Evaluation of CNN-Based Single-Image Depth Estimation Methods

Tobias Koch[1(✉)], Lukas Liebel[1], Friedrich Fraundorfer[2,3], and Marco Körner[1]

[1] Chair of Remote Sensing Technology,
Technical University of Munich, Munich, Germany
{tobias.koch,lukas.liebel,marco.koerner}@tum.de
[2] Institute of Computer Graphics and Vision,
Graz University of Technology, Graz, Austria
fraundorfer@icg.tugraz.at
[3] Remote Sensing Technology Institute,
German Aerospace Center (DLR), Oberpfaffenhofen, Germany

Abstract. While an increasing interest in deep models for *single-image depth estimation (SIDE)* can be observed, established schemes for their evaluation are still limited. We propose a set of novel quality criteria, allowing for a more detailed analysis by focusing on specific characteristics of depth maps. In particular, we address the preservation of edges and planar regions, depth consistency, and absolute distance accuracy. In order to employ these metrics to evaluate and compare state-of-the-art SIDE approaches, we provide a new high-quality RGB-D dataset. We used a *digital single-lens reflex (DSLR)* camera together with a *laser scanner* to acquire high-resolution images and highly accurate depth maps. Experimental results show the validity of our proposed evaluation protocol.

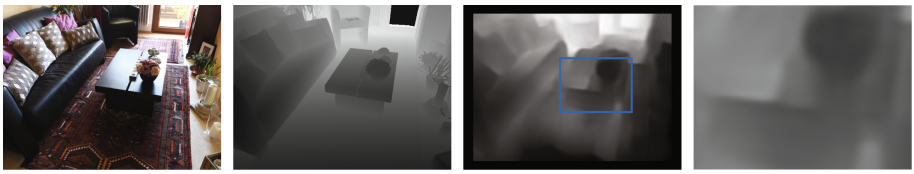
Keywords: Single-image depth estimation · Deep learning · CNN · RGB-D · Benchmark · Evaluation · Dataset · Error metrics

1 Introduction

With the emergence of *deep learning* methods within the recent years and their massive influence on the computer vision domain, the problem of SIDE got addressed as well by many authors. These methods are in high demand for manifold scene understanding applications like, for instance, autonomous driving, robot navigation, or augmented reality systems. In order to replace or enhance traditional methods, *convolutional neural network (CNN)* architectures have been most commonly used and successfully shown to be able to infer geometrical information solely from presented monocular RGB or intensity images, as exemplary shown in Fig. 1.

While these methods produce nicely intuitive results, proper evaluating the estimated depth maps is crucial for subsequent applications, *e.g.*, their suitability for further 3D understanding scenarios [30]. Consistent and reliable relative

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11131, pp. 331–348, 2019.
https://doi.org/10.1007/978-3-030-11015-4_25



(a) RGB image (b) Depth map (c) Prediction (d) Prediction detail

Fig. 1. Sample image pair from our dataset and depth prediction using a state-of-the-art algorithm [7]. Although the quality of the depth map seems reasonable, the prediction suffers from artifacts, smoothing, missing objects, and inaccuracies in textured image regions

depth estimates are, for instance, a key requirement for path planning approaches in robotics, augmented reality applications, or computational cinematography.

Nevertheless, the evaluation schemes and error metrics commonly used so far mainly consider the overall accuracy by reporting global statistics of depth residuals which does not give insight into the depth estimation quality at salient and important regions, like planar surfaces or geometric discontinuities. Hence, fairly reasonable reconstruction results, as shown in Fig. 1c, are probably positively evaluated, while still showing evident defects around edges. At the same time, the shortage of available datasets providing ground truth data of sufficient quality and quantity impedes precise evaluation.

As these issues were reported by the authors of recent SIDE papers [12,19], we aim at providing a new and extended evaluation scheme in order to overcome these deficiencies. In particular, as our main contributions, we

(i) present a new evaluation dataset acquired from diverse indoor scenarios containing high-resolution RGB images aside highly accurate depth maps from laser scans[1] *(ii)* introduce a set of new interpretable error metrics targeting the aforementioned issues *(iii)* evaluate a variety of state-of-the-art methods using these data and performance measures.

2 Related Work

In this section, we introduce some of the most recent learning-based methods for predicting depth from a single image and review existing datasets used for training and evaluating the accuracy of these methods.

2.1 Methods

Most commonly, stereo reconstruction is performed from multi-view setups, *e.g.*, by triangulation of 3D points from corresponding 2D image points observed by distinct cameras (*cf. multi-view stereo (MVS)* or *structure from motion*

[1] This dataset is freely available at www.lmf.bgu.tum.de/ibims1.

(SfM) methods) [27]. Nevertheless, for already many decades, estimating depth or shape from monocular setups or single views is under scientific consideration [2] in psychovisual as well as computational research domains. After several RGB-D datasets were released [4,5,11,25,28], data-driven learning-based approaches outperformed established model-based methods. Especially deep learning-based methods have proven to be highly effective for this task and achieved current state-of-the-art results [3,7,9,10,13,15–18,20–22,24,31–33]. One of the first approaches using CNNs for regressing dense depth maps was presented by Eigen et al. [8] who employ two deep networks for first performing a coarse global prediction and refine the predictions locally afterwards. An extension to this approach uses deeper models and additionally predicts normals and semantic labels [7]. Liu et al. [22] combine CNNs and *conditional random field (CRFs)* in a unified framework while making use of superpixels for preserving sharp edges. Laina et al. [15] tackle this problem with a fully convolutional network consisting of a feature map up-sampling within the network. While Li et al. [17] employ a novel set loss and a two-streamed CNN that fuses predictions of depth and depth gradients, Xu et al. [32] propose to integrate complementary information derived from multiple CNN side outputs using CRFs.

2.2 Existing Benchmark Datasets

In order to evaluate SIDE methods, any dataset containing corresponding RGB and depth images can be considered, which also comprises benchmarks originally designed for the evaluation of MVS approaches. Strecha et al. [29] propose a MVS benchmark providing overlapping images with camera poses for six different outdoor scenes and a ground truth point cloud obtained by a laser scanner. More recently, two MVS benchmarks, the `ETH3D` [26] and the `Tanks & Temples` [14] datasets, have been released. Although these MVS benchmarks contain high resolution images and accurate ground truth data obtained from a laser scanner, the setup is not designed for SIDE methods. Usually, a scene is scanned from multiple aligned laser scans and images acquired in a sequential matter. However, it cannot be guaranteed that the corresponding depth maps are dense. Occlusions in the images result in gaps in the depth maps especially at object boundaries which are, however, a key aspect of our metrics. Despite the possibility of acquiring a large number of image pairs, they mostly comprise only a limited scene variety and are highly redundant due high visual overlap. Currently, SIDE methods are tested on mainly three different datasets. `Make3D` [25], as one example, contains 534 outdoor images and aligned depth maps acquired from a custom-built 3D scanner, but suffers from a very low resolution of the depth maps and a rather limited scene variety. The `Kitti` dataset [11] contains street scenes captured out of a moving car. The dataset contains RGB images together with depth maps from a Velodyne laser scanner. However, depth maps are only provided in a very low resolution which furthermore suffer from irregularly and sparsely spaced points. The most frequently used dataset is the `NYU depth v2` dataset [28] containing 464 indoor scenes with aligned RGB and depth

images from video sequences obtained from a Microsoft Kinect v1 sensor. A subset of this dataset is mostly used for training deep networks, while another 654 image and depth pairs serve for evaluation. This large number of image pairs and the various indoor scenarios facilitated the fast progress of SIDE methods. However, active RGB-D sensors, like the Kinect, suffer from a short operational range, occlusions, gaps, and erroneous specular surfaces. The recently released `Matterport3D` [4] dataset provides an even larger amount of indoor scenes collected from a custom-built 3D scanner consisting of three RGB-D cameras. This dataset is a valuable addition to the `NYU-v2` but also suffers from the same weaknesses of active RGB-D sensors.

3 Error Metrics

This section describes established metrics and our new proposed ones allowing for a more detailed analysis.

3.1 Commonly Used Error Metrics

Established error metrics consider global statistics between a predicted depth map $\boldsymbol{Y}$ and its ground truth depth image $\boldsymbol{Y}^*$ with T depth pixels. Beside visual inspections of depth maps or projected 3D point clouds, the following error metrics are exclusively used in all relevant recent publications [7,8,15,19,32]:

Threshold: percentage of y such that $\max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \sigma < thr$
Absolute relative difference: $\text{rel} = \frac{1}{T}\sum_{i,j} \left|y_{i,j} - y_{i,j}^*\right| / y_{i,j}^*$
Squared relative difference: $\text{srel} = \frac{1}{T}\sum_{i,j} \left|y_{i,j} - y_{i,j}^*\right|^2 / y_{i,j}^*$
RMS (linear): $\text{RMS} = \sqrt{\frac{1}{T}\sum_{i,j} \left|y_{i,j} - y_{i,j}^*\right|^2}$
RMS (log): $\log_{10} = \sqrt{\frac{1}{T}\sum_{i,j} \left|\log y_{i,j} - \log y_{i,j}^*\right|^2}$

Even though these statistics are good indicators for the general quality of predicted depth maps, they could be delusive. Particularly, the standard metrics are not able to directly assess the planarity of planar surfaces or the correctness of estimated plane orientations. Furthermore, it is of high relevance that depth discontinuities are precisely located, which is not reflected by the standard metrics.

3.2 Proposed Error Metrics

In order to allow for a more meaningful analysis of predicted depth maps and a more complete comparison of different algorithms, we present a set of new quality measures that specify on different characteristics of depth maps which are crucial for many applications. These are meant to be used in addition to the traditional error metrics introduced in Sect. 3.1. When talking about depth maps, the following questions arise that should be addressed by our new metrics:

How is the quality of predicted depth maps for different absolute scene depths? Can planar surfaces be reconstructed correctly? Can all depth discontinuities be represented? How accurately are they localized? Are depth estimates consistent over the whole image area?

Distance-Related Assessment. Established global statistics are calculated over the full range of depth comprised by the image and therefore do not consider different accuracies for specific absolute scene ranges. Hence, applying the standard metrics for specific range intervals by discretizing existing depth ranges into discrete bins (*e.g.*, one-meter depth slices) allows investigating the performance of predicted depths for close and far ranged objects independently.

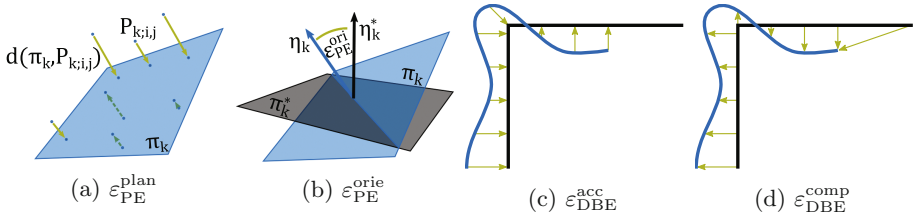


Fig. 2. Visualizations of the proposed error metrics for *planarity errors* (a and b) and *depth boundary errors* (c and d)

Planarity. Man-made objects, in particular, can often be characterized by planar structures like walls, floors, ceilings, openings, and diverse types of furniture. However, global statistics do not directly give information about the shape correctness of objects within the scene. Predicting depths for planar objects is challenging for many reasons. Primarily, these objects tend to lack texture and only differ by smooth color gradients in the image, from which it is hard to estimate the correct orientation of a 3D plane with three-degrees-of-freedom. In the presence of textured planar surfaces, it is even more challenging for a SIDE approach to distinguish between a real depth discontinuity and a textured planar surface, *e.g.*, a painting on a wall. As most methods are trained on large indoor scenes, like `NYU-v2`, a correct representation of planar structures is an important task for SIDE, but can hardly be evaluated using established standard metrics. For this reason, we propose to use a set of annotated images defining various planar surfaces (walls, table tops and floors) and evaluate the flatness and orientation of predicted 3D planes $\boldsymbol{\pi}_k = (\boldsymbol{\eta}_k, o_k)$ compared to ground truth 3D planes $\boldsymbol{\pi}_k{}^* = (\boldsymbol{\eta}_k^*, o_k^*)$. Each plane is specified by a normal vector $\boldsymbol{\eta}$ and an offset to the plane o. In particular, a masked depth map $\boldsymbol{Y}_k$ of a particular planar surface is projected to 3D points $\boldsymbol{P}_{k;i,j}$ where 3D planes $\boldsymbol{\pi}_k$ are robustly fitted to both the ground truth and predicted 3D point clouds $\mathcal{P}_k^* = \left\{\boldsymbol{P}_{k;i,j}^*\right\}_{i,j}$ and $\mathcal{P}_k = \{\boldsymbol{P}_{k;i,j}\}_{i,j}$, respectively. The planarity error

$$\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}\left(\boldsymbol{Y}_k\right) = \mathbb{V}\left[\sum_{\boldsymbol{P}_{k;i,j}\in\mathcal{P}_k} d\left(\boldsymbol{\pi}_k, \boldsymbol{P}_{k;i,j}\right)\right] \tag{1}$$

is then quantified by the standard deviation of the averaged distances d between the predicted 3D point cloud and its corresponding 3D plane estimate. The orientation error

$$\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}\left(\boldsymbol{Y}_k\right) = \operatorname{acos}\left(\boldsymbol{\eta}_k^{\top} \cdot \boldsymbol{\eta}_k^{*}\right) \tag{2}$$

is defined as the 3D angle difference between the normal vectors of predicted and ground truth 3D planes. Figures 2a and b illustrate the proposed planarity errors. Note that the predicted depth maps are scaled w.r.t. the ground truth depth map, in order to eliminate scaling differences of compared methods.

Location Accuracy of Depth Boundaries. Beside planar surfaces, captured scenes, especially indoor scenes, cover a large variety of scene depths caused by any object in the scene. Depth discontinuities between two objects are represented as strong gradient changes in the depth maps. In this context, it is important to examine whether predicted depths maps are able to represent all relevant depth discontinuities in an accurate way or if they even create fictitious depth discontinuities confused by texture. An analysis of depth discontinuities can be best expressed by detecting and comparing edges in predicted and ground truth depth maps. Location accuracy and sharp edges are of high importance for generating a set of ground truth depth transitions which cannot be guaranteed by existing datasets acquired from RGB-D sensors. Ground truth edges are extracted from our dataset by first generating a set of tentative edge hypotheses using *structured edges* [6] and then manually selecting important and distinct edges subsequently. In order to evaluate predicted depth maps, edges $\boldsymbol{Y}_{\mathrm{bin}}$ are extracted using structured edges and compared to the ground truth edges $\boldsymbol{Y}_{\mathrm{bin}}^{*}$ via *truncated chamfer distance* of the binary edge images. Specifically, an *Euclidean distance transform* is applied to the ground truth edge image $\boldsymbol{E}^{*} = DT\left(\boldsymbol{Y}_{\mathrm{bin}}^{*}\right)$, while distances exceeding a given threshold θ ($\theta = 10\,\mathrm{px}$ in our experiments) are ignored in order to evaluate predicted edges only in the local neighborhood of the ground truth edges. We define the *depth boundary errors (DBEs)*, comprised of an accuracy measure

$$\varepsilon_{\mathrm{DBE}}^{\mathrm{acc}}(\boldsymbol{Y}) = \frac{1}{\sum_i \sum_j y_{\mathrm{bin};i,j}} \sum_i \sum_j e_{i,j}^{*} \cdot y_{\mathrm{bin};i,j} \tag{3}$$

by multiplying the predicted binary edge map with the distance map and a subsequent accumulation of the pixel distances towards the ground truth edge. As this measure does not consider any missing edges in the predicted depth image, we also define a completeness error

$$\varepsilon_{\mathrm{DBE}}^{\mathrm{comp}}(\boldsymbol{Y}) = \frac{1}{\sum_i \sum_j y_{\mathrm{bin};i,j}^{*}} \sum_i \sum_j e_{i,j} \cdot y_{\mathrm{bin};i,j}^{*} \tag{4}$$

by accumulating the ground truth edges multiplied with the distance image of the predicted edges $\boldsymbol{E} = DT(\boldsymbol{Y}_{\text{bin}})$. A visual explanation of the DBEs are illustrated in Figs. 2c and d.

Directed Depth Error. For many applications, it is of high interest that depth images are consistent over the whole image area. Although the absolute depth error and squared depth error give information about the correctness between predicted and ground truth depths, they do not provide information if the predicted depth is estimated too short or too far. For this purpose, we define the *directed depth errors (DDEs)*

$$\varepsilon^{+}_{\text{DDE}}(\boldsymbol{Y}) = \frac{\left|\left\{y_{i,j} | d_{\text{sgn}}(\boldsymbol{\pi}, \boldsymbol{P}_{i,j}) > 0 \wedge d_{\text{sgn}}(\boldsymbol{\pi}, \boldsymbol{P}^{*}_{i,j}) < 0\right\}\right|}{T} \tag{5}$$

$$\varepsilon^{-}_{\text{DDE}}(\boldsymbol{Y}) = \frac{\left|\left\{y_{i,j} | d_{\text{sgn}}(\boldsymbol{\pi}, \boldsymbol{P}_{i,j}) < 0 \wedge d_{\text{sgn}}(\boldsymbol{\pi}, \boldsymbol{P}^{*}_{i,j}) > 0\right\}\right|}{T} \tag{6}$$

as the proportions of too far and too close predicted depth pixels $\varepsilon^{+}_{\text{DDE}}$ and $\varepsilon^{-}_{\text{DDE}}$. In practice, a reference depth plane $\boldsymbol{\pi}$ is defined at a certain distance (*e.g.*, at 3 m, *cf.* Fig. 7c) and all predicted depths pixels which lie in front and behind this plane are masked and assessed according to their correctness using the reference depth images.

4 Dataset

As described in the previous sections, our proposed metrics require extended ground truth which is not yet available in standard datasets. Hence, we compiled a new dataset according to these specifications.

4.1 Acquisition

For creating such a reference dataset, high-quality optical RGB images and depth maps had to be acquired. Practical considerations included the choice of suitable instruments for the acquisition of both parts. Furthermore, a protocol to calibrate both instruments, such that image and depth map align with each other, had to be developed. An exhaustive analysis and comparison of different sensors considered for the data acquisition was conducted, which clearly showed the advantages of using a laser scanner and a DSLR camera compared to active sensors like RGB-D cameras or passive stereo camera rigs. We therefore used the respective setup for the creation of our dataset.

In order to record the ground truth for our dataset, we used a highly accurate Leica HDS7000 laser scanner, which stands out for high point cloud density and very low noise level. We acquired the scans with 3 mm point spacing and 0.4 mm RMS at 10 m distance. As our laser scanner does not provide RGB images along

with the point clouds, an additional camera was used in order to capture optical imagery. The usage of a reasonably high-quality camera sensor and lens allows for capturing images in high resolution with only slight distortions and a high stability regarding the intrinsic parameters. For the experiments, we chose and calibrated a Nikon D5500 DSLR camera and a Nikon AF-S Nikkor 18–105 mm lens, mechanically fixed to a focal length of approximately 18 mm.

Using our sensor setup, synchronous acquisition of point clouds and RGB imagery is not possible. In order to acquire depth maps without parallax effects, the camera was mounted on a custom panoramic tripod head which allows to freely position the camera along all six degrees of freedom. This setup can be interchanged with the laser scanner, ensuring coincidence of the optical center of the camera and the origin of the laser scanner coordinate system after a prior calibration of the system. It is worth noting, that every single RGB-D image pair of our dataset was obtained by an individual scan and image capture with the aforementioned strategy in order to achieve dense depth maps without gaps due to occlusions.

4.2 Registration and Processing

The acquired images were undistorted using the intrinsic camera parameters obtained from the calibration process. In order to register the camera towards the local coordinate system of the laser scanner, we manually selected a sufficient number of corresponding 2D and 3D points and estimated the camera pose using EPnP [23]. This registration of the camera relative to the point cloud yielded only a minor translation, thanks to the pre-calibrated platform. Using this procedure, we determined the 6D pose of a virtual depth sensor which we use to derive a matching depth map from the 3D point cloud. In order to obtain a depth value for each pixel in the image, the images were sampled down to two different resolutions. We provide a high-quality version with a resolution of 1500×1000 px and a cropped `NYU-v2`-like version with a resolution of 640×480 px. 3D points were projected to a virtual sensor with the respective resolution. For each pixel, a depth value was calculated, representing the depth value of the 3D point with the shortest distance to the virtual sensor. It is worth highlighting that depth maps were derived from the 3D point cloud for both versions of the images separately. Hence, no down-sampling artifacts are introduced for the lower-resolution version. The depth maps for both, the high-quality and the `NYU-v2`-like version, are provided along with the respective images.

4.3 Contents

Following the described procedure, we compiled a dataset, which we henceforth refer to as the *independent Benchmark images and matched scans v1 (`iBims-1`)* dataset. The dataset is mainly composed of reference data for the direct evaluation of depth maps, as produced by SIDE methods. As described in the previous sections, pairs of images and depth maps were acquired and are provided in two different versions, namely a high-quality version and a `NYU-v2`-like version.

Example pairs of images and matching depth maps from `iBims-1` are shown in Figs. 1a and b and Figs. 3a and b, respectively.

(a) Camera image (b) Ground truth (c) Masks (d) Distinct edges

Fig. 3. Sample from the main part of the proposed `iBims-1` dataset with (a) RGB image, (b) depth map, (c) several masks with semantic annotations (*i.e.*, walls (■), floor (■), tables (■), transparent objects (■), and invalid pixels (■)), and (d) distinct edges (—) (Color figure online)

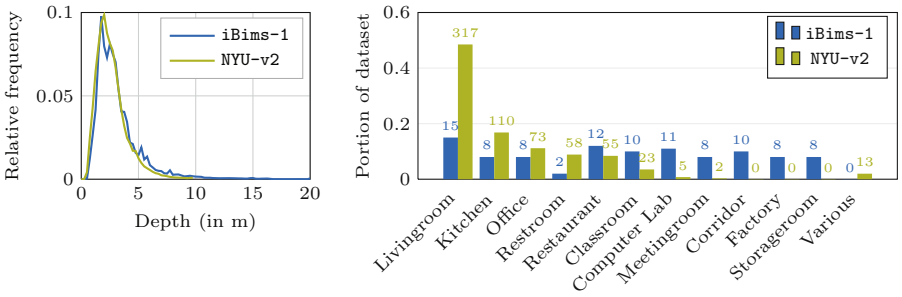
Additionally, several manually created masks are provided. Examples for all types of masks are shown in Fig. 3c, while statistics of the plane annotations are listed in Table 1. In order to allow for evaluation following the proposed DBE metric, we provide distinct edges for all images. Edges have been detected automatically and manually selected. Figure 3d shows an example for one of the scenes from `iBims-1`.

This main part of the dataset contains 100 RGB-D image pairs in total. So far, the `NYU-v2` dataset is still the most comprehensive and accurate indoor dataset for training data-demanding deep learning methods. Since this dataset has most commonly been used for training the considered SIDE methods, `iBims-1` is designed to contain similar scenarios. Our acquired scenarios include various indoor settings, such as office, lecture, and living rooms, computer labs, a factory room, as well as more challenging ones, such as long corridors and potted plants. A comparison regarding the scene variety between `NYU-v2` and `iBims-1` can be seen in Fig. 4b. Furthermore, `iBims-1` features statistics comparable to `NYU-v2`, such as the distribution of depth values, shown in Fig. 4a, and a comparable field of view.

Table 1. Number and statistics of manually labeled plane masks in `iBims-1`

Plane type	Images	Instances	Pixels (for NYUv2 res.)
Floor	47	51	1163499
Table	46	54	832984
Wall	82	140	6557108

Additionally, we provide an *auxiliary dataset* which consists of four parts: (1) Four outdoor RGB-D image pairs, containing vegetation, buildings, cars and



(a) Distribution of depth values (b) Distribution of samples for each scene type. Absolute numbers are given above

Fig. 4. `iBims-1` dataset statistics compared to the `NYU-v2` dataset. Distribution of depth values (a) and scene variety (b)

larger ranges than indoor scenes. (2) Special cases which are expected to mislead SIDE methods. These show 85 RGB images of printed samples from the `NYU-v2` and the `Pattern` dataset [1] hung on a wall. Those could potentially give valuable insights, as they reveal what kind of image features SIDE methods exploit. Figure 9a shows examples from both categories. No depth maps are provided for those images, as the region of interest is supposed to be approximately planar and depth estimates are, thus, easy to assess qualitatively. (3) 28 different geometrical and radiometrical augmentations for each image of our core dataset to test the robustness of SIDE methods. (4) Up to three additional handheld images for most RGB-D image pairs of our core dataset with viewpoint changes towards the reference images which allows to validate MVS algorithms with high-quality ground truth depth maps.

5 Evaluation

In this section, we evaluate the quality of existing SIDE methods using both established and proposed metrics for our reference test dataset, as well as for the commonly used `NYU-v2` dataset. Furthermore, additional experiments were conducted to investigate the general behavior of SIDE methods, *i.e.*, the robustness of predicted depth maps to geometrical and color transformations and the planarity of textured vertical surfaces. For evaluation, we compared several state-of-the-art methods, namely those proposed by Eigen and Fergus [8], Eigen et al. [7], Liu et al. [21], Laina et al. [15], and Li et al. [19]. It is worth mentioning that all of these methods were solely trained on the `NYU-v2` dataset. Therefore, differences in the results are expected to arise from the developed methodology rather than the training data.

5.1 Evaluation Using Proposed Metrics

In the following, we report the results of evaluating SIDE methods on both `NYU-v2` and `iBims-1` using our newly proposed metrics. Please note, that due to the page limit, only few graphical results can be displayed in the following sections.

Distance-Related Assessment. The results of evaluation using commonly used metrics on `iBims-1` unveil lower overall scores for our dataset (see Table 2). In order to get a better understanding of these results, we evaluated the considered methods on specific range intervals, which we set to 1 m in our experiments. Figure 5 shows the error band of the relative and RMS errors of the method proposed by Li et al. [19] applied to both datasets. The result clearly shows a comparable trend on both datasets for the shared depth range. This proves our first assumption, that the overall lower scores originate from the huge differences at depth values beyond the 10 m depth range. On the other hand, the results reveal the generalization capabilities of the networks, which achieve similar results on images from another camera with different intrinsics and for different scenarios. It should be noted that the error bands, which show similar characteristics for different methods and error metrics, correlate with the depth distributions of the datasets, shown in Fig. 4a.

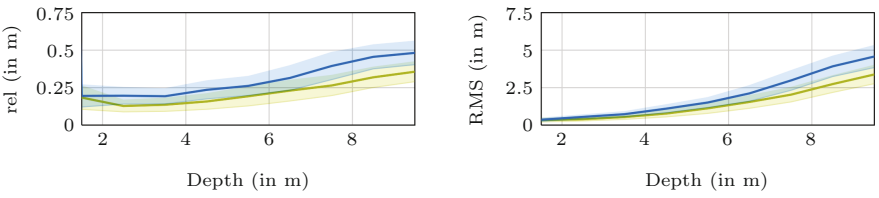


Fig. 5. Distance-related global errors (left: relative error and right: RMS) for `NYU-v2` (mean: —, ±0.5 std: ■) and `iBims-1` (mean: —, ±0.5 std: ■) using the method of Li et al. [19] (Color figure online)

Planarity. To investigate the quality of reconstructed planar structures, we evaluated the different methods with the planarity and orientation errors $\varepsilon_{\text{PE}}^{\text{plan}}$ and $\varepsilon_{\text{PE}}^{\text{orie}}$, respectively, as defined in Sect. 3.2, for different planar objects. In particular, we distinguished between horizontal and vertical planes and used masks from our dataset. Figure 6 and Table 2 show the results for the `iBims-1` dataset. Beside a combined error, including all planar labels, we separately computed the errors for the individual objects as well. The results show different performances for individual classes, especially orientations of floors were predicted in a significantly higher accuracy for all methods, while the absolute orientation error for walls is surprisingly high. Apart from the general performance of all methods, substantial differences between the considered methods can be determined. It is

notable that the method of Li et al. [19] achieved much better results in predicting orientations of horizontal planes but also performed rather bad on vertical surfaces.

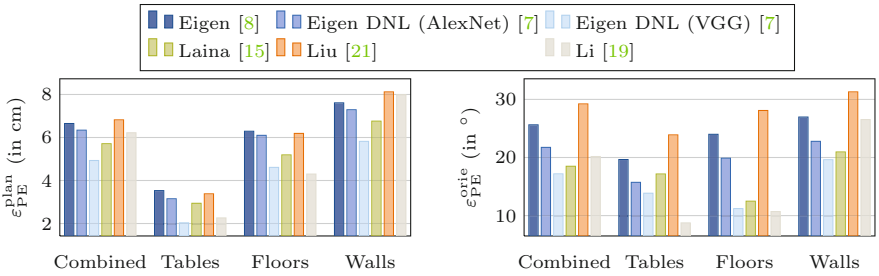


Fig. 6. Results for the planarity metrics $\varepsilon_{\text{PE}}^{\text{plan}}$ (left) and $\varepsilon_{\text{PE}}^{\text{orie}}$ (right) on `iBims-1`

Location Accuracy of Depth Boundaries. The high quality of our reference dataset facilitates an accurate assessment of predicted depth discontinuities. As ground truth edges, we used the provided edge maps from our dataset and computed the accuracy and completeness errors $\varepsilon_{\text{DBE}}^{\text{acc}}$ and $\varepsilon_{\text{DBE}}^{\text{comp}}$, respectively, introduced in Sect. 3.2. Quantitative results for all methods are listed in Table 2. Comparing the accuracy error of all methods, Liu et al. [21] and Li et al. [19] achieved best results in preserving true depth boundaries, while other methods tended to produce smooth edges losing sharp transitions which can be seen in Figs. 7a and b. This smoothing property also affected the completeness error, resulting in missing edges expressed by larger values for $\varepsilon_{\text{DBE}}^{\text{comp}}$.

(a) Ground truth (b) Predictions (c) Depth plane (d) Differences

Fig. 7. Visual results after applying DBE (a + b) and DDE (c + d) on `iBims-1`: (a) ground truth edge (—). (b) Edge predictions using the methods of Li et al. [19] (—) and Laina et al. [15] (—). (c) Ground truth depth plane at $d = 3\,\text{m}$ separating foreground from background (—). (d) Differences between ground truth and predicted depths using the method of Li et al. [19]. Color coded are depth values that are either estimated too short (—) or too far (—) (Color figure online)

Directed Depth Error. The DDE aims to identify predicted depth values which lie on the correct side of a predefined reference plane but also distinguishes between overestimated and underestimated predicted depths. This measure could be useful for applications like 3D cinematography, where a 3D effect is generated by defining two depth planes. For this experiment, we defined a reference plane at 3 m distance and computed the proportions of correct $\varepsilon^{0}_{\mathrm{DDE}}$, overestimated $\varepsilon^{+}_{\mathrm{DDE}}$, and underestimated $\varepsilon^{-}_{\mathrm{DDE}}$ depth values towards this plane according to the error definitions in Sect. 3.2. Table 2 lists the resulting proportions for `iBims-1`, while a visual illustration of correctly and falsely predicted depths is depicted in Figs. 7c and d. The results show that the methods tended to predict depths to a too short distance, although the number of correctly estimated depths almost reaches 85% for `iBims-1`.

Table 2. Quantitative results for standard metrics and proposed PE, DBE, and DDE metrics on `iBims-1` applying different SIDE methods

Method	Standard metrics ($\sigma_i = 1.25^i$)						PE (cm/°)		DBE (px)		DDE (%)		
	Rel	$\log_{10}$	RMS	σ_1	σ_2	σ_3	$\varepsilon^{\mathrm{plan}}_{\mathrm{PE}}$	$\varepsilon^{\mathrm{orie}}_{\mathrm{PE}}$	$\varepsilon^{\mathrm{acc}}_{\mathrm{DBE}}$	$\varepsilon^{\mathrm{comp}}_{\mathrm{DBE}}$	$\varepsilon^{0}_{\mathrm{DDE}}$	$\varepsilon^{-}_{\mathrm{DDE}}$	$\varepsilon^{+}_{\mathrm{DDE}}$
Eigen [8]	0.32	0.17	1.55	0.36	0.65	0.84	6.65	25.62	5.48	70.31	72.06	25.71	2.23
Eigen (AlexNet) [7]	0.30	0.15	1.38	0.40	0.73	0.88	6.34	21.74	4.57	46.52	78.24	17.86	3.90
Eigen (VGG) [7]	0.25	0.13	1.26	0.47	0.78	0.93	**4.93**	**17.18**	4.51	43.64	80.73	17.47	**1.80**
Laina [15]	0.25	0.13	1.20	0.50	0.78	0.91	5.71	18.49	6.89	40.48	81.65	15.91	2.43
Liu [21]	0.30	0.13	1.26	0.48	0.78	0.91	6.82	29.22	**3.57**	**31.75**	80.46	13.26	6.28
Li [19]	**0.22**	**0.11**	**1.07**	**0.59**	**0.85**	**0.95**	6.22	20.17	3.68	36.27	**84.13**	**12.49**	3.38

Table 3. Quantitative results on the augmented `iBims-1` dataset exemplary listed for the global relative distance error. Errors showing relative differences for various image augmentations towards the predicted original input image (Ref)

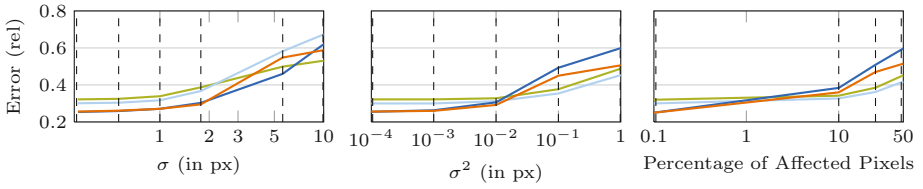
Method	Ref.	Geometric		Contrast			Ch. Swap		Hue		Saturation	
		LR	UD	$\gamma = 0.2$	$\gamma = 2$	Norm.	BGR	BRG	+9°	+90°	×0	×0.9
Eigen [8]	0.322	−0.003	0.087	0.056	0.015	0.000	0.017	0.018	0.001	0.021	0.003	0
Eigen (AlexNet) [7]	0.301	0.006	0.147	0.105	0.023	−0.002	0.017	0.008	0.002	0.017	0.007	0
Eigen (VGG) [7]	0.254	0.003	0.150	0.109	0.008	0.000	0.010	0.013	0.000	0.012	0.009	0
Laina [15]	0.255	−0.004	0.161	0.078	0.022	−0.001	0.007	0.009	0.000	0.007	0.003	0

5.2 Further Analyses

Making use of our *auxiliary dataset*, a series of additional experiments were conducted to investigate the behavior of SIDE methods in special situations. The challenges cover an augmentation of our dataset with various color and geometrical transformations and an auxiliary dataset containing images of printed patterns and `NYU-v2` images on a planar surface.

Data Augmentation. In order to assess the robustness of SIDE methods w.r.t. simple geometrical and color transformation and noise, we derived a set of augmented images from our dataset. For geometrical transformations we flipped the input images horizontally—which is expected to not change the results significantly—and vertically, which is expected to expose slight overfitting effects. As images in the `NYU-v2` dataset usually show a considerable amount of pixels on the floor in the lower part of the picture, this is expected to notably influence the estimated depth maps. For color transformations, we consider swapping of image channels, shifting the hue by some offset h and scaling the saturation by a factor s. We change the gamma values to simulate over- and under-exposure and optimize the contrast by histogram stretching. Blurred versions of the images are simulated by applying gaussian blur with increasing standard deviation σ. Furthermore, we consider noisy versions of the images by applying gaussian additive noise and salt and pepper noise with increasing variance and amount of affected pixels, respectively.

Table 3 shows results for these augmented images using the global relative error metric for selected methods. As expected, the geometrical transformations yielded contrasting results. While the horizontal flipping did not influence the results by a large margin, flipping the images vertically increased the error by up to 60%. Slight overexposure influenced the result notably, underexposure seems to have been less problematic. Histogram stretching had no influence on the results, suggesting that this is already a fixed or learned part of the methods. The methods also seem to be robust to color changes, which is best seen in the results for $s = 0$, *i.e.*, greyscale input images which yielded an equal error to the reference. The results for blurring the input images with a gaussian kernel of various sizes, as well as adding a different amount of gaussian and salt and pepper noise to the input images are depicted in Fig. 8.



(a) Gaussian Blur (b) Gaussian Noise (c) Salt & Pepper Noise

Fig. 8. Quality of SIDE results, achieved using the methods proposed by Eigen et al. [8] (—), Eigen and Fergus [7] (AlexNet —, VGG —), and Laina et al. [15] (—) for augmentations with increasing intensity. Vertical lines (- -) correspond to discrete augmentation intensities

Textured Planar Surfaces. Experiments with printed patterns and `NYU-v2` samples on a planar surface exploit which features influence the predictions of SIDE methods. As to be seen in the first example in Fig. 9, gradients seem

(a) Input RGB (b) Laina et al. [15] (c) Liu et al. [21] (d) Eigen [7]

Fig. 9. Predicted depth for a sample from the auxiliary part of the proposed `iBims-1` dataset showing printed samples from the Patterns [1] dataset (top) and the `NYU-v2` dataset [28] (bottom) on a planar surface

to serve as a strong hint to the network. All of the tested methods estimated incorrectly depth in the depicted scene, none of them, however, identified the actual planarity of the picture.

6 Conclusions

We presented a novel set of quality criteria for the evaluation of SIDE methods. Furthermore, we introduced a new high-quality dataset, fulfilling the need for an extended ground truth of our proposed metrics. Using this test protocol we evaluated and compared state-of-the-art SIDE methods. In our experiments, we were able to assess the quality of the compared approaches w.r.t. to various meaningful properties, such as the preservation of edges and planar regions, depth consistency, and absolute distance accuracy. Compared to commonly used global metrics, our proposed set of quality criteria enabled us to unveil even subtle differences between the considered SIDE methods. In particular, our experiments have shown that the prediction of planar surfaces, which is crucial for many applications, is lacking accuracy. Furthermore, edges in the predicted depth maps tend to be oversmooth for many methods. We believe that our dataset is suitable for future developments in this regard, as our images are provided in a very high resolution and contain new sceneries with extended scene depths.

The `iBims-1` dataset can be downloaded at www.lmf.bgu.tum.de/ibims1.

Acknowledgements. This research was funded by the German Research Foundation (DFG) for Tobias Koch and the Federal Ministry of Transport and Digital Infrastructure (BMVI) for Lukas Liebel. We thank our colleagues from the Chair of Geodesy for providing all the necessary equipment and our student assistant Leonidas Stöckle for his help during the data acquisition campaign.

References

1. Asuni, N., Giachetti, A.: TESTIMAGES: a large-scale archive for testing visual devices and basic image processing algorithms. In: Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference. The Eurographics Association (2014). https://doi.org/10.2312/stag.20141242
2. Bülthoff, H.H., Yuille, A.L.: Shape from X: psychophysics and computation. In: Computational Models of Visual Processing, pp. 305–330. MIT Press (1991)
3. Chakrabarti, A., Shao, J., Shakhnarovich, G.: Depth from a single image by harmonizing overcomplete local network predictions. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 2658–2666 (2016)
4. Chang, A., et al.: Matterport3D: learning from RGB-D data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1558–1570 (2015)
7. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2650–2658 (2015). https://doi.org/10.1109/ICCV.2015.304
8. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), vol. 2, pp. 2366–2374 (2014)
9. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2002–2011 (2018)
10. Garg, R., Vijay Kumar, B.G., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361. IEEE (2012)
12. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. arXiv preprint arXiv:1803.08673 (2018)
13. Kim, S., Park, K., Sohn, K., Lin, S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 143–159. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_9
14. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. **36**(4) (2017)
15. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: Proceedings of the Fourth International Conference on 3D Vision (3DV), pp. 239–248 (2016)
16. Lee, J.H., Heo, M., Kim, K.R., Kim, C.S.: Single-image depth estimation based on fourier domain analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 330–339 (2018)

17. Li, B., Dai, Y., Chen, H., He, M.: Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference. arXiv preprint arXiv:1705.00534 (2017)
18. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1119–1127 (2015)
19. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 22–29 (2017)
20. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: PlaneNet: piece-wise planar reconstruction from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2579–2588 (2018)
21. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5162–5170 (2015)
22. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2024–2039 (2016)
23. Moreno-Noguer, F., Lepetit, V., Fua, P.: Accurate non-iterative o(n) solution to the PnP problem. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
24. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5506–5514 (2016) https://doi.org/10.1109/cvpr.2016.594
25. Saxena, A., Sun, M., Ng, A.Y.: Make3D: learning 3D scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. **31**(5), 824–840 (2009)
26. Schöps, T., et al.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
27. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 519–528 (2006). https://doi.org/10.1109/CVPR.2006.19
28. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
29. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
30. Tateno, K., Tombari, F., Laina, I., Navab, N.: CNN-SLAM: real-time dense monocular slam with learned depth prediction. arXiv preprint arXiv:1704.03489 (2017)
31. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2800–2809 (2015)

32. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. arXiv preprint arXiv:1704.02157 (2017)
33. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3917–3925 (2018)