Improving Monocular Depth Estimation with Global Depth Histogram Matching

Anonymous CVPR submission

Paper ID ****

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

1.1. Introduction

Estimating depth from images is an important open problem with applications to robotics, autonomous driving, and medical imaging. Dense depth maps are useful precursors to higher-level scene understanding tasks such as pose estimation and object detection.

However, traditional approaches to depth estimation, such as stereo, suffer from lower performance when confronted with small angles or faraway objects. More exotic approaches use FMCW or time-of-flight LiDAR technologies, but these approaches are currently expensive and bulky.

The most promising solution to these issues uses deep learning and convolutional neural networks to perform *monocular depth estimation*, estimating dense depth maps from single RGB images. However, this problem is underconstrained due to *inherent scale ambiguity*, the unresolvable tradeoff between size and distance in single images. In practice, this issue commonly manifests itself in many monocular depth networks, and indeed, Wonka et. al. (cite) showed that if the method has oracle access to the ground truth median depth, then correcting the output of the CNN to match this median depth produces better depth maps both qualitatively and quantitatively.

In this paper, we go further and show that by augmenting the RGB image with a histogram of global image depths, we can achieve substantially improved performance (and generalizability) over state-of-the-art monocular depth estimators. By performing an exact, weighted histogram matching on the output depth map of the depth estimator, we can match the depth histogram of the scene to the depth histogram of our estimate. This histogram matching is described in (cite) and is flexible enough to accommodate different pixel reflectances in the RGB image. Finally, this histogram can be captured relatively inexpensively using only a single pixel single-photon avalanche diode (SPAD) and pulsed laser illumination diffused over the field of view, representing a significant improvement in cost and simplicity over multi-pixel LiDAR arrays with expensive scanning mechanisms. It is worth noting that SPADs of this type have already made their way into existing smartphones, such as the iPhone X, and will likely play a role in future mobile sensing platforms as well.

Our method is not without limitations. It still requires a laser and single-pixel LiDAR detector, and as such, is sensitive to ambient photons. Being a variant of histogram matching, our method is unable to transpose the values of pixels (i.e. if pixel a is farther than pixel b in the input, it will be farther than pixel b in the output). In other words, our method is not able to resolve ordinal depth errors (errors where an object is wrongly placed closer or farther relative to another object). Finally, our method is non-differentiable, and is therefore unsuitable for end-to-end optimization of multi-part networks.

- We introduce the idea of augmenting an RGB camera with a global depth histogram to address scale ambiguity error in monocular depth estimators.
- We analyze our approach on indoor scenes using the NYU Depth v2 dataset. We demonstrate that our approach is able to resolve scale ambiguity while being fast and easy to implement.
- We build a hardware prototype and evaluate the efficacy of our approach on real-world data, assessing both the the quality and the ability of our method to help generalization of monocular depth estimators across scene types.

2. Related Work

Monocular Depth Estimation Previous non-deep approaches used Markov Random Fields [22], geometric approaches [11], and non-parametric, SIFT-based methods [13]. More recently, deep neural networks have been applied to the problem of estimating depth from a single image. Eigen et. al. [2] use a multi-scale neural network to predict depth at multiple scales. Godard et. al. [5] use an unsupervised approach (i.e. that does not require ground truth depth) that trains a network using stereo pairs to produce disparity maps from single images, which can then be used to recover the depth. Fu et. al. [6] combined a logarithmic depth discretization scheme with a novel ordinal regression approach. Various experiments using different types of encoder networks (e.g. ResNet, DenseNet) [3] [9] have also been employed with some success, as have approaches mixing deep learning with conditional random fields [4], and attention-based approaches [8] [7].

Despite achieving remarkable success on the monocular depth estimation task, none of these methods are able to resolve inherent scale ambiguity in a principled manner (being monocular in nature). By combining a monocular depth estimator with a depth histogram, our method is able to do so.

Depth Imaging and Sensor Fusion with SPADs Previous work (see [12] for a survey) has been able to use single-pixel SPADs [14] and also 1D LinoSPADs in tandem with various scanning or DMD devices to capture 3D volumes of photon arrivals that can be used to reconstruct depth. Lindell et. al. [15] use a LinoSPAD and epipolar scanline and fuse the SPAD data with an RGB image to produce high-quality depth. Our approach uses a single pixel SPAD but does not require any scanning or DMD mechanism.

A parallel approach called 3D flash LiDAR uses a laser with an optical diffuser as the illumination source and a 2D array of SPADs to capture the 3D volume [19, 17]. Such arrays are capable of reconstructing high quality depth but remain relatively low resolution. Other arrays are able to achieve higher resolution, but suffer from low fill factor [21] or sacrifice per-pixel TDC [23].

Our approach uses flash LiDAR but requires only a single pixel sensor. (Still need to flesh out this section more with some higher resolution SPAD arrays)

Histogram Matching and Global Hints Histogram matching or histogram specification is a well-known image processing technique [10] for adjusting an image so that it's histogram matches some pre-specified histogram (often derived from another image). Nikolova et. al. [18] use optimization to recover a strict ordering of the image pixels that allows an exact histogram match to be obtained. Morovic et. al [16] provide a simple and concise method that also achieves an exact histogram match while being very fast. In the image reconstruction space, Swoboda and Schnörr [20] use a histogram to form an image prior based on the Wasserstein distance for image denoising and inpainting. Rother et. al. [1] use a histogram prior to create an energy function that penalizes foreground segmentations with dissimilar histograms. In a slightly different vein, Zhang et. al. [24] train a neural network to produce realistically colorized images given only a black-and-white image and a histogram of global color information.

Our method is essentially a modified form of the algorithm in [16], modified for our particular use case. Also worth noting is the fact that most algorithms compute histograms from existing images, whereas our method mesaures the depth histogram indirectly using photon arrivals.

3. Method

In this section, we begin by outlining some core ideas of monocular depth estimation. We then describe the measurement model for a single-pixel time-of-flight lidar sensor under diffuse, pulsed laser illumination. Finally, we explain our histogram matching procedure.

3.1. Monocular Depth Estimation

To acquire an initial depth estimate from our RGB image, we need a monocular depth estimator. In this paper, we investigate two networks, DORN [6] and DenseDepth [9]. We choose these for their high individual performance on the monocular depth estimation task on both NYU Depth v2 and KITTI. To clarify, however, for both networks, separate weights are learned for NYU Depth v2 and for the KITTI dataset, and no models are provided that have been trained on both datasets. We will use $f(\cdot)$ to refer to the neural network.

We pass our RGB image I through the network f to obtain z_{init}, our initial depth estimate. Let us leave this for now and return to it after discussing our measurement model.

3.2. Measurement Model

Consider a laser which emits a pulse at time $t = 0$ with time-varying intensity $g(t)$ uniformly illuminating some scene. We parameterize the geometry of the scene as a height map $z(x, y)$. Neglecting albedo and falloff effects, as well as radial corrections, an ideal detector counting photon events from a location (x, y) in the time interval $(n\Delta t, (n + 1)\Delta t)$ would record

$$\lambda_{x,y}[n] = \int_{n\Delta t}^{(n+1)\Delta t} (f * g)(t - 2z(x, y)/c)dt \quad (1)$$

where c is the speed of light, and f is a function that models the temporal uncertainty in the detector. Single-photon avalanche diodes (SPADs) are highly sensitive photodetectors which are able to record single photon events with high temporal precision [**?**]. Since the event corresponding to the detection of a photon can be described with a Bernoulli random variable, the total number of accumulated photons in this time interval follows a Poisson distribution according to

$$h[n] \sim \mathcal{P}\left(\sum_{x,y} \alpha_{x,y}\eta\lambda_{x,y}[n] + b\right) \tag{2}$$

where $\alpha_{x,y} = r_{x,y}/z(x,y)^2$ captures the attenuation of the photon counts due to the reflectance $r(x,y)$ of the scene and due to the inverse square falloff $1/z(x,y)^2$. In addition, η is the detection probability of a photon triggering a SPAD event, and $b = \eta a + d$ is the average number of background detections resulting from ambient photons a and erroneous "dark count" events d resulting from noise within the SPAD.

3.3. Monocular depth estimation with global depth hints

Given a single RGB image $I(x,y)$ and a vector of photon arrivals $h[n]$ described by equation 3, we seek to reconstruct the ground truth depth map $z(x,y)$. Our method has two parts. First, we **initialize** our estimate of the depth map from the single RGB image via a monocular depth estimator described below. Second, we **refine** this depth map using the captured measurements $h[n]$ via exact histogram matching.

Initialization The first step in our method is to produce an initial estimate of ground truth depth. Convolutional Neural Networks have been shown to produce accurate, if poorly-scaled, estimates of depth from only a single image. We therefore choose to initialize our depth map estimate $\hat{z}^{(0)}(x,y)$ using a CNN. However, any depth estimator reliant on only a single view may be used for this step. Furthermore, in the larger context of our algorithm, it is more important that the network predict the correct ordinal relationships between pixels - that is, to predict the correct relative ordering of pixels a and b, rather than to get all pixels exactly correct.

Exact Histogram Matching An image's *histogram* is a pair of vectors (h,b) where h_i is the number of pixels of the image whose value lies in the range $[b_i, b_i+1)$. Then, given a source image S with histogram (h_s,b) and a target histogram (h_t,b), histogram matching generates a new image M such that $h_m \approx h_t$ and the pixel values in M are in the same relative order as in S. The full details of the exact histogram matching algorithm can be found in [16].

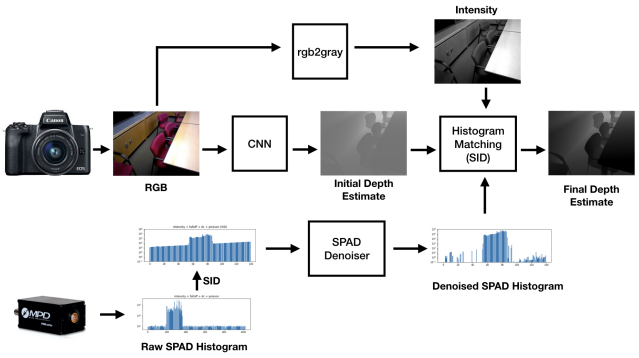

Figure 1: **Overview of the full pipeline** We use a CNN to get an initial per-pixel depth estimate. Then we perform exact histogram matching using intensity-weighted pixel values on the corrected SPAD data.

However, for our purposes, we need to modify our algorithm to accommodate differing per-pixel weights. We can account for squared depth falloff

SPAD Denoising

- Talk about histogram matching in the ideal case, jump straight to intensity
- Talk about histogram matching in our case, and how it approaches the ideal case. Discuss the following corrections
  - Ambient/DC - Use [**?**] to justify looking for large edges, then the ambient estimate to get rid of the noise floor.
  - Falloff
- Talk about how the histogram matching works with intensity considerations applied, briefly.
- We don't address jitter or poisson noise.

$$h[n] \sim \mathcal{P}\left(\sum_{x,y} \alpha_{x,y}\eta\lambda_{x,y}[n] + b\right) \tag{3}$$

Given a SPAD with histogram h according to the above equation, we first process the SPAD to remove the effects of some of the terms. First, we

3.4. Implementation Details

For the Monocular Depth Estimator, we use pretrained versions of the the Deep Ordinal Regression Network (DORN) [] and the DenseDepth Network. The exact histogram matching method is as described in [].

4. Simulation

4.1. Implementation Details

- Number of bins used, depth range, laser parameters, use of intensity image.
- Using

NYU Depth v2 The NYU Depth v2 Dataset consists of 249 training and 215 testing scenes of RGB-D data captured using a Microsoft Kinect. We used a version of DORN pre-trained according to [6] as our CNN.

5. Hardware Prototype

5.1. Setup

- Description of hardware used
- Images of scenes used

6. Discussion

References

[1] *Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs*, volume 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) 1. IEEE, 2006.

[2] *Depth map prediction from a single image using a multi-scale deep network*, volume Advances in neural information processing systems, 2014.

[3] *Deeper depth prediction with fully convolutional residual networks*, volume 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016.

[4] *Multi-scale continuous crfs as sequential deep networks for monocular depth estimation*, volume Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[5] *Unsupervised monocular depth estimation with left-right consistency*, volume Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[6] *Deep ordinal regression network for monocular depth estimation*, volume Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[7] *Detail Preserving Depth Estimation from a Single Image Using Attention Guided Networks*, volume 2018 International Conference on 3D Vision (3DV). IEEE, 2018.

[8] *Structured attention guided convolutional neural fields for monocular depth estimation*, volume Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[9] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv*, page 1812.11941v2, 2018.

[10] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. 2008.

[11] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. ACM transactions on graphics(TOG):577–584, 2005.

[12] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016.

[13] K Karsch, C Liu, and SB Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans Pattern Anal Mach Intell*, 36(11):2144–2158, 2014.

[14] Robert Lamb and Gerald Buller. Single-pixel imaging using 3d scanning time-of-flight photon counting. *SPIE Newsroom*, 2010.

	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$	$rel \downarrow$	$rmse \downarrow$	$log10 \downarrow$
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=10)	0.903	0.970	0.989	0.091	0.422	0.040
Proposed (SBR=50)	**0.906**	0.971	**0.990**	**0.089**	0.410	**0.039**
Proposed (SBR=100)	**0.906**	0.971	**0.990**	0.090	**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=10)	0.922	0.982	0.994	0.082	0.361	0.036
Proposed (SBR=50)	0.925	0.983	**0.995**	0.081	0.348	0.035
Proposed (SBR=100)	0.926	0.983	**0.995**	0.081	0.346	0.035

Table 1: Simulated Results on 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 even outperforms the median rescaling scheme, which has access to the true median depth value.

[15] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. Single-photon 3d imaging with deep sensor fusion. *ACM Trans. Graph.*, page 1812.11941v2, 2018.

[16] Jan Morovic, Julian Shaw, and Pei-Li Sun. A fast, non-iterative and exact histogram matching algorithm. *Pattern Recognition Letters*, 23(1-3):127–135, 2002.

[17] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon. Design and characterization of a cmos 3-d image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9):1847–1854, 2005.

[18] Mila Nikolova, You-Wei Wen, and Raymond Chan. Exact histogram specification for digital images using a variational approach. *Journal of Mathematical Imaging and Vision J Math Imaging Vis*, 46(3):309–325, 2013.

[19] David Stoppa, Lucio Pancheri, Mauro Scandiuzzo, Lorenzo Gonzo, Gian-Franco Dalla Betta, and Andrea Simoni. A cmos 3-d imager based on single photon avalanche diode. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(1):4–12, 2007.

[20] Paul Swoboda and Christoph Schnörr. Convex variational image restoration with histogram priors. *SIAM Journal of Imaging Sciences*, 6(3):1719–1735, 2013.

[21] Chockalingam Veerappan, Justin Richardson, Richard Walker, Day-Uey Li, Matthew W Fishburn, Yuki Maruyama, David Stoppa, Fausto Borghetti, Marek Gersbach, and Robert K Henderson. A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter. 2011 IEEE International Solid-State Circuits Conference:312–314, 2011.

[22] Y. Weiss, B. Schölkopf, and J. C. Platt, editors. *Learning depth from single monocular images*, volume Advances in neural information processing systems. MIT Press, 2006.

[23] C Zhang, S Lindner, IM Antolovic, M Wolf, and E Charbon. A cmos spad imager with collision detection and 128 dynamically reallocating tdcs for single-photon counting and 3d time-of-flight imaging. *Sensors (Basel)*, 18(11), 2018.

[24] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017.

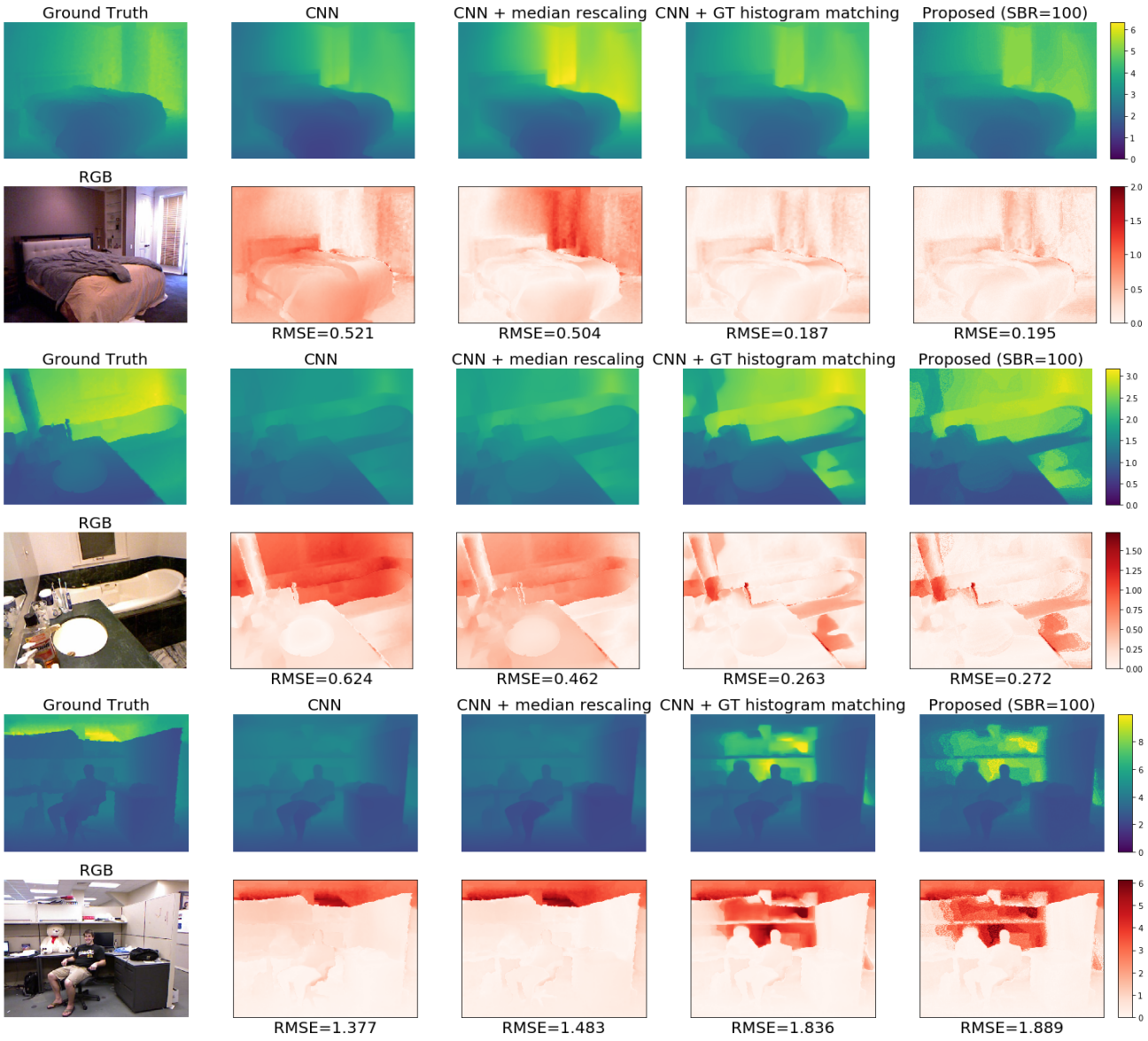

Figure 2: Selected Results on DenseDepth. First two examples demonstrate capability of proposed method to correct initial scaling/translation errors. Last example shows potential pitfall when ordinal depth is predicted incorrectly.

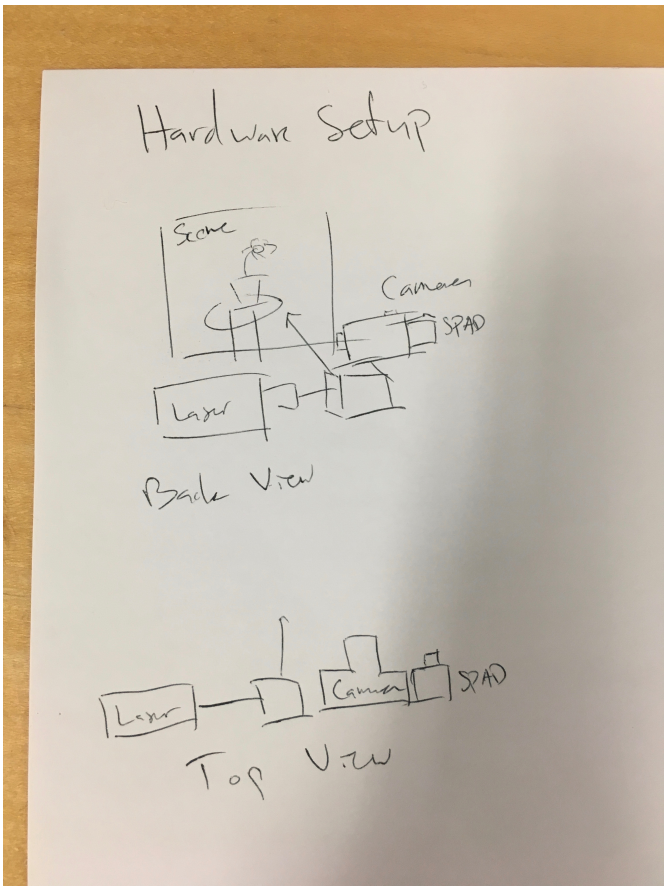

Figure 3: **Hardware setup**

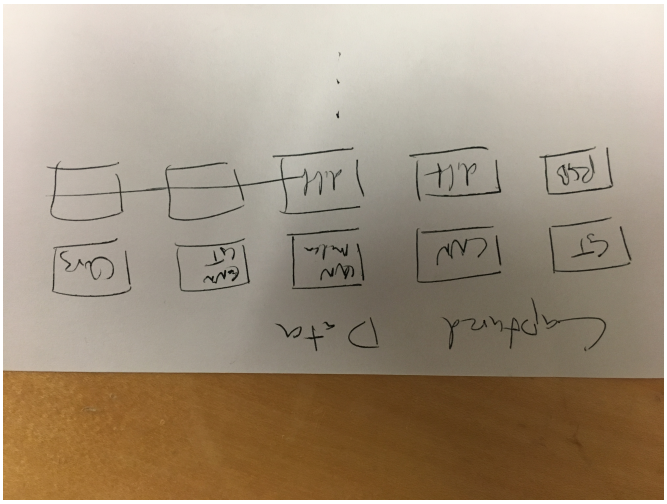

Figure 4: **Hardware results**