

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

L^AT_EX Author Guidelines for CVPR Proceedings

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

contributions

- We introduce the idea of augmenting an RGB camera with a single-pixel SPAD 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.

2. Related Work

Depth Imaging

- stereo and multiview
- structured illumination and random patterns (kinect, etc.), active stereo
- time of flight (continuous wave and pulsed)
- what we do: like pulsed but much simpler setup; no scanning, no spad array, ...

Monocular Depth Estimation

- summary of architectures and cost functions: u-net type architecture with reverse huber loss
- what we do: same thing, but augment with global hints (inspired by these approaches, we do ...)

Deep Sensor Fusion global hints for super-resolution, colorization, depth estimation

- colorization
- david’s 2018 paper for depth estimation and denoising (see david’s 2019 sig paper for related work)
- what we do: slightly different application

Histogram Matching Histogram matching as an image processing technique

- Exact histogram matching paper used in this work
- Wasserstein-based optimization techniques for histogram-based regularization

3. Method

In this section, we describe the measurement model for a single-pixel time-of-flight lidar sensor under diffuse, pulsed laser illumination.

3.1. Measurement Model

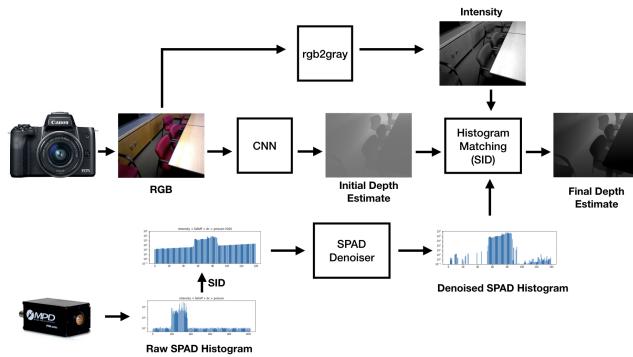
Consider a laser which emits a pulse at time $t = 0$ with time-varying intensity $g(t)$ uniformly illuminating some 3D scene. We parameterize the geometry of the scene as a height map $z(x, y)$. Neglecting albedo and falloff effects, an ideal detector counting photon events from a location (x, y) in the time interval $(n\Delta t, (n + 1)\Delta t)$ would record

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

108 where c is the speed of light, and f is a function that
 109 models the temporal uncertainty in the detector. Single-
 110 photon avalanche diodes (SPADs) are highly sensitive photo-
 111 detectors which are able to record single photon events
 112 with high temporal precision [?]. Since the detection of
 113 each photon can be described with a Bernoulli random variable,
 114 the total number of accumulated photons in this time
 115 interval follows a Poisson distribution according to
 116

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

120 where $\alpha_{x,y} = r_{x,y}/z(x,y)^2$ captures the attenuation of
 121 the photon counts due to the reflectance $r(x,y)$ of the scene
 122 and due to the inverse square falloff $1/z(x,y)^2$. In addition,
 123 η is the detection probability of a photon triggering a SPAD
 124 event, and $b = \eta a + d$ is the average number of background
 125 detections resulting from ambient photons a and erroneous
 126 “dark count” events d resulting from noise within the SPAD.
 127



128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161

Figure 1: **Overview of the full pipeline** We use a CNN to get an initial per-pixel depth estimate. We then perform gradient descent to optimize that estimate using the SPAD forward model and the dual-Sinkhorn distance .

3.2. Monocular depth estimation with global depth hints

162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 Given a single RGB image $I(x,y)$ and a vector of photon arrivals $h[n]$ described by equation 2, 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 a process we call Differentiable Histogram Matching (DHM). Differentiable histogram matching is a tool for post-processing the image to match the depth map to the statistics we capture from the SPAD.

196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
Initialization via CNN Convolutional Neural Networks have become increasingly capable of leveraging monocular depth cues to produce accurate 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 will be unable to resolve the inherent scale ambiguity in the scene resulting from the tradeoff between size of and distance to an object. The next step, differentiable histogram matching, will resolve this ambiguity using the depth information present in the SPAD histogram.

SPAD Denoising

- Discuss MLE for SPAD denoising
- Write optimization problem for SPAD denoising
- Show performance on a few examples

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231

232

233
234
235
236
237
238
239
240
241
242
243
244

Raw Depth Histogram

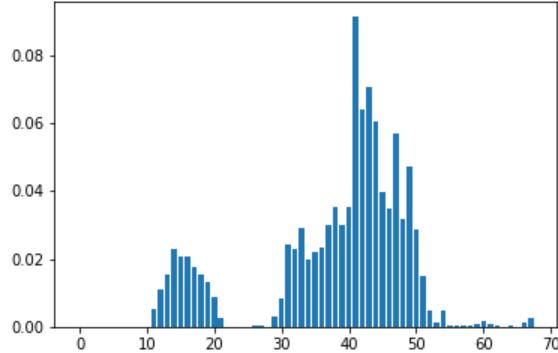
245
246
247
248
249
250

Table 1: Sample Image. Top Left is the RGB image. Top Right is ground truth depth. Bottom Left is Raw ground truth depth histogram. Bottom Right is simulated SPAD measurements. Notice how closer depths are magnified and far depths are attenuated.

251

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 .

259

260

261

262

263

264

265

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. Hardware Prototype

4.1. Setup

- Description of hardware used
- Images of scenes used

5. Assessment

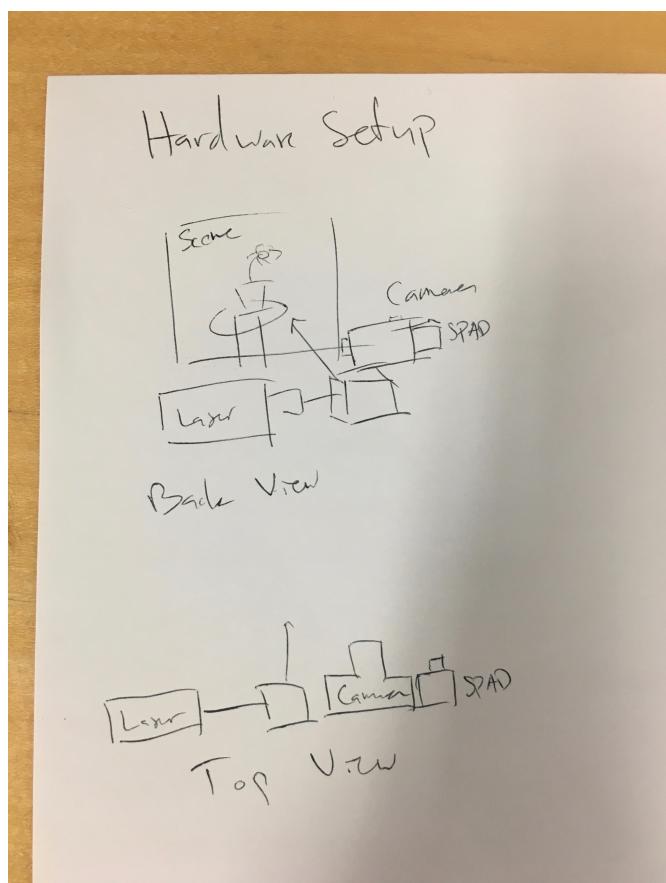
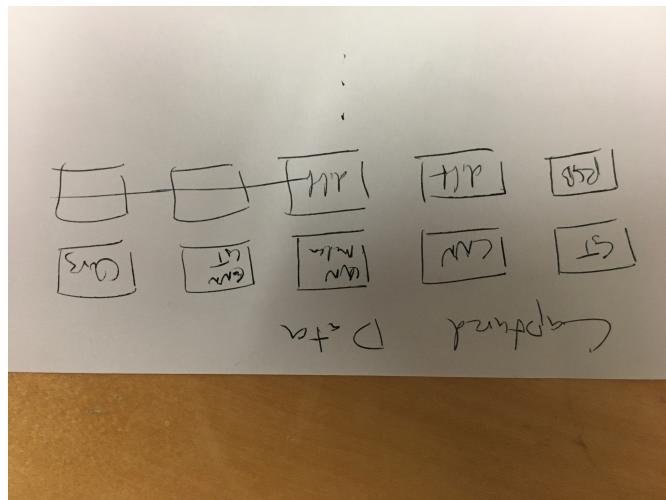
5.1. Implementation

hardware, calibration, etc.

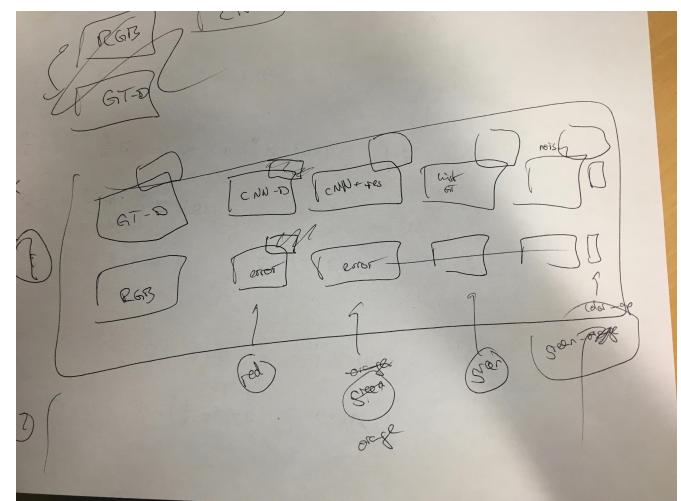
5.2. Results

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 pretrained according to [?] as our CNN.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351Figure 2: **Hardware setup**352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370Figure 3: **Hardware results**371
372
373
374
375
376
377

- Overlay spad histograms?
- Show RMSE/delta1/etc.
- Show colorbars in the last column.

Figure 4: **Comparing our results with other methods**

- Show inset squares and then show the zoomed regions in more detail?

6. Discussion

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$	RMSE \downarrow	rel \downarrow	$\log_{10} \downarrow$
Eigen et. al.	0.769	0.950	0.988	0.641	0.158	-
Laina et. al.	0.811	0.953	0.988	0.573	0.127	0.055
DORN	0.818	0.950	0.982	0.620	0.137	0.063
DORN (rescaled)	0.872	0.967	0.989	0.548	0.111	0.048
Alhashim, Wonka (2019)	0.847	0.973	0.994	0.548(0.461)	0.123	0.053
Alhashim, Wonka (2019) rescaled using GT depth"	0.888	0.978	0.995	0.499 (0.409)	0.106	0.045
Ours (raw depth counts)	0.899	0.970	0.990	0.529	0.199	0.055
Ours (DORN) (intensity/falloff)	0.835	0.953	0.984	0.521	0.129	0.060
Ours (DenseDepth) (intensity/falloff)	0.867	0.974	0.994	0.445	0.114	0.050

Table 2: Results on the NYU Depth v2 test set [?].

	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$	RMSE \downarrow	rel \downarrow	$\log_{10} \downarrow$
DORN (cite)	0.846	0.954	0.983	0.501	0.120	0.053
DenseNet(cite))	0.847	0.973	0.994	0.461	0.123	0.054
DORN (rescaled)	0.872	0.967	0.989	0.548	0.111	0.048
DORN (Wass)	0.847	0.953	0.983	0.499	0.117	0.053
DORN (Histogram Matching)	0.902	0.973	0.991	0.424	0.099	0.042
DenseNet (rescaled)	0.888	0.978	0.995	0.409	0.106	0.045
DenseNet (Wass)	-	-	-	-	-	-
DenseNet (Histogram Matching)	0.930	0.984	0.995	0.338	0.080	0.034
DORN (Median SPAD Rescaling)	-	-	-	-	-	-
DORN + Wasserstein (intensity/falloff)	0.835	0.953	0.984	0.521	0.129	0.060
DenseDepth (Median SPAD Rescaling)	-	-	-	-	-	-
DenseDepth + Wasserstein (intensity/falloff)	0.867	0.974	0.994	0.445	0.114	0.050

Table 3: Results on the NYU Depth v2 test set [?].