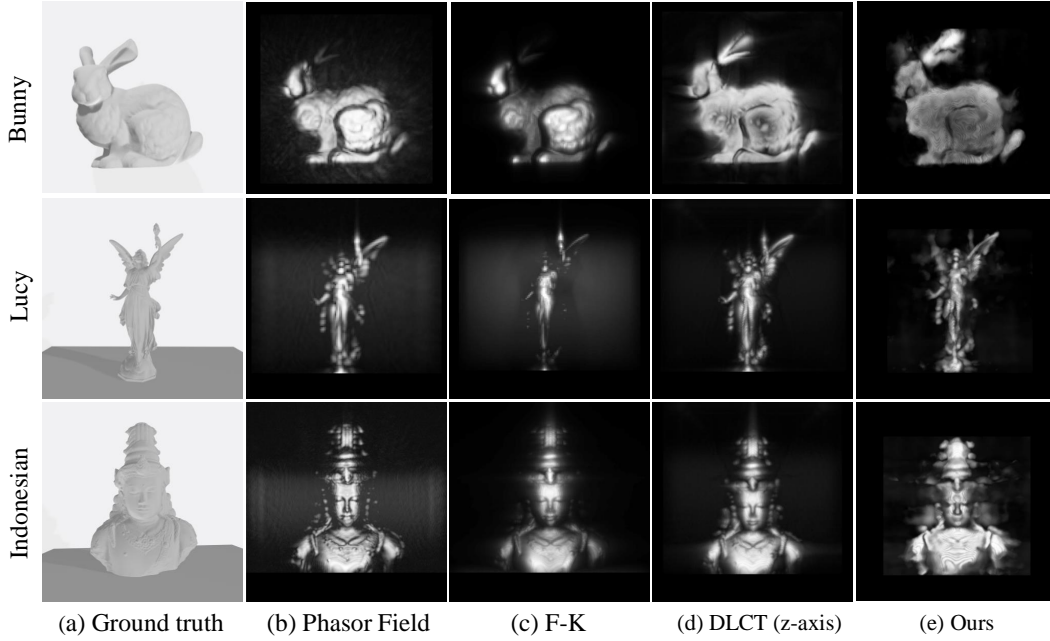


Fig. 8. Comparisons on simulated NLOS data. Our approach produces higher quality reconstructions on fine geometric details such as the ear of the bunny, the wing of Lucy and the crown of the Indonesian.

layers. We transform the spherical coordinates of sampling points into their Cartesian coordinates (x, y, z, θ, ϕ) , and feed them into the MLP. We predict the volume density as a function of only position, and the view-dependent reflectance as a function of both position and direction.

We first normalize the spatial coordinates (x, y, z) and the viewing direction (θ, ϕ) to range $[-1, 1]$. Next, we apply the positional encoding (PE) technique and map each input from 1 dimension onto a 10-dimensional Fourier domain to represent high-frequency variation in geometry and reflectance. Our MLP then processes the coordinates (x, y, z) as inputs with eight 256-channel layers and outputs a 256-dimensional feature vector. Note that we also concatenate (x, y, z) with the fourth layer for skip connection. This feature vector is passed to an additional 256-channel layer and pro-

duces σ . Simultaneously, the feature vector is concatenated with the direction (θ, ϕ) and passed to the 128-channel layer for reflectance ρ .

Under the NLOS setting, we consider a batch size of 1 to 4 transients and employ 16×16 or 32×32 samples for both uniform sampling N_c^2 and MCMC sampling N_f on the hemisphere. We adopt the Adam optimizer [33] with hyperparameters $\beta_1 = 0.9$, and $\epsilon = 1e^{-7}$. In our experiments, we use a learning rate that begins at $5e^{-4}$ and decays exponentially to $5e^{-5}$ through the optimization. On a single NVIDIA Tesla M40 GPU, consider a transient τ as a training epoch, the optimization with 256 transients converges at 1 to 3 epochs, and with 64 converges at 3 to 5 epochs. The training process takes around 10 to 48 h depending on batch sizes, epochs and samples.

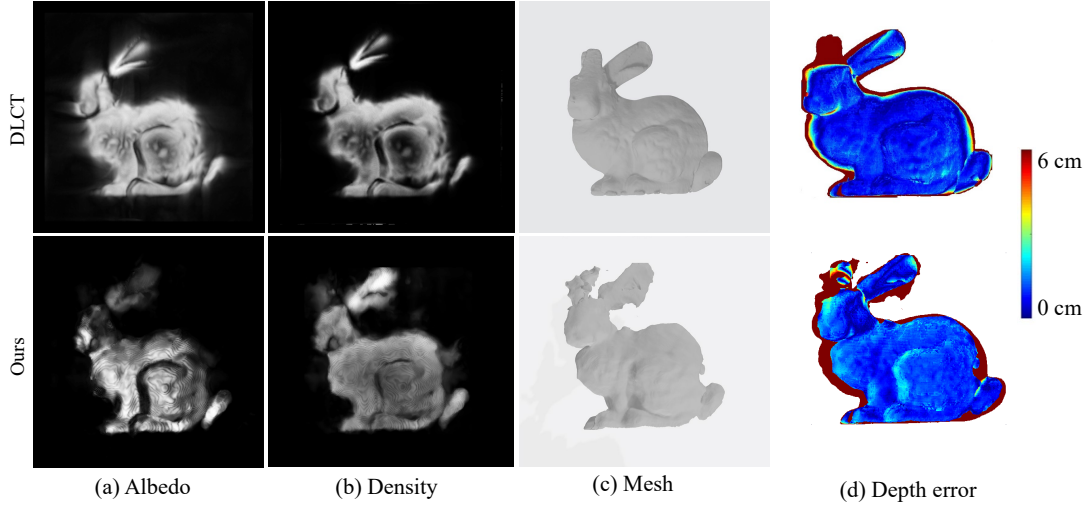


Fig. 9. Quantitative comparisons on our approach vs. the state-of-the-art DLCT technique.

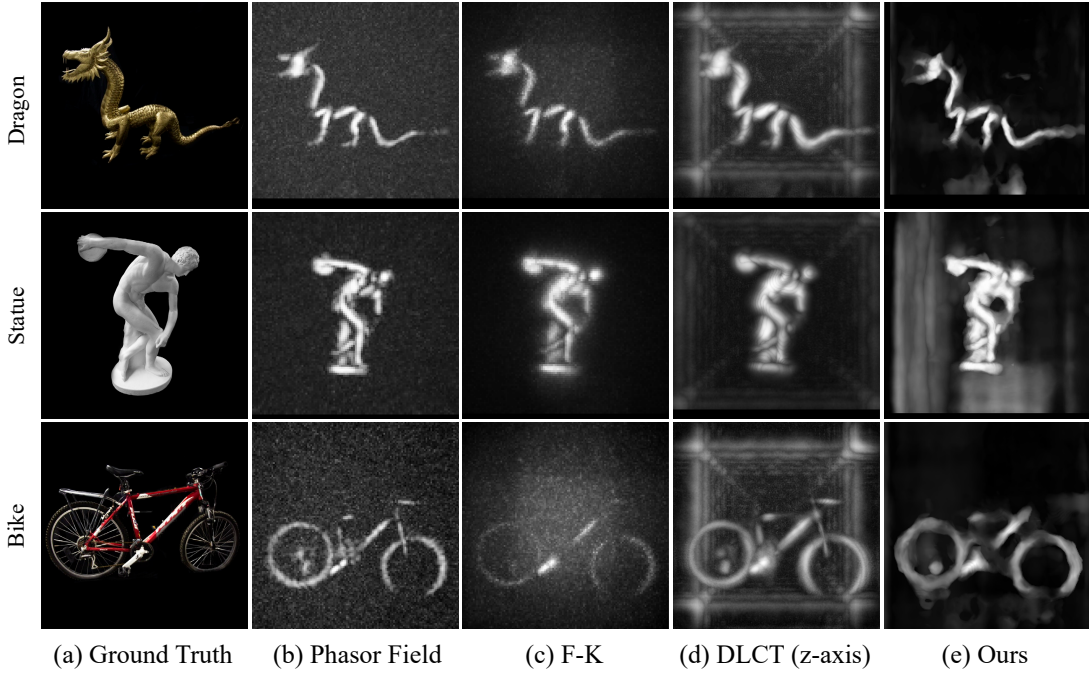


Fig. 10. Comparisons on real NLOS data. (a) The ground truth shows a photography of the hidden object. (b-c) show results using various techniques. NeTF overall produces sharper reconstructions compared with SOTA, except on Bike. This is because the Bike NLOS data contains strong noise due to heterogeneity in materials and complex topology in shape.

5.2 Validations

We have validated our approach on two public NLOS datasets: a simulated ZNLOS dataset [25], [26], and a real Stanford dataset [15]. ZNLOS consists of multi-bounce transients of synthetic objects that are 0.5 m from the relay wall. The transients have a temporal resolution of 512 time bins with a width of 10 ps and a spatial resolution of 256×256 pixels. The Stanford dataset captures transients measured in real scenes that are 1.0 m away from the relay wall. The transients in this dataset have a temporal resolution of 512 time bins with a width of 32 ps and a spatial resolution of 512×512 or 64×64 pixels.

On ZNLOS, we experiment on several simulated hidden

objects: Bunny, Lucy, and Indonesian at a spatial resolution of 256×256 pixels that correspond to an area of size $1 \text{ m} \times 1 \text{ m}$ on the relay wall. All three models are diffuse whereas Bunny does not contain a floor but Lucy and Indonesian do. On the Stanford dataset, we test on three real hidden objects with different materials: a diffuse Statue, a glossy Dragon, and a metal Bike. Their spatial resolution is of 512×512 spots and we down-sample them to 256×256 same as in [35] (128×128 in [14], [15]).

Fig. 7 illustrates our results. Our NeTF outputs a volume density map σ and a directional reflectance map ρ of ZNLOS Bunny and Stanford Statue. From these two maps we can produce volumetric albedo, and reconstruct a 3D mesh of

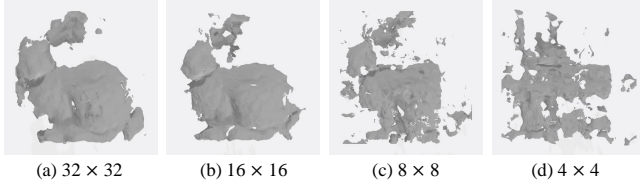


Fig. 11. Low Data Resolution: Mesh reconstructed from simulated transients of Bunny at 32×32 , 16×16 , 8×8 , and 4×4 spots on the wall. Even at a very low resolution of 8×8 , our NeTF produces reasonable reconstructions.

hidden objects. By sampling a 256×256 transients, our NeTF produces high quality reconstructions of objects with complex textures (e.g., Lucy). The density and reflectance maps of both the Statue and the Lucy, and the volumetric albedo produce much less error. We can then apply the Marching Cubes algorithm to further convert the volume to surfaces.

Figure 8 shows the comparisons with three most broadly adopted volume-based methods [14], [15], [35]. We compare the projected volumes to 2D maps of Indonesian, Lucy and Bunny. Note that the results from [14], [15] correspond to volumetric albedos whereas ours includes both the density and the albedo maps. We also show the normal volume on the Z-axis using [35]. The recovered volume maps on these hidden objects demonstrate NeTF outperforms SOTA by a large margin. Specifically, NeTF manages to recover many more details, e.g., the ear of Bunny, the wing of Lucy and the head of Indonesian that are partially or completely missing using prior art. The Phasor field technique achieves a similar performance as NeTF on Indonesian. However, it is wave-based and does not explicitly solve the inverse problem of volume reconstruction.

DLCT [35] produces comparable results to NeTF on Bunny and Indonesian. Yet, NeTF still manages to capture fine details on Lucy and Bunny where DLCT fails. DLCT, however, can produce better results by fusing results using the other two axes. Figure 9 shows in-depth comparisons between NeTF and DLCT on Bunny along all three axes on the recovered albedo, normal (density in our results) and mesh reconstruction. Our method preserves fine details but is slightly more noisy, as shown in the depth error. A similar phenomenon is observed on NeRF for multi-view 3D reconstruction where the noise can be potentially filtered.

On the real Stanford dataset, Figure 10 compares NeTF vs. SOTA on the glossy Dragon, diffuse Statue, and metal Bike. On Dragon and Statue where view-dependency is relatively small, NeTF and SOTA produce comparable results, although NeTF manages to better preserve high frequency features such as occluding edges. On the challenging Bike scene, NeTF achieves a similar performance as [14] while outperforming the other two. For DLCT, we only use the Z-axis and as a result, the reconstructed mesh exhibits adhesion between different parts whereas reconstruction produced by NeTF manages to separate these parts.

We have further tested the robustness and efficiency of NeTF on low resolution transient inputs as well as under non-confocal settings. Figure 11 shows the NeTF results that downsample the initial measures from 256×256 spots to

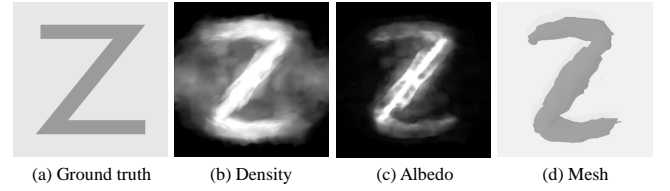


Fig. 12. NeTF reconstruction under the non-confocal setting on the density (b), albedo (c) and mesh (d) on a simulated scene composed of hidden letter Z.

32×32 , 16×16 , 8×8 and even 4×4 . Even with very sparse sampling spots (16×16 and 8×8), NeTF produces reasonable reconstructions.

To test our approach under the non-confocal setting, we use a hidden object from ZNLOS, i.e., a letter Z, and simulate transients. Figure 12 shows the density, albedo and mesh reconstruction. NeTF produces reasonable estimations to the ground truth.

6 CONCLUSION AND FUTURE WORK

We have presented a novel neural modeling framework called the Neural Transient Field (NeTF) for NLOS imaging. Similar to the recent Neural Radiance Field that seeks to use a multi-layer perception (MLP) to represent the 5D radiance function, NeTF recovers the 5D transient function in both spatial location and direction. Different from NeRF, the input training data are parametrized on the spherical wavefronts in NeTF rather than along lines (rays) as in NeRF. We have hence formulated the NLOS process under the spherical coordinates analogous to volume rendering under the Cartesian coordinates. Another unique characteristics of our NeTF solution is the use of Markov chain Monte Carlo (MCMC) to account for sparse and unbalanced sampling in NeTF. MCMC enables more reliable volume density estimation and produces more accurate shape estimation by recovering details caused by occlusions and non-uniform albedo. Our experiments on both synthetic and real data demonstrate the benefits of NeTF over SOTA in both the robustness and accuracy.

Same as NeRF, the final reconstruction of the hidden scene corresponds to a 3D density volume, implicitly represented by the MLP. Recovering the actual shape requires mapping the volume to surfaces, e.g., by thresholding followed by Marching Cubes. Such brute-force implementations may lead to noise on smooth surfaces. There are a number of emerging neural modeling techniques that can potentially provide smooth reconstructions, by imposing shape priors [40]. It is our immediate future work to investigate how to integrate such approaches into our NeTF framework. In addition, our current approach does not separately treat the confocal and non-focal setups. We observe though that the confocal setting resembles photometric stereo, or more precisely NeRF AA setting [41] and therefore we plan to study optimizing the network and the training process under the confocal setting.

REFERENCES

- [1] Otkrist Gupta, Thomas Willwacher, Andreas Velten, Ashok Veeraraghavan, and Ramesh Raskar. Reconstruction of hid-

- den 3d shapes using diffuse reflections. *Optics express*, 20(17):19096–19108, 2012.
- [2] Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Mounsi G Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications* 1747, 2012.
 - [3] Andreas Velten, Di Wu, Adrian Jarabo, Belen Masia, Christopher Barsi, Chinmaya Joshi, Everett Lawson, Mounsi Bawendi, Diego Gutierrez, and Ramesh Raskar. Femto-photography: Capturing and Visualizing the Propagation of Light. *ACM Transactions on Graphics*, 32(4):44:1–44:8, 2013.
 - [4] Mariko Isogawa, Ye Yuan, Matthew O’Toole, and Kris Kitani. Optical Non-Line-of-Sight Physics-based 3D Human Pose Estimation, *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
 - [5] Genevieve Gariepy, Nikola Krstajić, Robert Henderson, Chunyong Li, Robert R Thomson, Gerald S Buller, Barmak Heshmat, Ramesh Raskar, Jonathan Leach, and Daniele Faccio. Single-photon sensitive light-in-flight imaging. *Nature communications*, 6:6021, 2015.
 - [6] Donggeek Shin, Feihu Xu, Dheera Venkatraman, Rudi Lussana, Federica Villa, Franco Zappa, Vivek K Goyal, Franco NC Wong, and Jeffrey H Shapiro. Photon-efficient imaging with a single-photon camera. *Nature communications*, 7:12046, 2016.
 - [7] Matthew O’Toole, Felix Heide, David B Lindell, Kai Zang, Steven Diamond, and Gordon Wetzstein. Reconstructing transient images from single-photon sensors. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
 - [8] Victor Arellano, Diego Gutierrez, and Adrian Jarabo. Fast back-projection for non-line of sight reconstruction. *Optics Express*, 25(10):11574–11583, 2017.
 - [9] Feihu Xu, Gal Shulkind, Christos Thrampoulidis, Jeffrey H Shapiro, Antonio Torralba, Franco NC Wong, and Gregory W Wornell. Revealing hidden scenes by photonefficient occlusion-based opportunistic active imaging. *Optics Express*, 26(8):9945–9962, 2018.
 - [10] Tomohiro Maeda, Yiqin Wang, Ramesh Raskar, and Achuta Kadambi. Thermal Non-Line-of-Sight Imaging, *IEEE International Conference on Computational Photography (ICCP)*, 2019.
 - [11] David B Lindell, Gordon Wetzstein, and Vladlen Koltun. Acoustic non-line-of-sight imaging. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [12] Ioannis Gkioulekas, Anat Levin, Frédo Durand, and Todd Zickler. Micron-scale light transport decomposition using interferometry. *ACM Transactions on Graphics (TOG)*, 34(4):37, 2015.
 - [13] 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*, 8:11945, 2018.
 - [14] Xiaochun Liu, Ibon Guillen, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature*, 572:620–623, 2019.
 - [15] David B Lindell, Gordon Wetzstein, and Matthew O’Toole. Wave-based non-line-of-sight imaging using fast fk migration. *ACM Transactions on Graphics (TOG)*, 38(4):116, 2019.
 - [16] Shumian Xin, Sotiris Nousias, Kiriakos N Kutulakos, Aswin C Sankaranarayanan, Srinivasa G Narasimhan, and Ioannis Gkioulekas. A theory of fermat paths for non-line-of-sight shape reconstruction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [17] Adrian Jarabo, Belen Masia, Julio Marco, and Diego Gutierrez. Recent advances in transient imaging: A computer graphics and vision perspective. *Visual Informatics*, 1(1):65–79, 2017.
 - [18] Chia-Yin Tsai, Aswin C Sankaranarayanan, and Ioannis Gkioulekas. Beyond volumetric albedo—a surface optimization framework for non-line-of-sight imaging. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [19] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using ultrafast transient imaging. *International journal of computer vision*, 95(1):13–28, 2011.
 - [20] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using transient imaging. *International Conference on Computer Vision*, 2009.
 - [21] Yoann Altmann, Stephen McLaughlin, Miles J. Padgett, Vivek K. Goyal, Alfred O. Hero, and Daniele Faccio. Quantum-inspired computational imaging. *Science*, 361(6403): eaat2298, 2018.
 - [22] Felix Heide, Matthew O’Toole, Kai Zang, David B. Lindell, Steven Diamond, and Gordon Wetzstein. Non-line-of-sight imaging with partial occluders and surface normals. *ACM Transactions on Graphics*, 38(3):22:1–22:10, 2019.
 - [23] Syed Azer Reza, Marco La Manna, Sebastian Bauer, and Andreas Velten. Phasor field waves: A Huygens-like light transport model for non-line-of-sight imaging applications, *Optics Express*, 27(20): 29380-29400, 2019.
 - [24] Xiaochun Liu, Sebastian Bauer, Andreas Velten. Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems, *Nature communications*, 11(1): 1-13, 2020.
 - [25] Adrian Jarabo, Julio Marco, Adolfo Muñoz, Raul Buisan, Wojciech Jarosz, and Diego Gutierrez. A framework for transient rendering. *ACM Transactions on Graphics*. 33(6): 177, 2014.
 - [26] Miguel Galindo, Julio Marco, Matthew O’Toole, Gordon Wetzstein, Diego Gutierrez, and Adrian Jarabo. A dataset for benchmarking time-resolved non-line-of-sight imaging, *IEEE International Conference on Computational Photography (ICCP)*, 2019.
 - [27] Mariko Isogawa, Dorian Chan, Ye Yuan, Kris Kitani, and Matthew O’Toole. Efficient Non-Line-of-Sight Imaging from Transient Sinograms, *European Conference on Computer Vision (ECCV)*, 2020.
 - [28] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition, *International Conference on Learning Representations (ICLR)*, 2020.
 - [29] Marco La Manna, Fiona Kine, Eric Breitbach, Jonathan Jackson, Talha Sultan, and Andreas Velten. Error backprojection algorithms for non-line-of-sight imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7): 1615-1626, 2019.
 - [30] James T. Kajiya, Brian P Von Herzen. Ray tracing volume densities, *ACM Transactions on Graphics (TOG)*, 18(3):165-174, 1984.
 - [31] Mauro Buttavola, 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.
 - [32] Chia-Yin Tsai, Kiriakos N Kutulakos, Srinivasa G Narasimhan, and Aswin C Sankaranarayanan. The geometry of first-returning photons for non-line-of-sight imaging. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
 - [33] Diederik P. Kingma, Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2020.
 - [34] 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.
 - [35] Sean I. Young, David B. Lindell, Bernd Girod, David Taubman, Gordon Wetzstein. Non-Line-of-Sight surface reconstruction Using the directional light-Cone transform, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
 - [36] Julian Iseringhausen and Matthias B Hullin. Non-line-of-sight reconstruction using efficient transient rendering. *ACM Transactions on Graphics*, 39(1): 8, 2020.
 - [37] Daniele Faccio, Andreas Velten and Gordon Wetzstein. Non-line-of-sight imaging, *Nature Reviews Physics*, 2:318-327, 2020.
 - [38] Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, and Aswin C. Sankaranarayanan. Convolutional approximations to the general non-Line-of-sight imaging operator. *IEEE International Conference on Computer Vision*, 2019.
 - [39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *European Conference on Computer Vision (ECCV)*, 2020.
 - [40] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Neural Information Processing Systems (NeurIPS)*, 2020.
 - [41] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020.