Single-Photon 3D Imaging with Deep Sensor Fusion

DAVID B. LINDELL, Stanford University, USA
MATTHEW O'TOOLE, Stanford University, USA
GORDON WETZSTEIN, Stanford University, USA

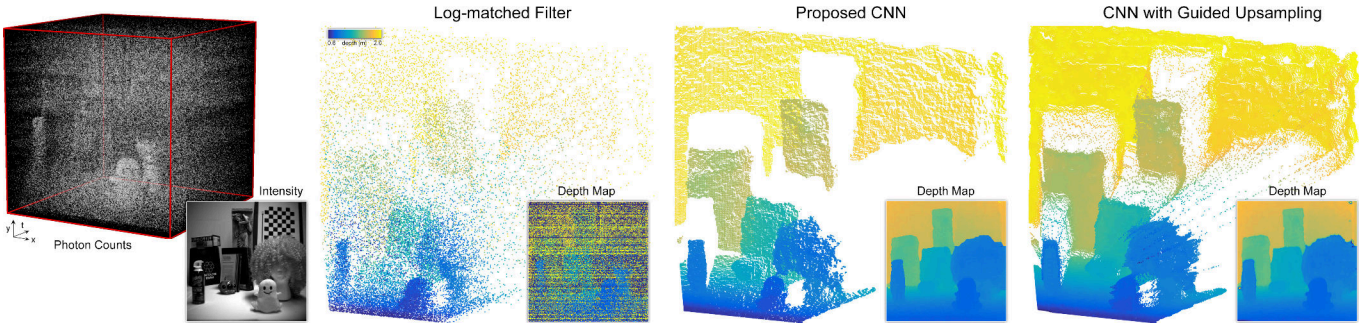


Fig. 1. Single-photon 3D imaging systems measure a spatio-temporal volume containing photon counts (left) that include ambient light, noise, and photons emitted by a pulsed laser into the scene and reflected back to the detector. Conventional depth estimation techniques, such as log-matched filtering (center left), estimate a depth map from these counts. However, depth estimation is a non-convex and challenging problem, especially for extremely low photon counts observed in fast or long-range 3D imaging systems. We introduce a data-driven approach to solve this depth estimation problem and explore deep sensor fusion approaches that use an intensity image of the scene to optimize the robustness (center right) and resolution (right) of the depth estimation.

Sensors which capture 3D scene information provide useful data for tasks in vehicle navigation, gesture recognition, human pose estimation, and geometric reconstruction. Active illumination time-of-flight sensors in particular have become widely used to estimate a 3D representation of a scene. However, the maximum range, density of acquired spatial samples, and overall acquisition time of these sensors is fundamentally limited by the minimum signal required to estimate depth reliably. In this paper, we propose a data-driven method for photon-efficient 3D imaging which leverages sensor fusion and computational reconstruction to rapidly and robustly estimate a dense depth map from low photon counts. Our sensor fusion approach uses measurements of single photon arrival times from a low-resolution single-photon detector array and an intensity image from a conventional high-resolution camera. Using a multi-scale deep convolutional network, we jointly process the raw measurements from both sensors and output a high-resolution depth map. To demonstrate the efficacy of our approach, we implement a hardware prototype and show results using captured data. At low signal-to-background levels, our depth reconstruction algorithm with sensor fusion outperforms other methods for depth estimation from noisy measurements of photon arrival times.

Authors' addresses: David B. Lindell, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, lindell@stanford.edu; Matthew O'Toole, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, matthew.otoole@gmail.com; Gordon Wetzstein, Department of Electrical Engineering, Stanford University, Stanford, CA, 94305, USA, gordon.wetzstein@stanford.edu.

0730-0301/2018/8-ART113 $15.00
https://doi.org/10.1145/3197517.3201316

CCS Concepts: • **Computing methodologies** → **3D imaging**; *Neural networks*; *Computational photography*;

Additional Key Words and Phrases: computational photography, single-photon imaging

ACM Reference Format:
David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. 2018. Single-Photon 3D Imaging with Deep Sensor Fusion. *ACM Trans. Graph.* 37, 4, Article 113 (August 2018), 12 pages. https://doi.org/10.1145/3197517.3201316

1 INTRODUCTION

3D imaging systems provide scene information crucial for a diverse range of applications, including autonomous vehicles, robotic vision, 3D object modeling, gesture recognition, pose tracking, scene understanding, segmentation, localization and mapping, and remote sensing. Depth sensors using the time-of-flight (ToF) principle are among the most popular choices with pulsed systems being most suitable for long-range outdoor applications [Horaud et al. 2016; Koskinen et al. 1992].

Pulsed ToF or light detection and ranging (LIDAR) systems often employ a co-axially aligned laser diode and single-photon detector. The laser emits a short pulse, usually on the order of hundreds of picoseconds or a few nanoseconds, and the detector timestamps the arrival of photons reflected back by the scene. Thus, for a single emitted pulse or a sequence of pulses, the detector records a temporal histogram of photon counts. A 3D spatio-temporal volume containing raw photon counts is then acquired by either scanning the laser-detector pair, using an array of these pairs, or using a hybrid approach of scanning and arraying.

Robustly estimating depth from raw photon counts acquired with pulsed ToF systems in low-flux scenarios is a major challenge [McCarthy et al. 2013; Pawlikowska et al. 2017]. When only a few signal

photons arrive at the detector, ambient photons and noise contaminate these measurements (see Fig. 1). With non-negligible amounts of ambient photons, the depth estimation problem becomes non-convex [Shin et al. 2015]. To address these challenges, a number of heuristic algorithms have recently been proposed to process noisy photon counts with as few as a single photon measured per pixel [Kirmani et al. 2014; Rapp and Goyal 2017; Shin et al. 2015, 2016]. However, such approaches make several restrictive simplifying assumptions and require many user-defined parameters, reducing their effectiveness when applied to harsh and diverse imaging conditions observed in the wild.

Inspired by recent successes of data-driven approaches in other communities, we adopt a deep learning approach to photon-efficient 3D imaging. We show that this approach is flexible enough to adapt to different imaging scenarios and that it intuitively allows for sensor fusion, for example with a high-resolution intensity image, to optimize the estimated scene depth. Specifically, we develop and train a convolutional neural network (CNN) that robustly estimates depth from raw photon counts measured with pulsed ToF systems.

Our contributions include

- a CNN architecture for estimating a depth map from single-photon sensor measurements;
- a sensor fusion model that improves the depth estimation by jointly processing raw photon counts and measurements of a regular camera with the CNN;
- an end-to-end approach for guided depth upsampling with the proposed sensor fusion model, improving the resolution of the estimated depth maps;
- validation of the proposed reconstruction techniques on a novel single-photon imaging system, which captures 256×256 single-photon measurements at 20 Hz.

Overview of System Tradeoffs. All 3D imaging systems that use active illumination make tradeoffs between image resolution, scanning speed, and light efficiency. Long-range scanning, for example, can be achieved by concentrating all available energy of the light source to a single point, which is sequentially scanned over the scene with slow acquisition rates [McCarthy et al. 2013; Pawlikowska et al. 2017]. The light source can also be diffused over the entire scene, enabling fast but short-range 3D imaging with array detectors [Kolb et al. 2009]. Hybrid systems use line sensors with line-scanned illumination to achieve a balanced tradeoff between range and scanning speed [Achar et al. 2017; O'Toole et al. 2015]. In principle, our algorithms apply to all of these systems, but we experimentally demonstrate them with a hybrid line-scanned system that achieves a moderate resolution and scanning speed for ranges up to a few meters.

2 RELATED WORK

Time-of-flight Imaging. 3D imaging systems based on the time-of-flight principle generally use either amplitude-modulated continuous wave (AMCW) or pulsed illumination. AMCW systems often suffer from phase-unwrapping artifacts and are typically limited to short or medium ranges, making them most suitable for indoor applications. Pulsed systems are often superior for long-range 3D imaging outdoors because, for a given power budget, they can concentrate all available energy of the light source both spatially and temporally, enabling more accurate ranging with faster acquisition times in the presence of strong ambient light. An overview of these technologies can be found in the surveys by Koskinen et al. [1992], Kolb et al. [2009], and Horaud et al. [2016].

Pulsed systems typically consist of a short duration light source paired with a detector, such as an avalanche photodiode (APD) or a single-photon avalanche diode (SPAD) [Dautet et al. 1993; Renker 2006]. Whereas linear-mode APDs have demonstrated robust performance in commercial LIDAR systems, SPADs are an emerging platform for photon-efficient 3D imaging, operating in a highly sensitive single-photon counting mode with improved timing precision [McCarthy et al. 2013; Pawlikowska et al. 2017; Tobin et al. 2017]. SPADs can also be used in low-flux regimes for intensity estimation [Altmann et al. 2017]. Our prototype uses a linear array of 256 SPADs [Burri et al. 2016] and a pulsed laser. We augment our system with a conventional high-resolution camera and explore novel sensor fusion algorithms that improve robustness, precision, and resolution of estimated depth maps compared to state-of-the-art algorithms [Kirmani et al. 2014; Rapp and Goyal 2017; Shin et al. 2015, 2016].

Sensor Fusion in 3D Imaging. Fusing depth and RGB (or monochrome) images has emerged as a popular method for improving depth estimates, especially due to the prevalence of cameras which can simultaneously capture RGB and depth (RGB-D) images. Sensor fusion approaches have been proposed to improve 3D mapping procedures with RGB-D cameras [Henry et al. 2012], perform segmentation and tracking [Bleiweiss and Werman 2009], and perform upsampling of the low-resolution depth image using a high-resolution intensity image captured from a similar viewpoint. Diebel and Thrun [2006] use Markov Random Fields to model the relationship between a depth and RGB image and perform upsampling on the depth image.

Image-guided depth upsampling has also been proposed using a joint bilateral filter [Kopf et al. 2007] and an iterative refinement approach with bilateral filtering [Yang et al. 2007]. Another approach extends the bilateral upsampling technique to use a multi-lateral filter which better accounts for noise in the depth data [Chan et al. 2008]. Park et al. [2011] demonstrate a framework for depth upsampling using nonlocal means filtering along with an edge-weighting scheme using high-resolution features from an RGB image. Finally, Ferstl et al. [2013] model depth as piecewise affine surfaces using Total Generalized Variation regularization; they then recover the upsampled depth by solving a convex optimization problem. An overview of additional sensor fusion approaches for AMCW systems can be found in the survey by Kolb et al. [2009].

More recently, image-guided depth upsampling has been discussed in the context of deep neural networks. In particular, Hui et al. [2016] use a multiscale approach for this task, where a learned downsampling operator is used to pass an intensity image to a depth-upsampling network at multiple scales. Li et al. [2016] use two subnetworks to extract features from the guidance image and the depth image. The features are then concatenated and mapped

to the output image by a third subnetwork. Their approach also extends to other guidance-based image reconstruction methods, such as flash and no-flash image reconstruction [Petschnigg et al. 2004]. Although Marco et al. [2017] and Su et al. [2018] do not use sensor fusion, they also recently showed that an end-to-end learning approach can optimize the quality of depth maps for AMCW systems. Similar to some of these methods, we employ a deep learning approach to depth estimation, where features from the intensity image are used at multiple scales to inform denoising and upsampling of the depth map; however, our approach is the first to model the full pipeline from raw measurements of single-photon detectors to estimated depth maps in an end-to-end fashion. Our methods are thus unique in tailoring deep sensor fusion to long-range pulsed time-of-flight imaging systems.

3 MODELING SINGLE-PHOTON IMAGING SYSTEMS

In this section, we outline an image formation model for the single-photon detectors and pulsed laser used in our prototype system. We also use this model to generate training data for the CNN discussed in the following sections.

3.1 Image Formation Model

Consider a light source that emits a short pulse at $t = 0$ with temporal shape $g(t)$, and an object at some distance z. The object reflects light back to an idealized detector, which counts photon arrivals over bins of duration Δt. Ignoring noise, ambient light, and radial falloff effects, the number of detected photons $\boldsymbol{\tau}$ at time interval n is

$$\boldsymbol{\tau}[n] = \int_{n\Delta t}^{(n+1)\Delta t} (g * f)\left(t - \frac{2z}{c}\right) dt, \tag{1}$$

where f models the temporal uncertainty of the detector and c is the speed of light. This model assumes single bounce light transport, and ignores multiply scattered light for simplicity.

Single-photon detectors, such as single-photon avalanche diodes (SPADs), approximate ideal photon detectors [O'Connor and Philips 1984; Renker 2006], but have a non-zero dark count d (number of false detections), and incoming photons are detected with probability $\eta \in [0, 1]$. For SPADs, f is related to the jitter of the underlying time-stamping mechanism and is usually on the order of tens to a few hundred picoseconds.

The number of photons measured by a SPAD in response to N illumination periods of a light pulse is represented by a temporal histogram

$$\mathbf{h_m}[n] \sim \mathcal{P}\left(N\left(\eta\gamma\boldsymbol{\tau}[n] + \eta a + d\right)\right), \tag{2}$$

where a is the number of ambient photons and the scalar γ models attenuation factors including radial falloff and reflectance. The measurements are thus modeled as a Poisson process $\mathcal{P}$ with a time-varying arrival function.

While this image formation model is insightful and widely used (e.g. [O'Toole et al. 2017; Rapp and Goyal 2017; Shin et al. 2015, 2016]), it makes several assumptions. First, after detecting an event, a SPAD must be reset or *quenched* before another event can be detected. We assume that this *dead time* is smaller or equal to the time between successively fired pulses. Second, the incident photon flux is assumed to be low, such that the effect of spurious charge carriers in the SPAD that could affect later events is minimized. This low-flux regime allows the detection of photon events to be modeled as being independent between pulses. Third, for a set of N illumination periods, the SPAD detects a random number $k \leq N$ of photons that follows a binomial distribution. Under the Poisson limit theorem, the measurements are well-approximated by a Poisson distribution as given by Equation (2).

4 ROBUST DEPTH ESTIMATION

4.1 Background

Using the probabilistic model for the measured photon counts, a maximum likelihood estimate for the target depth, given as the time-delay of the illumination pulse, can be derived [Shin et al. 2015]. The estimated time delay is calculated as $p^*\Delta t$, where p^* is given by

$$\underset{p\geq 0}{\operatorname{argmax}} \sum_n \log\left(\eta\gamma\boldsymbol{\tau}[n-p] + \eta a + d\right) \cdot \mathbf{h_m}[n], \tag{3}$$

and consists of determining the position of maximum correlation between the measurements $\mathbf{h_m}$ and a filter matched to the log of the photon arrival function. In practice, this estimate is unsuitable because it requires knowledge of γ, which may not be available. Further, the problem is non-convex when the noise terms a and d are non-zero. However, in the case of zero noise, the estimate reduces to applying a log-matched filter and determining the maximum correlation position:

$$\underset{p\geq 0}{\operatorname{argmax}} \sum_n \log\left(\boldsymbol{\tau}[n-p]\right) \cdot \mathbf{h_m}[n]. \tag{4}$$

This problem is convex as $\Delta t \longrightarrow 0$ in the common case when the measured illumination pulse $g * f$ can be approximated by a log-concave function, such as a Gaussian.

While the log-matched filter method for depth estimation is generally applicable when operating in a high signal-to-background (SBR) photon count regime, it produces poor results when operating at photon-efficient levels where only a few signal photons are detected among many background detections. Photon-efficient techniques for depth estimation thus produce an output $\hat{\mathbf{h}}$ which approximates a ground truth quantity $\mathbf{h}$: a set of measurement histograms where all detection events corresponding to background sources have been removed, and only detections corresponding to photon arrivals from the signal source remain. Once all background photons have been censored, a log-matched filter can be applied to estimate depth. A convex spatial prior may also be used to enforce spatial smoothness in the final depth estimate.

Background censoring techniques, however, rely on heuristic methods such as superpixel clustering [Rapp and Goyal 2017], identifying discrete, sparse depth clusters in a histogram of depth measurements [Shin et al. 2016], or using a rank-ordered mean filter [Abreu et al. 1996; Shin et al. 2015]. Such methods can also require scene-dependent parameter tuning or iterative computations to produce an acceptable result.

While recent photon censoring methods have enabled photon-efficient imaging at single photon levels, the difficulty of accounting for background photon counts in the analytical estimate and the use of heuristic methods motivates a data-driven approach. Learning to

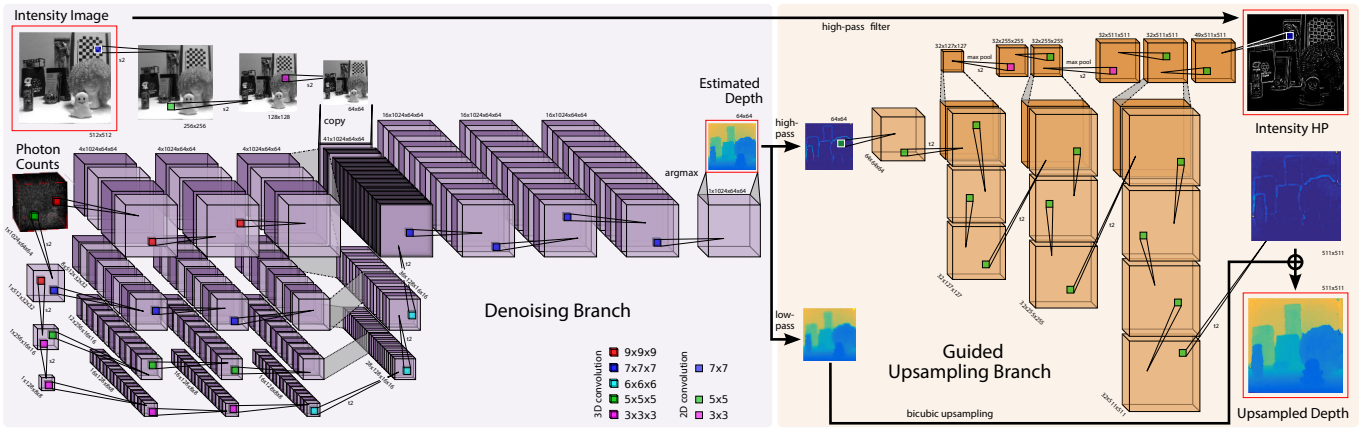


Fig. 2. Illustration of the convolutional neural network architecture. The denoising branch (left) takes as input the 3D volume of photon counts and processes it at multiple scales using a series of 3D convolutional layers. The resulting features from each resolution scale are concatenated together and optionally concatenated with additional features from an intensity image in a sensor fusion approach. A further set of 3D convolutional layers regresses a normalized illumination pulse, censoring the background photon events. A differentiable argmax operator is used to localize the time of flight of the estimated illumination pulse and determine the depth. In the image-guided upsampling branch (right), we adopt the approach of Hui et al. [2016]. The network predicts high-frequency differences between an upsampled low-frequency depth map and the high-resolution depth map using multi-scale guidance from high-frequency features of the intensity image. The entire network is trainable end-to-end for depth estimation and upsampling from raw photon counts and an intensity image.

censor noisy photons from the data yields improved non-iterative estimates of $\mathbf{h}$ without scene-dependent parameter tuning (given that the training data is consistent with the observation). As a result, significant improvements in depth estimates can be obtained at lower SBR levels as shown in the following.

4.2 Depth Estimation with a CNN

Motivated by the flexibility and strong performance of convolutional neural networks (CNNs) in denoising and reconstruction tasks, we design an architecture for the depth estimation problem. The CNN (illustrated in Figure 2) receives as input the 3D volume of raw photon counts $\mathbf{h_m}$ and produces a 2D depth image as output.

The structure of the network is inspired by state-of-the-art architectures for the task of semantic segmentation wherein object classifications are assigned to each pixel of an output image [Lin et al. 2017; Peng et al. 2017]. The depth estimation problem is analogous to the semantic segmentation problem in that the network should output one of many possible discrete depth values, as opposed to class values, for each pixel. Recent performance improvements for the semantic segmentation task have come through processing the input at multiple resolution scales, and then fusing the separate outputs together to produce the final estimate. We adopt a similar multi-scale approach.

Unlike the task of semantic segmentation, the input to the network consists of a large measurement volume ($256 \times 256 \times 1536$ for our hardware prototype), and we optimize a different objective function tailored to the depth estimation problem. Given the size of the input measurement volume, the number of convolutional layers and filters is limited in order to keep memory requirements practical during training and inference. We also consider the dimensionality of convolutional filters in the context of the depth estimation problem. A 2D convolutional filter would slide across the input spatial dimensions while spanning the sizeable length of the temporal dimension. To reduce the filter memory required and to better exploit temporal spatial correlations in the data, we use 3D convolutional filters.

The depth-estimation network incorporates an objective function which results in a two-step depth estimation process, similar to other approaches [Rapp and Goyal 2017; Shin et al. 2015, 2016]. The network first estimates the denoised histogram $\hat{\mathbf{h}}$, and the final depth is produced by a peak fitting step which reports the bin index of the maximum value of $\hat{\mathbf{h}}$. While a deeper network might be able to directly learn this two-step procedure, we find that our network achieves satisfactory performance with practical computation and memory requirements by directly modeling $\mathbf{h}$. Further, our objective function for depth estimation is fully differentiable, and so the process is trainable using conventional gradient backpropagation techniques.

The objective function for the denoising step is given as the Kullback-Leibler (KL) divergence at each spatial position between the output of the network and a normalized version of $\mathbf{h}$. This loss function can be written for each spatial position k as

$$D_{\mathrm{KL}}(\mathbf{h}^{(k)}, \hat{\mathbf{h}}^{(k)}) = \sum_n \mathbf{h}^{(k)}[n] \log \frac{\mathbf{h}^{(k)}[n]}{\hat{\mathbf{h}}^{(k)}[n]}, \tag{5}$$

where $\hat{\mathbf{h}}$ is the output of the final 3D convolutional layer after a softmax nonlinearity, which causes $\hat{\mathbf{h}}$ to sum to unity and values outside the location of the illumination pulse to tend to zero. Note that this expression is equivalent to cross-entropy up to an additive constant; however, the term cross-entropy sometimes implies that the ground truth class (or bin index) has probability equal to 1. In our problem, the network should have multiple non-zero outputs for bins within the estimated area of the illumination pulse.

We also introduce a term for total variation (TV) spatial regularization on the output depth image which improves denoising performance, especially in the low SBR case where few signal photons are detected. As the regularizer should be applied to the estimated 2D depth, we apply a differentiable, or "soft," argmax operator to $\hat{\mathbf{h}}$ to find the approximate value of the maximum bin index through a simple weighted sum calculation

$$\text{soft argmax}\left(\hat{\mathbf{h}}^{(k)}\right) = \sum_n n \cdot \hat{\mathbf{h}}^{(k)}[n]. \tag{6}$$

The complete loss function used to train the network for depth estimation is thus given as

$$\mathcal{L}(\mathbf{h}, \hat{\mathbf{h}}) = \sum_k D_{\text{KL}}(\mathbf{h}^{(k)}, \hat{\mathbf{h}}^{(k)}) + \lambda_{\text{TV}}\,\text{TV}\left(\text{soft argmax}(\hat{\mathbf{h}})\right), \tag{7}$$

where λ_{TV} is a non-negative scalar that determines the magnitude of the regularizer loss.

4.3 Evaluation

We evaluate the CNN on a set of measurements simulated from the Middlebury stereo dataset [Scharstein and Pal 2007]. Details on the training dataset and procedure can be found in Appendices A and B.

A summary of these simulations is shown in Table 1. We report the root-mean-square error (RMSE) values averaged across 8 Middlebury test scenes over a number of simulated signal and noise levels for log-matched filtering, the method of Shin et al. [2016], Rapp and Goyal [2017], and our CNN. For signal-to-background ratios greater or equal to 1, the reconstruction from Shin et al. provides accurate results, but as the SBR decreases further, the quality sharply decreases and the reconstruction exhibits a large depth bias due to the background photon censoring step. The log-matched filtering results consistently degrade with decreasing SBR. While Rapp and Goyal's method [2017] performs well across all levels, our methods (i.e. with the intensity image) show improvements in certain cases in the simulated depth estimation and upsampling results and for the captured results as shown in the following sections. In particular, Rapp and Goyal's method incorporates spatial regularization to encourage smoothness; however, in some circumstances, this comes at a cost of smoothing over edges or fine details which the CNN approaches correctly reconstruct. We also show that performance can be improved in certain cases by fine-tuning the network on specific signal and SBR levels as shown in Table 1. Additional quantitative and qualitative results are included in the supplemental document.

Our method also requires no user-defined parameters at runtime, resulting in improved performance where differing object depths and ambient illumination significantly alter spatially-local SBRs. This is particularly relevant for scenes captured with our hardware prototype and for longer-range scenes, where SBR levels change dramatically at different object distances. We characterize the sensitivity of Rapp and Goyal's method to an input SBR parameter in the supplementary document and find that for an SBR mismatch of a factor of 5 from the nominal value, RMSE on the Middlebury scenes becomes worse than our approach with the intensity image for SBRs < 0.5.

Table 1. Quantitative comparison of several depth estimation techniques for varying photon counts of signal and background (BG); the signal-to-background ratio (SBR) is indicated. All results are reported as average root-mean-square error (RMSE) over the test set containing 8 scenes. We show results for our CNN trained on a large range of signal and background ratios (A) and fine-tuned at each specific level (B). Using an intensity image with the CNN further improves the depth map. For the most challenging setting of 2 signal photons with 50 background photons, the RMSE of our method with the intensity image (A) outperforms other methods.

Avg. Photons	Avg. BG (SBR)	LM Filter	Shin [2016]	Rapp [2017]	Ours w/o Intensity (A)	Ours w/ Intensity (A)	Ours w/ Intensity (B)
10	2 (5)	0.513	0.0350	0.0170	0.0250	0.0189	**0.0143**
5	2 (2.5)	1.167	0.0662	**0.0177**	0.0317	0.0193	0.0218
2	2 (1)	2.416	0.1922	**0.0190**	0.0395	0.0248	0.0267
10	10 (1)	0.833	0.0392	**0.0172**	0.0293	0.0199	0.0176
5	10 (0.5)	1.879	0.0647	**0.0168**	0.0348	0.0206	0.0189
2	10 (0.2)	3.741	1.8484	**0.0209**	0.0464	0.0273	0.0336
10	50 (0.2)	1.757	0.5998	0.0177	0.0292	0.0217	**0.0169**
5	50 (0.1)	3.520	3.3772	**0.0195**	0.0375	0.0224	0.0438
2	50 (0.04)	5.763	4.5452	0.0384	0.0838	**0.0264**	0.1416

5 DEEP SENSOR FUSION OF RAW PHOTON COUNTS AND INTENSITY IMAGE

Features from a high-resolution intensity image can be used jointly with raw photon counts to improve the robustness of depth reconstruction and to inform upsampling of low-resolution depth maps. In the following sections, we explore using an intensity image for improved depth estimation and end-to-end training of the network for depth estimation and image-guided depth upsampling.

5.1 Sensor Fusion for Depth Estimation

Incorporating the intensity image into the depth estimation process gives distinct intuitive advantages: values of the intensity image indicate how much background noise can be expected at each spatial location, facilitating an adaptive denoising approach, and at high levels of background light, gradient information in the intensity image provides a strong signal for the structure of the reconstructed depth image.

We can therefore slightly modify the architecture described in the previous section to incorporate features from the intensity image, as shown in Figure 2. Our method is partly inspired by the general strategy of Li et al. [2016] who use a fusion network to process concatenated image features from two 2D images of differing imaging modalities. Our case, however, requires fusion of both 2D and 3D image features. The fusion step is performed by repeating the intensity image features along the temporal dimension of the 3D volume of photon counts, and concatenating it as an additional 3D feature volume. Intuitively, this allows the subsequent 3D convolutional filters to incorporate information from spatial structures in the intensity image features as they are translated across the temporal dimension. The combined features are processed with 3D convolutions in the last layers of the network. We demonstrate that this sensor fusion approach leads to improved depth estimation. The intensity image can also be used for upsampling the estimated depth image.

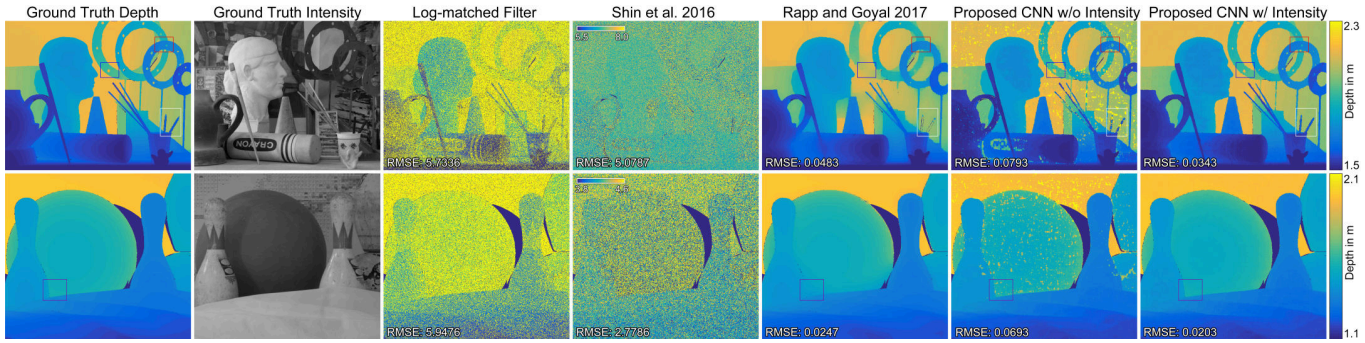


Fig. 3. Comparison of reconstruction techniques for measurements simulated with an average of 2 signal photons and 50 background photons per pixel. The proposed deep sensor fusion approach achieves the lowest error and more accurately reconstructs the geometry of thin structures and depth discontinuities. Depth estimation using the CNN without the intensity image exhibits artifacts in low-intensity regions and where large intensity variations occur. Note that the depth estimation errors of Shin's method are out of the bounds of the colormap used for the other subfigures; we colorize these two figures in a different depth scale as indicated.

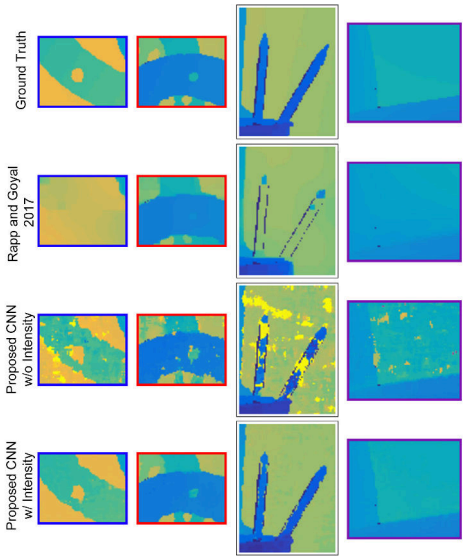


Fig. 4. Closeup views of ground truth depth and depth reconstructions from Rapp and Goyal [2017] and the proposed CNN with and without the intensity image (images cropped from indicated areas in Figure 3). The measurements are simulated with an average of 2 signal photons and 50 background photons per pixel.

5.2 Guided Depth Upsampling

Recent approaches for image-guided upsampling of depth images based on deep convolutional networks have exhibited state-of-the-art performance for this task [Hui et al. 2016; Li et al. 2016]. These approaches use a set of 2D convolutional layers to initially process the high-resolution intensity and low-resolution depth image, then process features from each image together while progressively upsampling the intermediate features to produce a high-resolution output depth map.

We explore the task of joint depth estimation and upsampling and adopt the approach of Hui et al. [2016] for image guided upsampling. However, instead of using clean depth images as input, we use the imperfect, low-resolution depth map from the depth-estimation network, given by soft argmax($\hat{\mathbf{h}}$). The resulting architecture is fully differentiable and can thus be jointly trained for depth estimation and upsampling.

To perform the upsampling, features from the intensity image are concatenated with features from the depth image at multiple resolution scales. The network predicts high-frequency differences between a bicubicly-upsampled low-pass-filtered depth image and the ground-truth high-resolution depth image. The final high-resolution image is reconstructed by adding the predicted high-frequency features to the low-pass-filtered depth image after bicubic upsampling. A diagram of the upsampling network is included in Figure 2 and illustrates the multi-scale approach. Intuitively, the approach of splitting the low and high-frequency components allows the network to focus explicitly on reconstructing high-resolution edge features.

The network is trained end-to-end for depth estimation and guided upsampling using the following loss function.

$$\mathcal{L}_{\text{up}}(\mathbf{h}, \hat{\mathbf{h}}, \mathbf{z}_{\mathbf{h}}, \hat{\mathbf{z}}_{\mathbf{h}}) = \sum_k \lambda_{\text{up}} \|\mathbf{z}_{\mathbf{h}}{}^{(k)} - \hat{\mathbf{z}}_{\mathbf{h}}^{(k)}\|_1 + D_{\text{KL}}(\mathbf{h}^{(k)}, \hat{\mathbf{h}}^{(k)}). \quad (8)$$

Here we follow Hui et al. [2016] and penalize differences between the predicted high-frequency features from the upsampling network, $\hat{\mathbf{z}}_{\mathbf{h}}$, and the ground truth high-frequency differences between the upsampled output of the depth-estimation network and the ground truth high-resolution depth image, $\mathbf{z}_{\mathbf{h}}$. We also retain the KL divergence loss from the depth-estimation network which we find improves performance. The relative weighting between the loss functions is tuned using λ_{up}.

5.3 Evaluation

We evaluate the performance of the trained depth-estimation network with sensor fusion and the jointly-trained upsampling network on the simulated measurements from the Middlebury test scenes. Details on the training procedure for each network can also be found in Appendix B.

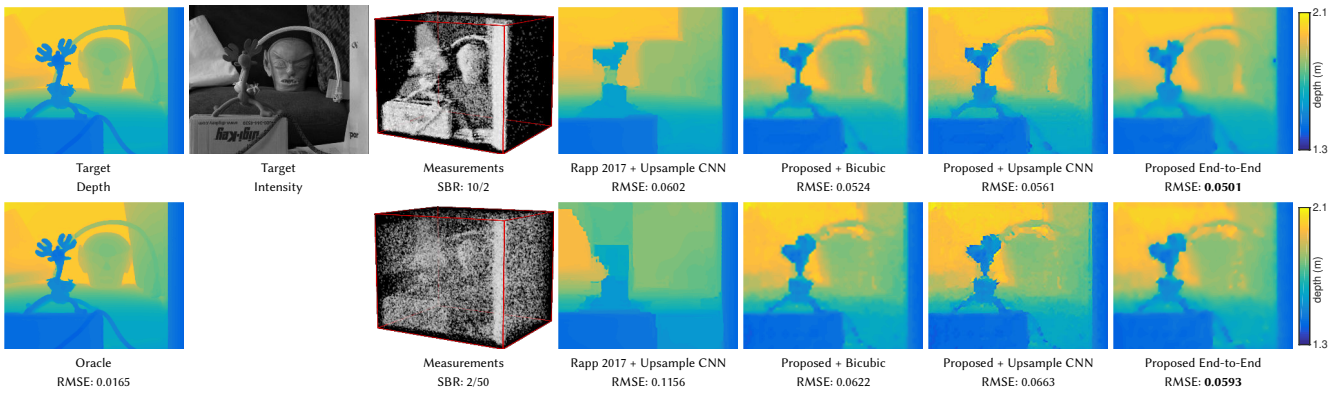


Fig. 5. Qualitative comparison of guided depth upsampling algorithms for "Reindeer" scene. In the columns, we show the measurements along with the estimated upsampled depth maps using different depth estimation and upsampling techniques: low-resolution depth estimation with Rapp and Goyal [2017] and a pre-trained network for guided upsampling [Hui et al. 2016], depth estimation with the proposed CNN and bicubic upsampling, depth estimation with the proposed CNN and pre-trained guided upsampling, and the proposed end-to-end trained network for depth estimation and guided upsampling. Ground truth depth and intensity images are shown as well as the output of an oracle, which is computed by downsampling the ground truth depth map and feeding that into the guided upsampling network. The rows compare the following settings of average signal photons / average background photons / signal-to-background ratio per pixel: 10/2/5 and 2/50/0.04.

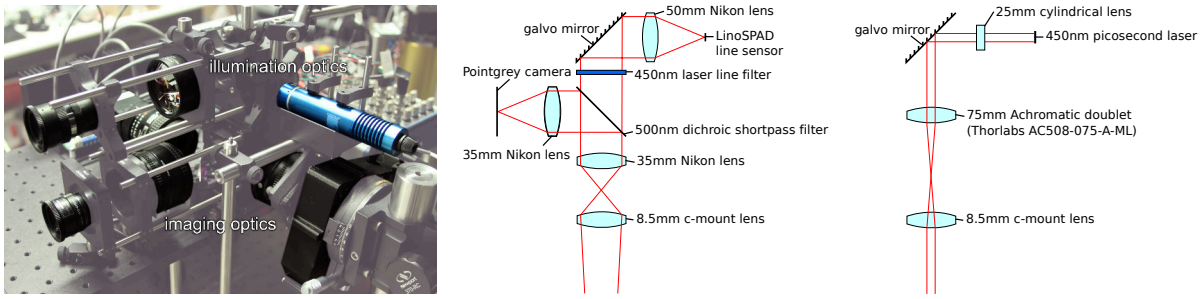
Average RMSE values for depth estimation with sensor fusion are shown in Table 1. Including the intensity image into the denoising process yields increased performance compared to the depth-estimation network without sensor fusion, suggesting that the sensor fusion provides significant quantitative gains in terms of the accuracy of depth reconstruction. We also show that fine-tuning on each individual noise level yields some improvement; however, training across a range of noise levels appears to improve in-painting ability. This is manifest by more accurate and spatially-consistent depth estimates in areas where few or no signal photons are recorded, resulting in better performance at the lowest SBR levels.

Qualitative results are shown in Figure 3 with closeup results shown in Figure 4 for two test scenes simulated with an average of 2 signal photons and 50 background photons per pixel. We also report the root-mean-square error (RMSE) values for each image. The conventional log-matched filtering approach produces a noisy result, and the method of Shin et al. [2016] produces out-of-range depth estimates for this extremely high noise level. Rapp and Goyal's approach [2017] demonstrates good performance, but depth values can be smeared at discontinuous boundaries or thin structures may not be completely reconstructed. The proposed depth estimation approach without the intensity image recovers thin structures better than other approaches, but exhibits artifacts, especially in regions with low intensity values. Finally, the proposed deep sensor fusion approach manages to accurately reconstruct the depth, even in challenging cases such as at object boundaries or for thin structures.

We also evaluate the jointly-trained upsampling method for the case of 8× upsampling (from 64×64 to 512×512) for a variety of signal and background photon counts. Results from the jointly-trained upsampling network are compared to low-resolution depth estimates generated with Rapp and Goyal [2017] and upsampled with the upsampling network, bicubic upsampling of the depth-estimation network output, and a naive case: using pre-trained depth estimation and upsampling networks without any joint training. Qualitative results are shown in Figure 5. For the method of Rapp and Goyal [2017], the initial low-resolution depth estimate fails to preserve many structures and edges, and these are not recovered after upsampling. The bicubic upsampling fails to reconstruct sharp edges in the high-resolution image. For the naive case, noisy edges remaining after the denoising and appear overly jagged after the upsampling. After jointly training the depth estimation and upsampling networks, the results show crisp edges and mitigate the artifacts exhibited by the naive approach. Finally, we show a comparison to an oracle, which shows the upsampling output from the ground-truth low-resolution depth image. Additional comparisons of denoising and upsampling using a retrained version of the model presented by Hui et al. [2016] are included in the supplement.

Table 2. Quantitative comparison of upsampling methods for varying signal and background photon counts. An average root-mean-square error (RMSE) across the 8 test scenes is reported. The end-to-end trained CNN demonstrates improved error figures compared to using a pre-trained upsampling network on the output of Rapp and Goyal [2017], bicubic upsampling of the depth-estimation network output, or using the pre-trained upsampling network on the output of the depth-estimation network. The RMSE for a signal oracle is also given, where a pre-trained upsampling network is applied to a clean, low-resolution depth map.

Avg. Photons	Avg. BG (SBR)	Rapp + Upsample CNN	Proposed + Bicubic	Proposed + Upsample CNN	Proposed End-to-end
10	2 (5)	0.0510	0.0407	0.0433	**0.0394**
5	2 (2.5)	0.0514	0.0435	0.0462	**0.0406**
2	2 (1)	0.0543	0.0503	0.0531	**0.0476**
10	10 (1)	0.0483	0.0407	0.0437	**0.0386**
5	10 (0.5)	0.0505	0.0431	0.0460	**0.0404**
2	10 (0.2)	0.0601	0.0488	0.0520	**0.0458**
10	50 (0.2)	0.0511	0.0403	0.0435	**0.0375**
5	50 (0.1)	0.0576	0.0430	0.0469	**0.0394**
2	50 (0.04)	0.0731	0.0508	0.0554	**0.0464**
Signal Oracle		0.0137			



(a) photo of setup (b) imaging optics (c) illumination optics

Fig. 6. Single-photon imaging prototype. (a) Photo of the prototype, showing both the imaging optics (bottom) and illumination optics (top). The illumination and imaging optics are aligned in a rectified setup to perform energy-efficient epipolar scanning [Achar et al. 2017; O'Toole et al. 2015]. (b) Illustration of the imaging optics (not shown to scale). A dichroic shortpass filter reflects light above 500 nm to a PointGrey vision camera, and transmits light of all remaining wavelengths through a 450 nm laser line filter and onto a 1D array of 256 SPAD pixels. The galvo mirror angle controls the scanline imaging the scene. (c) Illustration of the illumination optics (not shown to scale). A cylindrical lens creates a vertical laser line, and the galvo mirror determines the position of this laser line within the scene.

Quantitative results for the upsampling reconstructions are shown in Table 2, which reports RMSE values for a range of simulated signal and noise levels. Our approach outperforms Rapp and Goyal [2017] with upsampling, bicubic upsampling, and the naive approach in the joint depth estimation and upsampling task.

6 PROTOTYPE SYSTEM

6.1 Hardware

Our prototype 3D imaging system, shown in Figure 6(a), consists of synchronization electronics, off-the-shelf optical and optomechanical components, a standard vision camera, a picosecond laser, and a linear array of 256 SPADs (LinoSPAD [Burri et al. 2016]).

The system has two optical paths: one for generating a laser line source (Fig. 6(c)), and another for focusing the scene's response onto the LinoSPAD (Fig. 6(b)). We chose this design in order to combine single-photon sensing with epipolar scanning [Achar et al. 2017; O'Toole et al. 2015], an energy-efficient imaging technique that sequentially illuminates and images the scene one scanline at a time. We position the illumination optics directly above the imaging optics in a rectified stereo configuration as required by epipolar imaging. The laser line and sensor focus on a common vertical epipolar scanline, and a pair of galvo mirrors controls the lateral position of the scanline.

The illumination module consists of a 450 nm picosecond laser (ALPHALAS PICOPOWER-LD-450-50), a galvo mirror (Thorlabs GVS012), and a set of lenses (as specified in Figure 6(c)). A cylindrical lens spreads the collimated laser light along the vertical direction. The laser operates at a pulse repetition rate of 25 MHz with a peak power of 450 mW and average power of 0.5 mW.

The sensing module consists of a 1D SPAD array, a PointGrey camera (GS3-U3-23S6C-C), a second galvo mirror (Thorlabs GVS012) to control the imaging position of the LinoSPAD, and a set of lenses that simultaneously focus an image of the scene onto the camera sensor and LinoSPAD (see Fig. 6(b)). A dichroic shortpass filter (Edmund Optics #69-214) passes light above 500 nm to the camera sensor, and wavelengths below 500 nm to the LinoSPAD. Furthermore, a 450 nm laser line filter (Thorlabs FB450-10) reduces the amount of ambient light that reaches the LinoSPAD.

A National Instruments data acquisition device (NI-DAQ USB-6343) provides synchronization signals for the galvos, SPAD, and camera sensor. The NI-DAQ sends a 25 MHz clock signal to the SPAD, and the SPAD passes through the 25 MHz clock to trigger the laser. The NI-DAQ also provides acquisition trigger signals to the SPAD and camera, and sends a synchronized sawtooth waveform pattern to the galvos to scan the mirrors. The system captures results at 256×256 resolution at either a 20 Hz or 5 Hz scan rate. The scanning speed provides a tradeoff between exposure time or signal strength, and the acquisition speed of the system.

The LinoSPAD is limited to acquiring data with only 64 of 256 pixels at a time. As a result, by scanning at 256×256 resolution, the effective per-pixel exposure time is a maximum $\frac{64}{256^2} = \frac{1}{1024}$ of the total acquisition time of each frame. For example, at 20 Hz, the per-pixel acquisition time is approximately 50 µs. Additionally, memory latency limits the maximum number of time-stamped photon events that can be read out from the LinoSPAD for a given scan rate. A maximum of 8 time-stamped events can be recorded per pixel at 20 Hz, and 32 time stamps at 5 Hz. From the time stamps, we generate sparse histograms containing 1536 bins, where each bin has a resolution of 26 ps. We measure the full width at half maximum (FWHM) of the system to be approximately 440 ps.

6.2 Calibration

The calibration involves three steps: calibrating the time stamps generated by the LinoSPAD, aligning the laser line and LinoSPAD array of pixels, and computing the projective transformation to align the image captured by the camera and LinoSPAD sensors

The LinoSPAD generates time-stamped events that are non-uniformly distributed in time. Following the instructions from O'Toole et al. [2017] and Lindell et al. [2018], the raw photon time stamps from the LinoSPAD are processed such that all time bins are uniform in length and aligned so that the time bin corresponding to time 0 (the onset of the illumination pulse) is the same across all pixels.

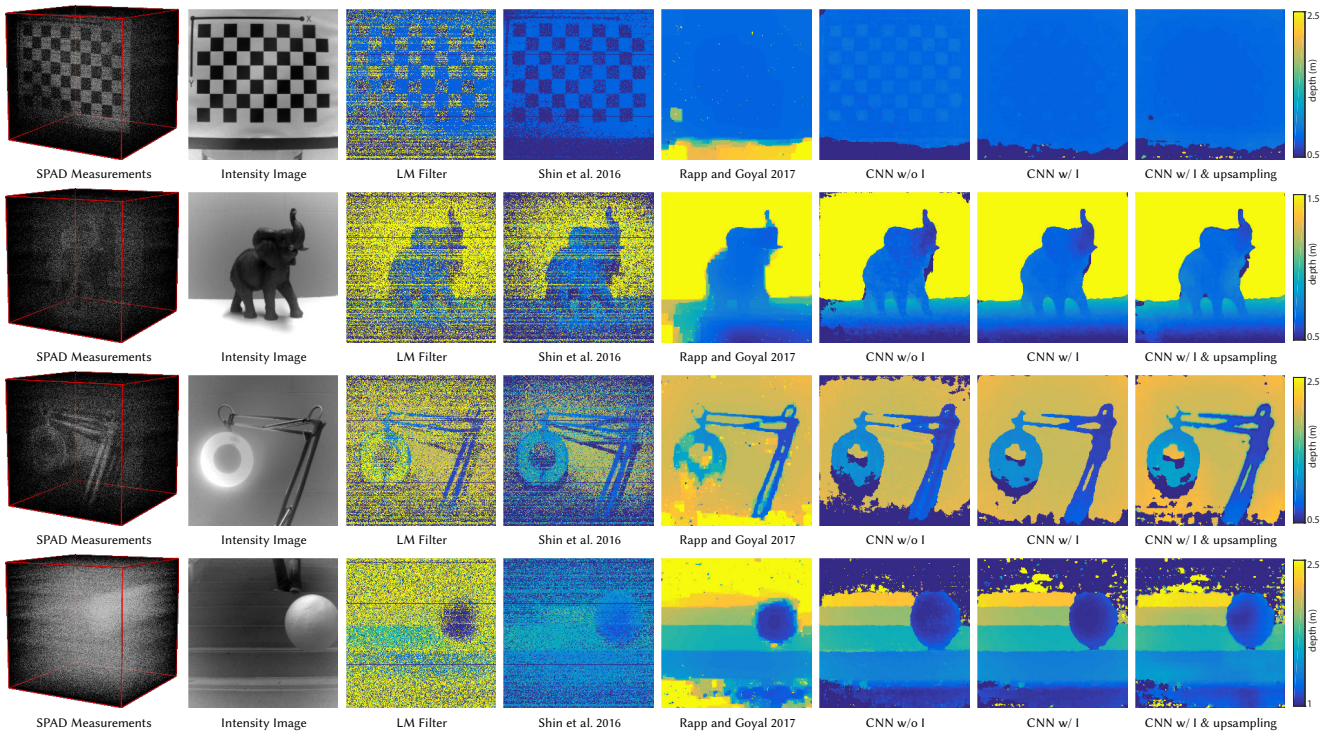


Fig. 7. Reconstruction results for four scenes: *checkerboard*, *elephant*, *lamp*, and *bouncing ball*. From left-to-right: the SPAD measurements consist of a 3D representation of the $256 \times 256 \times 1536$ spatio-temporal volume; the intensity image as captured by a regular camera sensor; a depth map reconstructed with the log-matched filter; a depth map recovered with the method proposed by Shin et al. [2016]; Rapp and Goyal [2017]; the CNN without an intensity image; the CNN with the intensity image; the CNN with the intensity image for guided upsampling.

Physically aligning the laser line with the LinoSPAD involves rotating the cylindrical lens, rotating the LinoSPAD, and adjusting the galvo mirrors such that all LinoSPAD pixels detect the laser illumination. We then choose mirror positions at the two extreme scanning positions that maximize the signal detected by the LinoSPAD, and linearly interpolate the mirror positions to determine intermediate scanning locations.

Registration of the high-resolution camera image to the LinoSPAD image involves acquiring and detecting the features of a checkerboard pattern. A projective transformation aligns the two images together. While this procedure does not explicitly account for radial distortions, the distortions are consistent for both images because the LinoSPAD and camera sensors image the scene through the same objective lens.

6.3 Implementation Details

The camera images and time-stamped measurements from the prototype are passed to the CNN in order to estimate depth. Measurements are processed with the same model parameters as used for the simulated results. Since the 3D volume of photon counts captured by the LinoSPAD is large ($256 \times 256 \times 1536$) and a limited amount of memory is available on a single GPU for processing, we process the input in 64×64 resolution patches with a stride of 32 to produce the full resolution output image. Each image patch takes approximately 5 s to process, and so the full frame (8×8 image patches) requires approximately 320 s of processing time.

7 RESULTS

We qualitatively test the performance of the CNN on measurements captured with the prototype system. Figure 7 includes four scenes referred to as *checkerboard*, *elephant*, *lamp*, and *bouncing ball*. We capture at 20 Hz for the first three scenes under office lighting, and the last scene at 5 Hz outdoors under indirect sunlight. The depth range for all scenes shown in Figure 7 is approximately 2 m; the supplementary document includes additional results that range up to approximately 4 m. All 3D measurement visualizations in the figures are generated using the Chimera renderer [Pettersen et al. 2004]. As suggested in Figure 8, the system tests the capabilities of the proposed CNN under extreme low-flux scenarios, where few signal photons ever reach the sensor.

In the checkerboard scene, the low reflectivity of the dark squares significantly reduces the number of photons detected by the SPAD. As a result, all methods that do not use the intensity image exhibit artifacts or fail to accurately recover depth in these regions. By taking the intensity image into account, we demonstrate more reliable depth estimation. For this scene, the average number of signal photons detected per pixel is 0.95, and the average SBR per pixel is 1.1. Note also that we use the calibration chart to determine the projective transformation that aligns the intensity image to the SPAD measurements.

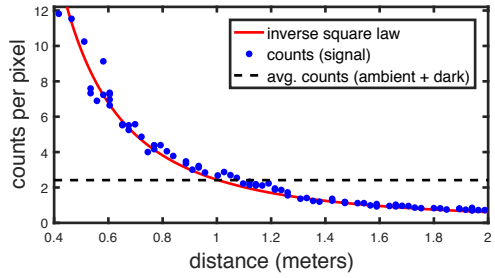


Fig. 8. Photon counts per pixel for the prototype system operating at 20 Hz. The plot shows the number of measured signal photons reflecting off a white, diffuse surface at different distances from the system, and the average counts associated with background sources (ambient light and dark noise) under office lighting. The number of signal photons detected by the SPAD closely approximates the inverse square law. Note that measurements above the black line correspond to measurements where the SBR is greater than 1.

The elephant has moderately-low reflectivity, but can be reliably reconstructed with the proposed CNN. The average signal photons detected per pixel is 0.43, and the average SBR is 0.50. This particular scene also highlights a limitation of the epipolar imaging setup for time-of-flight imaging; because the illumination and imaging optics are not co-located, certain regions within the field of view of the camera are not directly illuminated by the laser. For example, the manifestation of these shadows in the depth map can be seen in the result of Shin et al. [2016]. Our fusion technique demonstrates an ability to in-paint depth values in these regions of the scene.

The mechanical arm of the lamp contains a number of structures with high-frequency in depth. Note also that the lamp itself is turned on, significantly increasing the number of ambient photons detected by the SPAD sensor. The average signal photons detected per pixel is 0.40, and the average SBR per pixel is 0.35. While all three CNN reconstruction procedures reconstruct the lamp fairly reliably, taking the intensity image into account produces cleaner depth maps. Moreover, the upsampling procedure recovers additional high-frequency features (near the top of the lamp) within this lamp scene.

In the bouncing ball scene, a styrofoam ball bounces down the stairs under indirect sunlight. As indicated in the SPAD measurement volume, the strong contribution of sunlight results in saturated measurements (i.e., 32 time-stamped photons detected per pixel), resulting in an extremely low SBR value. The saturation also prevents a direct calculation of the SBR for the scene. Despite the strong ambient contribution, the CNN reconstruction procedures are capable of reconstructing accurate depth maps. The reconstruction procedure breaks down, however, for distances too far away from the camera (approaching 3 m).

Note that these scenes have some artifacts at the bottom and top of the image. This is a result of some misalignment between the laser illumination and sensor, causing fewer photons to be detected in this region.

8 DISCUSSION

In summary, we propose a deep sensor fusion approach to 3D imaging and show robust 3D reconstruction with very few photons returning from a laser. Compared to other methods of depth estimation from raw photon measurements, our approach demonstrates improved performance at the lowest SBR levels and qualitative improvements in reconstructing measurements captured with a hardware prototype.

The depth estimation model is trained only on simulated photon measurements from an RGB-D dataset. Although the simulated measurements ignore multipath effects and make other approximations, the trained network generalizes well to measurements captured with the hardware prototype without any alteration. While we demonstrate depth reconstruction in simulation and for captured results for a range of low SBR levels, the network may require retraining in the case of substantially different noise levels.

We apply the network directly to the output of the hardware prototype, and can accurately reconstruct depth for a number of indoor scenarios and for a limited range outdoors. The maximum range of the prototype is limited by the number of returning photons from the laser. A plot of the relationship between photons and distance is shown in Figure 8. For a distance of 2 m the SPAD records less than 1 photon per scan position on average. Non-negligible dark counts from the LinoSPAD also contribute to a decreased SBR at all ranges.

The laser used for the prototype has a relatively low average power (1 mW), and diffusing it into a line further reduces the photon flux returning to the sensor. Other systems using energy efficient epipolar scanning have used continuous-wave amplitude modulated lights sources with average powers of 500 mW, enabling considerably increased range [Achar et al. 2017]. Alternatively, concentrating the laser illumination on a single point could potentially improve range at the cost of scan speed. Indeed, recent systems have recovered depth maps at a range of over 200 m using a low average power (∼1 mW) laser, a low-noise SPAD sensor, long exposures, and a point-wise scanning approach [Tobin et al. 2017]. Pawlikowska et al. [2017] use a SPAD and pulsed laser at ∼800 m range under what appears to be direct sunlight. Their measurements capture 0.07 to 46 signal photons per pixel with SBR of 13 to 25. Our results contain similar signal photon levels and even lower SBR levels; in such long-range situations with low photon flux, our approach may be equally applicable.

The reconstructed depth estimates from our system are also upsampled with guidance from the captured high-resolution intensity image. Jointly training the depth estimation and upsampling networks shows an increase in performance over applying upsampling to state-of-the-art depth estimation techniques or using separately trained depth estimation and upsampling networks. This approach for upsampling may be especially useful in the case of SPAD sensor arrays, which capture images at lower resolutions of 32×32 or 64×64 pixels.

Finally, we comment on the tradeoffs of depth estimation with our data-driven approach compared to the approach of Rapp and Goyal [2017]. While our approach essentially relies on the neural network to learn the image formation model and noise statistics, these are explicitly incorporated into Rapp and Goyal's approach along with well-chosen heuristics. Our approach appears to better preserve fine structural details in simulation and for captured results, though Rapp and Goyal's method achieves improved RMSE in

simulation at high-resolutions with its spatially-aware smoothing and averaging techniques. However, incorporating sensor fusion into methods like that of Rapp and Goyal, is not straightforward. The proposed approach illustrates how data-driven models can be applied to depth estimation and we show a method of sensor fusion with an intensity image which improves results.

8.1 Future Work

Although our system currently operates at a range limited to within several meters, a number of methods could be used to improve the maximum range. Using a higher-power laser directly corresponds to an increase in range. Alternatively, a SPAD sensor with lower dark count rates could be used to increase the SBR. Another option is to focus the laser illumination to a point rather than a line, and to increase the scan speed of the laser. For the case where the laser point and laser line scan the same area over the same time interval, the pointwise scanning mechanism results in greater peak power with a shorter exposure time over which background photons are integrated. A disadvantage to such an approach is the more complicated alignment procedure between the fast scanning illumination source and the sensor. Finally, other imaging modalities, such as radar, could be incorporated into the sensor fusion algorithm.

Our learned approach to depth estimation can potentially also be extended for higher-level computer vision tasks. While we demonstrate coupling the depth estimation framework with image guided upsampling, additional tasks such as object classification or detection could also be trained end-to-end.

9 CONCLUSION

A fast-scanning, robust, photon-efficient depth imaging system has broad applications for 3D modeling, gesture and pose recognition, and for robotics and autonomous vehicles. In this work we demonstrate a depth sensing method that can potentially lead to improved 3D sensing through a marked increase in photon efficiency. Our robust, fast-scanning method demonstrates that using a photon-efficient method can alleviate limitations which force a tradeoff between range, frame rate, or resolution. We also demonstrate the value of a learned approach for depth estimation and sensor fusion for improving robustness and accuracy; such an approach may be useful for other low or high-level vision tasks involving 3D sensing.

ACKNOWLEDGMENTS

This project was supported by a Stanford Graduate Fellowship, a Banting Postdoctoral Fellowship, an NSF CAREER Award (IIS 1553333), a Terman Faculty Fellowship, a Sloan Fellowship, by the KAUST Office of Sponsored Research through the Visual Computing Center CCF grant, and by the DARPA REVEAL program.

REFERENCES

E. Abreu, M. Lightstone, S.K. Mitra, and K. Arakawa. 1996. A new efficient approach for the removal of impulse noise from highly corrupted images. *IEEE Trans. Image Process.* 5, 6 (1996), 1012–1025.

S. Achar, J.R. Bartels, W.L. Whittaker, K.N. Kutulakos, and S.G. Narasimhan. 2017. Epipolar time-of-flight imaging. *ACM Trans. Graph. (SIGGRAPH)* 36, 4 (2017), 37.

Y. Altmann, R. Aspden, M. Padgett, and S. McLaughlin. 2017. A Bayesian Approach to Denoising of Single-Photon Binary Images. *IEEE Trans. Computat. Imaging* 3, 3 (Sept 2017), 460–471. https://doi.org/10.1109/TCI.2017.2703900

A. Bleiweiss and M. Werman. 2009. Fusing time-of-flight depth and color for real-time segmentation and tracking. In *Dynamic 3D Imaging*. 58–69.

S. Burri, H. Homulle, C. Bruschini, and E. Charbon. 2016. LinoSPAD: A time-resolved 256×1 CMOS SPAD line sensor system featuring 64 FPGA-based TDC channels running at up to 8.5 giga-events per second. In *Proc. SPIE*, Vol. 9899. 98990D.

D. Chan, H. Buisman, C. Theobalt, and S. Thrun. 2008. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications*.

Q. Chen and V. Koltun. 2013. A simple model for intrinsic image decomposition with depth cues. In *Proc. ICCV*. 241–248.

H. Dautet, P. Deschamps, B. Dion, A.D. MacGregor, D. MacSween, R.J. McIntyre, C. Trottier, and P.P. Webb. 1993. Photon counting techniques with silicon avalanche photodiodes. *Applied optics* 32, 21 (1993), 3894–3900.

J. Diebel and S. Thrun. 2006. An application of Markov Random Fields to range sensing. In *Proc. NIPS*. 291–298.

D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. 2013. Image guided depth upsampling using anisotropic total generalized variation. In *Proc. CVPR*. 993–1000.

P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research* 31, 5 (2012), 647–663.

R. Horaud, M. Hansard, G. Evangelidis, and C. Ménier. 2016. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications* 27, 7 (2016), 1005–1020.

T. Hui, C.C. Loy, and X. Tang. 2016. Depth map super-resolution by deep multi-scale guidance. In *Proc. ECCV*. 353–369.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F.N.C. Wong, J.H. Shapiro, and V.K. Goyal. 2014. First-photon imaging. *Science* 343, 6166 (2014), 58–61.

A. Kolb, E. Barth, R. Koch, and R. Larsen. 2009. Time-of-flight sensors in computer graphics. In *Eurographics (STARs)*. 119–134.

J. Kopf, M.F. Cohen, D. Lischinski, and M. Uyttendaele. 2007. Joint bilateral upsampling. In *ACM Trans. Graph. (SIGGRAPH)*, Vol. 26. 96.

M. Koskinen, J.T. Kostamovaara, and R.A. Myllylae. 1992. Comparison of continuous-wave and pulsed time-of-flight laser range-finding techniques. In *Proc. SPIE 1614*. 296–305.

Y. Li, J. Huang, N. Ahuja, and M. Yang. 2016. Deep joint image filtering. In *Proc. ECCV*. 154–169.

G. Lin, A. Milan, C. Shen, and I. Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. CVPR*.

D.B. Lindell, M. O'Toole, and G. Wetzstein. 2018. Towards transient imaging at interactive rates with single-photon detectors. In *Proc. ICCP*.

J. Marco, Q. Hernandez, A. Muñoz, Y. Dong, A. Jarabo, M.H. Kim, X. Tong, and D. Gutierrez. 2017. DeepToF: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph. (SIGGRAPH Asia)* 36, 6 (2017), 219:1–219:12.

A. McCarthy, X. Ren, A. Della Frera, N.R. Gemmell, N.J. Krichel, C. Scarcella, A. Ruggeri, A. Tosi, and G.S. Buller. 2013. Kilometer-range depth imaging at 1550 nm wavelength using an InGaAs/InP single-photon avalanche diode detector. *Optics express* 21, 19 (2013), 22098–22113.

D. O'Connor and D. Philips. 1984. *Time-correlated single photon counting*. Academic Press.

M. O'Toole, S. Achar, S.G. Narasimhan, and K.N. Kutulakos. 2015. Homogeneous codes for energy-efficient illumination and imaging. *ACM Trans. Graph. (SIGGRAPH)* 34, 4, Article 35 (2015).

M. O'Toole, F. Heide, D.B. Lindell, K. Zang, S. Diamond, and G. Wetzstein. 2017. Reconstructing transient images from single-photon sensors. In *Proc. CVPR*.

J. Park, H. Kim, Y. Tai, M.S. Brown, and I. Kweon. 2011. High quality depth map upsampling for 3D-TOF cameras. In *Proc. ICCV*. 1623–1630.

A.M. Pawlikowska, A. Halimi, R.A. Lamb, and G.S. Buller. 2017. Single-photon three-dimensional imaging at up to 10 kilometers range. *Optics Express* 25, 10 (2017), 11919–11931.

C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. 2017. Large kernel matters- Improve semantic segmentation by global convolutional network. In *Proc. CVPR*. 1743–1751. https://doi.org/10.1109/CVPR.2017.189

G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. 2004. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph. (SIGGRAPH)* 23, 3 (2004), 664–672.

E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. 2004. UCSF Chimera-a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25, 13 (2004), 1605–1612.

J. Rapp and V.K. Goyal. 2017. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Trans. Computat. Imaging* 3 (2017), 445–459. Issue 3.

D. Renker. 2006. Geiger-mode avalanche photodiodes, history, properties and problems. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators,*

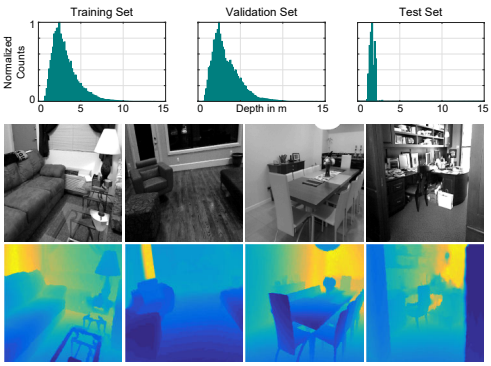


Fig. 9. Top: histograms of depth values demonstrated in the NYU v2 dataset (for training and validation) and the Middlebury dataset (for testing). Bottom: example intensity image and depth map pairs from the dataset.

Spectrometers, Detectors and Associated Equipment 567, 1 (2006), 48–56.
D. Scharstein and C. Pal. 2007. Learning conditional random fields for stereo. In *Proc. CVPR*. 1–8.
D. Shin, A. Kirmani, V.K. Goyal, and J.H. Shapiro. 2015. Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors. *IEEE Trans. Computat. Imaging* 1, 2 (2015), 112–125.
D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappa, V.K. Goyal, F.N.C. Wong, and J.H. Shapiro. 2016. Photon-efficient imaging with a single-photon camera. *Nature Communications* 7 (2016).
N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Proc. ECCV*.
S. Su, F. Heide, G. Wetzstein, and W. Heidrich. 2018. Deep end-to-end time-of-flight imaging. In *Proc. CVPR*.
R. Tobin, A. Halimi, A. McCarthy, X. Ren, K.J. McEwan, S. McLaughlin, and G.S. Buller. 2017. Long-range depth profiling of camouflaged targets using single-photon detection. *Optical Engineering* 57 (2017).
Q. Yang, R. Yang, J. Davis, and D. Nistér. 2007. Spatial-depth super resolution for range images. In *Proc. CVPR*. 1–8.

APPENDIX A SIMULATING MEASUREMENTS

To learn depth estimation and sensor fusion, we simulate SPAD measurements for a variety of scenes and illumination conditions using RGB-D images from the NYU v2 dataset captured with the Microsoft Kinect sensor [Silberman et al. 2012]. From the RGB-D images, we estimate the average photon detection parameters of Equation (2) and simulate the SPAD measurements by sampling the corresponding inhomogeneous Poisson process. For the intensity images used in sensor fusion, we estimate luminance from the RGB images.

The detection parameters of Equation (2) are described by the average number of photons arriving during each time bin from the illumination pulse (which depends on the time of flight and pulse width) plus the average number of photons arriving uniformly over time due to ambient illumination or dark counts. For each spatial location in an RGB-D image, the arrival rate function of the illumination pulse is given by calculating the time of flight and accounting for attenuation due to radial falloff and reflectance. The reflectance values of each scene are estimated using intrinsic decomposition [Chen and Koltun 2013], and we use the blue channel of the reflectance to be consistent with the blue (550 nm) laser used in the hardware prototype. Modeling the reflectance of the scene helps to account for non-Lambertian effects and spectral differences between the SPAD measurements and the intensity image.

Luminance calculated from the RGB image is used to simulate the number of background photons detected from ambient illumination throughout the scene. The luminance image also serves as the intensity image for sensor fusion. Dark counts are added to the number of background counts using a dark image captured from our hardware prototype. The detection parameters are scaled to achieve a given signal and background level. We simulate the SPAD measurement histograms with 1024 bins, a bin size of 80 ps, and a detected illumination pulse with a full width at half maximum (FWHM) of 400 ps.

To vary the signal and background levels across the dataset, we simulate an average of 2, 5, 10, and 20 signal photons detected per pixel, with 5, 10, 20, and 30 times as many background photons at each signal level. A total of 13,500 measurements are produced for training and 2,800 for validation using the NYU v2 dataset. Fine-tuning at specific noise levels, as described in Table 1, is accomplished by training on a re-generated version of this dataset where all measurements are simulated at the corresponding noise level. We also simulate measurements on a test set of 8 scenes from the Middlebury dataset [Scharstein and Pal 2007] for comparison to other reconstruction methods. Histograms of the depth distributions in training, validation, and test sets are shown in Figure 9.

APPENDIX B TRAINING THE CNN

We train the depth estimation CNNs with and without the intensity image each for 4 epochs using Adam [Kingma and Ba 2014], a learning rate of 10^{-4} and a learning rate decay of 0.9 applied after each epoch. We set the regularization strength to 10^{-5}. Training takes approximately 24 hours on an NVIDIA Titan X GPU. This network is used to process both the simulated and captured measurements. For depth estimation without guided upsampling, the intensity image is downsampled to be the same spatial resolution as the photon counts before input to the network. Results are also presented for fine-tuning the network on individual noise levels in Table 1. In this case, training is conducted for an additional 2 epochs with a learning rate of 10^{-4}.

For the simulated upsampling experiments, we train the network for 8× upsampling by initializing all layers with the pre-trained weights from the denoising network and the pre-trained upsampling network of Hui et al. [2016]. We fine-tune the network jointly for two epochs of our training dataset, upsampling random 32×32 resolution crops of the input depth image to 256×256 resolution. Training uses stochastic gradient descent with a momentum of 0.9 and a learning rate of 10^{-5}. To upsample the measurements captured with the hardware prototype, we truncate layers of the 8× upsampling network to form a 2× upsampling network [Hui et al. 2016] and downsample the intensity image by 2× for input into the denoising branch. The amount of upsampling is limited by the resolution of the vision camera (1920 × 1200), which is further reduced when the captured image is transformed and cropped to the field of view scanned by the LinoSPAD. The 2× upsampling network is trained on 32×32 image crops using the same procedure as the 8× upsampling network. For both upsampling networks, the regularization parameter used is $\lambda_{\text{up}} = 0.1$.