Measuring the Performance of Single Image Depth Estimation Methods

Cesar Cadena, Yasir Latif, and Ian D. Reid

***Abstract*— We consider the question of benchmarking the performance of methods used for estimating the depth of a scene from a single image. We describe various measures that have been used in the past, discuss their limitations and demonstrate that each is deficient in one or more ways. We propose a new measure of performance for depth estimation that overcomes these deficiencies, and has a number of desirable properties. We show that in various cases of interest the new measure enables visualisation of the performance of a method that is otherwise obfuscated by existing metrics. Our proposed method is capable of illuminating the relative performance of different algorithms on different kinds of data, such as the difference in efficacy of a method when estimating the depth of the ground plane versus estimating the depth of other generic scene structure. We showcase the method by comparing a number of existing single-view methods against each other and against more traditional depth estimation methods such as binocular stereo.**

I. Introduction

Single image depth estimation methods try to predict the 3D structure of a scene (i.e. the depth at each point in the scene) from a single photometric view, thus recovering the depth information that is lost during the imaging process. Inspired in part by the ability of humans to perform this task (albeit usually qualitatively), single-view depth estimation has become an active area within the computer vision community, with various methods proposed recently, including [1, 3, 4, 7, 9, 10, 11, 13]. To evaluate and compare the efficacy of these approaches, various metrics and methods have been proposed. However, as we show in this paper, each of these measures is deficient in one or more ways. To address this issue we propose a new measure of performance for depth estimation that has a number of desirable properties.

For most tasks, performance of a method to solve the task is measured by comparing the output of the method against some known ground truth. In our case, we are interested in depth estimation, and the "ground truth" here is typically a single depth image, normally acquired by a depth camera. Most existing metrics for benchmarking single-view depth estimation aim to capture, for each pixel in the predicted image, the closeness of the prediction to the corresponding pixel in the ground truth; i.e. they operate in the *image space*. A number of such metrics have been reported for image space comparisons, and we give an overview in Section II. An issue with these metrics is that they assume that the estimated and ground truth depth images have the same resolution, which is often not the case in practice. They also, typically address the issue of missing depth estimates by calculating the metric only over the set of pixels where both the prediction and the ground truth have values. However, as we will show later, problems arise in scenarios where a pixel by pixel comparison is not feasible against the full resolution ground truth especially in cases when: **a)** the resolution of the prediction does not match that of the ground truth, **b)** the density of the prediction and ground truth are not the same, or **c)** the coverage of the prediction is not the same as that of the ground truth (more on coverage and density later).

Fig. 1: RGB image and the corresponding ground truth depth obtained using a depth camera taken from the NYU-V2 dataset.

In this work, we advocate that comparisons should always be made against the given ground truth without up/down sampling, without recourse to in-painting (hallucination) of "ground truth", and should adequately portray how much of the ground truth is explained by each method. In particular, unlike most previous comparison methods, we propose to compare the predicted depths to ground truth in *3D-space* instead of the image space. Section III defines our proposed performance measure, and shows that it is agnostic to the differences in resolution, density, and coverage between the ground truth and the estimated depths.

Using the proposed measure, in Section IV we compare the performance of state of the art single image depth estimation methods for the NYU-V2 dataset [14]. This dataset comes with hand labelled semantic segmentation annotations. Then, we show how our performance measure takes advantages of the semantic classes to give more insights for each estimation method. Finally in Section V, we also show a comparison between classical stereo depth estimation and single image depth estimation in an outdoors setting using the KITTI dataset [5].

II. Current Metrics

Given a predicted depth image and the corresponding ground truth, with $\hat{d}_p$ and d_p denoting the estimated and

Cesar Cadena is with the Autonomous Systems Lab at ETH Zurich, Leonhardstrasse 21, 8092, Zurich, Switzerland. cesarc@ethz.ch

Yasir Latif and Ian D. Reid are with the Department of Computer Science, at the University of Adelaide, Adelaide, SA 5005, Australia. {yasir.latif, ian.reid} @adelaide.edu.au

We are extremely grateful to the Australian Research Council for funding this research through project DP130104413, the ARC Centre of Excellence for Robotic Vision C E140100016, and through a Laureate Fellowship FL130100102 to IDR.

ground-truth depths respectively at pixel p, and T being the total number of pixels for which there exist both valid ground truth and predicted depth, the following metrics have been reported in literature:

- Absolute Relative Error [13]

$$\frac{1}{T}\sum_{p}\frac{|d_p - \hat{d}_p|}{d_p}$$

- Linear Root Mean Square Error (RMSE) [8]

$$\sqrt{\frac{1}{T}\sum_{p}(d_p - \hat{d}_p)^2}$$

- log scale invariant RMSE (as proposed in [4])

$$\frac{1}{T}\sum_{p}\left(\log \hat{d}_p - \log d_p + \alpha(\hat{d}_p, d_p)\right)^2$$

  where $\alpha(\hat{d}_p, d_p)$ addresses scale alignment.
- Accuracy under a threshold [7]

$$\max\left(\frac{\hat{d}_p}{d_p}, \frac{d_p}{\hat{d}_p}\right) = \delta < th$$

  where th is a predefined threshold.

In the following, we present various cases which highlight the deficiencies of these metrics. In the scenarios presented here, the ground truth is taken from the NYU-V2 dataset [14]. We use the in-painted depth from the NYU-V2 as the estimation of the system since it is already close to the ground truth, therefore, any minor differences introduced by various transformations can be observed more easily.

A. Resolution

Single image depth estimation methods may predict a lower resolution depth image compared to the ground truth. Traditionally, comparisons for this case are done by down sizing the ground truth to the size of the estimation. However, if we insist on keeping the ground truth unchanged, the prediction can be upsampled using an appropriate scaling. A natural question is, which is a better way of comparing the two? Should the ground truth be scaled down or the prediction scaled up for comparison? Would the performance metrics be different in both cases?

To observe the effect of different resolutions, we use 5 different sized depth estimations, each derived from the in-painted depth using nearest-neighbour down-sampling: the full resolution, half, and down to one-sixteenth of the original image resolution. The results can be seen in Table I [In-painting], where the ground truth has been down sampled to match the resolution of the predicted depth. All the predictions have more or less the same performance based on these metrics.

Instead of down-sampling the ground truth, an alternative approach would be to up-sample the depth predictions using some form of interpolation. Here, we use the bilinear interpolation to up sample all the estimations to match the same resolution as the ground truth. Once again, we present a performance comparison using the traditional metrics (Table I [downscaling, then upscaling])

It is interesting to observe that in Table I, the performance at each resolution is now worse than the case in which we down scaled the ground truth [downscaling]. Up scaling creates information by interpolation, which may not always agree with the ground truth at that particular pixel position. The smaller the original resolution, the more we upscale the prediction, and the more error we introduce; this is reflected in the metrics. It should be noted that the initial depth estimation is very accurate since it is derived from the in-painted ground truth.

B. Density

Single image depth estimation involves predicting depth for each pixel of the input image; we refer to this as "dense estimation". On the contrary, in robotics it is common to track a set of sparse points while estimating their depth. We refer to that scenario as being "sparse estimation". In this case, depth is not available for each pixel of the image under consideration but only at a set of predetermined points which satisfy a certain criterion (dominant corner, gradient etc.). This scenario frequently arises in a robotics setting when carrying out Simultaneous Localization And Mapping (SLAM).

In the following, we show how the metrics behave with respect to the density of the points. We extract dominant corners, FAST keypoints [12], from the image and select the corresponding depth from the in-painted depth as the prediction for the extracted corner. This is repeated for a range of points from 10 to 2000 (as shown in Table I [keypoints]).

It can be seen from Table. I [keypoints] that these metrics do not capture the complexity of the estimation as they are all designed for dense estimation. Complexity in this case just refers to the density of the prediction – predicting the depth for fewer points is a less complex problem than predicting a full dense depth image. In the case of sparse prediction, the metrics are calculated over the intersection: taking into consideration only the points that are common to both the prediction and the ground truth. This introduces a favourable bias towards systems that predict sparse depths, but do so very accurately and in the extreme case, a single point predicted extremely accurately will lead to a very good score on most of these metrics.

C. Coverage

Another aspect of the problem is that of coverage, which is related to density but has a slight different meaning. A dense prediction that covers the whole image is said to have full coverage. However, prediction can be dense without covering the whole image (imagine the scenario when a system predicts depths for just planar areas in the scene such as ground or road). Evaluating this scenario has the same issue as that of sparse estimation: the metrics are evaluated over the intersection and therefore do not capture

TABLE I: Results on NYU-V2 dataset.

Method		Errors: Absolute Relative	Errors: RMSE linear [m]	Errors: RMSE log.sc.inv.	Accuracy $\delta <$ 1.25[%]	Accuracy $\delta <$ 1.25^2[%]	Accuracy $\delta <$ 1.25^3[%]
In-painting		1.2e-3	7.8e-2	11.4e-3	99.95	99.99	100.0
downscaling	1/2	1.5e-3	7.7e-2	11.4e-3	99.95	99.99	100.0
	1/4	1.6e-3	7.8e-2	11.2e-3	99.95	99.99	100.0
	1/8	1.6e-3	7.7e-2	11.2e-4	99.95	99.99	100.0
	1/16	1.6e-3	7.6e-2	11.2e-4	99.95	99.99	100.0
downscaling, then upscaling	1/2	2.5e-3	8.5e-2	15.6e-3	99.91	99.98	100.0
	1/4	3.7e-3	9.1e-2	19.2e-3	99.86	99.96	99.99
	1/8	6.7e-3	10.9e-2	27.9e-3	99.70	99.95	99.99
	1/16	12.8e-3	15.0e-2	44.1e-3	99.17	99.95	99.97
keypoints [#]	10	4.3e-3	12.6e-2	16.9e-3	99.9	99.9	100.0
	100	4.1e-3	13.3e-2	20.7e-3	99.9	99.9	100.0
	500	3.4e-3	11.5e-2	18.4e-3	99.9	100.0	100.0
	1000	3.0e-3	10.7e-2	17.4e-3	99.9	100.0	100.0
	2000	2.7e-3	10.3e-2	16.4e-3	99.9	100.0	100.0
coverage [%]	18	1.2e-3	1.0e-2	5.3e-3	100.0	100.0	100.0
	35	1.3e-3	3.1e-2	6.9e-3	100.0	100.0	100.0
	53	1.4e-3	7.3e-2	11.3e-3	100.0	100.0	100.0
Estimations		Full Scene Evaluation					
Mean		0.281	1.01	0.296	44.5	73.9	89.2
Eigen *et al*. [4]	coarse	0.221	0.81	0.213	63.2	90.1	97.4
	↑coarse	0.157	0.73	0.213	63.2	90.1	97.3
	fine	0.209	0.83	0.210	62.4	89.8	97.6
	↑fine	0.144	0.75	0.210	62.6	89.9	97.6
Liu *et al.*[10]		0.143	0.64	0.206	67.6	92.1	98.1
Eigen *et al.* [3]	multi	0.192	0.68	0.192	70.9	91.9	98.0
	↑multi	0.139	0.63	0.192	70.9	91.9	98.0
		Per Semantic Class					
Floor							
Floor Mean		0.144	0.51	0.059	80.6	98.4	99.8
Mean		0.175	0.71	0.090	62.3	92.4	98.8
Eigen *et al*. [4]	↑coarse	0.154	0.57	0.102	71.7	95.5	99.3
	↑fine	0.176	0.64	0.104	63.8	92.5	98.8
Liu *et al.*[10]		0.124	0.44	0.094	84.1	98.9	99.8
Eigen *et al.* [3]	↑multi	0.093	0.37	0.085	90.3	97.5	99.6
Structure							
Struct. Mean		0.371	1.22	0.261	43.2	71.9	88.2
Mean		0.376	1.22	0.250	43.6	71.9	87.2
Eigen *et al*. [4]	↑coarse	0.200	0.82	0.169	65.2	91.1	97.6
	↑fine	0.192	0.85	0.167	64.8	90.5	97.8
Liu *et al.*[10]		0.197	0.76	0.177	65.4	91.7	98.2
Eigen *et al.* [3]	↑multi	0.188	0.71	0.155	70.8	92.4	98.2
Furniture							
Furn. Mean		0.315	0.88	0.254	48.0	79.6	93.5
Mean		0.384	0.93	0.254	42.7	73.7	90.8
Eigen *et al*. [4]	↑coarse	0.224	0.64	0.189	62.8	90.1	97.7
	↑fine	0.202	0.64	0.187	64.1	91.4	98.2
Liu *et al.*[10]		0.199	0.55	0.180	68.1	92.4	98.3
Eigen *et al.* [3]	↑multi	0.198	0.58	0.175	69.3	92.0	98.2
Props							
Props Mean		0.379	1.02	0.258	39.8	71.2	90.1
Mean		0.467	1.09	0.256	38.4	65.8	83.8
Eigen *et al*. [4]	↑coarse	0.274	0.77	0.197	56.8	85.8	95.4
	↑fine	0.240	0.75	0.195	59.3	87.6	96.6
Liu *et al.*[10]		0.242	0.69	0.200	61.3	88.5	96.6
Eigen *et al.* [3]	↑multi	0.254	0.74	0.192	61.7	87.0	96.0

the complexity of the problem. When making comparisons, a system that does not predict at full coverage would therefore have an unfair advantage over those that make a full dense prediction.

To observe the effect of coverage on these metrics, we use cropped versions of the in-painted depth at various coverage levels as shown in Fig. 2. The evaluation is report in Table I [coverage], where the smallest coverage has the lowest RMSE, that is, it performs better by predicting just 18% of all the depths in the ground truth.

III. Proposed Performance Measure

We propose to compare the predicted depth against the ground truth in the 3D space instead of the image space. For a pixel p with image coordinates u_p and v_p and a predicted depth $\hat{d}_p$, the point in 3D is given by

$$\hat{\mathbf{x}} = \hat{d}_p \mathbf{K}^{-1} \mathbf{x_p}$$

where K contains the known camera intrinsics and $\mathbf{x_p} = [u_p \; v_p \; 1]^T$. Similarly, for each point in the ground truth, a corresponding point $\mathbf{x}$ in the 3D-space can be calculated.

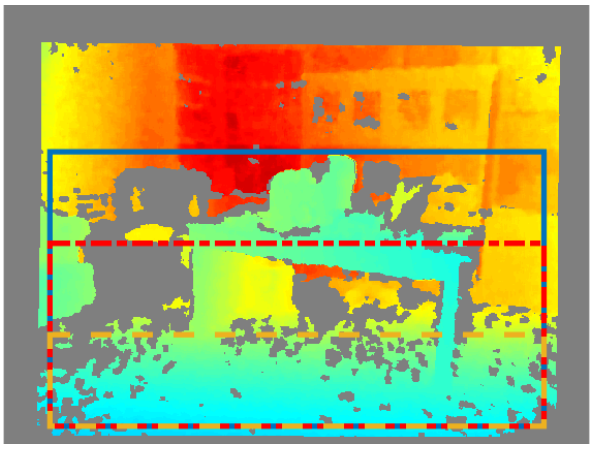


Fig. 2: We evaluate the effect of three different partial coverages of the scene. In the figure we show one depth image with rectangles covering the 53% (blue, solid line), 35% (red, dot-dashed line) and 18% (yellow, dashed line) of the full scene.

For each point $\mathbf{x_i}$ in the ground truth, we search for the nearest point in the estimated depth and form a set of all these nearest-neighbour distances:

$$\mathcal{S} = \{d_i | d_i = \min ||\mathbf{x_i} - \hat{\mathbf{x}}||\} \tag{1}$$

This set has the same cardinality as the number of pixels in the ground truth with valid depths. The objective function minimised by the Iterative Closest Point (ICP) algorithm is typically based on a sum or robust sum over $\mathcal{S}$, and so our measure naturally generalises to the case where there is a rigid misalignment between the ground-truth and the estimated depth (via application of ICP). In all the evaluations we report, however, the depth esimates and ground truth are already aligned.

For a given threshold D, we look for the distances in $\mathcal{S}$, that are less that D:

$$\mathcal{S}_\mathcal{D} = \{d_i | d_i \in \mathcal{S} \wedge d_i < D\} \tag{2}$$

and plot the ratio of cardinalities $|\mathcal{S}_\mathcal{D}|/|\mathcal{S}|$ which represents the fraction of ground truth that is explained by the estimation with a distance less that D. This threshold is increased until all of the ground truth is explained; that is, the ratio reaches 1.

Comparison using the closest point error allows for certain nice properties: it is no longer required to have the same resolution of the estimation and ground truth since a closest point always exists for each point in the ground truth. A visual representation of the metric, where each point in the ground truth is connected to its nearest neighbour can be see in Fig. 3. Moreover, now we have a systematic way of comparing estimations with different densities and different coverages. To illustrate, we revisit the scenarios presented in Section II.

Resolution: We first compare the performance using the proposed measure for the five down-sampled resolutions derived from the in-painted ground truth. The results are

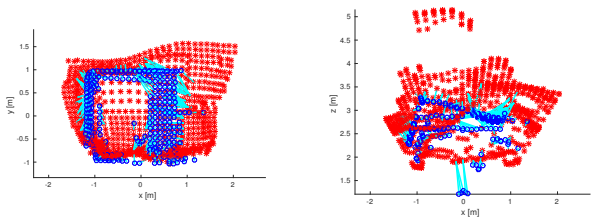


Fig. 3: A toy example to evaluate the accuracy of the estimation (red crosses) with respect to the ground truth (blue circles). The cyan lines denote the closest estimated point to each ground truth 3D point.

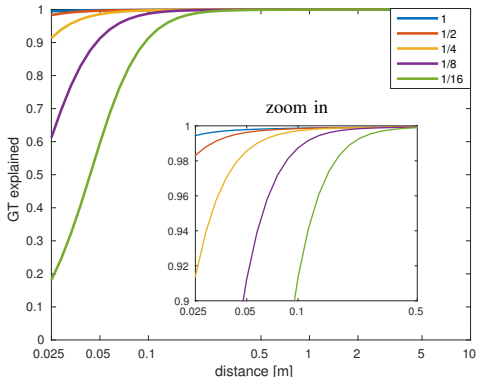
presented in Fig. 4a. It can been seen that under this measure, the difference in performance between various resolutions is prominent with the full resolution in-painted depth performing the best. Under older metrics, all the resolutions had more or less the same performance. The inset in Fig. 4a shows a zoom-in on the top-left corner to highlight the performance of the curves starting very close to 1. As we expect, higher resolution estimations perform better than lower resolutions.

The other case discussed in Section II is that of up-sampling the estimated depth to the resolution of the ground truth. Under the proposed performance measure, Fig. 4b shows a significant improvement for all curves. Up scaling, while introducing errors in the image space, leads to a reduction in the distance to the nearest neighbour (on average) for each ground truth point in 3D-space. This leads to the gain in performance reflected here. Later on, for real single image depth prediction methods, we show that the up scaling does not have a significant effect on the performance measure. This is because the initial low resolution estimation is far from the ground truth and then the interpolation during up sampling does not help.

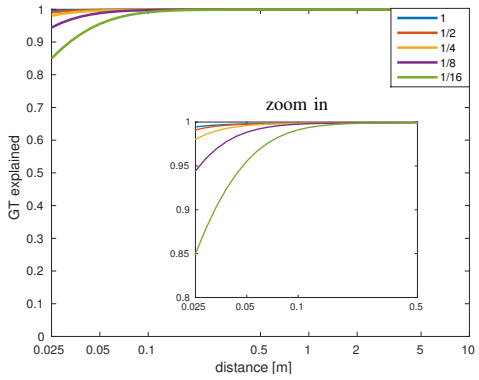
Density: It can be seen in Fig. 4c that by requiring that each ground truth point have a corresponding nearest-neighbour in the 3D-space, sparser predictions are ranked correctly, that is, in order of their complexity. This is true since the estimated depth has been derived from the ground truth and all the predictions for point depths are almost correct. However, methods are penalized when the points are sparse as the nearest neighbour for a ground truth lies farther away.

Coverage: The results for the proposed measure are shown in Fig. 4d, which paints a more complete picture than the scalar metrics in Table I. From the perspective of the possible explanation of ground truth, higher coverage performs better.

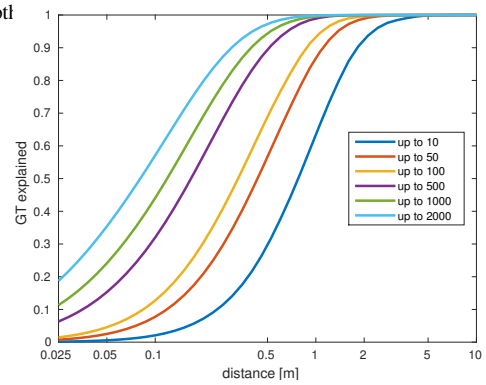
Finally, we show a side-by-side comparison for all the three previous cases in Fig. 5 where we show the full in-painted prediction, a down-sample one-sixteenth version, 1000 strongest key-points, and a partial coverage scenario. It can be seen that the partial coverage behaves the worst, even though it starts off higher than other curves. The reason behind this is that in order to explain the full ground truth with the partial coverage, the nearest neighbour for the



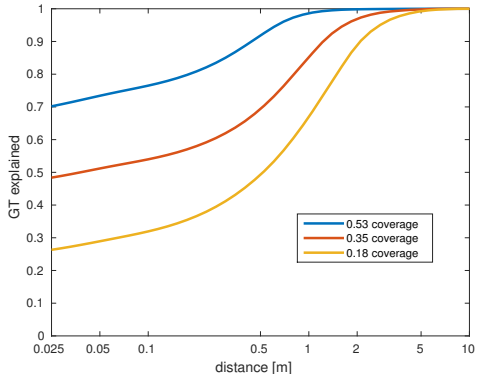
(a) Evaluation of low resolution version of inpainted depth.



(b) Bilinear upscaling of low resolution versions of inpainted deptl



(c) Keypoints.



(d) Partial coverages.

Fig. 4: Different cases evaluated in this section.

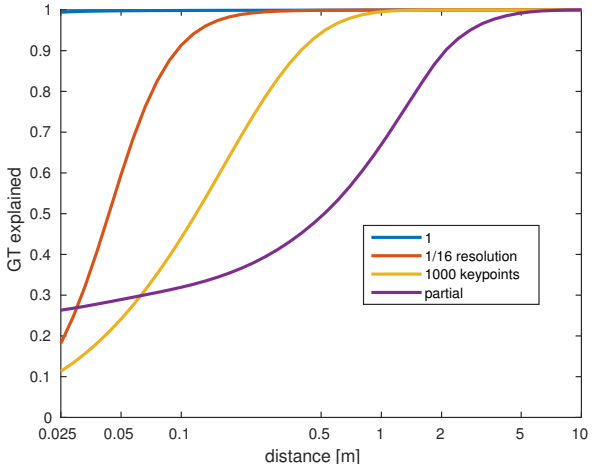


Fig. 5: Comparison between $1/16$ low resolution (40x30=1200 points), 1000 extracted keypoints, and a partial box of 540x100=54000 points (18% coverage).

ground truth point is further away than in all the other cases, leading to the slow rise of the curve. It can be said that the performance of the down-sampled prediction is better than the keypoints as the down-sampled prediction provides almost the same amount of point (1200) but does so over a regular grid in the the 3D space (induced by the regularity of the image grid), while the keypoints are not uniformly distributed and are defined by the structure of the scene. The proposed method allows us to compare these different kinds of depth estimates which would not be meaningful under traditional metrics.

IV. Comparing the State of the Art in Indoors

Previous sections presented some toy case studies of various situations in which using the proposed measure leads to a more meaningful comparison. This section provides the performance evaluation of some of the state-of-the-art algorithms using the proposed method; in particular we provide results for Eigen *et al.* [4] (coarse, fine), Liu *et al.* [10] and Eigen and Fergus [3].

The system of Liu *et al.* [10] makes the prediction at the same resolution as the ground truth, while the other two methods make depth prediction at lower resolution. For the latter, we use both the original predictions as well as upscaled version using bilinear interpolation. Table I [Full scene evaluation] reports the comparison for all these combinations as well as the mean of the training data.

We first consider the performance of the original (before upscaling) predictions where the ground truth has been scaled down to the same resolution as the prediction. Using the old metrics there is a difference in ordering of performance when we compare the upscaled versus the original predictions. The accuracy under a threshold measure is not effected by upscaling, but in general, simple upscaling leads to a better performance on the other metrics, as well as permuting the order of performance.

We now turn to the evaluation on the basis of the proposed method, in Fig. 7. The solid lines represent the original

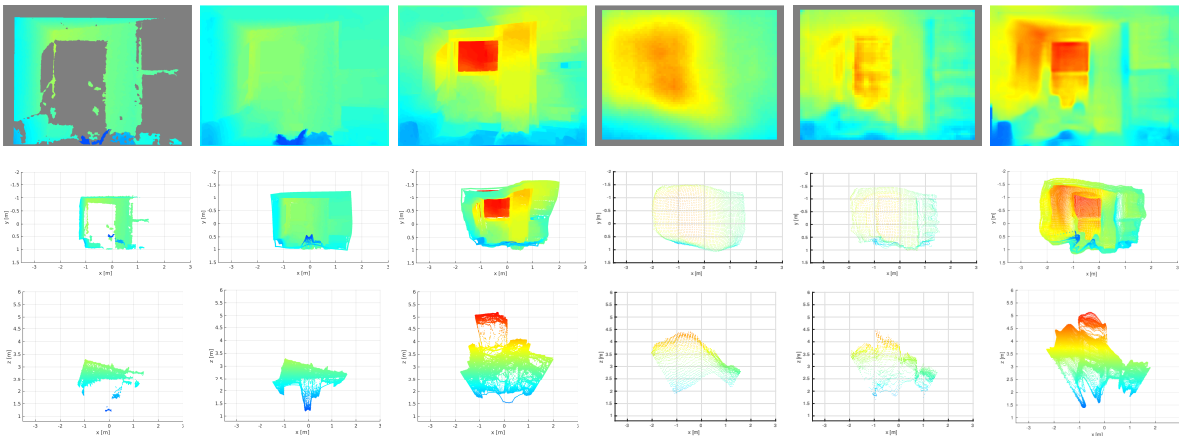


Fig. 6: One example from the NYU-V2 test set. *Left to Right*: Ground truth, Inpainted Ground Truth, Liu *et al*. [10], Eigen *et al*. [4] coarse, Eigen *et al*. [4] fine, and Eigen and Fergus [3]. *From top to bottom* : Depth Image, XY-view and XZ-view of the point cloud. The corresponding RGB image is shown in Fig. 1.

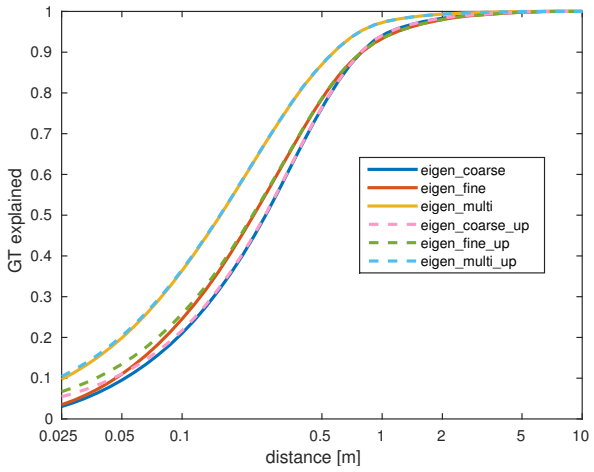


Fig. 7: Performance of orginal and full resolution depth estimates: eigen_coarse and eigen_fine [4], eigen_multi [3], **_up** indicates upscaling to the ground truth resolution

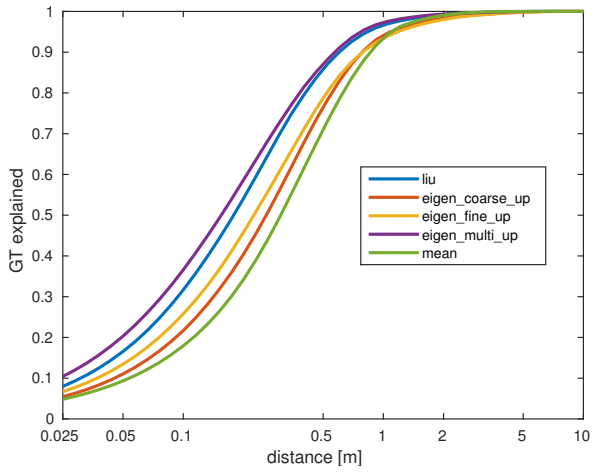


Fig. 8: Performance of full resolution depth estimates: liu [10], eigen_coarse and eigen_fine [4], eigen_multi [3]

estimates, while the dashed lines represent the upsampled versions. It can been seen that upsampling has no effect on the order of the curves.

We also present the output of these methods at a resolution matching the ground truth in Fig. 8. This gives an overview of the performance for each method. For a given allowed error, we can see the percentage of the ground truth that is associated with that distance. Additionally, the curve that remains higher than all the rest shows that is more accurate at every given distance compared to other methods. We can interpret this in terms of the distribution of the closest-point distance from the ground truth. The sharp rise in the middle indicates that the majority of distances lie in this area which are responsible for explaining most of the ground truth.

A. Break down by classes

The comparison presented so far has been over the full scenes in the NYU-V2 dataset without taking into account the semantic segmentation the dataset also provides. Since our proposed method is resolution, density, and coverage agnostic, we can use it to provide deeper insight into each method to evaluate their performance on each individual semantic class. The dataset provides ground truth semantics for categories such as Floor, Structure, Furniture, and Props.

To evaluate the per class performance, we use only the part of the ground truth with the particular class label and compare it against the whole prediction. This allows us to compare the semantic labels even when there is no prediction of them in the original methods. We also report the comparison of semantics against the mean of the dataset as well the mean over each semantic category. This allows us to have a "look under the hood" to find out why a given method performs better than others. Some semantic categories such as floor have lower complexity than others such as props, due to intra-class variations in depths.

Performance for each semantic category is given in Fig. 9. Curves that remain higher than the category mean have a better prediction of the semantic category on average for that class label. This can been seen in the Structure subplot,

where most of the methods are able to do a better job than the category mean. On the other hand, Props and Furniture are difficult categories for these methods. Eigen *et al.* [4] methods also seem to have some difficulty predicting Floor depths correctly. This novel insight comes from the ability to compare the depths class-wise without needing the prediction of class labels from the methods being evaluated.

V. Performance in Outdoors

This section provides the performance evaluation of Eigen *et al.* [4] and Cadena *et al.* [2] under our proposed measure for an outdoors setting, specifically on the KITTI dataset [5]. Authors of [4] already provide the single image depth estimations and the set of frames for testing.

This dataset provides 3D point clouds from a rotating LIDAR scanner. We project the 3D information on the left image for each frame to obtain the ground truth depth using the official toolbox provided with the dataset. This projection results in sparse depth images covering the bottom part of the RGB image, see Fig. 10 left. This ground truth is used for all the evaluations.

For comparison purposes and given that the dataset also provides the right images, we compute the depth from the stereo pair using the basic semi-global block matching algorithm [6], in a setting which gives almost dense coverage. We also compute the depth of around 1000 FAST extracted keypoints [12] in the stereo images. An example of these depth estimations from stereo is shown in Fig. 10 middle.

With the traditional metrics, shown in Table II, the stereo estimations perform the best. Even the "dense" and "sparse" estimation seems to perform equally well. This is another example of the flaws of the currently used metrics. It is clear, that the sparse stereo is giving us less information about the scene, for which should be penalized.

On the other hand, our measure of performance, shown in Fig. 11, tells a better story with a coherent ordering of the methods. The sparse stereo performs the worst, while a denser stereo performs in general better than single image depth estimation, as expected, since it provides more accurate estimations while covering the whole scene. To note, the estimations from [2] explain as much ground truth as the dense stereo method up to $\sim 15cm$ of error, and are better than [4] up to an error of $\sim 40cm$.

VI. Discussion and Conclusions

This paper has presented a better method for performance evaluation of single image depth estimation methods. The presented method is agnostic to the resolution, density and coverage of the depth estimation and we have shown both through case studies and real system evaluation that these are desirable properties. This allows a uniform comparison without altering the ground truth. Further, we have shown that it allows a deeper understanding of why a certain method performs better than the competition based on comparisons at the semantic category level. Although not presented here, the measure can be applied to methods that estimate both depth and semantics. In that case, it would capture the performance

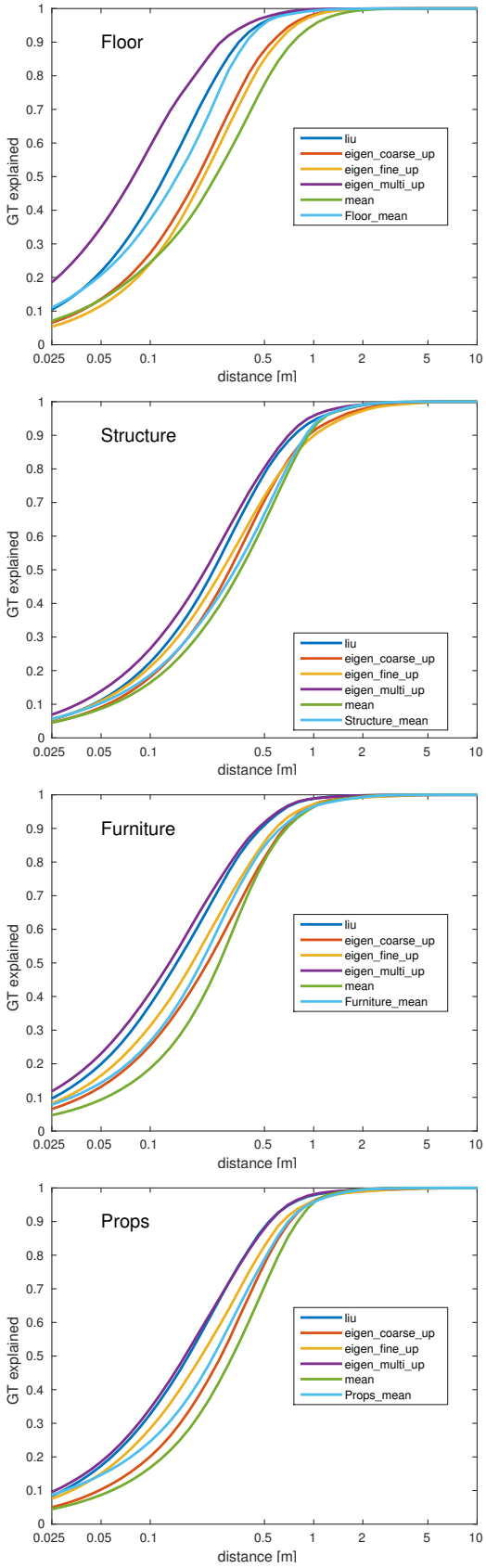


Fig. 9: Evaluation per semantic class on NYU-V2.

TABLE II: Results on KITTI dataset.

Method		Errors			Accuracy		
		Absolute	RMSE		$\delta <$		
		Relative	linear [m]	log.sc.inv.	1.25[%]	1.25^2[%]	1.25^3[%]
Stereo	dense	0.077	4.36	0.179	93.9	96.9	98.2
	sparse	0.073	4.53	0.180	93.2	96.1	97.8
Eigen *et al.* [4]	↑coarse	0.255	6.60	0.448	64.7	86.3	93.7
	↑fine	0.320	8.08	0.509	51.2	82.2	92.2
Cadena *et al.* [2]	rgb	0.291	8.65	0.363	59.7	79.1	89.4
	rgb-s	0.243	7.80	0.323	64.3	83.3	92.5

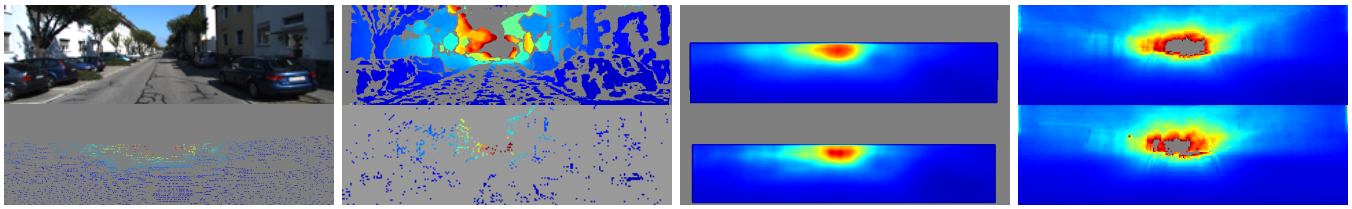


Fig. 10: One example from KITTI test set. *Left to Right*: RGB and ground truth depth from LIDAR, dense and sparse stereo depth, Eigen *et al.* [4] coarse and fine estimations, and Cadena *et al.* estimations from rgb-only and rgb and semantics [2].

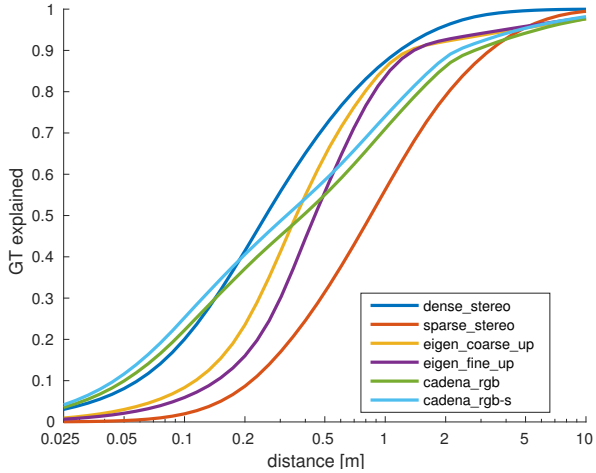


Fig. 11: Performance of full resolution for two stereo densities (dense and sparse), eigen_coarse and eigen_fine [4].

of both depth as well as semantic estimates in a single unified way.

One assumption that we make is that the estimation method actually tries to solve the problem rather than generating random depths over the image. To avoid those tricks it is important to always assess the estimations quantitatively and qualitatively. As we have already noted, our method also generalises to the case where the ground truth and estimated depth are measured in different coordinate frames; in this instance we would first apply ICP to align the scenes, and use the closest points found in that algorithm for our analysis.

References

[1] M. Baig, V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. Im2depth: Scalable exemplar based depth transfer. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 145–152, 2014.

[2] C. Cadena, A. Dick, and I. Reid. Multi-modal Auto-Encoders as Joint Estimators for Robotics Scene Understanding. In *Proc. Robotics: Science and Systems*, 2016.

[3] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Int. Conference on Computer Vision*, 2015.

[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems 27*, 2014.

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] K. Konolige. Small vision systems: Hardware and implementation. In *International Symposium on Robotics Research*, pages 203–212, 1998.

[7] L. Ladický, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 89–96, 2014.

[8] C. Li, A. Kowdle, A. Saxena, and T. Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *Advances in Neural Information Processing Systems 23*, pages 1351–1359, 2010.

[9] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260, 2010.

[10] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015.

[11] A. Owens, J. Xiao, A. Torralba, and W. Freeman. Shape anchors for data-driven multi-view reconstruction. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 33–40, 2013.

[12] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, pages 430–443, 2006.

[13] A. Saxena, M. Sun, and A. Ng. Make3d: Learning 3d scene structure from a single still image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):824–840, 2009.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012.