Data-Driven 3D Primitives for Single Image Understanding
Extended Data Analysis

David F. Fouhey, Abhinav Gupta, Martial Hebert

We present some additional analysis of our results, answering the following questions: (a) how much does the performance of our technique vary? (Section 1) (b) how does incorporating the Manhattan-world assumption change our results? (Section 2) (c) how well can our algorithm predict its own accuracy? (Section 3). Finally, we include a complete set of precision-vs-coverage curves in Section 4.

1 Statistical Analysis of Results

One important consideration in any technique is its variability. This section addresses how much our performance varies.

Recall that our methodology is to compute the angle between the predicted surface normal and the ground truth and then summarize the results over the dataset. Since we have enormous amounts of data (one error per pixel), checking the the null hypothesis that two methods have the same error is insufficient. Specifically, it is likely that there is some difference, albeit minute, between the errors of different models or even instances of the same model trained on slightly different data. Statistical significance tests with an errors-equal null hypothesis (the usual one) necessarily conflate small and large performance gains. Thus with a complex model and a large test set, it seems likely that a t-test would reject the null hypothesis of errors-equal. The appropriate tests for each statistic[1] all yield p-values orders of magnitude smaller than 0.01 when testing whether 3DP and another method have the same performance. Frequently, the p-value is close to machine precision, suggesting that errors-equal is too low a bar to surpass to be a good test.

A better way to do this is bootstrapped confidence intervals, which simultaneously give an idea of how much performance gains and rankings (i.e., whether there is a gain) vary. The bootstrap draws a large number of replicates, or equally sized datasets sampled with replacement from the original dataset, and uses these to compute confidence intervals for some statistic (e.g., the mean or the % Good Pixels at 11.25°). This approach makes mild assumptions, and notably does not assume the statistic follows any particular distribution. We compute confidence intervals with the Bias Corrected and Accelerated Bootstrap [1]. Because errors within images are correlated, we use block resampling and sample entire images all at once (i.e., each replicate is created by sampling images with replacement). Intuitively, our procedure can be viewed as seeing how performance varies as we sampled datasets.

We report 95% confidence intervals in the below tables. Results where another method's estimated statistic overlaps 3DP's 95% CI are indicated with a $*$. We observe only two overlaps on the B3DO dataset for methods that are substantially worse on the other metrics. The only confidence interval overlap on the NYU Dataset is a negligible overlap when comparing with Singh et al. in RMSE: their RMSE is right at the upper limit of 3DP's RMSE confidence interval. This strongly suggests that 3DP is capable of consistently outperforming the baselines.

[1] Paired t-tests for mean and RMSE; wilcoxon rank-sum test for median; and equality of proportions test for % Good Pixels

NYU Depth v2 Dataset

With Manhattan-world constraints

	Summary Stats. (°)		
	(Lower Better)		
	Mean	Median	RMSE
Lee	42.9 (41.6,44.1)	34.6 (31.4,37.3)	54.8 (53.7,55.9)
Hedau	41.2 (40.0,42.5)	25.5 (23.3,28.0)	55.1 (54.0,56.2)
3DP	**33.5** (32.5,34.5)	**18.0** (16.5,19.2)	**46.6** (45.7,47.6)

	% Good Pixels		
	(Higher Better)		
	11.25°	22.5°	30°
Lee	24.8 (22.9,26.8)	40.5 (38.4,42.6)	46.7 (44.7,48.7)
Hedau	33.2 (31.4,35.0)	47.7 (45.9,49.5)	53.0 (51.4,54.7)
3DP	**37.4** (35.6,39.3)	**55.0** (53.3,56.7)	**61.2** (59.5,62.6)

Without Manhattan-world constraints

	Summary Stats. (°)		
	(Lower Better)		
	Mean	Median	RMSE
Depth Transfer	40.8 (40.0,41.4)	38.2 (37.4,39.3)	46.7 (46.0,47.4)
Make 3D	47.1 (46.1,48.1)	42.3 (40.6,44.1)	56.3 (55.3,57.2)
Geometric Context	41.1 (40.2,42.0)	34.9 (33.4,36.6)	49.2 (48.3,50.0)
Singh et al.	35.0 (34.4,35.6)	32.5 (31.8,33.3)	∗ 40.4 (39.9,41.0) ∗
RF + SIFT	36.0 (35.4,36.5)	33.5 (32.9,34.3)	41.5 (41.0,42.1)
SVR + SIFT	36.2 (35.7,36.8)	33.1 (32.5,33.8)	42.1 (41.5,42.7)
3DP	**32.7** (32.0,33.4)	**27.6** (26.6,28.4)	**39.7** (39.0,40.4)

	% Good Pixels		
	(Higher Better)		
	11.25°	22.5°	30°
Depth Transfer	7.6 (6.9,8.3)	25.0 (23.7,26.4)	37.5 (36.1,39.0)
Make 3D	11.2 (10.3,12.2)	28.0 (26.5,29.4)	37.4 (35.8,39.0)
Geometric Context	8.9 (7.8,10.3)	31.3 (29.3,33.5)	43.5 (41.4,45.5)
Singh et al.	11.5 (10.8,12.2)	32.0 (30.8,33.2)	45.7 (44.4,47.0)
RF + SIFT	11.2 (10.7,11.7)	30.9 (30.0,31.8)	44.2 (43.1,45.2)
SVR + SIFT	10.8 (10.4,11.2)	31.0 (30.1,31.9)	44.4 (43.4,45.5)
3DP	**19.0** (17.9,20.0)	**41.4** (40.0,42.8)	**53.4** (52.0,54.8)

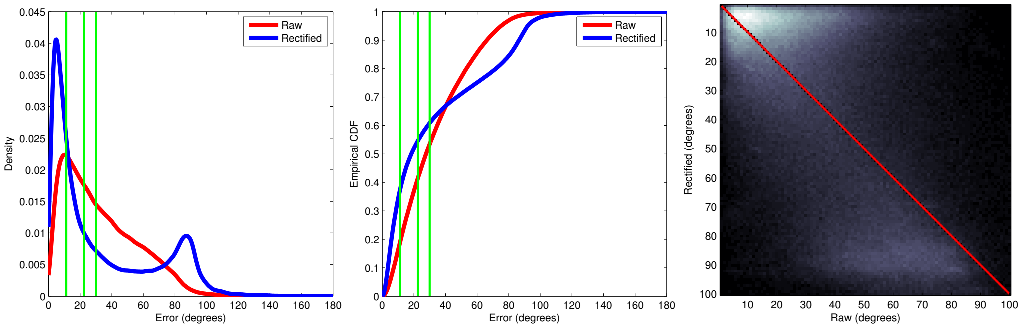
Berkeley 3D Object Dataset

With Manhattan-world constraints

		Summary Stats. (°)	
		(Lower Better)	
	Mean	Median	RMSE
Lee	41.9 (40.7,43.2)	28.4 (25.5,31.1)	56.6 (55.4,57.9)
Hedau	43.5 (42.3,44.9)	30.0 (27.2,33.2)	58.1 (56.8,59.6)
3DP	**38.0** (37.0,39.0)	**24.5** (22.6,26.7)	**51.2** (50.3,52.1)
		% Good Pixels	
		(Higher Better)	
	11.25°	22.5°	30°
Lee	∗ 32.7 (30.9,34.5) ∗	45.7 (43.8,47.5)	50.8 (49.1,52.5)
Hedau	∗ 32.8 (31.1,34.5) ∗	45.0 (43.3,46.7)	50.0 (48.4,51.6)
3DP	**33.6** (31.9,35.3)	**48.5** (46.8,50.2)	**54.5** (52.9,56.0)

Without Manhattan-world constraints

		Summary Stats. (°)	
		(Lower Better)	
	Mean	Median	RMSE
Make 3D	45.6 (44.8,46.3)	41.2 (40.0,42.4)	53.5 (52.8,54.3)
Geometric Context	41.9 (41.1,42.8)	37.2 (35.9,38.2)	49.7 (48.9,50.5)
Singh et al.	36.7 (36.2,37.2)	34.0 (33.3,34.6)	42.2 (41.7,42.7)
RF + SIFT	36.8 (36.4,37.3)	34.1 (33.4,34.6)	42.5 (42.0,42.9)
SVR + SIFT	36.9 (36.5,37.3)	33.9 (33.3,34.5)	42.6 (42.1,43.0)
3DP	**34.5** (34.0,35.1)	**30.5** (29.8,31.4)	**41.0** (40.5,41.6)
		% Good Pixels	
		(Higher Better)	
	11.25°	22.5°	30°
Make 3D	8.4 (7.7,9.0)	25.5 (24.1,26.8)	36.1 (34.6,37.6)
Geometric Context	8.3 (7.3,9.6)	25.8 (24.0,27.7)	38.3 (36.3,40.3)
Singh et al.	9.9 (9.5,10.4)	29.4 (28.5,30.4)	43.0 (42.0,44.1)
RF + SIFT	10.2 (9.8,10.6)	29.6 (28.9,30.4)	43.0 (42.1,43.9)
SVR + SIFT	9.7 (9.3,10.1)	29.4 (28.6,30.2)	43.1 (42.2,44.1)
3DP	**14.5** (13.7,15.2)	**36.0** (34.8,37.2)	**49.2** (48.0,50.4)



(a) Distribution (thresholds used in green)

(b) Empirical CDF (thresholds used in green)

(c) Joint histogram (non-linear intensity)

Figure 1: Characterization of the distribution of errors before and after rectification.

2 How does the Manhattan-world assumption change the error distribution?

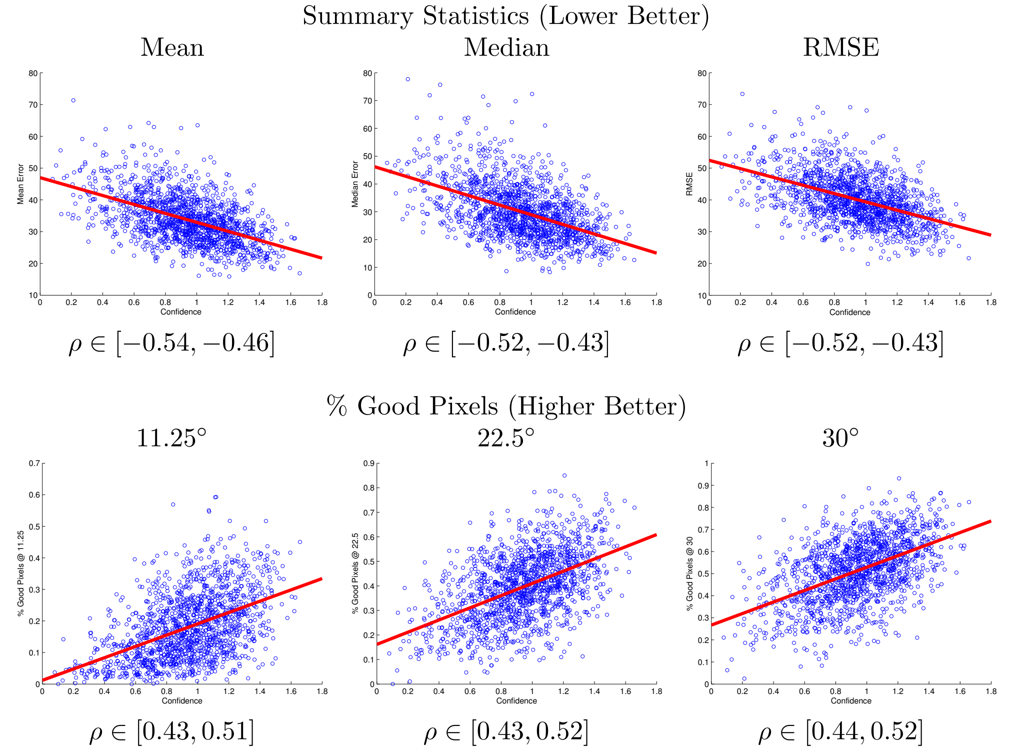
Another natural question is – how does incorporating the Manhattan world assumption by rectifying the predictions (i.e., snapping them to the nearest vanishing point) change the distribution of errors? Looking at the tables, the mean error increases mildly and RMSE increases dramatically; the median goes down equally dramatically. The percent-good-pixels goes up, although the increase seems to go down quite a bit as we increase the threshold.

We show the distribution of errors viewed from a variety of lenses in Fig. 1. Fig. 1(a), shows that rectification clears out moderate errors and pushes them towards either accurate or inaccurate. Thus, the median goes down dramatically, while the RMSE, which stronly penalizes large errors, is dramatically increased. The mean error, nonetheless goes down a small amount. The CDF $F(x)$, shown in Fig. 1(b), can also be interpreted as the percent-good-pixels at each threshold x; the explosive initial growth for the rectified approach levels off, leading to decreased gains in percent-good-pixels; eventually at about 41°, the rectified approach is overtaken by the raw output. Thus, if one only cares about 45° accuracy, it is not clear which approach would be better.

Finally, Fig. 1(c) confirms the interpretation of pushing moderate errors towards larger or small. This shows the joint histogram of raw and rectified errors (with the square root of the bin shown), with raw being the rows and rectified the columns. A red line divides the histogram in half: energy below the line are results where rectification degrades performance; energy above where it improves. Each row shows the distribution of raw errors for that particular rectified error. For instance, in the top rows (ones below 10°), the raw errors (columns) are largely from above 10°. The other salient mode is towards the bottom rows, where there are frequent instances of results with high rectified error (e.g., the rows between 85° and 100°) but with much lower raw error ($60° - 80°$).

3 Relationship between confidence and performance

In other parts of the supplementary material, we show results ranked according to their confidence. Specifically, for an image, we use the per-pixel mean normalization Z used during



Summary Statistics (Lower Better)

Mean

Median

RMSE

$\rho \in [-0.54, -0.46]$

$\rho \in [-0.52, -0.43]$

$\rho \in [-0.52, -0.43]$

% Good Pixels (Higher Better)

11.25°

22.5°

30°

$\rho \in [0.43, 0.51]$

$\rho \in [0.43, 0.52]$

$\rho \in [0.44, 0.52]$

Figure 2: Performance vs. algorithm confidence. Here we plot the metric against the algorithm confidence Z and a line of best fit. We also report a 95% CI for the spearman rank correlation.

dense-transfer as a confidence measure. This is the average sum of calibrated detector scores. This qualitatively seems to select good results, which we confirm quantitatively here. We show scatter plots of per-image scores or errors in Fig. 2. These show a relationship between the normalization and the performance of the approach. We also report the Spearman rank-correlation ρ for each variable in the form of a bootstrapped 95% confidence intervals. In all cases, confidence is negatively correlated with errors (i.e., higher confidence is lower error) and postively correlated with % Good Pixels.

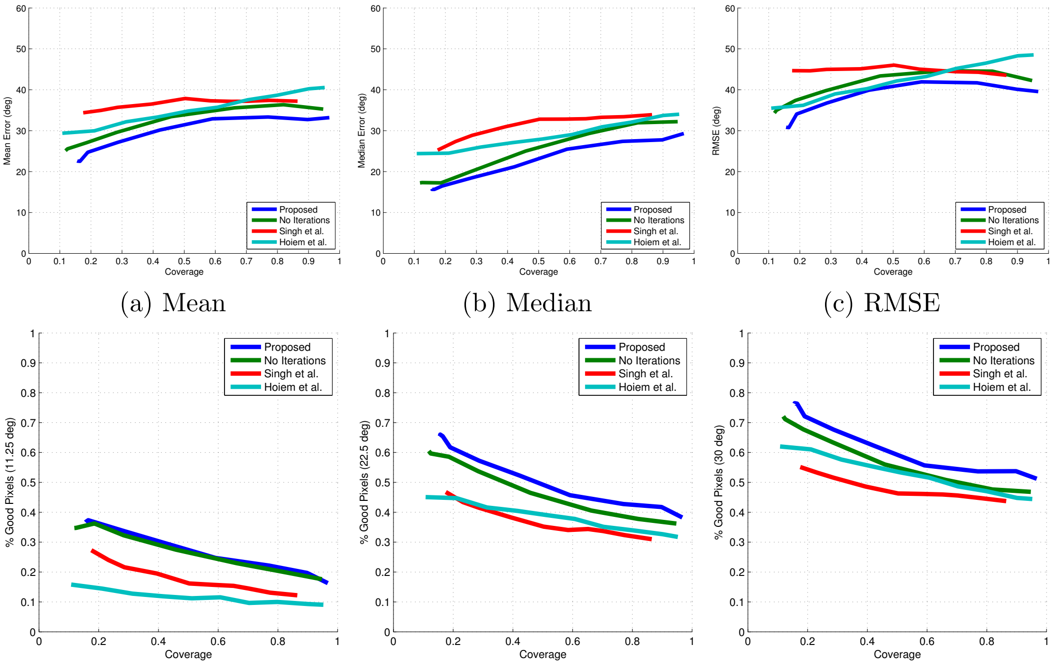
4 Additional Graphs

We now present complete precision-vs-coverage curves for techniques capable of providing confidence measures (thus enabling a sliding threshold). We again separate techniques into non-Manhattan-world (Fig. 3) and Manhattan-world (Fig. 4) techniques. In the case of the room fitting approach of Hedau et al., we show the operating point if pixels predicted as clutter are ignored.

3DP does as well or beats each method at all considered operating points. The next-best non-Manhattan technique is the 3DP algorithm with no repeated iterations (i.e., no update of membership assignment $\mathbf{y}$). Among the Manhattan-world techniques, the room-fitting approach does competitively with 3DP; however, it can provide only a single operating point in the very low-recall regime, and does considerably worse in the dense, 100% recall case.

References

[1] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.



(a) Mean

(b) Median

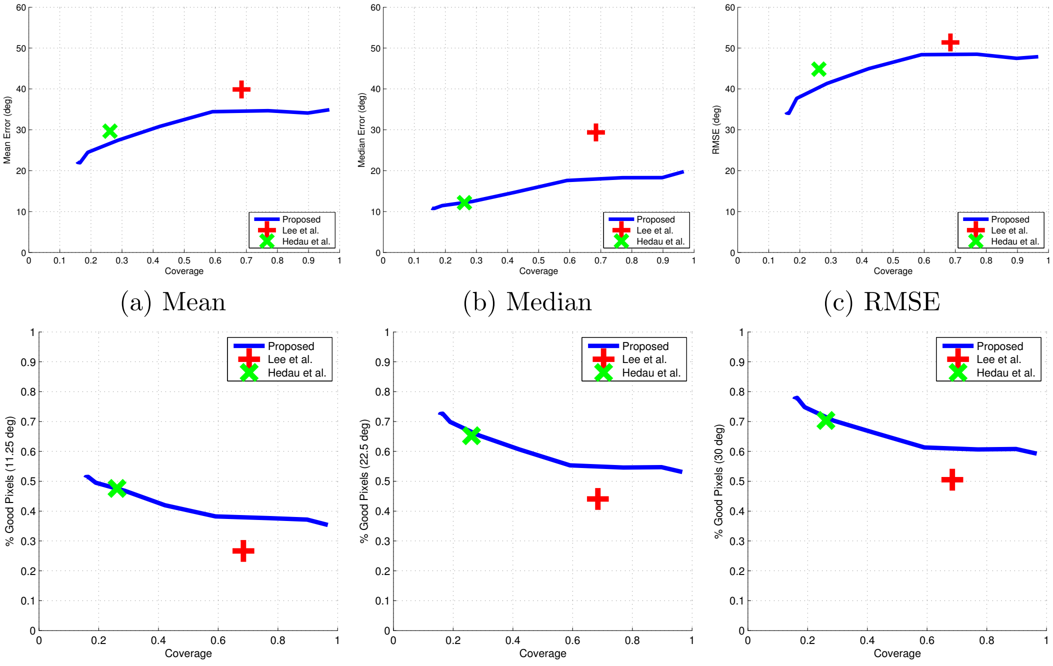
(c) RMSE

(d) % Good Pixels 11.25°

(e) % Good Pixels 22.5°

(f) % Good Pixels 30°

Figure 3: Complete precision-vs-coverage curves for all performance metrics used. **Non-Manhattan-world techniques**



(a) Mean

(b) Median

(c) RMSE

(d) % Good Pixels 11.25°

(e) % Good Pixels 22.5°

(f) % Good Pixels 30°

Figure 4: Complete precision-vs-coverage curves for all performance metrics used. **Manhattan-world techniques**