Disambiguating Monocular Depth Estimation with a Single Transient

Mark Nishimura[0000−0003−3976−254X], David B. Lindell[0000−0002−6999−3958], Christopher Metzler[0000−0001−6827−7207], and Gordon Wetzstein[0000−0002−9243−6885]

Stanford University, Stanford, CA
{markn1, lindell, cmetzler, gordon.wetzstein}@stanford.edu

Abstract. Monocular depth estimation algorithms successfully predict the relative depth order of objects in a scene. However, because of the fundamental scale ambiguity associated with monocular images, these algorithms fail at correctly predicting true metric depth. In this work, we demonstrate how a depth histogram of the scene, which can be readily captured using a single-pixel time-resolved detector, can be fused with the output of existing monocular depth estimation algorithms to resolve the depth ambiguity problem. We validate this novel sensor fusion technique experimentally and in extensive simulation. We show that it significantly improves the performance of several state-of-the-art monocular depth estimation algorithms.

Keywords: depth estimation, time-of-flight imaging

1 Introduction

Estimating dense 3D geometry from 2D images is an important problem with applications to robotics, autonomous driving, and medical imaging. Depth maps are a common representation of scene geometry and are useful precursors to higher-level scene understanding tasks such as pose estimation and object detection. Additionally, many computer vision tasks rely on depth sensing, including navigation [11], semantic segmentation [16, 43, 49], 3D object detection [17, 27, 48, 50, 51], and 3D object classification [32, 41, 58].

Traditional depth sensing techniques include those based on stereo or multiview, active illumination, camera motion, or focus cues [55]. However, each of these techniques has aspects that may make their deployment challenging in certain scenarios. For example, stereo or multiview techniques require multiple cameras, active illumination techniques may have limited resolution or require time-consuming scanning procedures, and other techniques require camera motion or multiple exposures at different focus distances.

One of the most promising approaches to overcoming these challenges is monocular depth estimation (MDE), which requires only a single RGB image

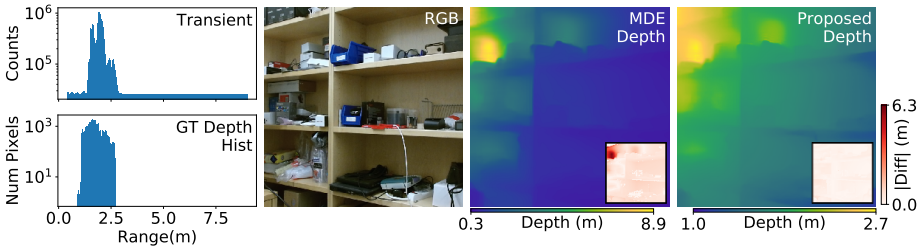

Fig. 1: Monocular depth estimation predicts a depth map (second from right) from a single RGB image (second from left). The ill-posedness of the problem prevents reliable absolute depth estimation, resulting in large errors (inset images). The proposed method uses a single transient measurement aggregating the time-of-flight information of the entire scene (leftmost) to correct the output of the depth estimation and optimize the quality of the estimated absolute depth (rightmost).

from a conventional camera to recover a dense depth map [2, 7, 9, 23, 45]. Recent approaches to MDE employ neural networks that learn to predict depth by exploiting pictorial depth cues such as perspective, occlusion, shading, and relative object size. While such models have significantly improved over recent years, MDE approaches to date are incapable of reliably estimating absolute distances in a scene due to the inherent scale ambiguities of monocular image cues. Instead, these models excel in predicting ordinal depth, or the relative ordering of objects in a scene [7, 9]. Interestingly, Alhashim and Wonka [2] recently showed that if the median ground truth depth of the scene is known, the initial output of a MDE network can be corrected to produce accurate absolute depth.

Although access to the median ground truth depth is impossible in a realistic scenario, low-cost sensors capable of capturing aggregated depth information from a scene are readily available. For example, the proximity sensor on recent generation Apple iPhones uses a low-power pulsed light source and a single-pixel time-resolved detector to sense distance to an object directly in front of the phone. Time-resolved detectors, such as avalanche photon diodes (APDs) or single-photon avalanche diodes (SPADs), can measure the full waveform of time-resolved incident radiance at each pixel (Fig. 1). These detectors form the backbone of modern LiDAR systems [22, 26, 40]. However, single-photon sensor arrays have not yet been used for 3D imaging on consumer electronics, primarily because the requirement for ultra-fast timing electronics makes it difficult to produce high-resolution arrays at low cost and because the scanning requirement for single-pixel systems introduces a point of mechanical failure and complicates high-resolution, high-framerate imaging.

Here, we propose to use a single-pixel time-resolved detector and pulsed light source in an unconventional way: rather than optically focusing them to record the distance to a single scene point, we diffuse the emitted light and aggregate the reflected light over the entire scene with the detector. The resulting transient

measurement resembles a histogram of the scene's depth and can be used to achieve accurate absolute depth in conjunction with a monocular depth estimate (Fig. 1).

To this end, we develop a sensor fusion strategy that processes the ordinal depth computed by a monocular depth estimator to be consistent with the measurements captured by the aggregated time-resolved detector. We demonstrate in extensive simulations that our approach achieves substantial improvements in the quality of the estimated depth maps, regardless of which specific depth estimator is used. Moreover, we build a camera prototype that combines an RGB camera and a single-pixel time-resolved detector and use it to validate the proposed depth estimation technique.

In summary, we make the following contributions:

- We propose augmenting an RGB camera with a global depth transient aggregated by a time-resolved detector to address scale ambiguity in MDE.
- We introduce a depth reconstruction algorithm that uses the detector's image formation model in conjunction with a modified version of histogram matching, to produce a depth map from a single RGB image and transient. The algorithm can be applied instantly to any existing and future MDE algorithms.
- We analyze this approach on indoor scenes using the NYU Depth v2 dataset and demonstrate that our approach is able to resolve scale ambiguity while being fast and easy to implement.
- We build a prototype camera and evaluate its efficacy on captured data, assessing both the quality and the ability of our method to improve generalization of monocular depth estimators across scene types.

2 Related Work

Monocular Depth Estimation Estimating a depth map from a single RGB image has been approached using Markov Random Fields [45], geometric approaches [20], and non-parametric, SIFT-based methods [21]. More recently, deep neural networks have been applied to this problem, for example using a multi-scale neural network to predict depth maps [7], using an unsupervised approach that trains a network using stereo pairs [12], and using a logarithmic depth discretization scheme combined with an ordinal regression loss function [9]. Various experiments using different types of encoder networks (*e.g.*, ResNet, DenseNet) [2, 23] have also been employed with some success, as have approaches mixing deep learning with conditional random fields [60], and attention-based approaches [18, 61]. Recently, Lasinger et al. [25] improved the robustness of monocular depth estimation using cross-dataset transfer.

Despite achieving remarkable success on estimating ordinal depth from a single image, none of these methods is able to resolve inherent scale ambiguity in a principled manner. We introduce a new approach that leverages existing monocular depth estimation networks and disambiguates the output using

depth histogram–like measurements obtained from a single time-resolved detector. Other approaches to disambiguating monocular depth estimation use optimized freeform lenses [6, 57] or dual-pixel sensors [10], but these approaches require custom lenses or sensors and specialized image reconstruction methods. In contrast, our approach adds minimal additional hardware to a single RGB camera, and may leverage sensors currently deployed in consumer electronics.

Depth Imaging and Sensor Fusion with Time-resolved Detectors Emerging LiDAR systems use avalanche photon diodes (APDs) or single-photon avalanche diodes (SPADs) to record the time of flight of individual photons. These time-resolved detectors can be fabricated using standard CMOS processes, but the required time-stamping electronics are challenging to miniaturize and fabricate at low cost. For this reason, many LiDAR systems, especially those using SPADs, use a single or a few detectors combined with a scanning mechanism [22, 24, 26, 40, 15]. Unfortunately, this makes it challenging to scan dynamic scenes at high resolution and scanners can also be expensive, difficult to calibrate, and prone to mechanical failure. To reduce the scanning complexity to one dimension, 1D detector arrays have been developed [3, 4, 38], and 2D SPAD arrays are also an active area of research [35, 52, 56, 62]. Yet, single-pixel time-resolved detectors remain the only viable option for low-cost consumer devices today.

The proposed method uses a single-pixel APD or SPAD and pulsed light source that are diffused across the entire scene instead of aimed at a single point, as with proximity sensors. This unique configuration captures a measurement that closely resembles the depth histogram of the scene. Our sensor fusion algorithm achieves reliable absolute depth estimation by combining the transient measurement with the output of a monocular depth estimator using a histogram matching technique. While other recent work also explored RGB-SPAD sensor fusion [28, 53, 1], the RGB image was primarily used to guide the denoising and upsampling of measurements from a SPAD array.

Histogram Matching and Global Hints Histogram matching is a well-known image processing technique for adjusting an image so that its histogram matches some pre-specified histogram (often derived from another image) [13, 14]. Nikolova et al. [36] use optimization to recover a strict ordering of the image pixels, yielding an exact histogram match. Morovic et al. [34] provide an efficient and precise method for fast histogram matching which supports weighted pixel values. In the image reconstruction space, Swoboda and Schnörr [54] use a histogram to form an image prior based on the Wasserstein distance for image denoising and inpainting. Rother et al. [44] use a histogram prior to create an energy function that penalizes foreground segmentations with dissimilar histograms. In the area of non-line-of-sight imaging [8, 29–31, 39], Caramazza et al. [5] use a single non-line-of-sight transient to recover the identity of a person hidden from view. In a slightly different application area, Zhang et al. [63] train a neural network to produce realistically colorized images given only a black-and-white image and a histogram of global color information.

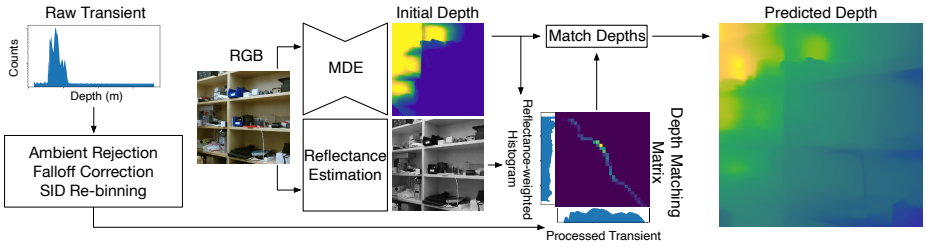

Fig. 2: Overview of processing pipeline. The processing pipeline uses the input transient measurement and an RGB image to produce an accurate depth map. The transient is preprocessed to adjust for ambient photon detections, radiometric falloff factors, and to calibrate the bin widths. From the RGB image, an MDE estimates an initial depth map and the scene reflectance is estimated. A reflectance-weighted depth histogram is compared to the processed transient to calculate a histogram matching matrix which is used to output the corrected depth.

In our procedure, the transient measurements closely resemble a histogram of the depth map where the histogram values are weighted by spatially varying scene reflectances and inverse-square falloff effects. We therefore adapt the algorithm in Morovic et al. [34] in order to accommodate general per-pixel weights during histogram matching.

3 Method

In this section, we describe the image formation of a diffused pulsed laser and time-resolved detector. Although our model is derived for the specific case of imaging with a single-photon avalanche diode (SPAD), the resulting image formation model equally applies to other time-resolved detectors. We also describe an approach for correcting a depth map generated with a monocular depth estimator to match the global scene information captured by the transient.

3.1 Image Formation Model of a Diffused SPAD

Consider a diffused laser that emits a pulse at time $t = 0$ with time-varying intensity $g(t)$ illuminating some 3D scene. We parameterize the geometry of the scene as a distance map $z(x, y)$, where each of the 3D points has also some unknown reflectivity α at the wavelength of the laser. Ignoring interreflections of the emitted light within the scene, a single-pixel diffused SPAD integrates light scattered back from the scene onto the detector as

$$s\left(t\right) = \int_{\Omega_x} \int_{\Omega_y} \frac{\alpha\left(x, y\right)}{z(x, y)^2} \cdot g\left(t - \frac{2z(x, y)}{c}\right) dxdy, \tag{1}$$

where c is the speed of light, $\Omega_{x,y}$ is the spatial extent of the diffused light, and we assume that the light is diffused uniformly over the scene points. Each time

such a light pulse is emitted into the scene and scattered back to the detector, the single-pixel SPAD time-stamps up to one of the returning photons with some probability. The process is repeated millions of times per second with the specific number of emitted pulses being controlled by the repetition rate of the laser. As derived in previous work, the resulting measurement can be modeled as an inhomogeneous Poisson process $\mathcal{P}$ [22, 46, 47]. Each detected photon arrival event is discretized into a histogram h of the form

$$h[n] \sim \mathcal{P}\left(\eta \int_{n\Delta t}^{(n+1)\Delta t} (f * s)(t)\, dt + b\right), \tag{2}$$

where $[n\Delta t, (n+1)\Delta t)$ models the n^{th} time interval or bin of the temporal histogram, η is the photon detection probability of the SPAD, f is a function that models the temporal uncertainty in the detector, and b represents background detections from ambient light and false positive detections known as *dark count*. Like previous work, we neglect scene interreflections and confine ourselves to the low-flux condition (where the number of photon detections is controlled to be much smaller than the number of emitted pulses) to avoid pileup [47]. Finally, we adopt the term *transient* for the histogram $h[n]$ [59].

3.2 Ambient Rejection and Falloff Correction

Before performing histogram matching, we apply three preprocessing steps to (1) remove background counts from the transient, (2) compensate for distance falloff effects, and (3) re-bin the transient to improve relative accuracy with increasing distance. An overview of the processing pipeline, including these preprocessing steps and the histogram matching procedure is depicted in Figure 2.

Background Subtraction. In the first step, we remove the background counts from the transient by initially estimating the average amount of background counts in each time bin. For nearly all natural scenes, the closest objects to the camera are a finite distance away, and so the first bins of the SPAD measurement contain only background counts without any backscattered signal. We can therefore estimate the average number of background and noise counts $\hat{b}$ as

$$\hat{b} = \frac{1}{N} \sum_{n=0}^{N} h[n]. \tag{3}$$

where we choose the number of bins N to correspond to time values before the backscattered signal arrives.

While simply subtracting $\hat{b}$ from the measurements would remove many of the background counts, a large number of bins containing only background counts would still have non-zero values, resulting in a skewed estimate after applying histogram matching. Instead, we estimate the temporal support of transient bins containing signal photons (*i.e.*, the range of depths in the scene) and only subtract $\hat{b}$ from these bins (clipping negative bin values to 0). We assume that other transient bins contain only background counts that can be discarded.

Specifically, we identify the first and last bins that record backscattered signal photons by locating discontinuities in the recorded counts [59]. An initial spike in the measurements at bin n_{first} results from the onset of backscattered signal from the closest object, and a steep dropoff occurs after bin n_{last} after backscattered photons from the furthest object are recorded. We estimate n_{first} and n_{last} by calculating first order differences of the transient $d[n] = |h[n] - h[n+1]|$. For a moderate number of background counts, each background bin $h[n]$ can be approximated as a Gaussian with mean and variance b, and thus $h[n] - h[n+1]$ can be approximated as a Gaussian with mean 0 and variance $2b$. We identify candidate discontinuities $\mathcal{E}$ with a threshold on the measured differences:

$$\mathcal{E} = \left\{ n : d[n] > \beta\sqrt{2\hat{b}} \right\}. \tag{4}$$

We find that $\beta = 5$ yields good results across both simulated and captured data.

Initial estimates n'_{first} and n'_{last} are set to the minimum value in $\mathcal{E}$ and the maximum value, incremented by one bin. Then, we refine these estimates by selecting the closest bins that remain above a threshold τ such that

$$\begin{aligned} \hat{n}_{\text{first}} &= \min\{n : h[n] > \tau, h[n+1] > \tau, \cdots, h[n'_{\text{first}}] > \tau\} \\ \hat{n}_{\text{last}} &= \max\{n : h[n'_{\text{last}}] > \tau, \cdots, h[n-1] > \tau, h[n] > \tau\}. \end{aligned} \tag{5}$$

The remaining ambient counts are discarded by setting the recorded counts to zero for all bins where $n < \hat{n}_{\text{first}}$ and $n > \hat{n}_{\text{last}}$. We use $\tau = \hat{b} + \sqrt{\hat{b}}$ in all of our experiments.

Falloff Compensation. In the second step, we compensate for distance falloff effects by multiplying the transient by the distance-dependent scaling factor,

$$h'[n] = h[n] \cdot z_n^2. \tag{6}$$

Here, $z_n = \left(n + \frac{1}{2}\right)\left(\frac{c\Delta t}{2}\right)$ is the distance corresponding to bin n, and this radiometric falloff model is consistent with measurements captured with our prototype.

Transient Re-binning. Last, we re-bin the transient so that the bin widths increase for increasingly distant objects. We select the Spacing-Increasing Discretization (SID) method of [9], which changes the bin widths according to an exponential function, allocating more bins to closer distances and fewer bins to farther distances for a fixed number of bins. The bin edges t_i are given by the following equation, parameterized by the number of bins K and the range of distances $[\ell, u]$:

$$t_i = e^{\log(\ell) + \frac{\log(u/\ell) \cdot i}{K}} \qquad \text{for} \qquad i = 0, \ldots, K. \tag{7}$$

This rebinning procedure allows us to use a reduced number of bins in the histogram matching, reducing computation time while maintaining accuracy. For the simulated results we use $K = 140$ bins with (ℓ, u) corresponding to the depth values of bins $\hat{n}_{\text{first}}$ and $\hat{n}_{\text{last}}$ respectively. The output of the rebinning procedure is the target histogram h_{target} which we use for histogram matching.

3.3 Histogram Matching

Histogram matching is a procedure that adjusts pixel values from an input image so that the image histogram matches a target histogram. We apply this procedure to match the histogram of an input depth map, obtained from a monocular depth estimator, to the post-processed target histogram h_{target} from the SPAD. This initialize-then-refine approach allows us to swap out the monocular depth estimator to deal with different scene types without requiring end-to-end retraining.

The input depth map cannot be directly histogram-matched to the target histogram because the target histogram incorporates the spatially varying reflectance of the scene. To account for reflectance in the histogram matching procedure, we use the normalized image color channel closest to the laser wavelength as an estimate of the reflectance and compute a reflectance-weighted depth histogram h_{source}; instead of incrementing a bin in the depth histogram by one for every pixel in the MDE at the corresponding depth, we add the estimated reflectance value of the pixel to the histogram bin. We also re-bin this histogram, following Fu et al. and using $K = 140$ with $(\ell, u) = (0.657, 9.972)$ [9].

We match the re-binned histogram h_{source} to h_{target} using the method of Morovic et al. [34]. The method involves computing a pixel movement matrix T such that $T[m, n]$ is the fraction of $h_{\text{source}}[m]$ that should be moved to $h_{\text{target}}[n]$. We refer the reader to the supplement for pseudocode. Intuitively, the procedure starts from the first bin of the source histogram and distributes its contents to the first bins of the target histogram, with successive source histogram bins being shifted to successive target bins in sequence.

Finally, we use the movement matrix T to shift the pixels of the input depth map to match the global depth of the target histogram. For a depth map pixel with depth bin k, we select the corrected bin by sampling from the distribution $T[k, :] / \sum_{n=1}^{N} T[k, n]$. This sampling procedure handles the case where a single input depth bin of the MDE is mapped to multiple output bins [34].

Pseudo-code for this procedure is included in the supplement; we will make source code and data available.

4 Evaluation and Assessment

4.1 Implementation Details

We use the NYU Depth v2 dataset to evaluate our method. This dataset consists of 249 training and 215 testing RGB-D images captured with a Kinect.

To simulate a transient, we take the provided depth map and calculate a weighted depth histogram by weighting the pixel contributions to each depth bin by the luminance of each pixel. To model radiometric falloff, we multiply each bin by $1/z^2$, and convolve with a modeled system temporal response, which we approximate as a Gaussian with a full-width at half-maximum of 70 ps. We scale the histogram by the total number of observed signal photon counts (set to 10^6) and add a fixed number of background photons $b \in \{2 \times 10^5, 10^5, 2 \times 10^4, 10^4\}$.

	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$	$rel \downarrow$	$rmse \downarrow$	$log10 \downarrow$
DORN	0.846	0.954	0.983	0.120	0.501	0.053
DORN + median rescaling	0.871	0.964	0.988	0.111	0.473	0.048
DORN + GT histogram matching	0.906	**0.972**	**0.990**	0.095	0.419	0.040
Proposed (SBR=5)	0.902	0.970	0.989	0.092	0.423	0.040
Proposed (SBR=10)	0.905	0.971	**0.990**	0.090	0.413	**0.039**
Proposed (SBR=50)	0.906	0.971	**0.990**	**0.089**	**0.408**	**0.039**
Proposed (SBR=100)	**0.907**	0.971	**0.990**	**0.089**	**0.408**	**0.039**
DenseDepth	0.847	0.973	0.994	0.123	0.461	0.053
DenseDepth + median rescaling	0.888	0.978	**0.995**	0.106	0.409	0.045
DenseDepth + GT histogram matching	**0.930**	**0.984**	**0.995**	**0.079**	**0.338**	**0.034**
Proposed (SBR=5)	0.922	0.981	0.994	0.083	0.361	0.036
Proposed (SBR=10)	0.924	0.982	**0.995**	0.082	0.352	0.035
Proposed (SBR=50)	0.925	0.983	**0.995**	0.081	0.347	0.035
Proposed (SBR=100)	0.926	0.983	**0.995**	0.081	0.346	0.035
MiDaS + GT histogram matching	**0.801**	**0.943**	**0.982**	**0.149**	**0.558**	**0.062**
Proposed (SBR=5)	0.792	0.937	0.978	0.153	0.579	0.064
Proposed (SBR=10)	0.793	0.937	0.979	0.152	0.572	0.064
Proposed (SBR=50)	0.794	0.938	0.979	0.151	0.570	0.063
Proposed (SBR=100)	0.794	0.938	0.979	0.151	0.570	0.064

Table 1: Quantitative evaluation using NYU Depth v2. Bold indicates best performance for that metric, while underline indicates second best. The proposed scheme outperforms DenseDepth and DORN on all metrics, and it closely matches or even outperforms the median rescaling scheme and histogram matching with the exact depth map histogram, even though those methods have access to ground truth. Metric definitions can be found in [7].

The background counts are evenly distributed across all bins to simulate the ambient and dark count detections, and the different background levels correspond to signal-to-background ratios (SBR) of $5, 10, 50$ and 100 respectively. Finally, each bin is Poisson sampled to produce the final simulated transient.

4.2 Simulated Results

We show an extensive quantitative evaluation in Table 1. Here, we evaluate three recent monocular depth estimation CNNs: DORN [9], DenseDepth [2], and MiDaS [25]. To evaluate the quality of DORN and DenseDepth, we report various standard error metrics [7]. Moreover, we show a simple post-processing step that rescales their outputs to match the median ground truth depth [2]. We also show the results of histogram matching the output of the CNNs with the ground truth depth map histogram. Note that we do not report the quality of the direct output of MiDaS as this algorithm does not output metric depth. However, we do show its output histogram matched with the ground truth depth

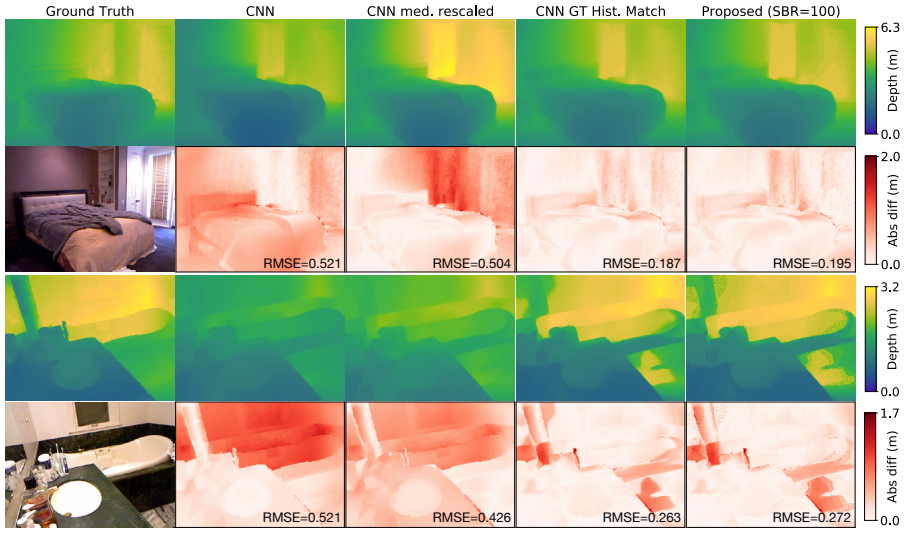

Fig. 3: Simulated results from NYU Depth v2 computed with the DenseDepth CNN [2]. The depth maps estimated by the CNN are reasonable, but contain systematic error. Oracle access to the ground truth depth maps, either through the median depth or the depth histogram, can remove this error and correct the depth maps. The proposed method uses a single transient measurement and does not rely on ground truth depth, but it achieves a quality that closely matches the best-performing oracle.

map histogram. In all cases, post-processing the estimated depth maps either with the median depth or depth histogram significantly improves the absolute depth estimation, often by a large margin compared to the raw output of the CNNs. Unfortunately, ground truth depth is typically not accessible so neither of these two post-processing methods are viable in practical application scenarios.

Instead, our method uses the simulated measurements from a single aggregated transient to correct the depth map. In Table 1, results are shown for several different signal-to-background ratios (SBRs). We see that the proposed method achieves high-quality results for correcting the raw depth map estimated by the respective CNNs for all cases. The quality of the resulting depth maps is almost as good as that achieved with the oracle ground truth histogram, which can be interpreted as an approximate upper bound on the performance, despite a relatively high amount of noise and background signal. These results demonstrate that the proposed method is agnostic to the specific depth estimation CNN applied to get the initial depth map and that it generally achieves significant improvements in the estimated depth maps, clearly surpassing the variation in performance between depth estimation CNNs.

In Figure 3, we also show qualitative results of our simulations. For each of these scenes, we show the RGB reference image, the ground truth depth map,

the raw output of the DenseDepth CNN, the result of rescaling the CNN output with the median ground truth depth, the result of histogram-matching the CNN output by the ground truth depth map histogram, and the result achieved by the proposed method for an SBR of 100. Error maps for all the depth estimation methods are shown. As expected, the CNN outputs depth maps that look reasonable but that have an average root mean squared error (RMSE) of about 50–60 cm. Rescaling this depth map to match the median ground truth depth value slightly improves the quality and histogram-matching with the ground truth depth histogram shows a large amount of improvement. The quality of the proposed method is close to using the oracle histogram, despite relying on noisy transient measurements. Additional simulations using DenseDepth and other depth estimation CNNs for a variety of scenes are shown in the supplement.

5 Experimental Demonstration

5.1 Prototype RGB-SPAD Camera Hardware

As shown in Figure 4, our prototype comprises a color camera (Microsoft Kinect v2), a single-pixel SPAD (Micro Photon Devices 100 μm PDM series, free-running), a laser (ALPHALAS PICOPOWER-LD-450-50), and a two-axis galvanometer mirror system (Thorlabs GVS012). The laser operates at 670 nm with a pulse repetition rate of 10 MHz with a peak power of 450 mW and average power of 0.5 mW. The ground truth depth map is raster-scanned at a resolution of 512×512 pixels, and the single transient is generated by summing all of these measurements for a specific scene. This allows us to validate the accuracy of the proposed histogram matching algorithm, which only uses the integrated single histogram, by comparing it with the captured depth. To verify that our digitally aggregated scanned SPAD measurements match measurements produced by an optically diffused SPAD (see Figure 4(b,c)), we set up a slightly modified version of our prototype consisting of both scanned and optically diffused SPADs side-by-side. Additional details about the hardware prototype can be found in the supplement.

We determined camera intrinsics and extrinsics for the Kinect's RGB camera and the scanning system using MATLAB's camera calibration toolbox. The SPAD histogram and RGB image were captured from slightly different viewpoints; we account for this in the SPAD histogram by shifting the 1D transient according to the SPAD's offset from the RGB camera. We re-bin the captured 1D transient for the indoor captured results using Equation 7 with $K = 600$ bins, and $(\ell, u) = (0.4, 9.)$. For the outdoor captured result, we use $K = 600$ and $(\ell, u) = (0.4, 11)$.

5.2 Experimental Results

Using the hardware prototype, we captured a number of scenes as shown in Figures 5, 6, and in the supplement. We crop the RGB image to have dimensions that are multiples of 32. For DORN only, we further downsample the image to

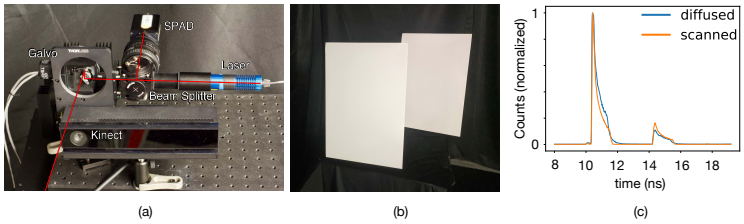

Fig. 4: (a) Prototype scanning setup. The pulsed light from the laser travels through a beam splitter before being guided by the galvo to the scene. Returning light is measured by the single-pixel SPAD. The Kinect v2 RGB camera is used to capture the image used to generate the monocular depth estimate (the depth camera is not used). (b) Scene and (c) measurements for diffused and summed scanned mode (10s capture, aligned peaks). The observed counts in the diffuse mode match closely with the sum of the raster-scanned measurements.

a resolution of 353×257. We then feed this RGB image into the monocular depth estimation algorithm. In Figure 5 we show a subset of the scenes we captured and processed with MiDaS [25], which achieved the best results among the depth estimators we tested. Additional scenes, also processed with other MDE approaches, including DenseDepth [2] and DORN [9], are included in the supplement. The ground truth depth is captured with the scanned SPAD, as described above, and regions with low signal-to-noise ratio are masked out (shown in black).

In the first two examples, the "Hallway" and "Conference Room" scenes, we see that the monocular depth CNN estimates the ordinal depth of the scene reasonably well. However, the root mean squared error (RMSE) for these two scenes is relatively high ranging from 2.6–3.2 m (see red/white error maps in Fig. 5). The proposed method using a single diffused SPAD measurement corrects this systematic depth estimation error and brings the RMSE down to 0.6–0.9 m. The "Poster" scene is meant to confuse the CNN—it shows a flat poster with a printed scene. As expected, the CNN predicts that the statue is closer than the arches in the background, which is incorrect in this case. The proposed method uses the SPAD histogram to correctly flatten the estimated depth map.

Figure 6 shows the RGB image of a scene along with the monocular depth estimate computed by MiDaS and depth maps corrected by our method using both the digitally aggregated transients from the scanned SPAD and the single optically diffused measurement, which are very similar.

6 Discussion

In summary, we demonstrate a method to greatly improve depth estimates from monocular depth estimators by correcting the scale ambiguity errors inherent with such techniques. Our approach produces depth maps with accurate absolute depth, and helps MDE neural networks generalize across scene types, including

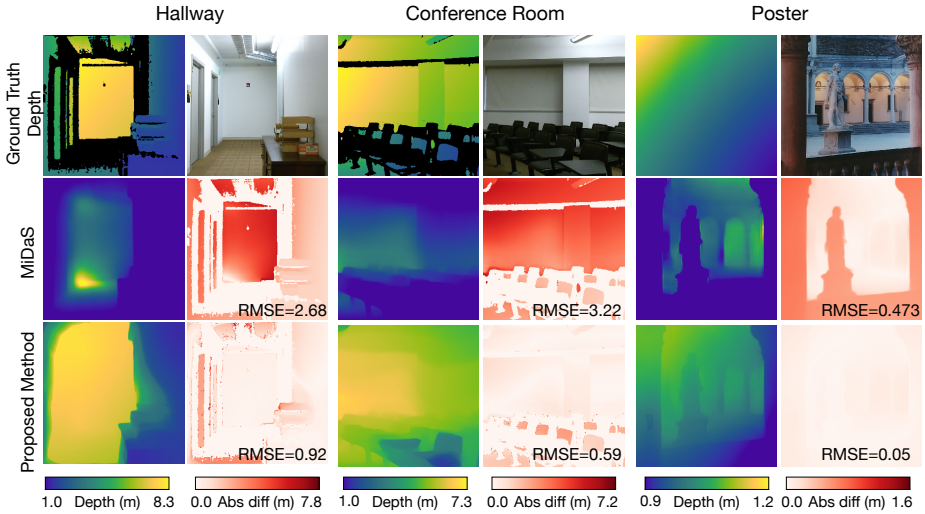

Fig. 5: Experimental results. For each scene, we record a ground truth depth map that is raster-scanned with the SPAD (upper left subimages), and an RGB image (upper right subimages). A monocular depth CNN predicts an initial depth map (top middle left subimages), which is corrected with the digitally aggregated SPAD histogram using the proposed method (bottom left subimages), as shown by the error maps and root mean squared error (RMSE) for each example (middle left, bottom subimages). The CNN is confused when we show it a photograph of a poster (rightmost scene); it incorrectly predicts the depth of the scene depicted on the flat print. Our method is able to correct this error.

on data captured with our hardware prototype. Moreover, we require only minimal additional sensing hardware; we show that a single measurement histogram from a diffused SPAD sensor contains enough information about global scene geometry to correct errors in monocular depth estimates.

The performance of our method is highly dependent on the accuracy of the initial depth map of the MDE algorithm. Our results demonstrate that when the MDE technique produces a depth map with good ordinal accuracy, where the ordering of object depths is roughly correct, the depth estimate can be corrected to produce accurate absolute depth. However, if the ordering of the initial depths is not correct, these errors may propagate to the final output depth map.

In the optically diffused configuration, the laser power is spread out over the entire scene. Accordingly, for distant scene points very little light may return to the SPAD, making reconstruction difficult (an analogous problem occurs with dark objects). Thus, our method is best suited to short- to medium-range scenes. On the other hand, in bright environments, pileup will ultimately limit the range of our method. However, this can be mitigated with optical elements to reduce the amount of incident light, with pileup correction [19, 42], or even by taking two transient measurements, one with and one without laser illumination, and using

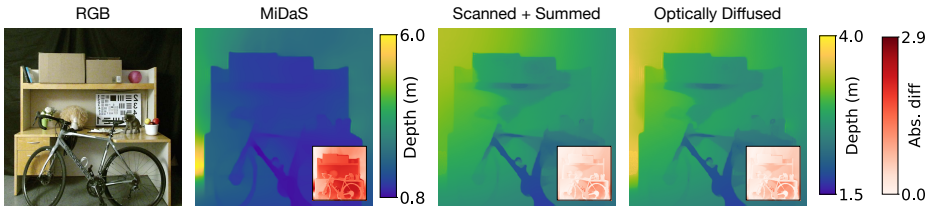

Fig. 6: Captured result comparing the direct output of the MiDaS MDE and depth maps corrected by our method using the digitally aggregated transient of the scanned SPAD (center right) and a single transient captured by an optically diffused SPAD and laser (right). Both of these approaches result in very similar results and both are significantly better than the output of the MDE, as shown by the error maps in the insets. The diffused SPAD results are captured at ∼25mW laser power indoors.

their difference to approximate the background-free transient. Finally, under normal indoor conditions, it is theoretically possible to achieve an SBR of 5 at a range of 3 meters with a laser of only 21 mW while remaining in the low-flux regime. We confirm this empirically with our diffused setup, which operates without significant pileup effects while using approximately 25 mW of laser power (see Figure 6 and the supplement for details).

Future Work Future work could implement our algorithm or similar sensor fusion algorithms on smaller platforms such as existing cell phones with single-pixel SPAD proximity sensors and RGB cameras. Necessary adjustments, such as pairing near-infrared (NIR) SPADs with NIR sensors, could be made. The small baseline of such sensors would also mitigate the effects of shading and complex BRDFs on the reflectance estimation step. More sophisticated intrinsic imaging techniques could also be employed.

Conclusions Since their introduction, monocular depth estimation algorithms have improved tremendously. However, recent advances, which have generally relied on new network architectures or revised training procedures, have produced only modest performance improvements. In this work we dramatically improve the performance of several monocular depth estimation algorithms by fusing their estimates with transient measurements. Such histograms are easy to capture using time-resolved single-photon detectors and are poised to become an important component of future low-cost imaging systems.

Acknowledgments

D.L. was supported by a Stanford Graduate Fellowship. C.M. was supported by an ORISE Intelligence Community Postdoctoral Fellowship. G.W. was supported by an NSF CAREER Award (IIS 1553333), a Sloan Fellowship, by the KAUST Office of Sponsored Research through the Visual Computing Center CCF grant, and a PECASE by the ARL.

References

1. Ahmad Siddiqui, T., Madhok, R., O'Toole, M.: An extensible multi-sensor fusion framework for 3d imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1008–1009 (2020)
2. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv:1812.11941v2 (2018)
3. Burri, S., Bruschini, C., Charbon, E.: Linospad: A compact linear SPAD camera system with 64 FPGA-based TDC modules for versatile 50 ps resolution time-resolved imaging. Instruments **1**(1), 6 (2017)
4. Burri, S., Homulle, H., Bruschini, C., Charbon, E.: 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: Optical Sensing and Detection IV. vol. 9899, p. 98990D. International Society for Optics and Photonics (2016)
5. Caramazza, P., Boccolini, A., Buschek, D., Hullin, M., Higham, C.F., Henderson, R., Murray-Smith, R., Faccio, D.: Neural network identification of people hidden from view with a single-pixel, single-photon detector. Sci. Rep. **8**(1), 11945 (2018)
6. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3D object detection. In: Proc. ICCV (2019)
7. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Proc. NeurIPS (2014)
8. Faccio, D., Velten, A., Wetzstein, G.: Non-line-of-sight imaging. Nature Reviews Physics pp. 1–10 (2020)
9. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proc. CVPR (2018)
10. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proc. ICCV (2019)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
12. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proc. CVPR (2017)
13. Gonzales, R., Fittes, B.: Gray-level transformations for interactive image enhancement. Mech. Mach. Theory **12**(1), 111–122 (1977)
14. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2008)
15. Gupta, A., Ingle, A., Velten, A., Gupta, M.: Photon-flooded single-photon 3D cameras. In: Proc. CVPR. IEEE (2019)
16. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: Proc. CVPR (2013)
17. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Proc. ECCV (2014)
18. Hao, Z., Li, Y., You, S., Lu, F.: Detail preserving depth estimation from a single image using attention guided networks. In: Proc. 3DV (2018)
19. Heide, F., Diamond, S., Lindell, D.B., Wetzstein, G.: Sub-picosecond photon-efficient 3D imaging using single-photon sensors. Sci. Rep. **8**(17726) (2018)
20. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM Trans. Graph. **24**(3), 577–584 (2005)
21. Karsch, K., Liu, C., Kang, S.: Depth transfer: Depth extraction from video using non-parametric sampling. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2144–2158 (2014)

22. Kirmani, A., Venkatraman, D., Shin, D., Colaço, A., Wong, F.N., Shapiro, J.H., Goyal, V.K.: First-photon imaging. Science **343**(6166), 58–61 (2014)
23. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: Proc. 3DV. IEEE (2016)
24. Lamb, R., Buller, G.: Single-pixel imaging using 3D scanning time-of-flight photon counting. SPIE Newsroom (2010)
25. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv:1907.01341 (2019)
26. Li, Z.P., Huang, X., Cao, Y., Wang, B., Li, Y.H., Jin, W., Yu, C., Zhang, J., Zhang, Q., Peng, C.Z., et al.: Single-photon computational 3D imaging at 45 km. arXiv preprint arXiv:1904.10341 (2019)
27. Lin, D., Fidler, S., Urtasun, R.: Holistic scene understanding for 3D object detection with RGBD cameras. In: Proc. ICCV (2013)
28. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3D imaging with deep sensor fusion. ACM Trans. Graph. (SIGGRAPH) **37**(4), 113 (2018)
29. Lindell, D.B., Wetzstein, G., O'Toole, M.: Wave-based non-line-of-sight imaging using fast f–k migration. ACM Trans. Graph. **38**(4), 1–13 (2019)
30. Liu, X., Bauer, S., Velten, A.: Phasor field diffraction based reconstruction for fast non-line-of-sight imaging systems. Nat. Commun. **11**(1), 1–13 (2020)
31. Liu, X., Guillén, I., La Manna, M., Nam, J.H., Reza, S.A., Le, T.H., Jarabo, A., Gutierrez, D., Velten, A.: Non-line-of-sight imaging using phasor-field virtual wave optics. Nature **572**(7771), 620–623 (2019)
32. Maturana, D., Scherer, S.: Voxnet: A 3D convolutional neural network for real-time object recognition. In: Proc. IROS (2015)
33. McManamon, P.: Review of ladar: a historic, yet emerging, sensor technology with rich phenomenology. Optical Engineering **51**(6), 060901 (2012)
34. Morovic, J., Shaw, J., Sun, P.L.: A fast, non-iterative and exact histogram matching algorithm. Pattern Recognit. Lett. **23**(1-3), 127–135 (2002)
35. Niclass, C., Rochas, A., Besse, P.A., Charbon, E.: Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. IEEE J. Solid-State Circuits **40**(9), 1847–1854 (2005)
36. Nikolova, M., Wen, Y.W., Chan, R.: Exact histogram specification for digital images using a variational approach. J. Math. Imaging. Vis. **46**(3), 309–325 (2013)
37. O'Connor, D.V., Phillips, D.: Time-correlated single photon counting. Academic Press (1984)
38. O'Toole, M., Heide, F., Lindell, D.B., Zang, K., Diamond, S., Wetzstein, G.: Reconstructing transient images from single-photon sensors. In: Proc. CVPR (2017)
39. O'Toole, M., Lindell, D.B., Wetzstein, G.: Confocal non-line-of-sight imaging based on the light-cone transform. Nature **555**(7696), 338–341 (2018)
40. Pawlikowska, A.M., Halimi, A., Lamb, R.A., Buller, G.S.: Single-photon three-dimensional imaging at up to 10 kilometers range. Opt. Express **25**(10), 11919–11931 (2017)
41. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view CNNs for object classification on 3D data. In: Proc. CVPR (2016)
42. Rapp, J., Ma, Y., Dawson, R.M.A., Goyal, V.K.: Dead time compensation for high-flux depth imaging. In: Proc. ICASSP (2019)
43. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. In: Proc. CVPR (2012)

44. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In: Proc. CVPR (2006)
45. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Proc. NeurIPS (2006)
46. Shin, D., Kirmani, A., Goyal, V.K., Shapiro, J.H.: Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors. IEEE Trans. Computat. Imag. **1**(2), 112–125 (2015)
47. Shin, D., Xu, F., Venkatraman, D., Lussana, R., Villa, F., Zappa, F., Goyal, V., Wong, F., Shapiro, J.: Photon-efficient imaging with a single-photon camera. Nat. Commun. **7**, 12046 (2016)
48. Shrivastava, A., Gupta, A.: Building part-based object detectors via 3D geometry. In: Proc. ICCV (2013)
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Proc. ECCV (2012)
50. Song, S., Xiao, J.: Sliding shapes for 3D object detection in depth images. In: Proc. ECCV (2014)
51. Song, S., Xiao, J.: Deep sliding shapes for amodal 3D object detection in RGB-D images. In: Proc. CVPR (2016)
52. Stoppa, D., Pancheri, L., Scandiuzzo, M., Gonzo, L., Dalla Betta, G.F., Simoni, A.: A CMOS 3-D imager based on single photon avalanche diode. IEEE Trans. Circuits Syst. I, Reg. Papers **54**(1), 4–12 (2007)
53. Sun, Z., Lindell, D.B., Solgaard, O., Wetzstein, G.: Spadnet: Deep RGB-SPAD sensor fusion assisted by monocular depth estimation. Opt. Express **28**(10), 14948–14962 (2020)
54. Swoboda, P., Schnörr, C.: Convex variational image restoration with histogram priors. SIAM J. Imaging Sci. **6**(3), 1719–1735 (2013)
55. Szeliski, R.: Computer vision: Algorithms and applications. Springer Science & Business Media (2010)
56. Veerappan, C., Richardson, J., Walker, R., Li, D.U., Fishburn, M.W., Maruyama, Y., Stoppa, D., Borghetti, F., Gersbach, M., Henderson, R.K.: A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter. In: Proc. ISSCC (2011)
57. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: PhaseCam3D—Learning phase masks for passive single view depth estimation. In: Proc. ICCP (2019)
58. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: A deep representation for volumetric shapes. In: Proc. CVPR (2015)
59. Xin, S., Nousias, S., Kutulakos, K.N., Sankaranarayanan, A.C., Narasimhan, S.G., Gkioulekas, I.: A theory of Fermat paths for non-line-of-sight shape reconstruction. In: Proc. CVPR (2019)
60. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In: Proc. CVPR (2017)
61. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proc. CVPR (2018)
62. Zhang, C., Lindner, S., Antolovic, I., Wolf, M., Charbon, E.: A CMOS SPAD imager with collision detection and 128 dynamically reallocating TDCs for single-photon counting and 3D time-of-flight imaging. Sensors **18**(11) (2018)

63. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. **9**(4) (2017)