Bayesian Time-of-Flight for Realtime Shape, Illumination and Albedo

Amit Adam, Christoph Dann, Omer Yair, Shai Mazor, Sebastian Nowozin

Abstract—We propose a computational model for shape, illumination and albedo inference in a pulsed time-of-flight (TOF) camera. In contrast to TOF cameras based on phase modulation, our camera enables general exposure profiles. This results in added flexibility and requires novel computational approaches. To address this challenge we propose a generative probabilistic model that accurately relates latent imaging conditions to observed camera responses. While principled, realtime inference in the model turns out to be infeasible, and we propose to employ efficient non-parametric regression trees to approximate the model outputs. As a result we are able to provide, for each pixel, at video frame rate, estimates and uncertainty for *depth, effective albedo, and ambient light intensity*. These results we present are state-of-the-art in depth imaging. The flexibility of our approach allows us to easily enrich our generative model. We demonstrate this by extending the original single-path model to a two-path model, capable of describing some multipath effects. The new model is seamlessly integrated in the system at no additional computational cost. Our work also addresses the important question of optimal exposure design in pulsed TOF systems. Finally, for benchmark purposes and to obtain realistic empirical priors of multipath and insights into this phenomena, we propose a physically accurate simulation of multipath phenomena.

Index Terms—Time-of-flight, Bayes, depth cameras, intrinsic images, multipath

1 Introduction

The commercial success of depth cameras in recent years has enabled numerous computer vision applications. Notable applications are human pose estimation [1, 2], dense online 3D reconstruction of an environment [3], and other uses—an overview is available in a recent special issue [4] and in the review article [5].

Broadly speaking we may differentiate between depth cameras based on triangulation and cameras which estimate depth based on time of flight (TOF) [6, 7]. Furthermore, while in the context of TOF the cameras often operate using modulated illumination and sensing, and the computational methods usually employ phase-space reasoning [7], in this paper we take a different approach which we now describe.

Figure 1 describes the inputs and outputs of our system. We start with n concurrently captured intensity images obtained under active illumination of the scene. Each of the n images is captured using a different exposure profile as will be described in Section 2. Using these n observations at every pixel, we infer the depth, reflectivity, and ambient lighting conditions. We achieve this by using a generative probabilistic model that relates the unknown imaging conditions—*shape*, *ambient illumination* and *albedo*—to the per-pixel camera observations. To perform inference we use either Bayesian inference or maximum likelihood estimation.

However, achieving realtime video rate by direct application of these inference methods is infeasible under practical constraints on computation. Therefore we use an approach inspired by model compression [8] and approximate the accurate but slow inference methods using regression trees, a fast non-parametric regression method [9].

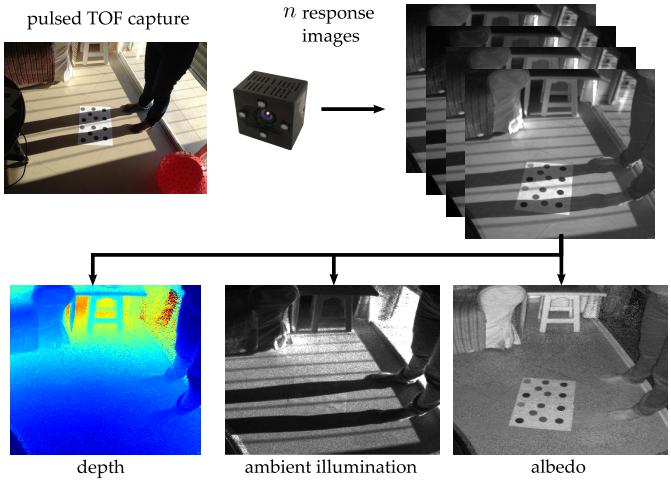


Fig. 1: System overview: the inputs are n pulsed TOF response images, obtained concurrently using different exposure profiles. In realtime (30fps) we separate *depth*, *ambient illumination* and *effective albedo* at every pixel.

The regression approach has two advantages; first, it allows to approximate inference in principled probabilistic models under tight compute and memory constraints; second, it decouples the model from the runtime implementation, allowing continuing improvements in the model without requiring changes to the test-time implementation.

We demonstrate this important advantage in Section 5 where we extend our generative model to a richer model which considers multipath effects. Our decoupling of model+inference from runtime regression allows us seam-

- *A. Adam, O. Yair, and S. Mazor were with Microsoft AIT, Haifa, Israel. {email.amitadam, omeryair, smazor.shai}@gmail.com*
- *C. Dann is with the School of Computer Science, Carnegie Mellon University, USA. cdann@cmu.edu*
- *S. Nowozin is with Microsoft Research, Cambridge, UK. Sebastian.Nowozin@microsoft.com*

less switching between different generative models, at no additional computational cost. To the best of our knowledge no other depth cameras have used a statistical regression approach for online depth inference.

No matter which model we use for inference, at times there will be pixels that the model fails to explain. Common reasons are mixed pixels (imaging a depth discontinuity), sensor saturation, complex multipath, interference from another active device, or extreme image noise. We propose a robust fit-to-model score that can be used to detect and invalidate affected pixels from further processing.

Having described inference and pixel invalidation, we address an orthogonal but important question in pulsed TOF systems: exposure profile design. We are flexible to choose exposure profiles and we directly optimize the expected accuracy of inferred depth using Bayesian decision theory. This yields a challenging optimization problem and we propose an approximate solution.

Finally, we introduce an accurate TOF simulation procedure based on physically accurate light transport simulation. We use this capability for both exposure design and for synthetic but physically accurate benchmarking.

A video demonstration of the live system is available on the authors' homepages.

1.1 Related Work

Most commercially available time-of-flight cameras (as of early 2015) work using *modulated* time of flight [10, 11], also known as *phase-based* time-of-flight. They generate a sinusoidal illumination signal and measure correlation of the reflected signal with a sinusoidal-profiled exposure function of the same frequency, delayed by a phase shift [12]. For a fixed frequency and phase shift a recorded frame does not contain sufficient information to reconstruct depth. Therefore, modern systems typically record a sequence of frames at multiple frequencies and multiple phase shifts and use the combined set of frames to unambiguously infer depth using so called *phase unwrapping* algorithms [7, 12].

In contrast, our camera uses *pulsed* TOF, also known as *gated* TOF. This technology has differentiators in terms of hardware-related aspects (size, power, resolution) which are not relevant here, but let us highlight an important computational aspect of this camera: in contrast with the sine-like exposure functions used in modulated TOF, we are allowed to choose from a large space of possible exposure functions. Hence more general inference methods are required.

Our work on optimizing the exposure profiles has not been addressed in pulsed TOF systems, but for modulated TOF prior work [13] has attempted to optimize the illumination profile to improve depth accuracy.

Shape, Illumination, and Reflectance. Recovering the imaging conditions leading to a specific image—the inverse problem of imaging, is a long standing goal of computer vision. A recent modern treatment of this problem has been given in [14, 15, 16, 17], with a comprehensive historical review. Conceptually the approach in these works is similar to ours: find the most likely shape, illumination and albedo to give rise to the observed image. In contrast with these works, we do the inference at the pixel level and not the image level, being able to do so due to the unique imaging process we employ. Additionally our regression approach allows this inference to be done in realtime. Moreover, our shape output actually gives the full posterior depth distribution. This allows direct usage of our depth in incremental estimators or integrators such as [3] that specifically take care to maintain the state distribution at all times [18, 19].

In the context of illumination estimation, we remark that there have been specific works on shadow removal [20, 21], which is a nice byproduct of our approach (see Figure 1).

Multipath Interference. Multiple reflections (multipath) commonly occur in real scenes [6]. There is now a solid body of work on handling multipath in modulated TOF systems, but to the best of our knowledge there is no published work on handling multipath in pulsed TOF cameras.

We briefly discuss work that exists for modulated TOF and relate it to our proposed solution. The work of [22, 23] and [24] model the light reflections in the scene globally to improve depth inference. To do this, they assume planar Lambertian surfaces and iteratively minimize an energy function. The methods work in important settings but the expensive minimization procedure precludes a realtime implementation. The work [25, 26, 27] assumes two-path interference from close-to-specular surfaces. The resulting methods are practical and efficient and our approach in Section 5 makes similar model assumptions. However, we work with different signals (pulsed TOF) and also provide a probabilistic model with uncertainty estimates. The work [28] generalizes the above two-path models to signals which arise from either two-path specular or two-path Lambertian reflectors; these signals are "compressible" and can hence be described with few parameters; the resulting method can be implemented in real time.

Transient imaging is a recent research discipline where light is captured "in flight" (e.g. [29, 30, 31, 32]). Recent work inspired by this discipline (i.e. [33]) uses modulated TOF imaging with Fourier-based reconstruction of the time-dependent light density. The work of [34, 35] reconstructs the transient light density for each pixel from a large number of modulated TOF images, each with a different modulation frequency. While this line of work could inspire practical multipath techniques and is computationally efficient, currently the large number of required frequencies (several dozen) and the large acquisition time precludes realtime applications in dynamic scenes. The recent Structured Light Transport framework extends the performance envelope of these approaches to include dynamic scenes ([36] and see also [37]).

The robust invalidation of observations seems to have not been considered before with the exception of [28] who provide an adhoc method for invalidation. Because we use a sound probabilistic model we can leverage and adapt standard methods in Bayesian practice [38] for this purpose.

1.2 Contributions

To summarize and as an aide in following the paper, our novel contributions are:

Principled Framework

– A probabilistic generative model for pulsed TOF imaging;

– Principled inference of all latent imaging conditions, given camera observations;

– Accurate depth uncertainty estimates;

– Robust Bayesian per-pixel invalidation for outlier observations;

Practicalities

– A novel use of regression to enable realtime inference under tight compute and memory constraints;

– Complete decoupling of runtime mechanism from model and inference;

Extensibility and Multipath

– A probabilistic model for depth inference in the presence of simple multi-path;

Results

– Experimental results showing robust video-rate inference of shape, illumination and reflectance, both indoors and outdoors at direct sunlight;

Computational Photography and Tools

– Design of exposure profiles to directly optimize depth accuracy under task-derived imaging conditions;

– A novel physically-based renderer for TOF simulation, exposure design, and benchmarking.

2 Modeling the Imaging Process

We start with our camera's principle of operation, then formulate a generative model relating the unknown imaging conditions to the observable camera outputs. The imaging model we use below is similar to the ones used in modulated TOF (e.g [31, 32, 39]). Our derivation below is more explicit and is presented for completeness.

Assume that a specific pixel images a point at a certain distance L and denote by t the time it takes light to travel twice this distance ($t = 2L/c$ where c is the speed of light). The reflected signal is integrated at the pixel using an exposure determined by an exposure profile $S(u)$. It is helpful to imagine the camera has a mechanical shutter, and the function $S(u)$ denotes the amount of opening of the shutter as a function of time. If $P(u)$ is the emitted light pulse, the reflected pulse arriving after time t is $P(u-t)$. The observed response due to the reflected light pulse is

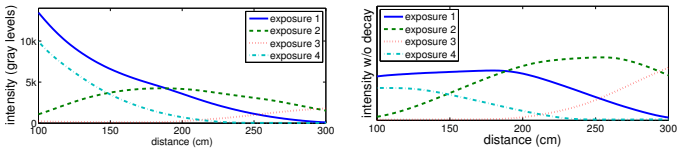
$$R_{\text{active}} = \int S(u)\ \rho\ P(u-t)\ d(t)\,\mathrm{d}u. \tag{1}$$

Here ρ denotes the effective reflectivity[1] of the imaged point, and $d(t) = \frac{4}{c^2t^2}$ denotes decay of the reflected pulse due to (one way) distance. Therefore, the reflected pulse is downscaled by a factor of $\rho\, d(t)$. The quantity $\rho P(u-t)d(t)$ is integrated with an exposure function $S(\cdot)$.

Let us now consider the effect of ambient illumination. We denote by λ the ambient light level falling on the imaged point. Then the reflected light level is $\rho\lambda$, and we assume that during the integration period, this level of ambient light is constant. Therefore, the observed response due to ambient light is $R_{\text{ambient}} = \int S(u)\ \rho\ \lambda\,\mathrm{d}u$. The actual observed response is the sum of the responses due to active illumination and due to ambient light,

$$R = \int S(u)\ \left(\rho\, P(u-t)\, d(t) + \rho\, \lambda\right)\,\mathrm{d}u. \tag{2}$$

1. We use both the terms albedo and reflectivity. The quantity ρ we use in the model actually contains the effect of foreshortening and therefore we refer to effective reflectivity/albedo.



(a) Response curve $\vec{C}$ (b) $\vec{C}/d$, without decay

Fig. 2: A typical response curve. (a) The actual curve $\vec{C}(\cdot)$. As distance grows the response decays, as per equation (5). (b) Decay-compensated response where we plot $\vec{C}(t)/d(t) = t^2\vec{C}(t)$ for more details (from now on we use decay-compensated curves for visualization).

Equation (2) specifies the relationship between the unknown imaging conditions (t, ρ, λ) (depth, albedo, and ambient light level), and the observation we obtain at the pixel, when using the exposure profile $S(\cdot)$. We *concurrently* use n different exposure profiles $S_1(\cdot), S_2(\cdot), \ldots, S_n(\cdot)$, and obtain n observations as

$$\begin{aligned}\begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix} &= \rho \begin{bmatrix} \int S_1(u)P(u-t)d(t)\,\mathrm{d}u \\ \vdots \\ \int S_n(u)P(u-t)d(t)\,\mathrm{d}u \end{bmatrix} + \rho\lambda \begin{bmatrix} \int S_1(u)\,\mathrm{d}u \\ \vdots \\ \int S_n(u)\,\mathrm{d}u \end{bmatrix} \\ &= \rho\vec{C}(t) + \rho\lambda\vec{A}. \end{aligned} \tag{3}$$

In short, we have the observed response vector

$$\vec{R} = \rho\vec{C}(t) + \rho\lambda\vec{A}. \tag{4}$$

Here $\vec{C}(t)$ is the expected response from a point at distance equivalent to time t, assuming unit reflectivity and no ambient light. This response is scaled by the reflectivity ρ and shifted in the ambient light direction $\vec{A}$, the magnitude of the shift being the product of albedo and ambient light level. Equation (4) is the model describing our imaging process.

We remark that $\vec{C}(\cdot)$ and $\vec{A}$ are determined by the illumination and exposure signals and are estimated using a simple camera calibration process which is outside the scope of this paper.

The hardware system enabling concurrent capture of n images under n different exposure profiles is based on fast manipulation of the photosensor reset mechanism (the substrate). It is fully described in [40] and [41, 42].

Figure 2 shows the curve $\vec{C}(\cdot)$ as a function of depth t, for a typical exposure profile design. The four colored curves denote the specific response curves of four exposure profiles $S_1(\cdot), \ldots, S_4(\cdot)$, namely

$$C_i(t) = \int S_i(u)\ P(u-t)\ d(t)\,\mathrm{d}u. \tag{5}$$

Looking at Figure 2, consider the response vector $\vec{R}$ we may expect from depth $t = 150$cm. We see that the first (blue) coordinate should be high, the second and fourth coordinates should be approximately equal and the third coordinate (red) should be the lowest. In contrast, at depth $t = 190$cm the first and second entries of $\vec{R}$ should be approximately equal (blue and green). Thus we see that by suitable design of the curve $\vec{C}(\cdot)$, we may expect to be able to infer depth accurately using the responses we observe.

Since the response we observe is scaled by the albedo ρ, it

may be tempting to normalize the response vector. However, as we discuss in the next section, the noise *does* depend on the magnitude and therefore the unnormalized response contains relevant information for depth inference and for predicting depth uncertainty that would be lost upon normalization.

3 A Probabilistic Model

We now rephrase (4) as a probabilistic model, relating the imaging conditions (t, ρ, λ) to a distribution over responses $\vec{R}$. Specifically we model $\vec{R}$ given t, ρ, λ as

$$\vec{R} \sim \Pr(\vec{R} \mid t, \rho, \lambda), \tag{6}$$

where we assume that $\Pr(\vec{R} \mid t, \rho, \lambda)$ is a multivariate Gaussian distribution with mean vector as in (4),

$$\mathbb{E}[\vec{R} \mid t, \rho, \lambda] = \vec{\mu}(t, \rho, \lambda) = \rho\, \vec{C}(t) + \rho\, \lambda\, \vec{A}, \tag{7}$$

and with a diagonal covariance matrix

$$\Sigma(\vec{\mu}) = \begin{pmatrix} \eta\mu_1 + K & & \\ & \ddots & \\ & & \eta\mu_n + K \end{pmatrix}. \tag{8}$$

Here K is related to *read noise*—noise that is part of the system even when no light is present. η is related to unit conversion between photo-electrons and image gray levels. This affine relationship between the magnitude of the response and its variance is due to *shot noise* and is well known [43, 44]. We validate this noise model experimentally in the supplementaries.

The generative model (6) is the distribution of the observed $\vec{R}$ at a pixel given the imaging conditions. We would like to *infer* the imaging conditions depth t, reflectivity ρ and ambient light level λ given the observation $\vec{R}$. There are three mainstream approaches for doing so, namely Bayesian posterior inference, maximum likelihood estimation (MLE), and maximum aposteriori (MAP) estimation.

3.1 Bayesian Inference

We assume certain priors on depth, reflectivity and ambient light level, denoted by $p(t)$, $p(\rho)$, and $p(\lambda)$. In addition we assume independence between these factors. Let us focus on inferring depth t, the most relevant unknown for depth cameras. Bayes rule gives

$$\begin{aligned} \Pr(t \mid \vec{R}) &\propto \Pr(\vec{R} \mid t)\, p(t) \\ &= p(t) \iint \Pr(\vec{R} | t, \rho, \lambda)\, p(\rho)\, p(\lambda)\, \mathrm{d}\rho\, \mathrm{d}\lambda. \end{aligned} \tag{9}$$

Equation (9) gives the posterior distribution over the true unknown depth. We get the posterior density up to a normalization factor which may be extracted by integrating over every possible t. The posterior density is the ideal input to higher level applications which use probabilistic models [18, 19]. For other applications, it may be sufficient to summarize this posterior distribution by a point estimate, for example the posterior mean $\hat{t}_{\text{Bayes}}(\vec{R}) = \mathbb{E}[t \mid \vec{R}]$ or the MAP depth $\hat{t}_{\text{map}}(\vec{R}) = \operatorname{argmax}_t \Pr(t|\vec{R})$, together with a measure of the dispersion such as the posterior variance.

Computationally, we have to solve the integration problem (9) at every pixel. Doing this at frame rate under low compute resources is currently not feasible.

A second issue with (9) is that it requires the specification of priors $p(t)$, $p(\rho)$, and $p(\lambda)$. While using uniform priors on depth and reflectivity is physically plausible, specifying the prior on ambient light level is harder. For example, operating the camera in a dark room versus a sunlit room, would require very different priors. If the used prior deviates too much from the actual situation our estimates of depth could be biased, that is, suffer from systematic errors.

3.2 Maximum Likelihood Inference (MLE)

Alternatively we use maximum likelihood estimation for the imaging conditions,

$$(\hat{t}_{\text{mle}}, \hat{\rho}_{\text{mle}}, \hat{\lambda}_{\text{mle}}) = \operatorname*{argmax}_{t,\rho,\lambda} \Pr(\vec{R} \mid t, \rho, \lambda). \tag{10}$$

Instead of considering the depth that accumulates the most probability over all reflectivity and ambient light explanations, we consider the single combined imaging conditions $(\hat{t}_{\text{mle}}, \hat{\rho}_{\text{mle}}, \hat{\lambda}_{\text{mle}})$ which have the highest probability of producing the observed response $\vec{R}$.

3.3 Maximum Aposteriori Inference (MAP)

This method is the most likely point estimate taking into account prior preferences. We obtain it similar to the MLE estimate as

$$(\hat{t}_{\text{map}}, \hat{\rho}_{\text{map}}, \hat{\lambda}_{\text{map}}) = \operatorname*{argmax}_{t,\rho,\lambda}\ p(t)\, p(\rho)\, p(\lambda) \Pr(\vec{R} \mid t, \rho, \lambda). \tag{11}$$

The optimization problems (10) and (11) are non-linear because $\vec{\mu}(\cdot)$ is non-linear and because our noise model (8) has a signal-dependent variance. An iterative numerical optimization is required and a frame rate solution at every pixel is infeasible. We discuss further details of the inference procedures for MLE, MAP, and Bayesian inference in the supplementary materials.

4 A Regression Tree Approach

All inference methods we propose, MLE $\hat{t}_{\text{mle}}$, MAP $\hat{t}_{\text{map}}$, and Bayesian inference $\hat{t}_{\text{Bayes}}$ produce reliable depth estimates. However the computation of these estimates is expensive and impractical for a realtime camera system. To perform realtime inference we use a regression approach to approximate the model as follows.

1) **Offline:** Sample imaging conditions (t_i, ρ_i, λ_i) from the prior and responses $\vec{R}_i$ from the model (6). Then use one of the slow inference methods to generate labeled training data $(\vec{R}_i, \hat{t}(\vec{R}_i))$.
2) **Offline:** Train a regression tree/forest using the training data set, to obtain a predictor $\hat{t}_{\text{RF}}$.
3) **Online:** Given an observed response $\vec{R}$ predict the inferred depth $\hat{t}_{\text{RF}}(\vec{R})$.

Why would this procedure be a good idea?

– First, $\hat{t}_{\text{mle}}$, $\hat{t}_{\text{map}}$, and $\hat{t}_{\text{Bayes}}$ are smooth functions from the response space to depth and are simple to learn.

– Second, the regression tree $\hat{t}_{\text{RF}}$ has small performance requirements in terms of memory and computation.

– Third, it decouples the runtime from future changes to the probabilistic model and inference procedures.

In principle it would be desirable to train directly on a large and diverse corpus of ground truth data captured from the real world; however, capturing ground truth depth data is challenging [45], expensive, and ensuring the diversity in imaging conditions is difficult. Training on our forward model instead allows us to represent a wide variety of imaging conditions. Likewise, while we could train directly on samples $(\vec{R}_i, t_i)$ from the model this would incur additional variance because the noise makes $\vec{R}$ stochastic even for a fixed depth value. By training on the estimator $(\vec{R}, \hat{t}_i)$ instead we effectively remove this variance from the regression task.

For learning the regression tree we use the standard CART sum-of-variances criterion in a greedy depth first manner [9]. For the interior nodes of the trees we perform binary comparisons on the individual responses, $R_i \leq a$. At each leaf node b we store a linear regression model,

$$\hat{t}_b(\vec{R}) = \theta_b^T \cdot \left[1, R_1, \ldots, R_n, R_1^2, R_1 R_2, \ldots, R_n^2\right]^T, \quad (12)$$

where we use a quadratic expansion of the responses. We estimate the parameters θ_b of each leaf model using least squares on all training samples that reach this leaf [46].

We cannot over emphasize the practical importance of a flexible and decoupled-from-model regression scheme, in handling unexpected or new phenomena. An example is detailed in the supplementaries.

Approximation Tradeoffs. Because of its non-parametric nature the regression tree or forest can approach the quality of the exact inference output if given sufficient training data and expressive power. However, the key limiting factor in our actual implementation are specific constraints on available memory and compute. Basically the depth of the tree and the structure of the leaf predictor, determine the memory requirements. In Section 9 we example the accuracy vs memory tradeoffs experimentally.

4.1 Additional Regression Outputs

In addition to the estimated depth we output several other quantities per pixel. These outputs too are produced using trained regression trees. Specifically, we produce the following additional outputs:

– Reflectivity, $\hat{\rho}$, using $\mathbb{E}[\rho|\vec{R}]$ or via (10), (11).

– Ambient light level, $\hat{\lambda}$, using $\mathbb{E}[\lambda|\vec{R}]$ or via (10), (11).

– Depth uncertainty, as described below.

– Fit-to-model invalidation score γ, for detection of irregular imaging conditions, described in Section 6.

4.2 Computing Depth Uncertainty

In many applications of depth cameras to computer vision problems the estimated depth is used as part of a larger system; in these applications it is useful to know the uncertainty of the depth estimate. One example would be surface reconstruction [3], where uncertainty can be used to weight individual estimates and to average them over time.

We use the variance of the depth, in the form of a *standard deviation* $\hat{\sigma}(\vec{R})$, as a measure of uncertainty. Depending on whether we use $\hat{t}_{\text{Bayes}}$ or $\hat{t}_{\text{mle}}$, we compute the standard deviation as follows.

For $\hat{t}_{\text{Bayes}}$ we use the posterior distribution (9), and directly compute $\hat{\sigma}_{\text{Bayes}}(\vec{R}) = \sqrt{\mathbb{V}_{t\sim\Pr(t|\vec{R})}[t]}$.

For $\hat{t}_{\text{mle}}$ in (10), we employ the approach described in [47]. A first order Taylor expansion of the gradient (wrt imaging conditions) of the likelihood function in (10) is used to relate a perturbation $\Delta\vec{R}$ in the response to the resulting perturbation of the estimator $\hat{t}_{\text{mle}}(\vec{R} + \Delta\vec{R})$. This analysis leads to the covariance matrix of the maximum likelihood estimator and an approximation to the standard deviation, $\hat{\sigma}_{\text{mle}}(\vec{R}) = \sqrt{\mathbb{V}[\hat{t}_{\text{mle}}]}$.

In Section 9 and Figure 8(b) we demonstrate the accuracy of our uncertainty estimates by comparing them with the actual observed uncertainty. In the context of phase-based TOF, previous work [48] used random forests to output depth confidence scores for measured phase signals; their regressor was trained using laser scans. Here we instead obtain uncertainty directly from our probabilistic model.

5 Two-Path Model for Simple Multipath

The generative model (4) we used so far assumed a single direct response from the point being imaged. In order to account for multipath, this model needs to be extended as to describe the additional multipath light being integrated at the pixel. We demonstrate a simple extension of the model and inference as follows.

Consider a two-path model as proposed in [25, 26, 28]. In addition to the three unknowns t, ρ, and λ, we now also assume a second contribution having travelled depth $t_2 > t$, from a patch with reflectivity ρ_2. We extend the generative model (7) to

$$\vec{\mu}_2(t, \rho, \lambda, t_2, \rho_2) = \rho \left(\vec{C}(t) + \lambda\vec{A} + \rho_2\vec{C}(t_2)\right), \quad (13)$$

where ρ scales both the direct and indirect response, and ρ_2 scales only the indirect response. The model is exact for a second specular surface, but becomes an approximation in case the second surface is diffuse. For inference we extend the inference procedures to this model in a straightforward manner (details in the supplementaries).

For the prior of t_2 we select a uniform prior relative to t, such that $t_2 - t$ is uniform between 0cm and Δ (typically $\Delta = 150$cm), that is,

$$p(t_2|t) = \mathcal{U}(t_2; t, t + \Delta). \quad (14)$$

For the second reflectivity we allow $\rho_2 > 1$. This allows us to approximate the aggregated response from a larger surface patch. After studies of simulation data select a Beta distribution on the interval $[0; 2]$.

$$p(\rho_2) = \mathcal{B}(\rho_2/2; \alpha = 1, \beta = 5). \quad (15)$$

This prior specifies that values up to $\rho_2 = 2$ are possible, but that low values of ρ_2 are more likely. Both priors are visualized in Figure 3 and we will evaluate this model on real and simulated data in the experiments section.

It is important to emphasize that our regression-decoupled-from-model approach allows us to seamlessly use this extended model in the camera, just by plugging it in the offline step 1 of the procedure outlined at the beginning of Section 4. The runtime process and its computational cost do not change at all.

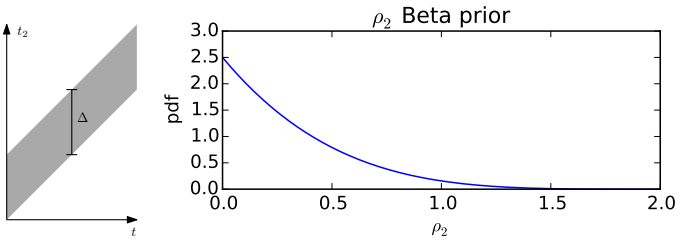


Fig. 3: Two-path model priors for the additional latent variables t_2 and ρ_2. Left: The prior for the second bounce depth t_2 is uniform over the shaded polygon. Right: the prior for ρ_2 is defined over $[0; 2]$ in order to handle large diffuse reflectors.

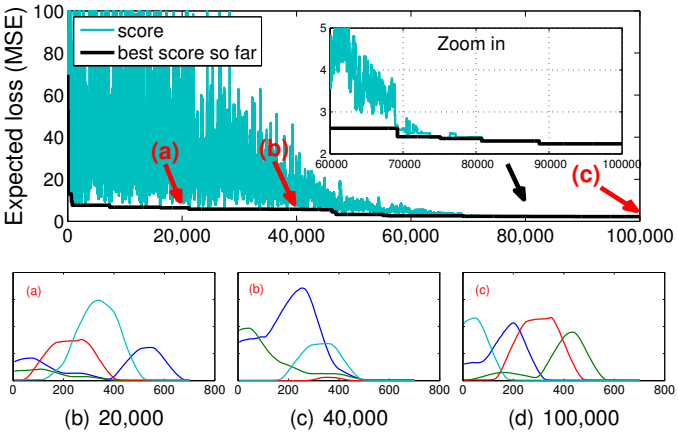


Fig. 4: Exposure profile optimization. **Top**: Simulated annealing over 100k iterations, finding response curves to minimize (19), the expected error (MSE) in depth estimation. **Bottom**: snapshots of the response curves after 20k, 40k, and after all 100k iterations. The x-axis is depth (cm).

6 Bayesian Model Invalidation Score γ

Our imaging model is an idealization of the real world and in each frame a certain number of pixels will have measured responses $\vec{R}$ which do not conform to this model. The main reasons for this are systematic errors such as multipath [22, 28], pixels of mixed depth, sensor saturation, as well as statistical extremes in imaging noise. In Section 5 we extended our model to explain some multipath effects, but even this extended model may fail to explain some of the responses.

When our model fails to accurately explain the observed response vector $\vec{R}$ we would like to detect such a deviation from the model assumptions and exclude affected observations from further processing. The strict Bayesian paradigm cannot detect deviation from model assumptions because it only provides the calculus to go from assumptions and observations to conclusions and no mechanism to falsify the assumptions themselves [49]. However, in Bayesian modelling practice [50] a common method to assess deviations from model assumptions is to perform so called *posterior predictive checks*.

We use the posterior predictive p-value [38, 51, 52] for our purposes. Intuitively our particular p-value will measure the total probability mass of all observations which have a smaller likelihood than the likelihood of the observed response. Therefore the score is always between zero and one and a value close to zero indicates that the observation is unlikely under the assumed model. This intuition is helpful but the controversy around p-values and model checking more generally is deep and we give a brief discussion in the supplementary materials.

To formalize this problem, let us first unify notation by writing $\theta = (t, \rho, \lambda)$ or $\theta = (t, \rho, \lambda, t_2, \rho_2)$, depending on whether we use the single path model (7) or the two path model (13), so that θ are all the unknown imaging conditions to be inferred. Given an observed response vector $\vec{R}$ and using the model $P(\vec{R}|\theta)$ and the prior $P(\theta)$ we can use Bayesian inference to infer the posterior distribution $P(\theta|\vec{R})$. Following the above intuition the invalidation score γ is defined as

$$\gamma(\vec{R}) = \mathbb{E}_{\theta\sim P(\theta|\vec{R})}\left[\mathbb{E}_{\vec{R}'\sim P(\vec{R}'|\theta)}\left[1_{\{P(\vec{R}'|\theta)\leq P(\vec{R}|\theta)\}}\right]\right], \quad (16)$$

Here we used the notation $1_{\{\text{predicate}\}}$ which evaluates to one if the predicate is true and to zero otherwise. The above equation integrates all probability mass of less likely observations, weighted by the posterior $P(\theta|\vec{R})$. If we have many repetitions of the experiment $\theta \sim P(\theta)$ and $\vec{R} \sim P(\vec{R}|\theta)$ the scores $\gamma(\vec{R})$ would be uniformly distributed. The computation of (16) is essentially free during our approximate Bayesian inference procedure.

The value $\gamma(\vec{R})$ can be used to reject the null hypothesis of the assumed model: if $\gamma(\vec{R}) \leq \tau$ for some threshold τ we reject the assumed model for this observation. The score (16) is also applicable to MLE and MAP inference if we replace the outer expectation by a unit point mass at the inferred imaging conditions $\hat{\theta}_{\text{mle}}$ or $\hat{\theta}_{\text{map}}$, respectively. We evaluate the invalidation score experimentally in more detail in the supplementary materials and in Section 9.7.

7 Exposure Profiles Design

So far we covered our imaging model, our realtime regression approach and how we can invalidate responses unlikely to have been generated by our model. We now turn to an orthogonal question of designing a suitable response curve $\vec{C}$ for use in (4) ($\vec{A}$ is closely related to $\vec{C}$ and are both derived from $\mathbf{Z}$ which will immediately be defined). Recall from (5) that $\vec{C}$ is the integral of the illumination pulse P with the exposure profile S. In the camera, a laser diode and driver produce the illumination pulse P, and its design is fixed. The exposure profile $S(u)$, however, has a flexible design space parameterized by linear basis functions. We would like to design response curves $\vec{C}$ that will produce observations from which low-error estimates of the imaging conditions could be inferred.

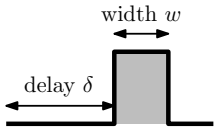


Fig. 5: A basis element $j = (\delta, w)$ defining $B_j(\cdot)$.

The hardware generates the exposure function $S()$ from a combination of basic gain profiles in the form of a boxcar function, as shown in Fig. 5. Each basic exposure profile has two parameters: a delay δ, and a width w. Each possible pair $j = (\delta, w)$ specifies one possible profile B_j from a fixed discrete set of choices J. Typically the set J contains several hundred possible combinations. With (5) we now get the

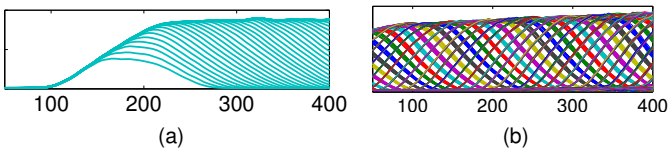


Fig. 6: **Left**, (a): basis functions $\{Q_j/d\}$ for a fixed delay δ and varying width w. **Right**, (b): all basis functions $\{Q_j/d\}_{j\in J}$, defined by Eq. (17).

basis function Q_j associated with B_j as convolution with the pulse,

$$Q_j(t) = \int B_j(u)\, P(u-t)\, d(t)\, \mathrm{d}u. \tag{17}$$

Fig. 6 shows a set of basis functions for all possible $j \in J$. We represent the basis response curves as vectors $Q_j \in \mathbb{R}^T$, for a time discretization with T values. By stacking the $m = |J|$ vectors horizontally we obtain a matrix $\mathbf{Q} \in \mathbb{R}^{T\times m}$ containing all possible basis response curves. The possible design space represents each curve as linear combination of basis curves, that is $S(\cdot) = \sum z_j B_j(\cdot)$. The coefficients z_j have to be positive integers because each unit value of z_j is actually a single firing of the shutter driver (More details are provided below). With (17) we then obtain the combined response curve $C(\cdot) = \sum z_j Q_j(\cdot)$. To design not just one but n response curves $S_i(\cdot)$ for $i = 1,\ldots,n$, we represent the design space using a matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$ as

$$\mathbf{C} = \mathbf{Q}\,\mathbf{Z}, \tag{18}$$

where in $\mathbf{C} \in \mathbb{R}^{T\times n}$ the k'th column contains the response for the k'th exposure sequence.

For the design objective we utilize *statistical decision theory* [53] to select $\mathbf{Z}$ to optimize the expected quality of depth inference. There are two components to this idea: the quality measure, and the expectation. The quality of depth inference is measured by means of a *loss function* which compares an estimated depth $\hat{t}$ with a known ground truth depth t to yield a quality score $\ell(\hat{t}, t)$. One possible loss function which we use is the squared error, $\ell(\hat{t}, t) = (\hat{t} - t)^2$, but we can also use other functions, for example $\ell_t(\hat{t}, t) = \ell(\hat{t}, t)/t$. For the expectation, as for the Bayesian depth inference, we devise priors, typically uniform, $p(t)$, $p(\rho)$, and $p(\lambda)$ over the unknowns. Then the design problem is

$$\underset{\mathbf{Z}}{\operatorname{argmin}} \quad \mathbb{E}_{t,\rho,\lambda}\, \mathbb{E}_{\vec{R}\sim \Pr(\vec{R}|t,\rho,\lambda,\mathbf{Z})} \left[\ell(\hat{t}(\vec{R}), t)\right] \tag{19}$$

$$\text{sb.t.} \quad \sum_{j=1}^{m}\sum_{i=1}^{n} Z_{ji} \le K_{\text{shutter}}, \tag{20}$$

$$\sum_{j=1}^{m} 1_{\{Z_{ji}>0\}} \le K_{\text{sparsity}}, \quad i = 1,\ldots,n, \tag{21}$$

$$Z_{ji} \in \mathbb{N}, \quad j = 1,\ldots,m,\ i = 1,\ldots,n,$$

where the notation $1_{\{\text{pred}\}}$ evaluates to one if the predicate is true and to zero otherwise.

The constraints (20) and (21) deserve some comments. Each captured frame contains a fixed number K_{shutter} of light pulses, each of which is associated with a basic exposure signal B_j. Therefore the variables Z_{ji} are positive integers. The total number of basis functions that may be used is constrained by K_{sparsity} due to various shutter driver restrictions. Because in each pulse a single basis function is selected, this makes the effective response curve $\mathbf{C}$ a non-negative linear combination of the basis functions (with integer coefficients).

Solving (19) is a challenging combinatorial problem on three levels: first, computing $\hat{t}(\vec{R})$ is the inference problem, which has no closed form solution. Second, as a result, computing the expectations also has no closed form solution. Third, more than just merely evaluating it, we would like to optimize the objective function over $\mathbf{Z}$.

The approximate solution which we adopt is as follows (more details in the supplementary materials). We approximate the objective function by a Monte Carlo evaluation for both expectations (imaging conditions, and responses): for $i = 1,\ldots,K$ we draw t_i, ρ_i, λ_i, then draw $\vec{R}_i$, then perform inference to obtain $\hat{t}_i = \hat{t}(\vec{R}_i)$ and evaluate $\ell_i = \ell(\hat{t}_i, t_i)$. Finally we approximate the objective (19) as empirical mean $\frac{1}{K}\sum_{i=1}^{K} \ell_i$. For $K = 8192$ samples this computation takes around one second. For optimization of (19) we use simulated annealing [54] on a custom-designed Markov chain which respects the structure induced by (20) and (21).

Figure 4 shows the progress of the optimization process. We start the optimization at a completely closed exposure profile with zero value, that is $Z_{ji} = 0$ for all j, i.

We remark that the optimization scheme just described outperforms all our previous attempts to manually design the exposure profiles.

8 Multipath Modeling and Design

In this section we describe our method for simulating realistic multipath images together with ground truth. Having a realistic simulation enables several important goals:

– exposure design for reduced multipath artifacts
– learning/obtaining realistic priors for multipath effects
– benchmarking

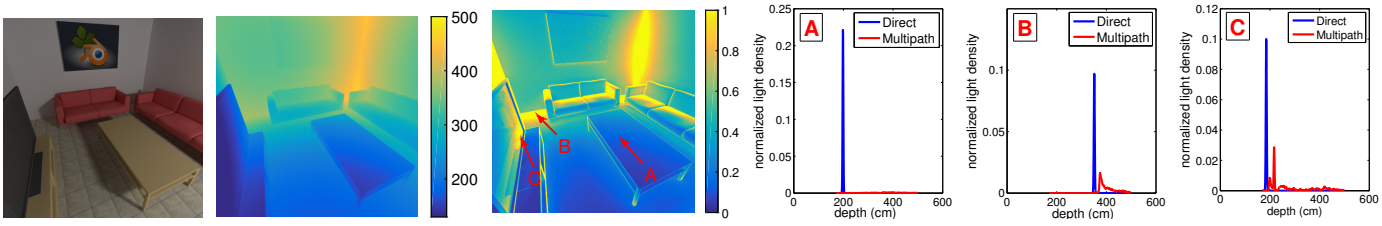
We show results for the first and last goal in section 9.

8.1 Time of Flight Simulation

In computer graphics physically-accurate renderers are mature technology that are readily available. We adapt the open source *Mitsuba renderer* [55]. Mitsuba supports, based on physical modeling of light scattering, light transport simulation, integrating paths of light at every pixel, thereby producing a highly realistic rendered image. We adapt the code so that we obtain the total light path length and the number of segments of the light trajectory.

In more detail, we modify two rendering algorithms, the bidirectional path tracer algorithm [56] and the Metropolis light transport (MLT) [57] algorithm; normally both algorithms are used to render the intensity of a pixel by means of approximating an integral over light pathes connecting light sources to surfaces to camera pixels [58].

Our modification is to record for each pixel a weighted set of light path samples $\{(w_i, L_i, t_i)\}_{i=1,\ldots,N}$, typically a few thousand, say $N = 4096$. For each light path we store the intensity weight $w_i \ge 0$, the number of straight path segments $L_i \in \mathbb{N}$, and the total length of the path t_i. The segment count allows us to distinguish direct responses ($L_i = 2$, emitter-to-surface and surface-to-camera) from indirect responses ($L_i > 2$, multipath). Together with a fixed



(a) Scene (b) Ground truth depth (c) Multipath ratio (d) Pixel A (e) Pixel B (f) Pixel C

Fig. 7: Insights into multipath using physically accurate light transport simulation. (a) A scene created in Blender. (b) Ground truth depth. (c) Normalized measure of multipath intensity compared to direct contributions. (please see the main text). (d)-(f) Normalized light densities for three selected pixels; pixel A has no multipath component, pixel B has one multipath component from 30–50cm further away, and pixel C, where a specular component gives rise to a narrow multipath response.

ambient value τ measuring light intensity without active illumination, for example from a regular rendering pass, the path lengths and weights now permit us to simulate a realistic mean response vector $\vec{\mu}$ as

$$\vec{\mu} = \tau \vec{A} + \sum_{i=1}^{N} \frac{w_i}{d(t_i)} \vec{C}(t_i). \tag{22}$$

The sum in the second term approximates the time-of-flight integral $\int_{\mathbb{R}_+} \vec{C}(t)\, \mathrm{d}\nu(t)$, where ν is an intensity measure over time. The division by $d(t_i)$ is due to both w_i and $\vec{C}$ containing the distance decay function $d(t)$; see (5). Once we have $\vec{\mu}$ we can optionally simulate sensor noise as specified in (8). We provide details in the supplementaries.

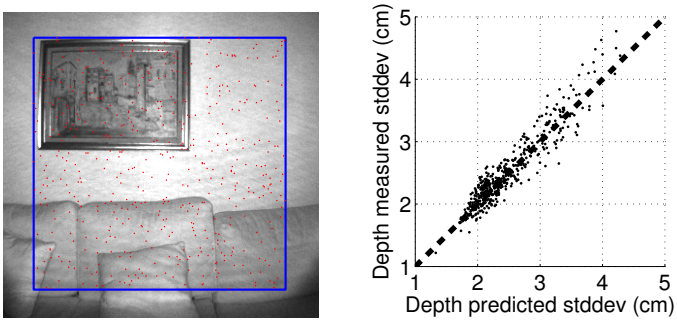
We remark that additional relevant work on light transport is considered in [59, 60], published independently and concurrently with our work.

8.2 Simulation Results

In part (a) of Figure 7 we show a synthetic scene. Part (b) shows the ground truth depth map corresponding to the scene. We marked three points (A, B and C shown in part 7(c)) at which we have different amounts of multipath. In parts 7(d), 7(e), and 7(f) we show the depth histograms we obtain from our modified Mitsuba renderer. For every point, the histogram shows the distribution of distances travelled by the photons integrated at this pixel. This distribution is properly weighted to account for both distances and reflectivity of materials along the pathes. Furthermore we show the distribution of distances travelled over a direct path in blue (this essentially corresponds to a delta function), and distances travelled over multiple pathes in red. We see that at point A (part 7(d)) there is no multipath, while at point B (part 7(e)) there is multipath due to the wall. We may see from the histogram the dominant additional path lengths - 30 to 50 cm in this case. Finally, in part 7(c) we show a normalized measure of the percentage of intensity integrated from multipath (as opposed to intensity integrated from a single direct light path), for every pixel in the image. We see that corners and just in front of the wall or other vertical surfaces actually return more multipath signals than direct path signals.

8.3 Multipath-Robust Exposure Profile Design

In the exposure profile design objective (19) we take two expectations: the first over prior imaging conditions (prior p)



(a) R_1 response and 500 random pixels.

(b) Predicted vs actual depth uncertainty.

Fig. 8: Predicted uncertainty versus actual uncertainty; the model is well-calibrated in that it accurately predicts depth uncertainty.

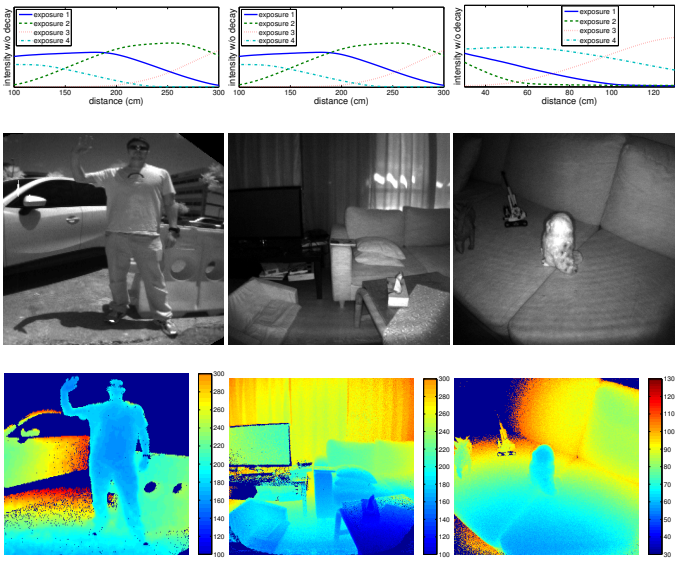


Fig. 9: Sample scenes. **Top**: exposure profile used. **Middle**: first response image R_1. **Bottom**: inferred depth image using the SP-MLE model. The left and middle column are scenes with a far-range design, the right column is a scene with a near-range design. The designs were obtained using different priors $p(t)$ in (19).

and the second over the assumed forward model (forward model P, equation (6)). This indeed is the way to minimize the loss when responses come from our basic generative model, which does not include multipath.

We now want to design an exposure profile that will be

more resistant to multipath. Therefore we should measure the loss over responses that also include multipath. We use our realistic simulator for that as follows. Given one or multiple 3D scenes and their realistic light transport pathes, we sample responses from these scenes. Formally, the scenes and the simulator are a more complex generative model G. We denote the sampling from this complex model by $(\vec{R}, t) \sim G$, but do keep in mind that the model G uses multiple reflectivity values and ambient lighting along the pathes to generate the response $\vec{R}$. Both P and G depend on the design $\mathbf{Z}$ through (7) and (22), respectively.

We combine both generative models in a mixture: a fraction $\beta \in [0; 1]$ are samples from our assumed model prior p and P, and a fraction $1 - \beta$ are samples from the physical simulation prior Q. Then the expectation (19) becomes

$$\beta\, \mathbb{E}_{t,\rho,\lambda\sim p}\mathbb{E}_{\vec{R}\sim P}[\ell(\hat{t}(\vec{R}), t)] + (1-\beta)\, \mathbb{E}_{(\vec{R},t)\sim G}[\ell(\hat{t}(\vec{R}), t)]. \quad (23)$$

We see that the design objective (19) can, by a simple change as in (23), accommodate richer priors over scenes and effects such as multipath. We demonstrate this in section 9.5.

9 Experimental Results

We use a prototype camera as shown in Figure 1. In our experiments we avoid reference and comparison with other depth cameras in terms of noise characteristics and variance of depth estimates because the validity of such comparison is affected by hardware configurations such as power used, field of illumination, resolution, thermal design constraints, and sensor sensitivity. Instead we focus on demonstrating the validity of our model, inference procedures, and regression approximations.

Throughout the experiments we will use the abbreviations SP and TP to refer to the single-path model (6) and the two-path model (13), respectively. Depending on the inference method we use MAP, MLE, and Bayes, so that TP-Bayes for example means the two-path model with full Bayesian inference.

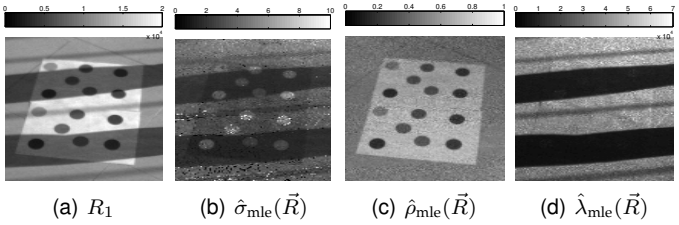
9.1 Sample scenes

We start with a few sample scenes shown in Figure 9. We designed two exposure profiles using two different uniform priors on depth $p(t)$ in (19). The first prior focused on larger depths while the second prior focused on smaller ranges. The two left columns show outdoor and indoor scenes using the far range exposure profile, and the right column shows a scene captured with the short range profile. The middle row shows the first response image, and the bottom row shows the inferred depth (obtained using the regression tree).

9.2 Accurate Depth Uncertainty

Next we show that by accurately modeling the noise present in the observed response our model is able to assess its own uncertainty in the inferred depth. To demonstrate this we capture 200 frames of a static scene as shown in Figure 8(a) and sample 500 pixel locations in the shown box.

Since the camera is static, we can obtain the empirical standard deviation of the depth estimators for each of the 500 points. We plot this empirical depth uncertainty, against the predicted uncertainty obtained in the first frame as described in Section 4.2 (the predicted uncertainty is nearly identical over all 200 frames). Figure 8(b) shows the good agreement between the predicted uncertainty and the actual uncertainty. This provides empirical data about how well the model is *calibrated* [61], that is, how accurately it judges the uncertainty in its own predictions.



(a) R_1 (b) $\hat{\sigma}_{\text{mle}}(\vec{R})$ (c) $\hat{\rho}_{\text{mle}}(\vec{R})$ (d) $\hat{\lambda}_{\text{mle}}(\vec{R})$

Fig. 10: Posterior inference results under different illuminations and albedos. (a) First response image, exhibiting varying ambient light levels and albedos, including strong shadows. (b) Posterior depth uncertainty (cm), higher under either stronger ambient light or lower albedo. (c) Inferred albedo map, in $[0; 1]$. (d) Inferred ambient illumination map.

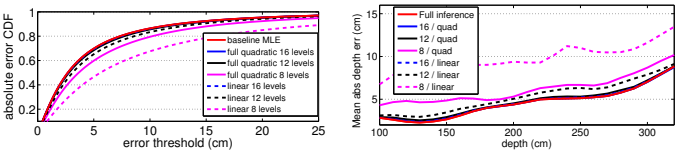


Fig. 11: Regression tree errors compared to full inference. **Left**: Cumulative error distribution over test set. **Right**: mean absolute error over prior albedo and ambient levels.

To gain some insight on what determines depth standard deviation, we turn to Figure 10, showing a part of the scene shown in Figure 1. In Fig. 10(a), we see one of the input responses, showing the combined effect of different albedos and shadows. In Fig. 10(b) we see how imaging conditions affect the variance of the depth estimates. In the shadowed regions the ratio between the active illumination and ambient light is higher, and this generally leads to a tighter posterior. On materials with higher albedo (the white page vs the dark circles) the amount of reflected light is higher and this also leads to smaller variances (as compared with the variances on dark circles which reflect less light). In addition, the depth itself affects the measure of uncertainty but this is not illustrated in this zoomed scene.

9.3 Ambient and Effective Reflectivity

In our model, the inferred albedo image is illumination-invariant and therefore does not contain shadows. Therefore we can perform realtime shadow removal [20, 21], providing illumination-invariant inputs to computer vision algorithms. This is illustrated in Fig. 10(c). In Fig. 10(d) we show the estimated ambient light level at each pixel.

For more results on realtime extraction of illumination, reflectivity and shape please view the enclosed video.

9.4 Regression Tree Approximation Quality

As discussed we use regression trees to regress depth, thus approximating full inference which is infeasible in realtime. An optimized implementation running on a *Intel HD Graphics*

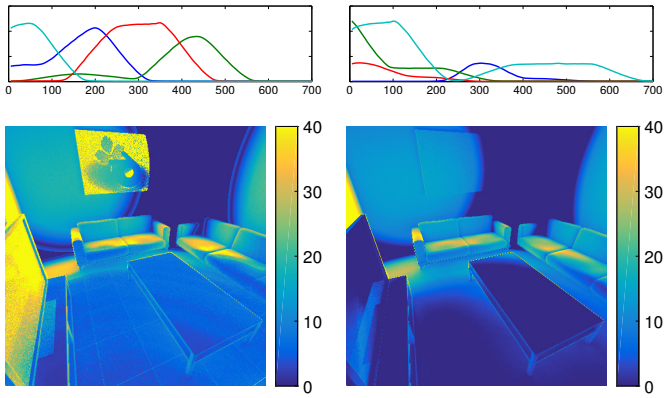


Fig. 12: Multipath-robust exposure profile design. **Left column**: original exposure profile. **Right column**: multipath-optimized exposure profile. **From top-to-bottom**: *Top*; regular exposure profile and robust-to-multipath profile. *Bottom*; bias magnitude (cm) of resulting depth inference images.

4400 GPU, evaluates a regression tree of depth 12 with full quadratic polynomials on a 200-by-300 pixel frame in 2.5ms. This means we are able to run four trees (depth, illumination, albedo and depth std) effectively at $\approx$ 100fps. The enclosed video shows this implementation running (the std was computed but not shown in the output windows).

We now quantify the additional errors incurred due to the use of regression trees instead of full inference. The added error depends on the tree structure, which determines required memory resources as described in section 4. We tested two types of trees at three depths, yielding six possible tree structures. The two types of trees used either a linear polynomial or a quadratic polynomial on the leafs. The depths we used were $8, 12$ and 16 (full binary trees).

After training the trees, we generate test data by sampling from the prior imaging conditions and our generative model (6). We compute the baseline error by running full inference (in this case MLE, but similar results hold for Bayes) on the test data, then run the various tree predictors. Fig. 11 shows the results. On the left we plot the cumulative distribution of errors over the test set. On the right we partition the test set by depths, and show the mean absolute error for each depth, averaged over all albedos and illumination levels. We see that as the trees get deeper the quality approaches that of the full inference. At 16 levels they essentially match.

9.5 Design for Multipath Robustness

Now we demonstrate how we may use our simulator for obtaining exposure profiles designed to be robust to multipath. In section 8.3 we allowed for more complex generative models during exposure profile design, one that will generate responses that are "contaminated" by multipath. We took two simple scenes of an object in front of highly reflective wall (scene provided in supplementaries) and used them as the model Q as in section 8.3. For the mixture model we used $\beta = 0.5$. We then ran our design optimization scheme to obtain a new exposure profile. Let us call this exposure profile the MP-resistant design. We compare it with the standard design obtained using $\beta = 1$. The two designs are tested on the scene shown in Fig. 7. We emphasize that this test scene is different than the scene used in design. Fig. 12 compares the results we obtain. The top row shows the regular design on the left, and the MP-resistant design on the right (obtained using two different values of β in (23)). The second row shows the magnitude of the resulting depth bias. On the left we see significant biases due to multipath (compare with the multipath map in Fig. 7(c)). With the MP-resistant design we see a significant reduction in bias.

9.6 Experimental Verification of Simulation Accuracy

Our light transport simulation is based on an accurate physical model of light and the simulation results should agree with the real camera. However, a real time-of-flight camera is a complex system with many components and potentially unaccounted for interactions between them. In this section we verify that our simulation serves as a good proxy for the real system.

To this end, we take images from two real scenes with a box and an optional reflector, shown in Figure 13(a) and 13(h), keeping the camera static between captures. We then perform *camera mapping* using the known camera intrinsics and reconstruct a matching 3D scene for our simulator. The synthetic scene allows us to assess the agreement between qualitative effects in the real capture and in the synthetic image. In our comparison we mask the results to the area occupied by the box and reflector, and in addition mask the bottom third of the sensor array because this particular camera has no active illumination design in this region.

The top row in Figure 13 shows the scene with only the box, the bottom row shows the multipath-corrupted scene with a large diffuse reflector added.

The following important observations can be made: 1. Comparing each of the four pairs of real and synthetic results the qualitative and quantitative error agree between the actual recording and the simulation; 2. The multipath corruption is clearly visible for the single-path (SP) model in Figure 13(k) and 13(l) and to a smaller extend in the two-path (TP) models, Figure 13(m) and 13(n).

Overall the simulation agrees very well with the real camera system. We remark that beyond this single experiment we describe here, the simulator is in daily use in our group and we have seen excellent agreement between simulated results and live tests over many months of using it.

9.7 Benchmarking using Simulation Data

In this experiment we leverage the ability of our simulator to provide ground truth depth. This allows us to assess the depth inference performance quantitatively. We use five scenes adapted from blendswap.com for this purpose. The depth range in each of these scenes is within 50cm to 500cm and the scene surfaces represent a good variety in materials and convex and concave geometries.

For each scene we obtain the IR responses and then run two inference engines. The first is Bayesian inference using the single path model, and the second is Bayesian inference using the two-path model. We emphasize that these inference procedures were run on exactly the same IR responses. Therefore any difference in results is due solely to a change in the model (both engines used Bayesian inference).

The results are visualized in Figure 14 and we report quantitative results for depth reconstruction in Table 1. Four

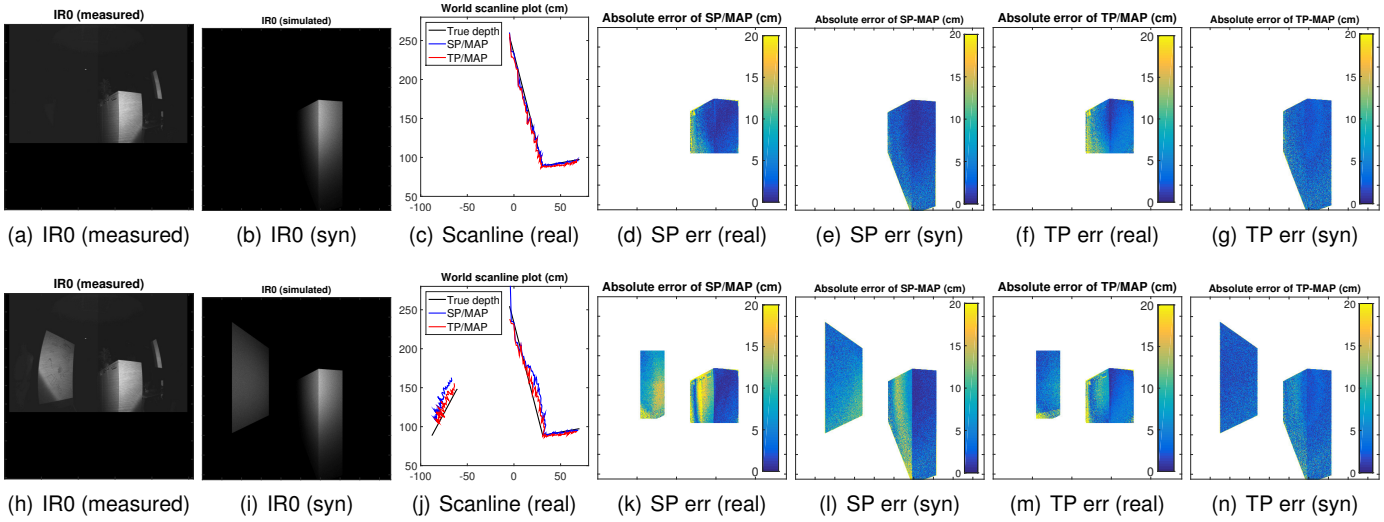


Fig. 13: Validation of the accuracy of light transport simulation. From the measured IR0 frames we use camera mapping to approximately reconstruct the 3D scene geometry and surface properties manually. The top row shows the no-multipath setting and we mask the frame so that only the target object is shown. The bottom row shows the multipath setting with a large white reflector added to the scene (left side). In the error results, each column corresponds to either real errors from the measured IR frames or is entirely simulated. Qualitatively there is an excellent agreement between real measurements and synthetic simulations. Small quantitative differences remain, for example in the no-multipath setting with the two-path model.

additional visualizations are provided in the supplementary materials. As error metric we use the 25/50/75 quantiles of absolute depth errors because these approximately correspond to easy/medium/difficult surfaces. We mask pixels in white for which no direct single-path response is created during rendering. These pixels typically correspond to either infinite rays or to perfectly specular surfaces. In both cases the time-of-flight operating principle does not apply.

We again re-emphasize that the absolute magnitude of the errors is not material and incomparable to other cameras for two reasons: first we report raw-depth errors with absolutely no spatial or temporal filtering that is usually present. Second, the jitter error highly depends on camera power, sensor size and other hardware characteristics which are of no concern in this paper.

From the results we make the following observations:

1) Surfaces with low reflectivity have large depth errors but the model is aware of this through a large inferred $\hat{\sigma}$ value. For example, the black floor in Figure 14(d) and 14(e). Improving on these regions would require increasing the light output or sensor sensitivity.
2) Areas affected by strong multipath also have large depth errors; for example the ceiling in Figure 14(d). The SP model $\hat{\sigma}$ does not indicate a potential error, but the γ score can invalidate these observations, for example the ceiling in Figure 14(f).
3) The two-path model (TP) improves depth accuracy in every scene but also reports increased model $\hat{\sigma}$ compared to the single-path model. For example, compare the absolute errors between Figures 14(d) and 14(j), and the $\hat{\sigma}$ maps in Figure 14(e) and 14(k).
4) Less invalidation happens in the two-path model. In all scenes, the TP-Bayes γ invalidates less pixels compared to the SP-Bayes γ map, because as a model it is a better representation for the physical simulator.
5) In Table 1 the errors are significantly reduced by the two-path model, typically by 40 percent.

Scene	Model	Absolute error quantile (cm) 25%	50%	75%
Sitting Room	SP-Bayes	7.23	13.46	21.20
Fig. 14	TP-Bayes	3.03	6.40	11.82
Breakfast Room	SP-Bayes	3.13	6.17	11.79
(supp. mat.)	TP-Bayes	1.85	4.18	8.75
Kitchen Nr 2	SP-Bayes	5.85	10.23	17.98
(supp. mat.)	TP-Bayes	2.59	5.80	13.32
Country Kitchen	SP-Bayes	4.86	10.22	18.74
(supp. mat)	TP-Bayes	3.10	6.71	14.30
Wooden Staircase	SP-Bayes	3.95	8.49	14.51
(supp. mat.)	TP-Bayes	2.17	4.80	9.43

TABLE 1: Predictive performance of the Bayesian single-path (SP) and two-path (TP) models on realistic data obtained from physically-accurate light transport simulation. Across all scenes the 25/50/75 error quantiles are significantly reduced by the two-path model. (raw-depth errors - no spatial or temporal filtering whatsoever)

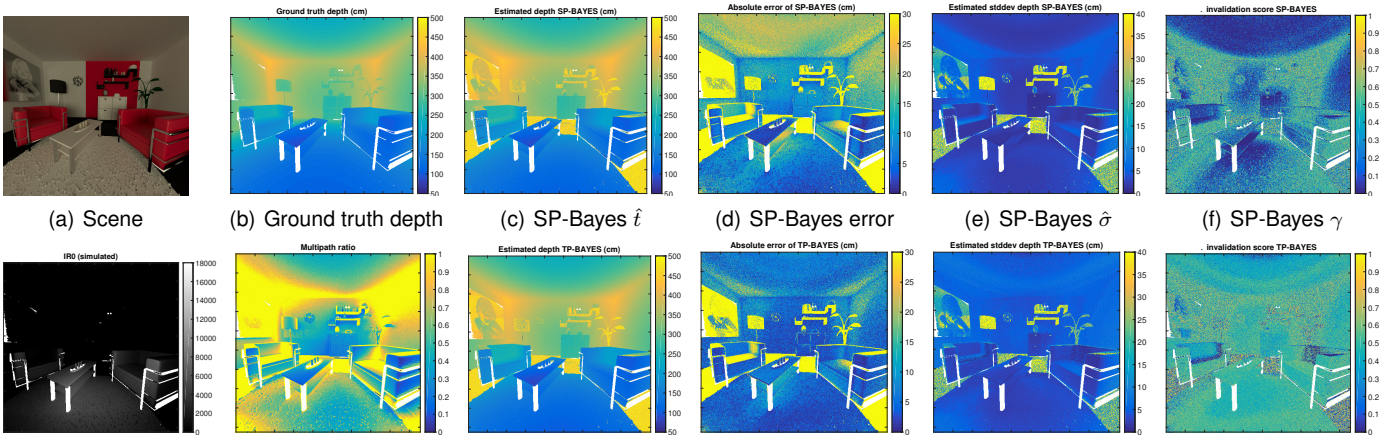
These results and insights agree with extensive live tests performed in the process of productizing our system.

9.8 Comparison with phase-modulated TOF

Lastly, we demonstrate the strength of our computational approach by comparing the results obtained by our general inference mechanism, with those obtained by using classic phase-based derivation of depth.

In a classic four-exposures phase-based TOF approach (e.g. [39]) four measurements are captured using equally spaced $\pi/2$ phase shifts. The phase associated with these measurements is computed using an analytic formula. In principle a linear relationship should hold between the phase and depth, but in practice a lookup table is used to correct for deviations from this linear relationship.

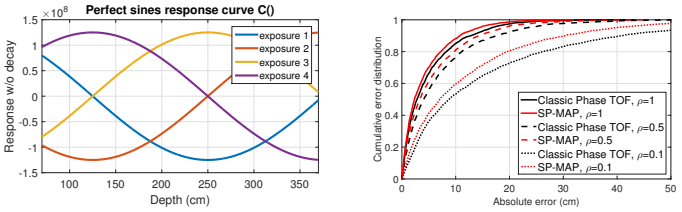
We simulated the responses as coming from the ideal perfect sines response curve, as shown in part (a) of Figure 15



(a) Scene (b) Ground truth depth (c) SP-Bayes $\hat{t}$ (d) SP-Bayes error (e) SP-Bayes $\hat{\sigma}$ (f) SP-Bayes γ

(g) IR0 (h) Multipath ratio (i) TP-Bayes $\hat{t}$ (j) TP-Bayes error (k) TP-Bayes $\hat{\sigma}$ (l) TP-Bayes γ

Fig. 14: Rendered simulation (scene adapted from "sitting room" by cenobi, licensed CC-BY from blendswap.com). High errors are present due to low reflectivity surfaces and multipath; the multipath errors are reduced by the two-path model (ceiling, wall, floor). The uncertainty estimate $\hat{\sigma}$ is higher for the two-path model, reflecting the multipath awareness (compare the $\hat{\sigma}$ values at the ceiling).



(a) ideal sines response curve (b) Cumulative distribution of absolute errors

Fig. 15: Our general inference method outperforms the classic phase-based TOF formula, even on a perfect sines response curve (see text)

(the modulation frequency we used is 30MHz leading to an unambiguous range of 5 meters). To these responses we added shot noise using the standard model [44] and then ran depth inference using both the analytic phase-based TOF formula [39], and our single path MAP inference as described in Section 3.3. The cumulative distribution of the absolute errors is presented in part (b) of Figure 15.

The responses corresponded to the ones expected from the following imaging conditions: depth is uniform between 70cm and 370cm, ambient light level is uniform between $\lambda = 0$ and $\lambda = 20000$ (covering a wide range of lighting conditions[2].) The reflectivity was chosen to be one of the three values $100\%, 50\%$ or 10%. We see that in each of the three cases, the errors of our probabilistic method were lower than the errors of the classical phase-based formula (the red CDF curves dominate the black ones).

We emphasize the fact that our inference method is general and may be used with any response curve $\vec{C}()$, while this is not the case for the classic phase-based formula. Indeed, an example of a response curve deviating from the sines curve is given in the top-right part of Figure 12.

2. This light level corresponds to about $10\frac{\text{mW}}{\text{cm}^2\text{nm}}$ - which fully covers indoor conditions (where the maximal light level is usually below $5\frac{\text{mW}}{\text{cm}^2\text{nm}}$)

10 Conclusion

Our presented approach is based on sound probabilistic modelling given our understanding of the physical reality. Bayesian inference naturally provides a powerful formal calculus to perform depth inference given our modelling assumptions. We have shown that even a simple model of multipath enables significant reductions in the depth error. However, both parts of our approach—the *prior* and *model*—are general and open to future extensions. For the prior we plan to develop scene- and task-specific priors to be able to improve performance in the presence of strong multipath and ambient light. We envision more refined models of multipath, for example by replacing the two-path pulse response by a more accurate analytic model of diffuse Lambertian multipath. This would require adding further latent variables related to multipath responses and creating suitable priors for them; this may be challenging but our simulation framework will likely enable us to make progress in this direction in the future. Our statistical view on time-of-flight enables all these extensions within a principled framework.

Acknowledgments

We thank Michael Baltaxe, Yair Sharf, Sahar Vilan and Giora Yahav for their contributions, long term support and commitment to this work.

References

[1] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE, 2011, pp. 1297–1304.

[2] J. Shotton, R. B. Girshick, A. W. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, 2013.

[3] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges,

and A. W. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2011, pp. 127–136.

[4] L. Shao, J. Han, D. Xu, and J. Shotton, "Computer vision for RGB-D sensors: Kinect and its applications," *IEEE T. Cybernetics*, vol. 43, no. 5, pp. 1314–1317, 2013.

[5] H. Jungong, S. Ling, X. Dong, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, Oct 2013.

[6] D. Lefloch, R. Nair, F. Lenzen, H. Schäfer, L. Streeter, M. J. Cree, R. Koch, and A. Kolb, "Technical foundation and calibration methods for time-of-flight cameras," in *Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications*. Springer, 2013, pp. 3–24.

[7] M. E. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras - Principles, Methods and Applications*, ser. Springer Briefs in Computer Science. Springer, 2013.

[8] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *12th ACM SIGKDD*, 2006.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[10] R. Schwarte, Z. Xu, H.-G. Heinol, J. Olk, R. Klein, B. Buxbaum, H. Fischer, and J. Schulte, "New electro-optical mixing and correlating sensor: facilities and applications of the photonic mixer device (PMD)," vol. 3100, 1997, pp. 245–253.

[11] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *IEEE Journal of Quantum Electronics*, vol. 37, no. 3, pp. 390–397, 2001.

[12] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin, "Phasor imaging: A generalization of correlation-based time-of-flight imaging," Tech. Rep., 2014.

[13] A. D. Payne, A. A. Dorrington, and M. J. Cree, "Illumination waveform optimization for time-of-flight range imaging cameras," in *SPIE Optical Metrology*. International Society for Optics and Photonics, 2011, p. 80850D.

[14] J. T. Barron and J. Malik, "Shape, albedo, and illumination from a single image of an unknown object," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 334–341.

[15] ——, "Intrinsic scene properties from a single RGB-D image," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 17–24.

[16] ——, "Shape, illumination, and reflectance from shading," EECS, UC Berkeley, Tech. Rep. UCB/EECS-2013-117, May 2013.

[17] P. V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf, "Recovering intrinsic images with a global sparsity prior on reflectance," in *NIPS*, 2011.

[18] M. Isard and A. Blake, "CONDENSATION - conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[19] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*, 2005.

[20] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," in *ECCV*, 2002.

[21] Y. Xiao, E. Tsougenis, and C.-K. Tang, "Shadow removal from single RGB-D images," in *CVPR*. IEEE, 2014.

[22] S. Fuchs, "Multipath interference compensation in time-of-flight camera images," in *ICPR*, 2010.

[23] S. Fuchs, M. Suppa, and O. Hellwich, "Compensation for multipath in ToF camera measurements supported by photometric calibration and environment integration," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science, M. Chen, B. Leibe, and B. Neumann, Eds. Springer, 2013, vol. 7963, pp. 31–41.

[24] D. Jiménez, D. Pizarro, M. Mazo, and S. Palazuelos, "Modeling and correction of multipath interference in time of flight cameras," *Image and Vision Computing*, vol. 32, no. 1, 2014.

[25] A. Dorrington, J. Godbaz, M. Cree, A. Payne, and L. Streeter, "Separating true range measurements from multi-path and scattering interference in commercial range cameras," vol. 7864, 2011.

[26] J. P. Godbaz, M. J. Cree, and A. A. Dorrington, "Closed-form inverses for the mixed pixel/multipath interference problem in AMCW lidar," vol. 8296, 2012.

[27] A. Kirmani, A. Benedetti, and P. A. Chou, "SPUMIC: simultaneous phase unwrapping and multipath interference cancellation in time-of-flight cameras using spectral methods," in *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[28] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt, "SRA: fast removal of general multipath for ToF sensors," in *13th European Conference on Computer Vision*. Springer, 2014, pp. 234–249.

[29] A. Smith, J. Skorupski, and J. Davis, "Transient rendering," School of Engineering, University of California, Santa Cruz, Tech. Rep. UCSC-SOE-08-26, February 2008.

[30] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar, "Looking around the corner using transient imaging," in *IEEE 12th Intl. Conference on Computer Vision*, 2009, pp. 159–166.

[31] F. Heide, M. B. Hullin, J. Gregson, and W. Heidrich, "Low-budget transient imaging using photonic mixer devices," *ACM Transactions on Graphics*, vol. 32, no. 4, 2013.

[32] 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)*, vol. 32, no. 6, p. 167, 2013.

[33] D. Wu, M. O'Toole, A. Velten, A. Agrawal, and R. Raskar, "Decomposing global light transport using time of flight imaging," in *CVPR*, 2012.

[34] A. Bhandari, A. Kadambi, R. Whyte, C. Barsi, M. Feigin, A. Dorrington, and R. Raskar, "Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization," *Optics Letters*, vol. 39, no. 6, pp. 1705–1708, 2014.

[35] J. Lin, Y. Liu, M. B. Hullin, and Q. Dai, "Fourier analysis on transient imaging with a multifrequency time-of-flight camera," in *CVPR*, 2014.

[36] M. O'Toole, J. Mather, and K. N. Kutulakos, "3d shape and indirect appearance by structured light transport," *Proc. CVPR*, 2014.

[37] 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)*, vol. 33, no. 4, p. 87, 2014.

[38] M. J. Bayarri and J. O. Berger, "P values for composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1127–1142, 2000.

[39] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight sensors in computer graphics," *Eurographics*, vol. 6, 2009.

[40] E. Tadmor, I. Bakish, S. Felzenshtein, E. Larry, G. Yahav, and D. Cohen, "A fast global shutter image sensor based on the VOD mechanism," in *2014 IEEE Sensors*, 2014.

[41] S. Felzenshtein, G. Yahav, and E. Larry, "Fast gating photosurface," US Patent 8717469, 2014.

[42] G. Yahav, S. Felzenshtein, and E. Larry, "Capturing gated and ungated light in the same frame on the same photosurface," US Patent Application 20120154535, 2010.

[43] G. E. Healey and R. Kondepudy, "Radiometric CCD camera calibration and noise estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 3, pp. 267–276, 1994.

[44] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1737–1754, 2008.

[45] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition*. Springer, 2014, pp. 31–42.
[46] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2012.
[47] R. M. Haralick, "Propagating covariance in computer vision," *IJPRAI*, vol. 10, no. 5, pp. 561–572, 1996.
[48] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in *CVPR*, 2011.
[49] A. Gelman and C. R. Shalizi, "Philosophy and the practice of Bayesian statistics," *British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 1, pp. 8–38, 2013.
[50] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press, 2012.
[51] X.-L. Meng, "Posterior predictive p-values," *The Annals of Statistics*, pp. 1142–1160, 1994.
[52] J. M. Robins, A. van der Vaart, and V. Ventura, "Asymptotic distribution of p values in composite null models," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1143–1156, 2000.
[53] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.
[54] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, 1983.
[55] W. Jakob, "Mitsuba renderer," 2010, http://www.mitsuba-renderer.org.
[56] E. Veach and L. Guibas, "Bidirectional estimators for light transport," in *Fifth Eurographics Workshop on Rendering*, 1994.
[57] E. Veach and L. J. Guibas, "Metropolis light transport," in *Proceedings of the ACM SIGGRAPH Conference*, 1997, pp. 65–76.
[58] M. Pharr and G. Humphreys, *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2010.
[59] P. Pitts, A. Benedetti, M. Slaney, and P. Chou, "Time of flight tracer," Microsoft Research, Tech. Rep. MSR-TR-2014-142, November 2014.
[60] A. Jarabo, J. Marco, A. Muñoz, R. Buisan, W. Jarosz, and D. Gutierrez, "A framework for transient rendering," *ACM Trans. on Graphics (SIGGRAPH Asia 2014)*, vol. 33, no. 6, 2014.
[61] P. A. Dawid, "The well-calibrated Bayesian," *Journal of the American Statistical Association*, vol. 77, no. 379, pp. 605–610, 1982.

Amit Adam Amit Adam received the PhD degree from the Technion-Israel Institute of Technology in 2001 for a thesis on vision-based navigation. Since his graduation, he has been working as an applied computer vision researcher in various application areas such as medical navigation, video surveillance, and recognition. Joining Microsoft's Advanced Imaging Technologies Group (AIT) in 2012, he has since worked on computational problems related to time-of-flight depth cameras.

Christoph Dann Christoph Dann obtained his B.Sc. and M.Sc. degree in Computer Science from the Technical University of Darmstadt, Germany, in 2011 and 2014, respectively. He is currently working toward a PhD degree in the Machine Learning Department at Carnegie Mellon University, USA. In the past, Christoph worked as an undergraduate researcher at the Max-Planck Institute for Informatics, the Intelligent Autonomous Systems group at the Technical University of Darmstadt, the Aerospace Controls Laboratory at MIT and as a research intern at Microsoft Research, Cambridge, UK. His research primarily focuses on sequential decision making under uncertainty including reinforcement learning as well as applications in computer vision.

Omer Yair received the BSc degree in Electrical Engineering (summa cum laude) and a BSc in Physics (summa cum laude) from the Technion-Israel Institute of Technology in 2011. He is currently with the Advance Imaging Technology Group at Microsoft and pursuing a MSc degree in Physics at the Technion.

Shai Mazor Shai received the B.Sc. and M.Sc. degrees in Electrical Engineering from the Technion-Israel Institute of Technology. After graduating he worked as a developer and later program manager, gaining experience both in startup companies and large corporations. He joined Microsoft's Advanced Imaging Technologies Group (AIT) in 2013, where he is now a senior program manager responsible for incubation of new applications for AIT technology. In addition to his engineering education, Shai holds an MBA from the IDC, where he was also an exchange student at the Wharton Business School.

Sebastian Nowozin is a senior researcher in the Machine Learning and Perception group at Microsoft Research Cambridge. He received his Master of Engineering degree from the Shanghai Jiaotong University (SJTU) and his diploma degree in computer science with distinction from the Technical University of Berlin in 2006. He received his PhD degree summa cum laude in 2009 for his thesis on learning with structured data in computer vision, completed at the Max Planck Institute for Biological Cybernetics, Tübingen and the Technical University of Berlin. His research interest is at the intersection of computer vision and machine learning. He is associate editor for TPAMI and JMLR and regularly serves as PC-member and reviewer for machine learning (NIPS, ICML, AISTATS, UAI, ECML, JMLR) and computer vision (CVPR, ICCV, ECCV, PAMI, IJCV) conferences and journals.