

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

# Disambiguating Monocular Depth Estimation with a Single Transient: Supplemental Information

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

Anonymous ECCV submission

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

Paper ID 3668

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

**Abstract.** TODO: Abstract

## 1 Hardware prototype details

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

The monocular depth estimate is calculated using the RGB image captured by the Kinect v2. The SPAD records temporal histograms with 4096 bins, each corresponding to a time window of 16 ps. The SPAD and laser are co-axially aligned using a beam splitter (Thorlabs PBS251). The full width at half maximum (FWHM) of the combined laser pulse width and SPAD jitter is about 70 ps, allowing the system to record depth maps with an accuracy of about 1 cm. A National Instruments data acquisition device (NI-DAQ USB-6343) provides synchronization signals for the galvos, SPAD, and laser. The ground truth depth map is raster-scanned at a resolution of  $512 \times 512$  pixels, and the single-pixel, diffused SPAD measurement 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—such validation would not be possible if we were to capture measurements with an optically diffused SPAD.

## 2 Comparison of diffused vs. scanned imaging

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

In our experiments, we capture measurements by scanning the scene with a single-pixel SPAD detector whose optical path is aligned with a laser. This arrangement allows us to capture a reference “ground truth” depth map for quantitative validation of our method. To emulate measurements captured using a system where the laser and detector are diffused over the scene, we digitally sum the measurements to obtain a single transient.

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

In order to verify that digital summation of scanned measurements yields results that are similar to those captured by a diffused laser and detector, we capture an example scene using a modified hardware prototype in both scanned and diffused modes. This hardware prototype (shown in Fig. 1) is less mobile than our unmodified prototype, but allows us to use a more powerful laser (Katana 05HP, 532 nm) operated at approximately 25 mW average power. We also use

045 two single-pixel SPAD detectors, where one SPAD is aligned with the optical  
046 path of the laser, and the other SPAD is operated without a lens to integrate  
047 light from the entire scene. Both SPADs are fitted with a 10 nm bandpass filter  
048 centered at 532 nm, which reduces the amount of integrated ambient light. We  
049 attach a holographic diffuser (Thorlabs ED1-S50) to the laser output in order  
050 to diffuse light onto the scene. Alternatively, we remove the diffuser and use a  
051 pair of scanning mirrors to scan the scene.

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

The modified hardware setup is used to capture an example scene in both scanned and diffused modes, and the resulting transients are used to refine an initial depth estimate from the Kinect RGB image. We illustrate the results of this procedure in Fig. 2. The reconstructions from the scanned and diffused measurements are similar in reconstruction quality and also show similar quantitative improvement in terms of error over the initial depth estimate. The unnormalized photon counts are also shown in Fig. 2, and we note that the counts show similar trends. The number of recorded photons in these experiments is shown in Table 1. In both cases, the rate of detected photons is far less (<5%) than the number of emitted laser pulses, and so we conclude that the measurements are captured in the low-flux regime where pileup effects are negligible. We attribute most of the differences between the scanned and diffused transients to the approximately 16 cm vertical baseline between the positions of the diffused and scanned SPADs (see Fig. 1).

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

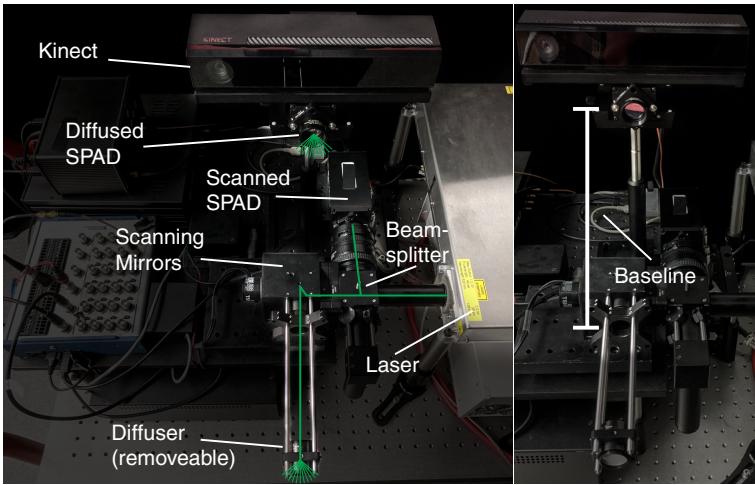


Fig. 1: Modified hardware setup. The setup is used to compare scanned and diffused measurements and employs two SPAD detectors and two laser configurations. In the first configuration, the scene is illuminated by sending the laser light through a holographic diffuser and a lensless SPAD integrates light from the entire scene. In the second, the SPAD is aligned with the optical path of the laser and the scene is scanned using a pair of scanning mirrors. The baseline between the two SPADs (right) results in some observed differences in the recorded transients.

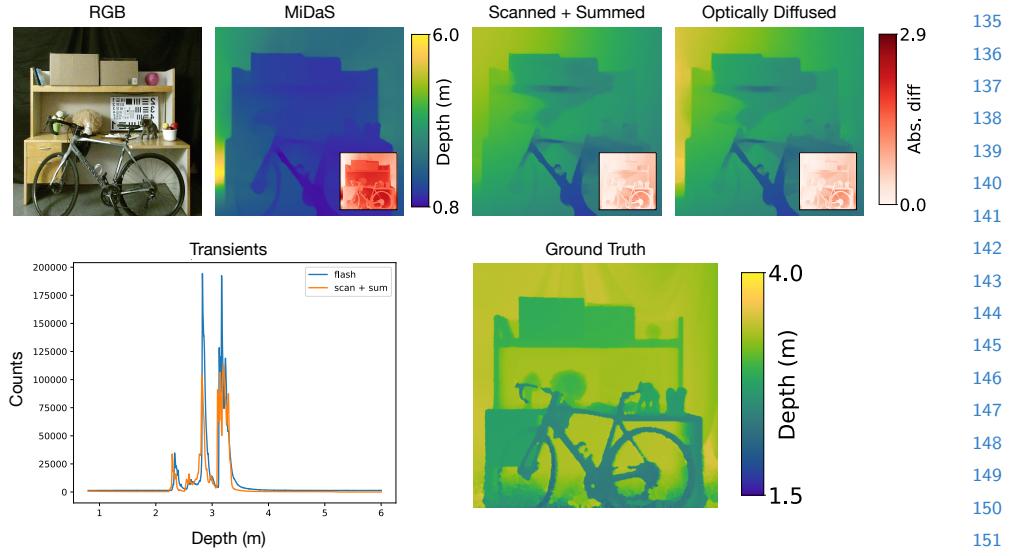


Fig. 2: Comparison of scan + sum and diffused SPAD. The transients are captured with the same total exposure time and are qualitatively similar without noticeable pileup effects. We Use  $K = 300$  bins for the reconstruction and a depth range of  $[0.8, 6]$  meters. MiDaS [3] does not produce globally-scaled depth, so we scale it to fit this depth range. We inpaint the depth map from the Kinect's depth camera to acquire ground truth depth.

Experiment	Detected Photons	Laser Pulses	Detection Rate
Scanned	$1.4 \times 10^7$	$6 \times 10^8$	2.3%
Diffused	$2.4 \times 10^7$	$6 \times 10^8$	4.0%

Table 1: Recorded photons for diffused vs. scanned scene. In each capture mode, scanned or diffused, the number of detected photons does not exceed 5% of the number of emitted laser pulses, placing the capture within the low-flux regime where pileup effects are negligible.

### 3 Nominal values for radiometric calculation

Here we give the values used in the equation for received power, repeated here for convenience:

$$P_R = P_T \cdot \rho \cdot \frac{A_{rec}}{\pi R^2} \cdot \eta. \quad (1)$$

For our scene, we assume a vertical, planar, perfectly Lambertian surface. The following table gives the values used for this calculation.

Symbol	Description	Nominal Value
$\rho$	Albedo of lambertian surface	0.3
$P_T$	Total irradiance at wavelength (W/m <sup>2</sup> )	0.026
$R$	Distance to surface (m)	3
$A_{rec}$	Area of detector (m <sup>2</sup> )	$1.96 \times 10^{-9}$
$\eta$	Quantum efficiency of detector	0.3
$P_R$	Received power at detector (W)	$1.62 \times 10^{-13}$

Fig. 3: Table of nominal values for radiometric calculation.

Once  $P_R$  is determined, we compute the number of photons using the laser wavelength  $\lambda = 532$  nm as

$$N = \frac{P_R \lambda}{hc} \quad (2)$$

where  $h \approx 6.626 \times 10^{-34}$  J · s is Planck's constant and  $c \approx 3 \times 10^8$  m/s is the speed of light. Using the fact that our laser runs at 10 MHz, we get the number of photons per pulse as 0.043 or 4.3%, which puts us in the low-flux regime (where photons per pulse is < 5%).

## 4 Ablation study on number of SID bins

We conducted an ablation study on the effect of the number of SID bins [2] on both runtime and RMSE. We performed this analysis using SPAD data with a signal-to-background (SBR) of 100, simulated on the test set of NYU Depth v2. We used DenseDepth [1] for our MDE CNN. Only the histogram matching portion was timed, not the CNN nor the denoising pipeline.

# of sid bins	RMSE Approx.	Time/image (sec)
70	0.351	0.24
140	0.346	0.63
210	0.345	1.12
280	0.345	1.84

Fig. 4: Effect of number of SID bins on RMSE and runtime. The marginal improvement in RMSE is offset by the increase in runtime as the number of bins grows.

## 225 5 Ablation study on effect of reflectance estimation

226  
227  
228  
229 We conducted an ablation study on whether the use of a reflectance estimate  
230 has an impact on the runtime and quality of the solution. We performed this  
231 analysis using SPAD data with a signal-to-background (SBR) of 100, simulated  
232 on the test set of NYU Depth v2 and using DenseDepth [1] for our MDE CNN.  
233 Only the histogram matching portion was timed, not the CNN nor the denoising  
234 pipeline. Using the intensity to produce the initial weighted histogram  $h_{\text{source}}$   
235 provides noticeable improvements in RMSE, but intensity may safely be ignored  
236 during the pixel movement step, resulting in noticeable speed improvements.

Intensity-weighted histogram	Intensity-aware pixel movement	Avg. RMSE	Time per image (sec)
Yes	Yes	0.346	4.6
	No	0.346	0.6
No	Yes	0.444	4.7
	No	0.444	0.6

240  
241 Fig. 5: Effect of reflectance modeling on RMSE and runtime. When the SPAD is  
242 simulated with the reflectance info but no reflectance estimate is used to generate  
243 a weighted histogram from the CNN depth map, the results are significantly  
244 worse. Furthermore, once the pixel movement matrix has been computed, the  
245 pixel movement procedure need not take into account the weights of the pixels  
246 being moved, since doing so provides no improvement and can take appreciably  
247 longer than a vectorized implementation that does not take pixel weights into  
248 account.

## 258 6 Pseudocode, pixel shifting, and dither artifacts

259  
260  
261  
262 We give pseudocode for our algorithm here. In the first part of our algorithm we  
263 compute the pixel shifting matrix mapping the histogram  $h_s$  (computed from the  
264 initial depth map and reflectance estimate) to  $h_t$  (computed from the captured  
265 transient).

266 Given this pixel movement matrix  $T$ , we apply the appropriate movements  
267 to the initial depth map  $I$ . The pixels of the image  $I$  take depth bin values in  
268  $\{0, \dots, K - 1\}$ .

---

**Algorithm 1** Find Pixel Movement
 

---

```

270
271 procedure FINDPIXELMOVEMENT( $h_s$  of length  $M$ ,  $h_t$  of length  $N$ )
272   Initialize  $T$  as an  $M \times N$  array of zeros.
273   for  $m$  in  $1, \dots, M$  do
274     for  $n$  in  $1, \dots, N$  do
275        $p_s \leftarrow \sum_{i=1}^{n-1} T[m, i]$ 
276        $p_t \leftarrow \sum_{i=1}^{m-1} T[i, n]$ 
277        $T[m, n] \leftarrow \min(h_s[m] - p_s, h_t[n] - p_t)$ 
278     end for
279   end for
280   return  $T$ 
281 end procedure
282
283
284
285
286
287
  
```

---

**Algorithm 2** Move Pixels
 

---

```

288
289 procedure MOVEPIXELS(input image  $I$  size  $M \times N$ , pixel movement matrix  $T$  of
290   size  $K \times K$ )
291   for  $k$  in  $0, \dots, K - 1$  do
292      $p[k, :] \leftarrow T[k, :] / \sum_{i=1}^K T[k, i]$ 
293   end for
294   for  $m$  in  $1, \dots, M$  do
295     for  $n$  in  $1, \dots, N$  do
296       Sample  $k'$  according to  $p[I[m, n], :]$ .
297        $I[m, n] \leftarrow k'$ .
298     end for
299   end for
300   return  $I$ 
301 end procedure
302
303
304
305
306
307
308
  
```

---

Because the pixel shifting process in Algorithm 2 contains a sampling step, it is possible for *dither artifacts* to appear in the output image  $I$ , as shown in figure 6. Specifically, when there are multiple possible output depth bins for a given input depth bin, and a large region of equal depth in the input image, the randomness in the pixel shifting algorithm will distribute the pixels of large, equal-depth region in the input across the multiple possible output depth bins in a random fashion.

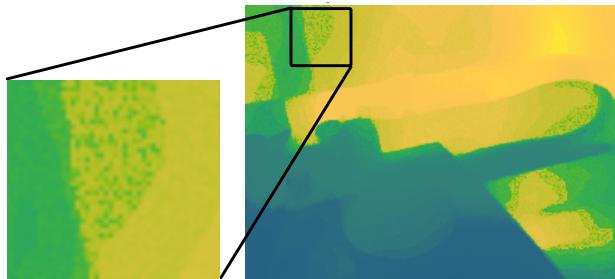


Fig. 6: Example of dither artifacts. Sometimes, when our histogram matching is applied to images with large regions of similar depths, dither artifacts will occur.

## 360 7 Additional results on NYU Depth v2

361  
362 Figures 7–15 show additional results for our method on the NYU Depth v2  
363 dataset when the depth estimate is initialized with the DenseDepth [1] (Figures  
364 7–9), DORN [2] (Figures 10–12) and MiDaS [3] (Figures 13 – 15) monocular  
365 depth estimators.

366 We compare the output of the network  $z_0$ , the median-rescaled network out-  
367 put (where the depth map  $z_0$  is scaled pixel-wise by a scalar  $\frac{\text{median}(z_{GT})}{\text{median}(z_0)}$ ,  $z_{GT}$   
368 being the ground truth depth map), the network output matched to the ground  
369 truth depth histogram, and the output of our histogram matching method under  
370 a signal-to-background ratio (SBR) of 100. We use the luminance of the RGB  
371 image as our reflectance map for both SPAD simulation and histogram match-  
372 ing. We show absolute difference maps and also give the root-mean-square error  
373 (RMSE) for each example.

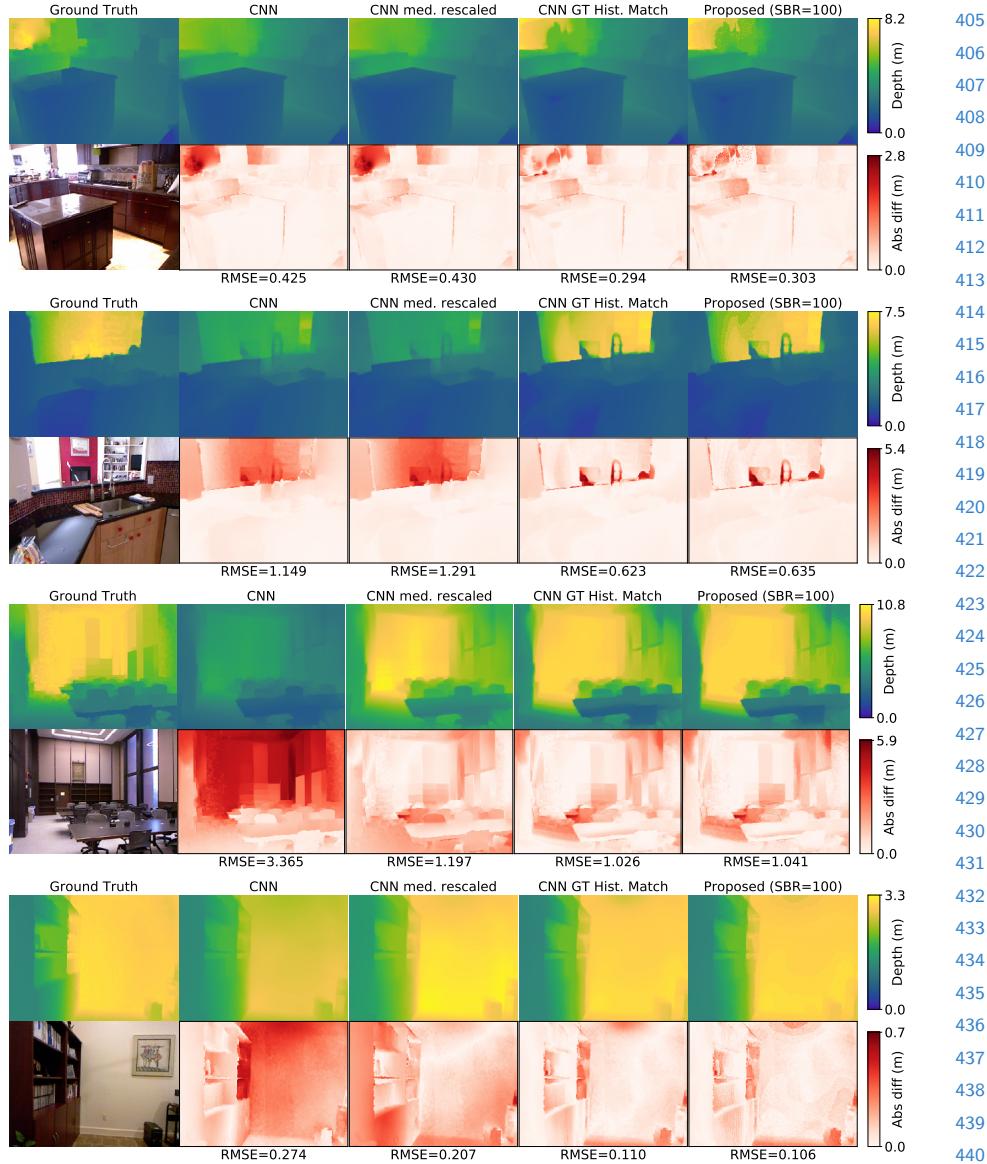


Fig. 7: Results with DenseDepth as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.

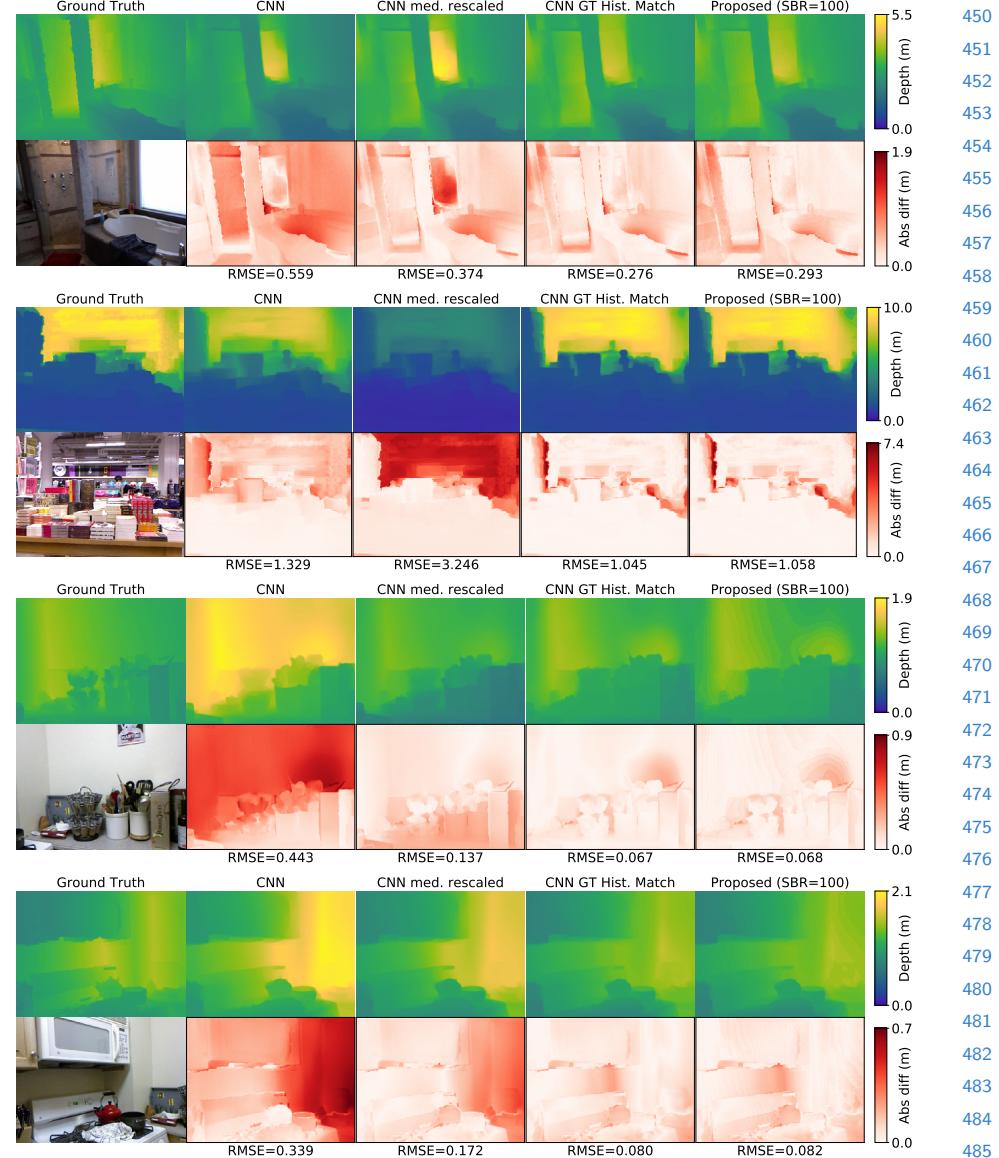


Fig. 8: Results with DenseDepth as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.

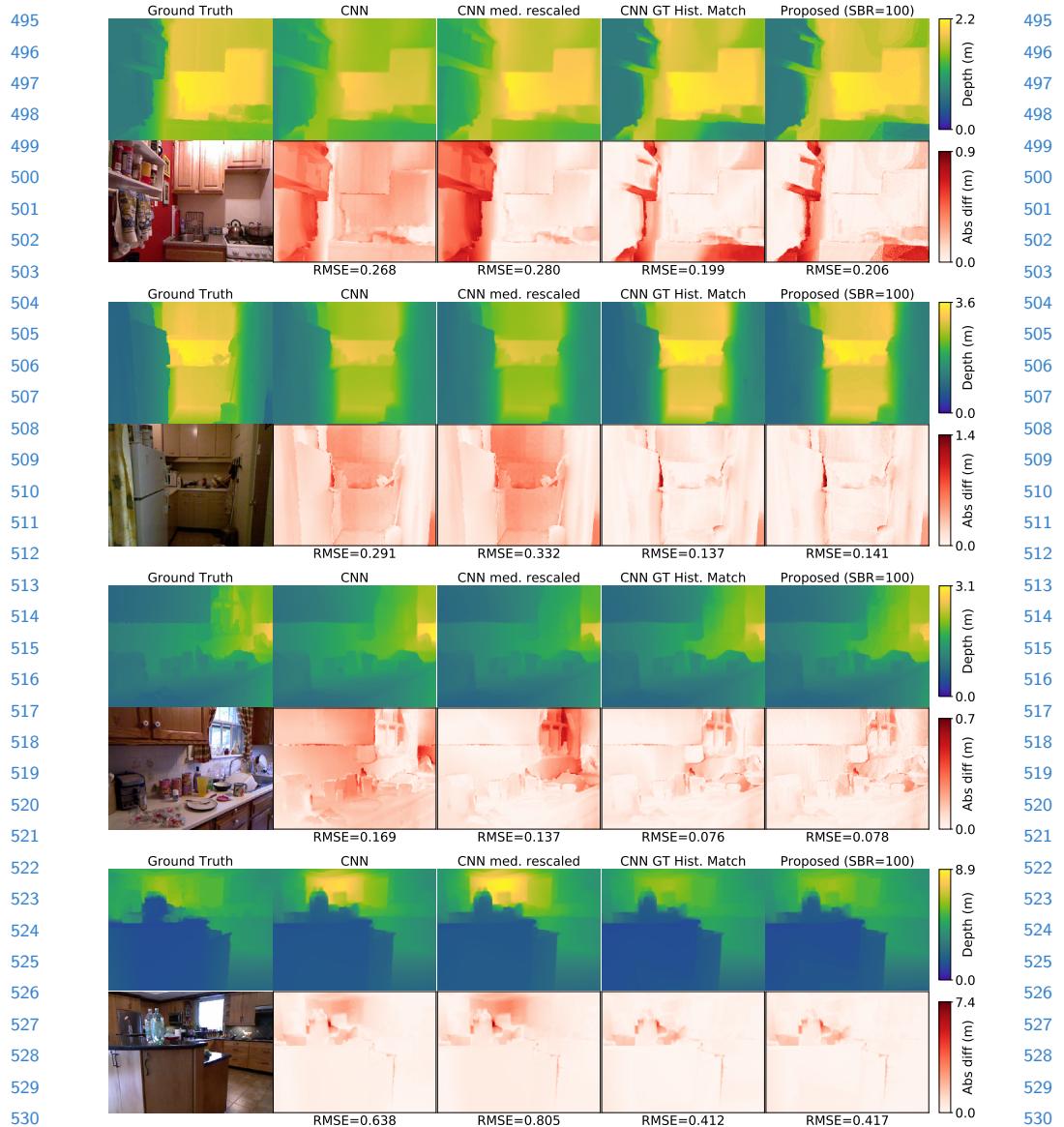
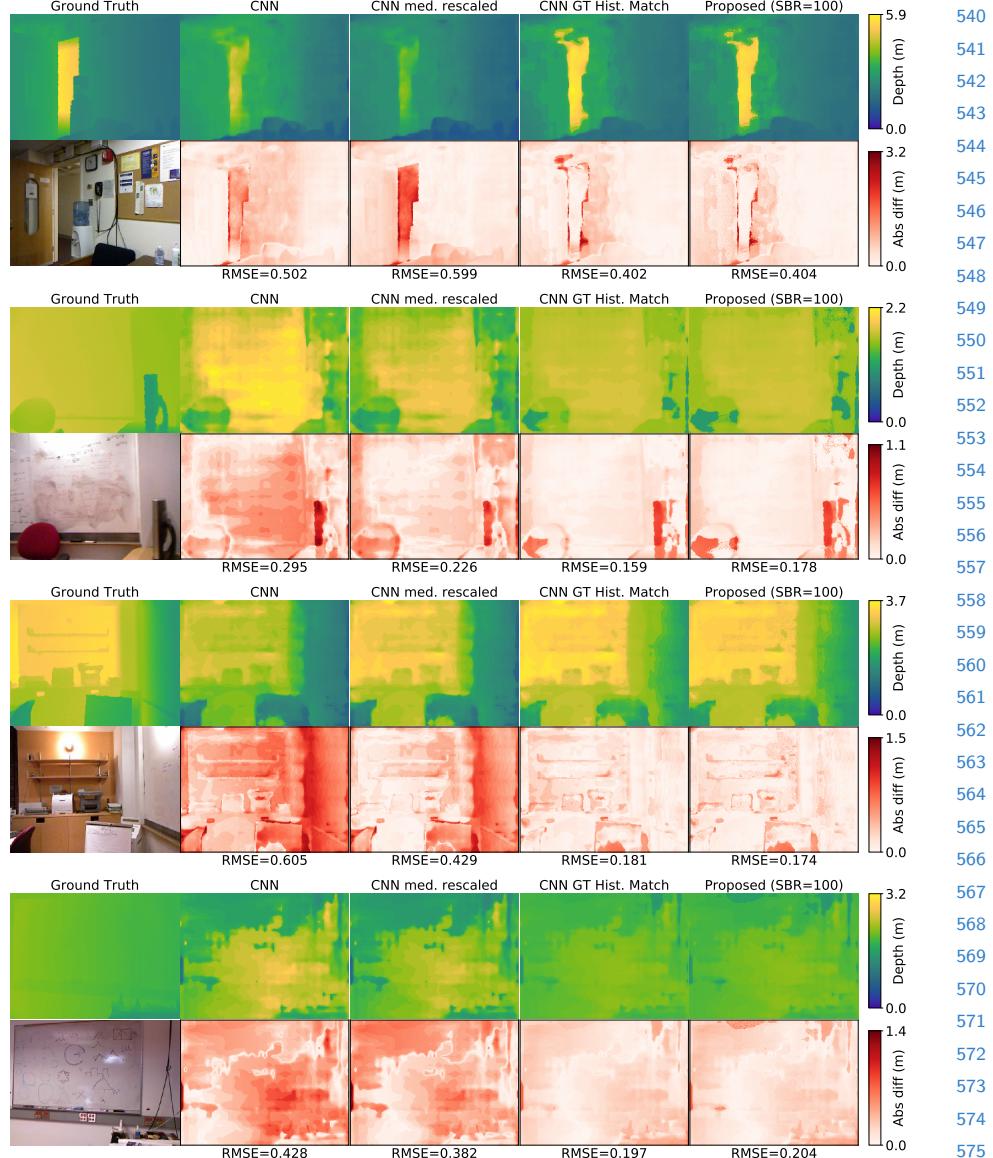


Fig. 9: Results with DenseDepth as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.



576 Fig. 10: Results with DORN as the monocular depth estimator. Our method is  
 577 able to scale and shift the depth maps to mitigate gross errors in depth scaling.  
 578

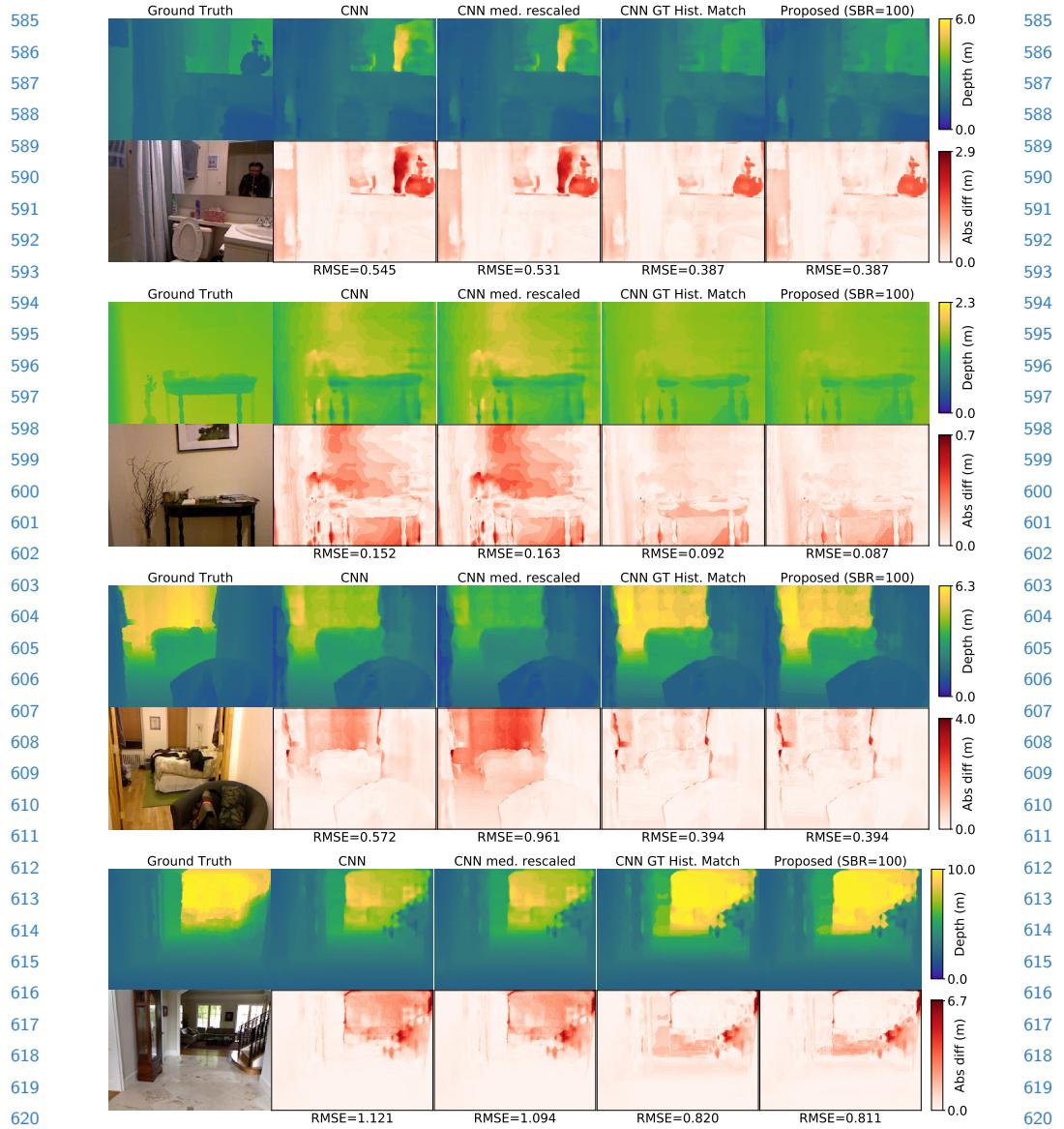
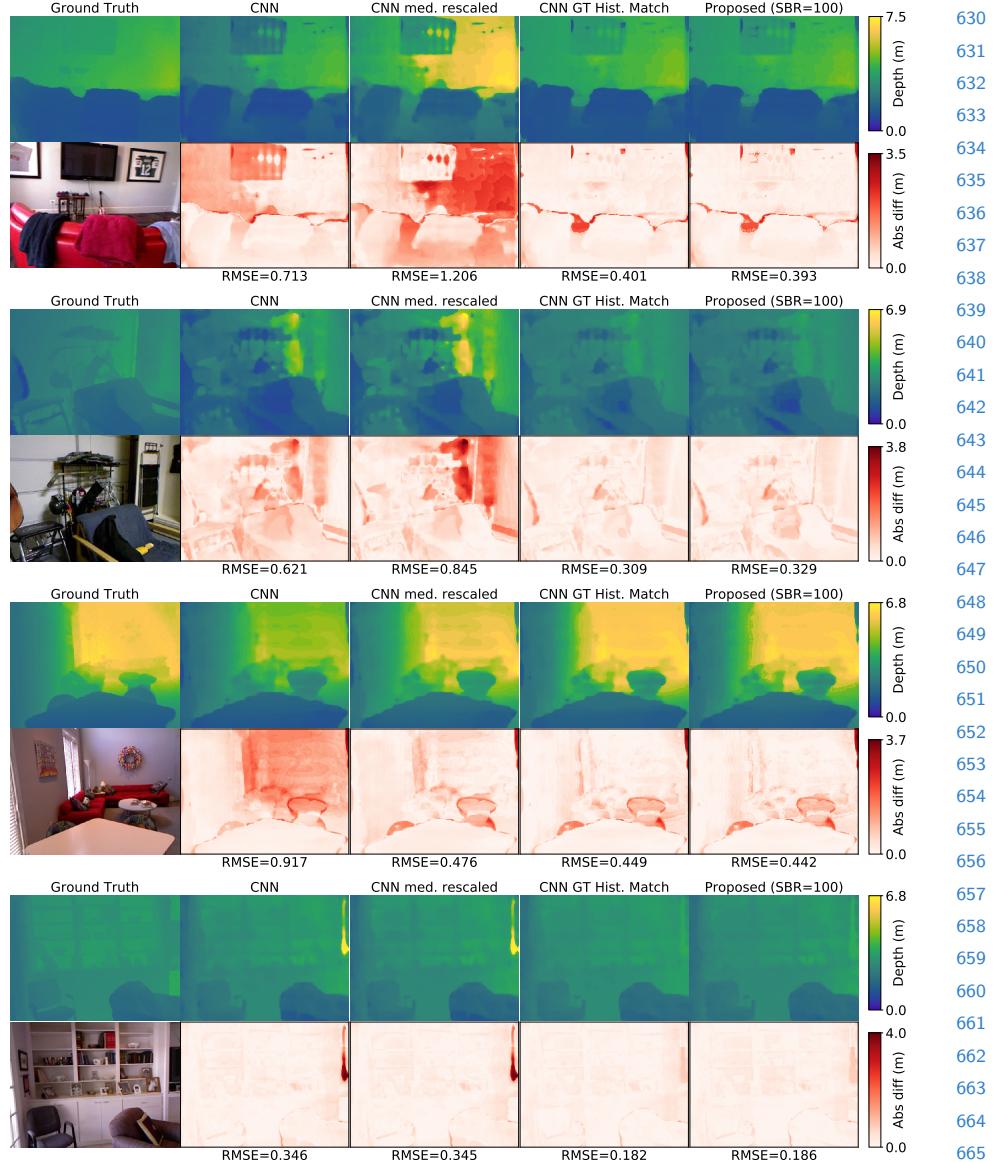


Fig. 11: Results with DORN as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.



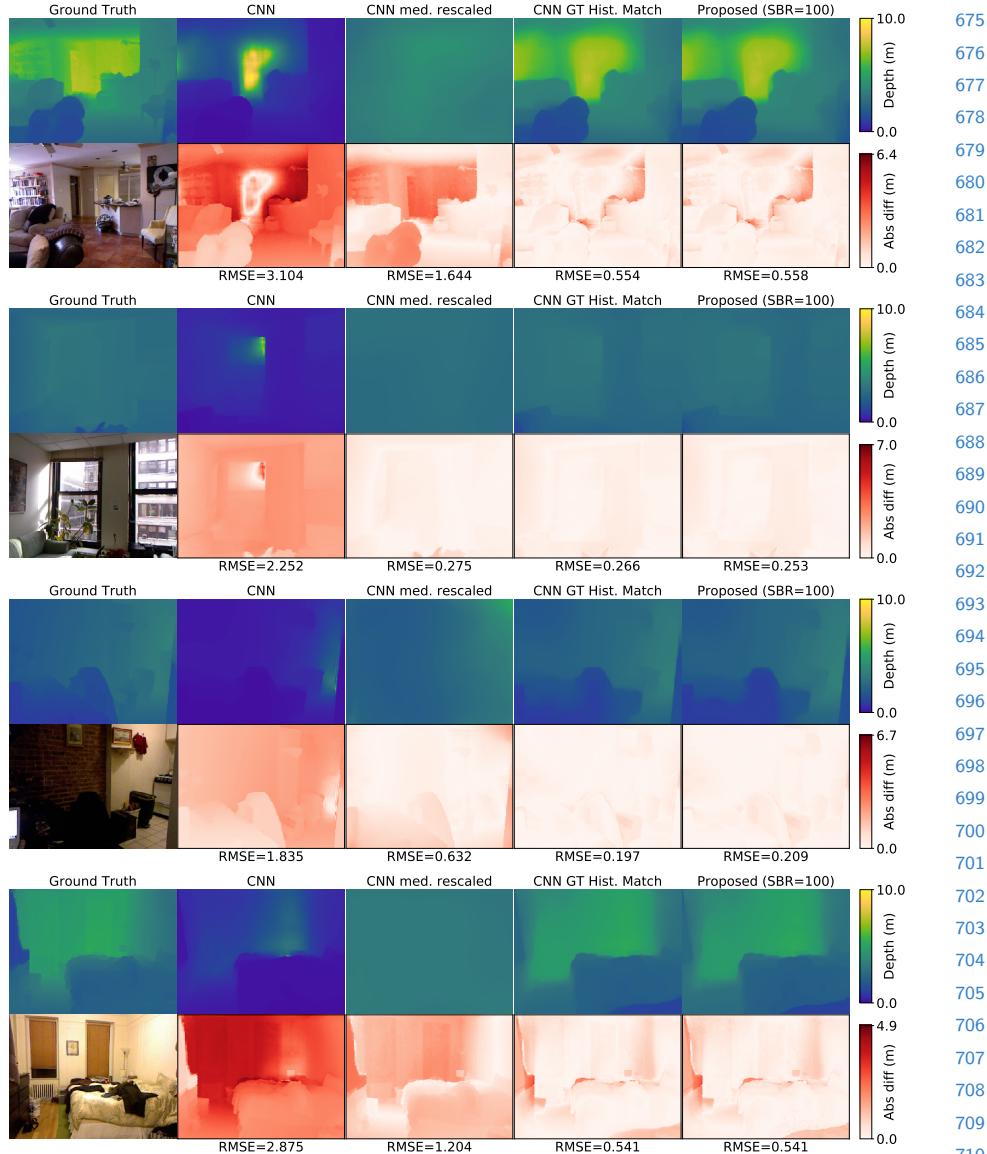


Fig. 13: Results with MiDaS as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.

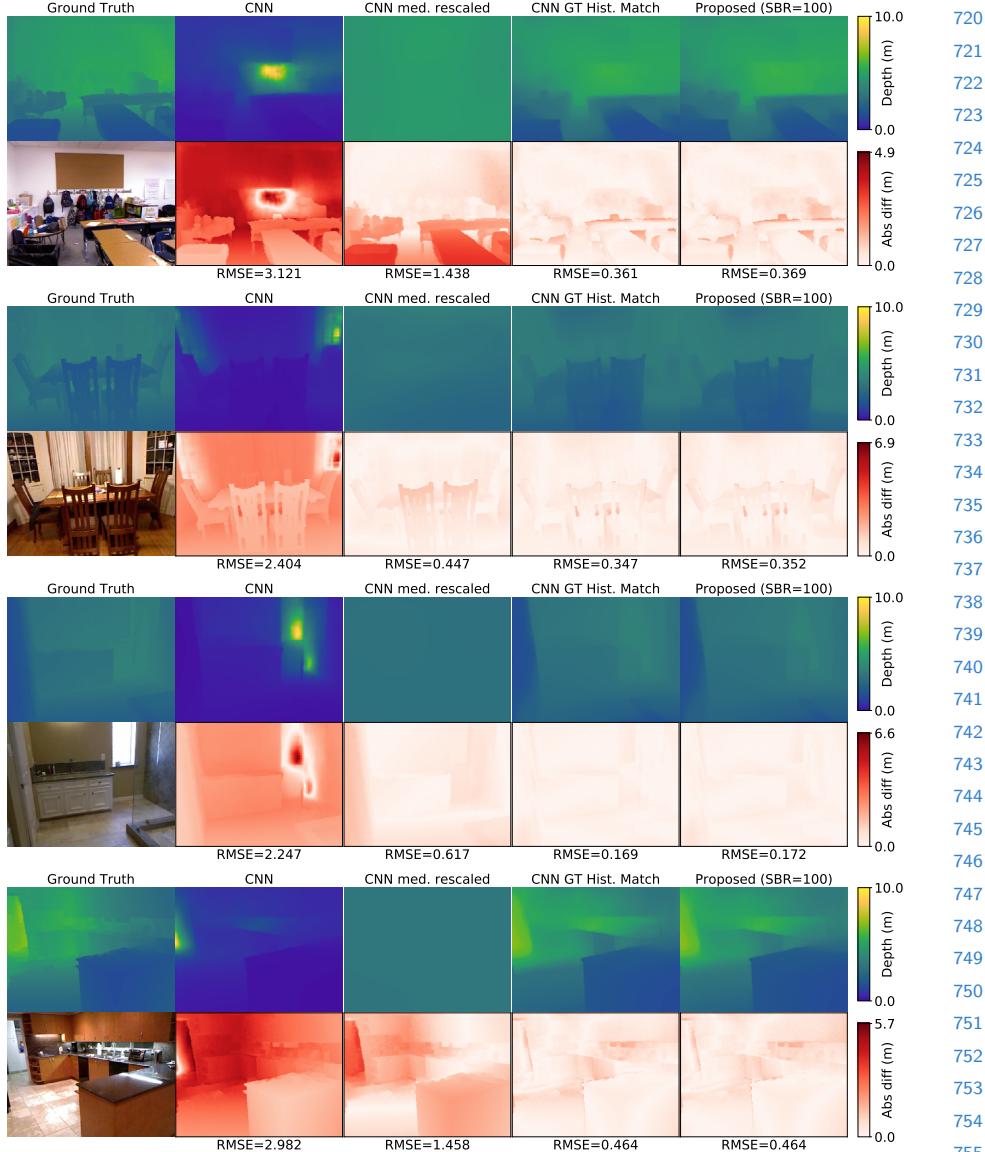


Fig. 14: Results with MiDaS as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.

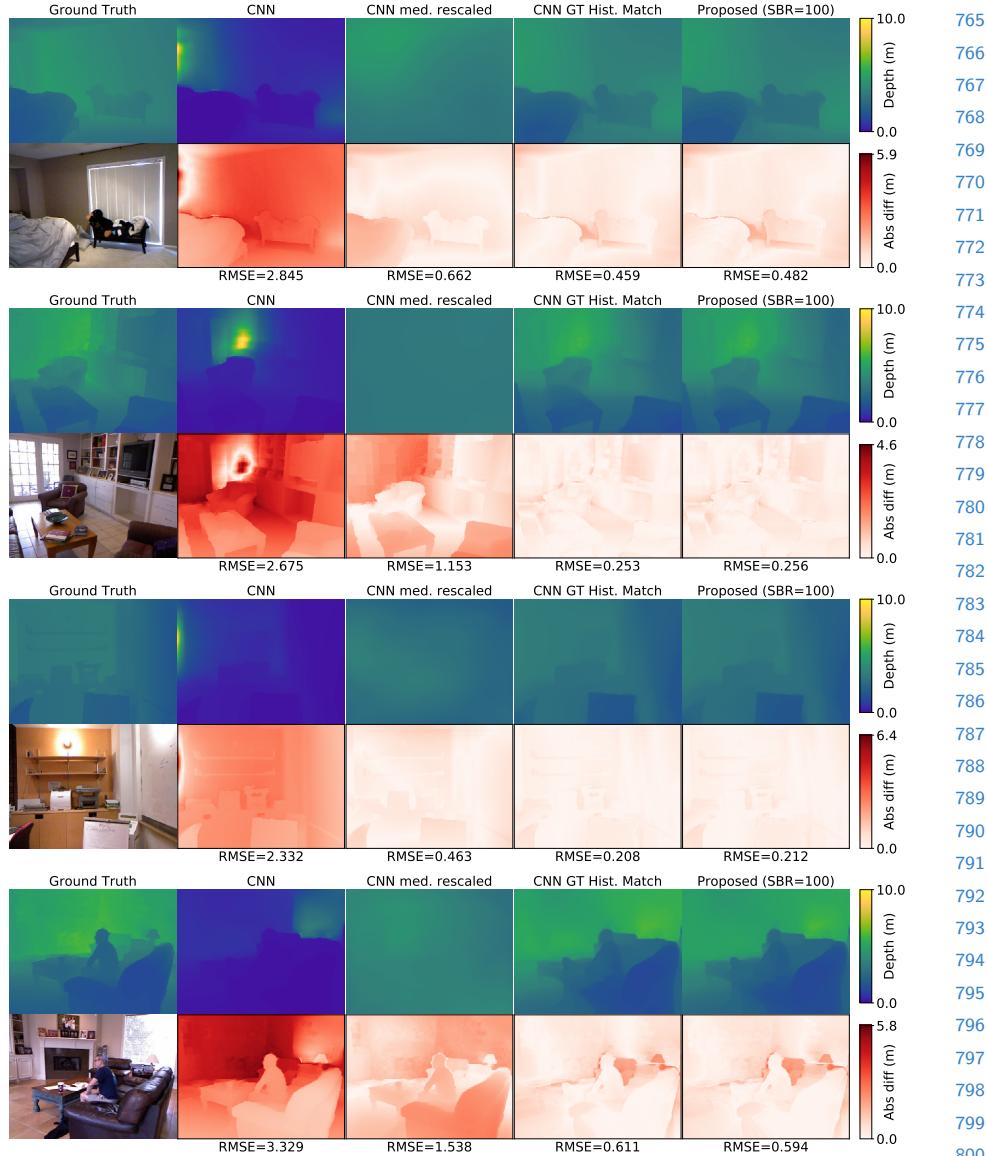


Fig. 15: Results with MiDaS as the monocular depth estimator. Our method is able to scale and shift the depth maps to mitigate gross errors in depth scaling.

## 8 Additional results for hardware prototype

Figures 16–24 show all the captured results when the depth estimate is initialized with the MiDaS [3] (Figures 16–18), DenseDepth (Figures 19–21), and DORN (Figures 22–24). We compare the output of the network  $z_0$ , the mean-rescaled network output where the depth map  $z_0$  has been scaled pixel-wise by the scalar  $\frac{\text{median}(h_{\text{target}})}{\text{median}(z_0)}$  ( $h_{\text{target}}$  is the processed SPAD transient), and the output of our method. As our laser is red, we use the R channel of the RGB image as our reflectance map. We show absolute difference maps and also give the root-mean-square-error (RMSE) for each example.

Black pixels in the ground truth depth correspond to locations where our scanner was unable to produce an accurate depth estimate (this can occur for a variety of reasons including dark albedo and surface specularity). These pixels are masked off and not used in the RMSE calculation, and appear as an absolute difference of 0 in the difference maps.



Fig. 16: Captured results initialized using the MiDaS CNN. Second row shows absolute difference between above estimates and ground truth. MiDaS does not output metric depth, so the CNN depth maps are scaled to be in the range (0.494, 9.094) by default. However, MiDaS does produce accurate ordinal depth, leading to stronger performance of our histogram matching compared to other methods.

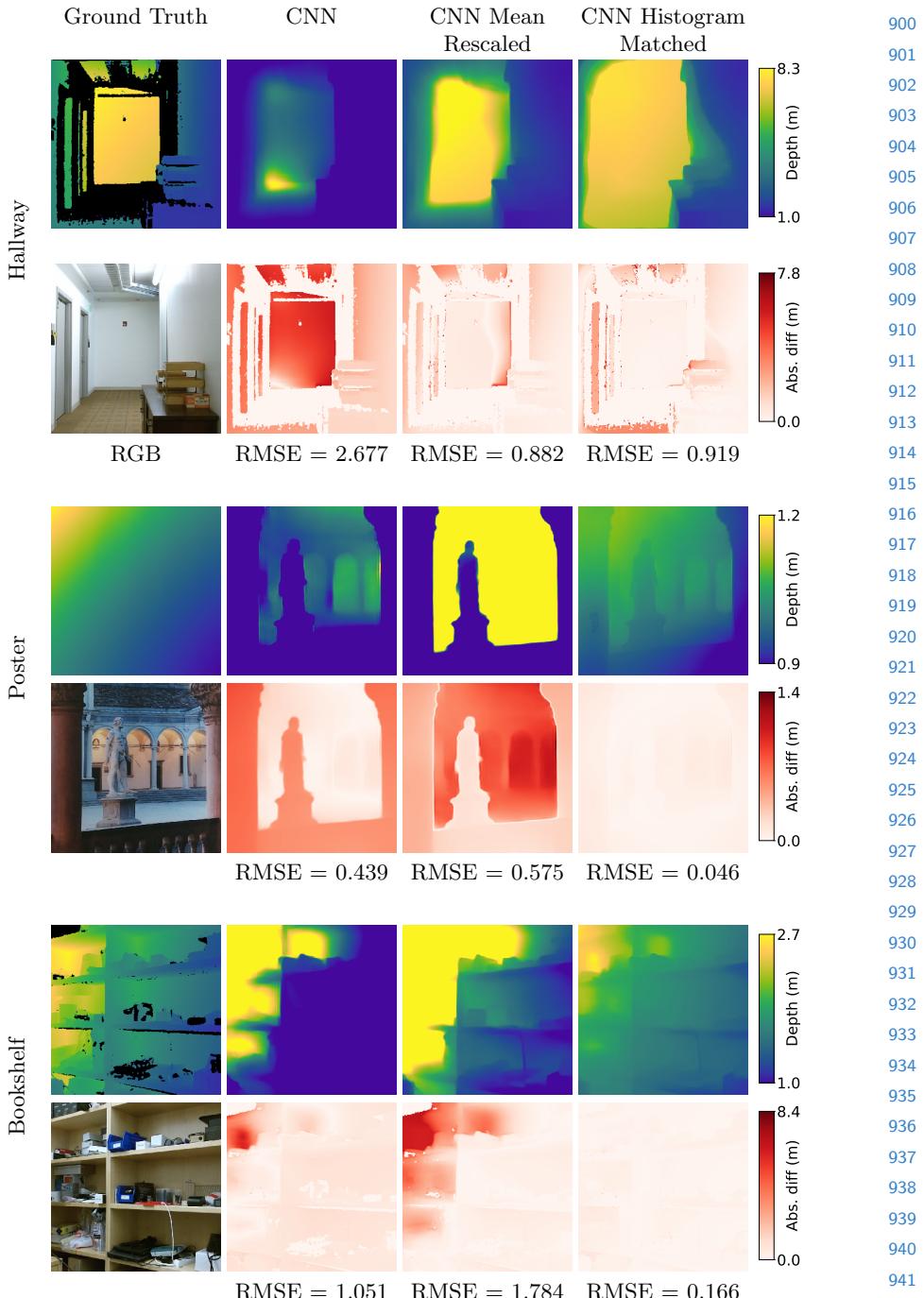


Fig. 17: Captured results initialized using the MiDaS CNN. Second row shows absolute difference between above estimates and ground truth. MiDaS does not output metric depth, so the CNN depth maps are scaled to be in the range (0.494, 9.094) by default. However, MiDaS does produce accurate ordinal depth, leading to stronger performance of our histogram matching compared to other methods.

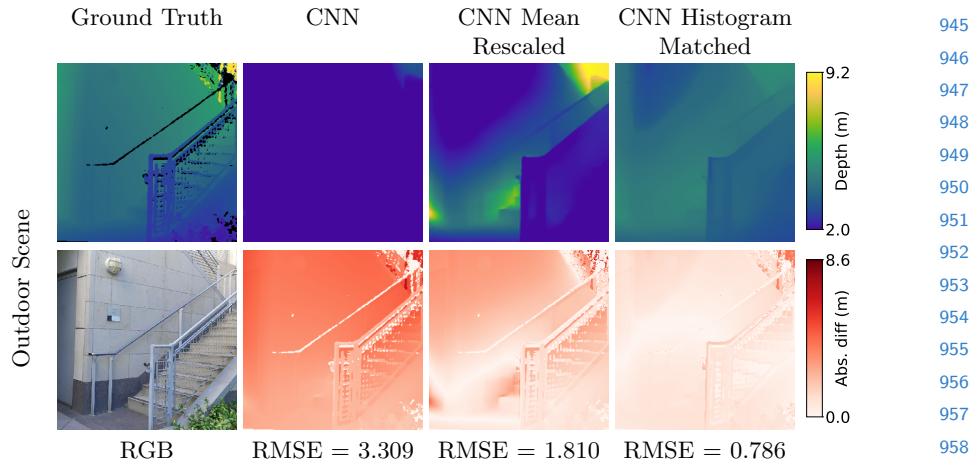


Fig. 18: Captured results initialized using the MiDaS CNN on an outdoor scene. Second row shows absolute difference between above estimates and ground truth. MiDaS does not output metric depth, so the CNN depth map is scaled to be in the range (0.494, 11.094) by default. However, MiDaS does produce accurate ordinal depth, leading to stronger performance of our histogram matching compared to other methods.

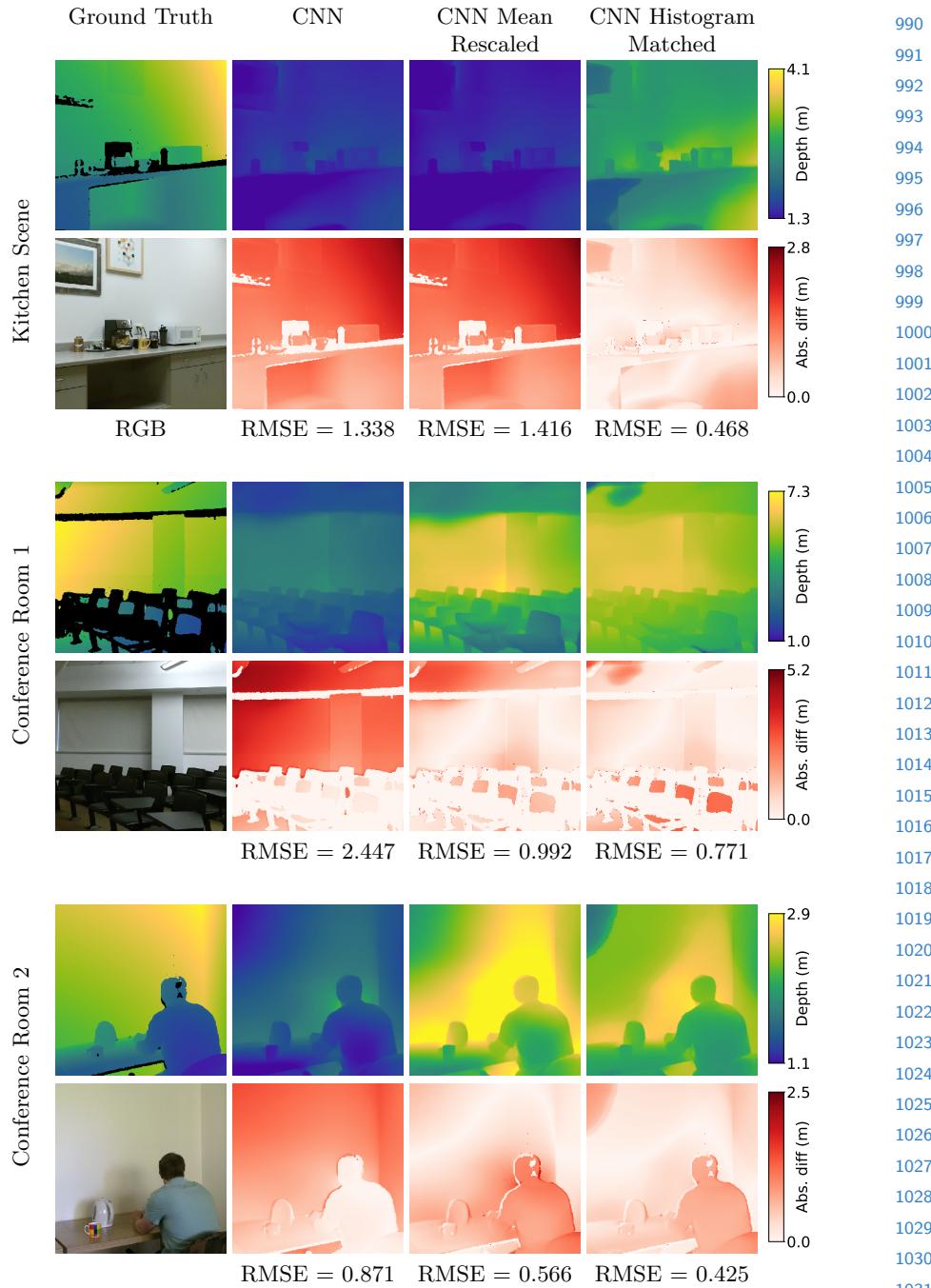


Fig. 19: Captured results initialized using the DenseDepth CNN. Second row shows absolute difference between above estimates and ground truth.

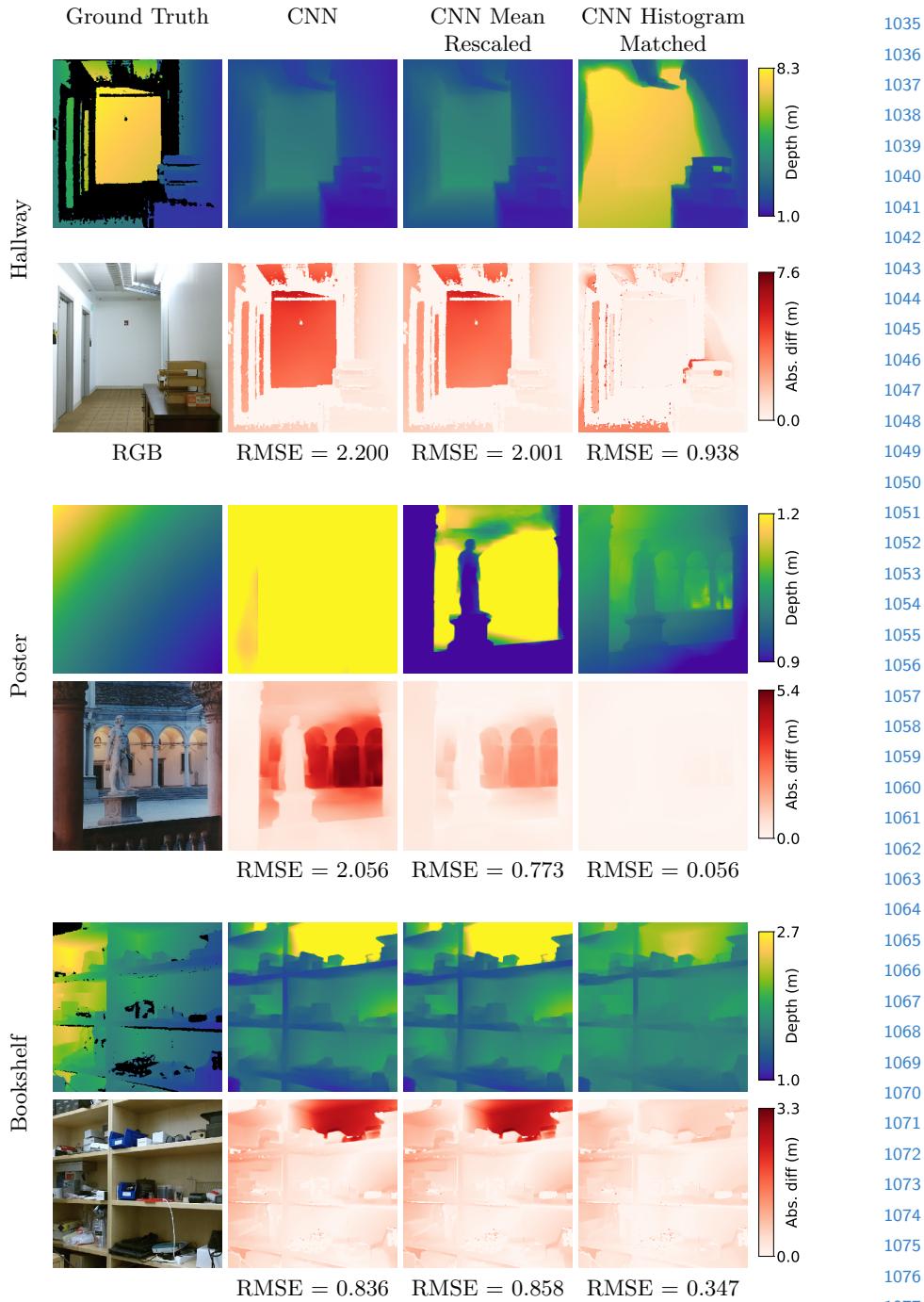


Fig. 20: Captured results initialized using the DenseDepth CNN. Second row shows absolute difference between above estimates and ground truth.

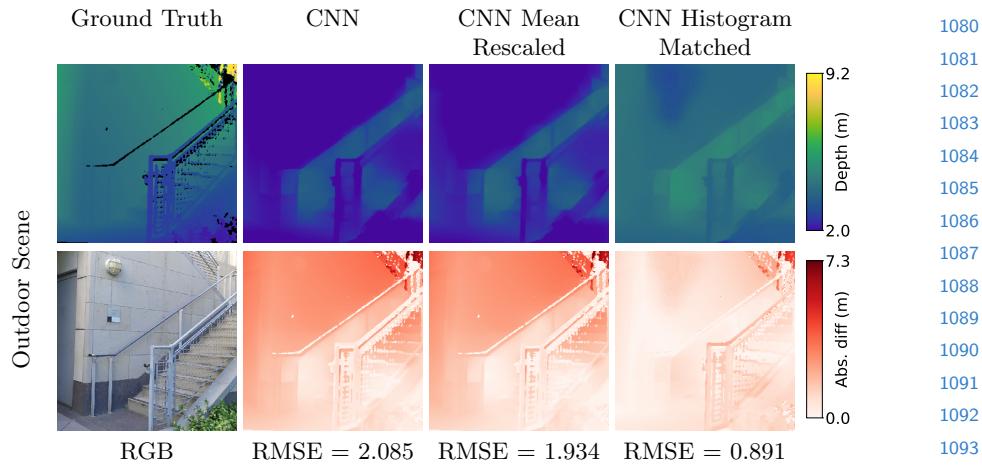


Fig. 21: Captured results initialized using the DenseDepth CNN on an outdoor scene. Second row shows absolute difference between above estimates and ground truth.

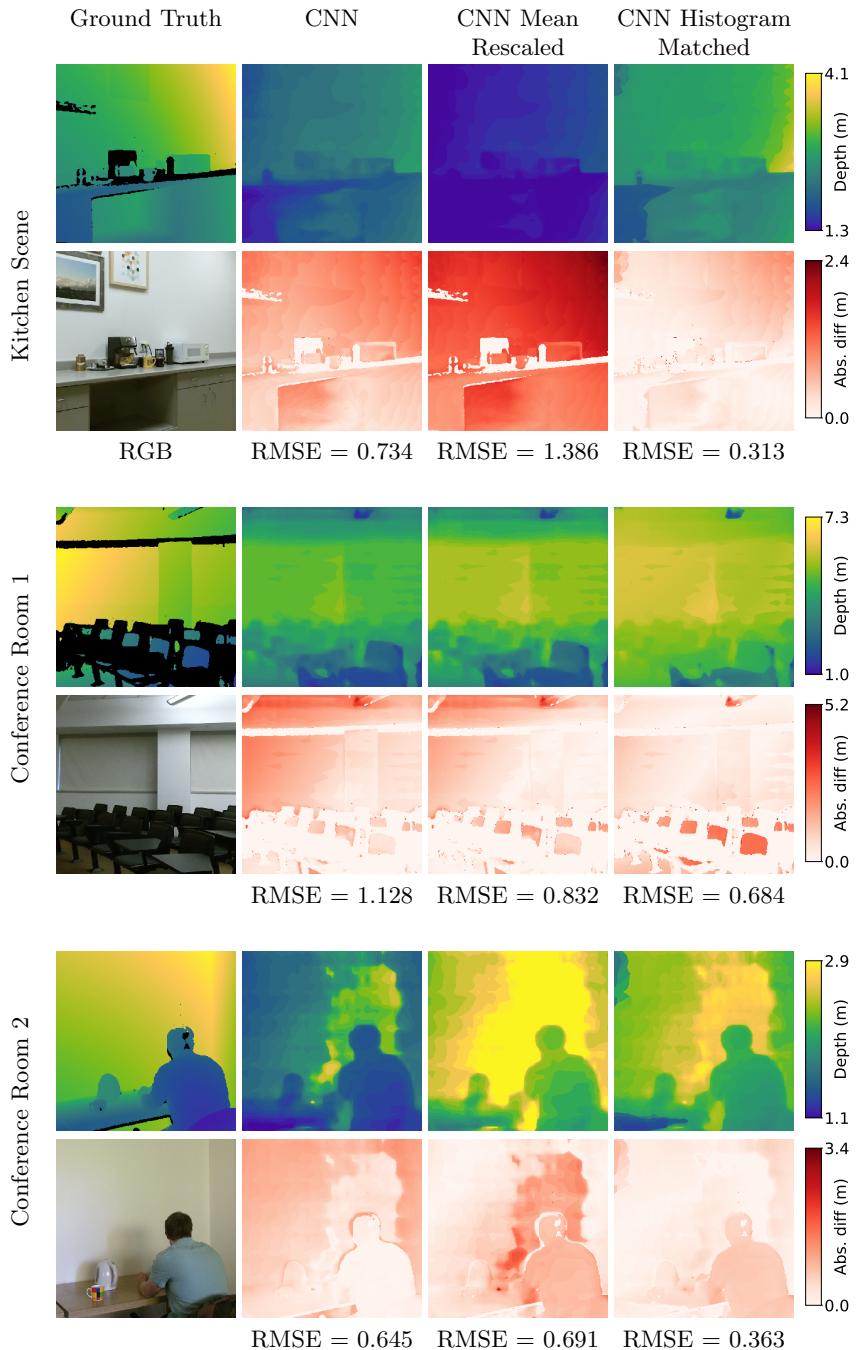
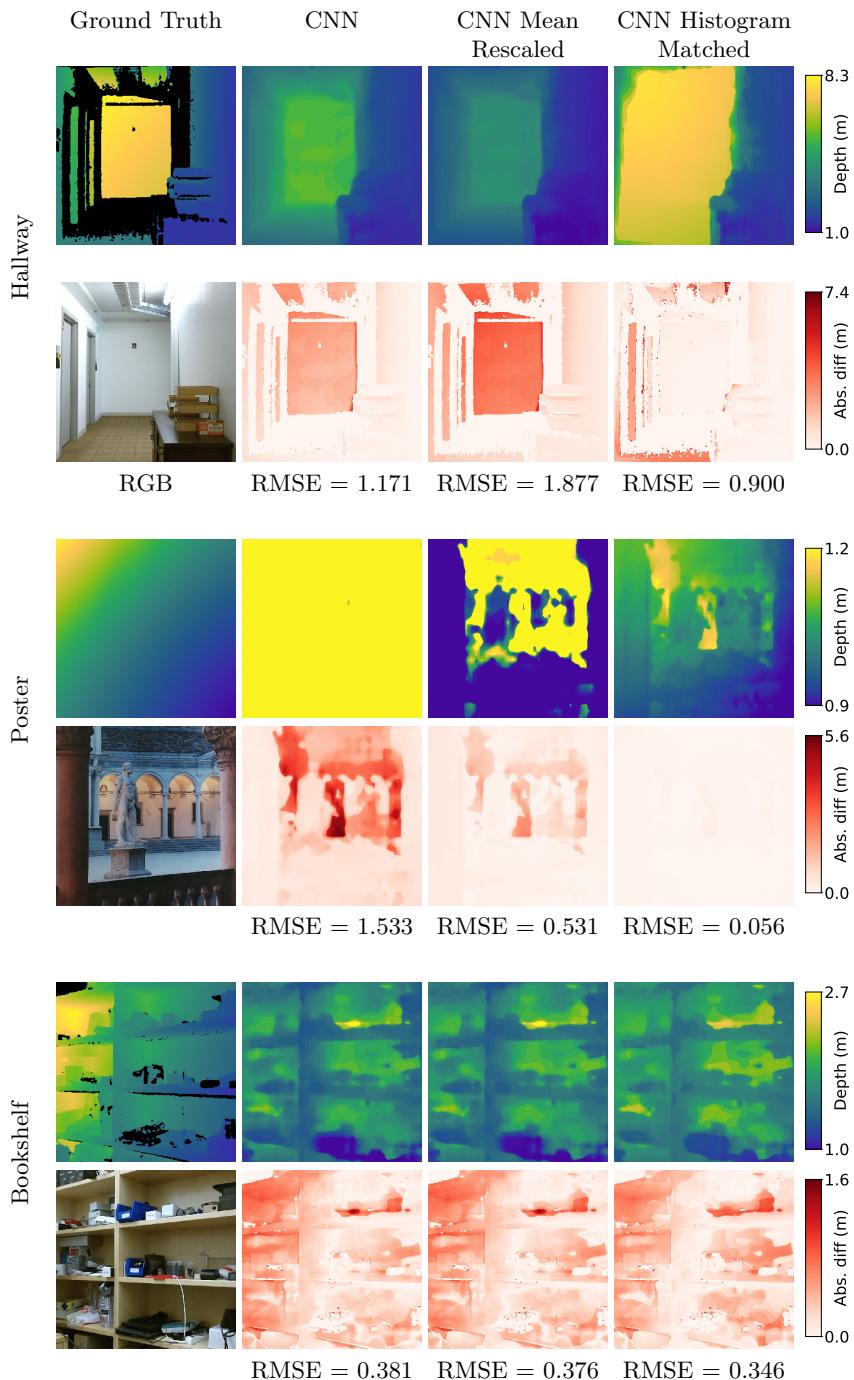


Fig. 22: Captured results initialized using the DORN CNN. Second row shows absolute difference between above estimates and ground truth.



1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259

## 1260 References

- 1261  
1262 1. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learn-  
1263 ing. arXiv:1812.11941v2 (2018)  
1264 2. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression  
1265 network for monocular depth estimation. In: Proc. CVPR (2018)  
1266 3. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth  
1267 estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv:1907.01341  
1268 (2019)

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304