Computer Vision and Image Understanding 191 (2020) 102877

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset☆

Tobias Koch [a,*], Lukas Liebel [a], Marco Körner [a], Friedrich Fraundorfer [b,c]

[a] Chair of Remote Sensing Technology, Technical University of Munich, Germany
[b] Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[c] Remote Sensing Technology Institute, German Aerospace Center, Germany

ARTICLE INFO

Communicated by Joao Pedro Barreto

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

The task of predicting a dense depth map from a monocular RGB image, commonly known as single-image depth estimation (SIDE) or monocular depth estimation (MDE), is an active research topic in computer vision for decades. With the significant progress of deep models in recent years, new standards were set yielding remarkable results in capturing the 3D structure from a single image. However, established evaluation schemes of predicted depth maps are still limited, as they only consider global statistics of the depth residuals. In order to allow for a geometry-aware analysis, we propose a set of novel quality criteria addressing the preservation of depth discontinuities and planar regions, the depth consistency across the image, and a distance-related assessment. As current datasets do not fulfill the requirements of all proposed error metrics, we provide a new high-quality indoor RGB-D test dataset, acquired by a digital single-lens reflex (DSLR) camera together with a *laser scanner*. New insights into the performance of current state-of-the-art SIDE approaches, as well as subtle differences among them, could be unveiled by employing the proposed error metrics on our reference dataset. Additionally, investigations on the real-world applicability of SIDE methods by a series of experiments regarding different image augmentations, illumination changes and textured planar regions have shown current limitations in this research field.

1. Introduction

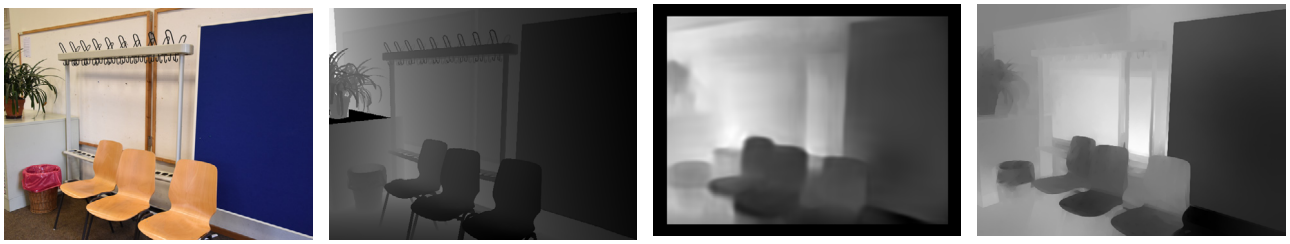
Capturing the 3D structure of a scene from a single image is a fundamental question in computer vision and enables manifold scene reconstruction and understanding applications, such as 2D-to −3D conversion (Xie et al., 2016), 3D modeling (Hassner and Basri, 2006), room layout estimation (Izadinia et al., 2017), image refocusing (Shi et al., 2015), foreground–background segmentation (Dhamo et al., 2019), computational cinematography (Devernay and Beardsley, 2010; Phan and Androutsos, 2013), robot navigation (Mancini et al., 2018), autonomous driving (Godard et al., 2017), or augmented reality systems (Liu et al., 2018). The process of predicting a depth map of a scene using one or more images is commonly known as *depth estimation* and is usually derived from correspondences across stereo images or motion sequences which provide relatively rich information for understanding 3D structures. In contrast, a broad range of research has dealt with the task of predicting pixel-wise depth maps from monocular images, which is generally referred to as MDE. Among the multitude of different approaches, SIDE addresses depth prediction from a single view without prior knowledge and, thus, constitutes the most challenging scenario of this discipline. However, recent years have witnessed the fast development of *deep learning* methods and their massive impact on the computer vision domain, which has also affected the progress of SIDE by implicitly learning relevant scene priors to cope with this task. Current state-of-the-art methods replace traditional handcrafted methods and employ convolutional neural network (CNN) architectures to address the problem of SIDE as a pixel-level regression task. The remarkable results of such methods, exemplary shown in Fig. 1, demonstrate the power of such deep networks by inferring geometrical information solely from monocular RGB or grayscale images.

While these methods produce nicely intuitive results, proper evaluating the estimated depth maps is crucial for subsequent analysis and improvement of the methods, as well as their usability for further 3D understanding scenarios. Consistent and reliable relative depth estimates are, for instance, a key requirement for path planning approaches in robotics (Mancini et al., 2018), augmented reality applications (Liu et al., 2018), or computational cinematography (Devernay and Beardsley, 2010), while preserving the planarity of predicted walls and floors of a room plays a decisive role in room layout estimation applications (Zhuo et al., 2015).

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.102877.

* Corresponding author.
E-mail address: tobias.koch@tum.de (T. Koch).

https://doi.org/10.1016/j.cviu.2019.102877
Received 17 December 2018; Received in revised form 24 August 2019; Accepted 19 November 2019
Available online 26 November 2019



(a) RGB input image (b) Ground truth depth (c) Prediction using method of Eigen and Fergus (2015) (d) Prediction using method of Liu et al. (2015)

Fig. 1. Depth maps produced by different methods, scoring similar errors using standard metrics.

Nevertheless, the evaluation schemes and error metrics commonly used so far mainly consider the overall accuracy by reporting global statistics of depth residuals which do 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 Figs. 1c and 1d, are evaluated with similar errors, although they apparently show different characteristics in terms of ordinal relations, smoothness of planar regions, and defects at object boundaries.

For this reason, we provide a set of new geometrically interpretable error metrics targeting the aforementioned issues allowing for a precise analysis of the performance of depth estimation methods under different perspectives. At the same time, we present a new evaluation dataset[1] acquired from diverse indoor scenarios containing high-resolution RGB images aside highly accurate depth maps from laser scans to overcome the shortage of available datasets providing ground truth data of sufficient quality and quantity.

This work extends our previous work on the evaluation of SIDE (Koch et al., 2018) by providing a more detailed description of our dataset and error metrics, further information on our acquisition procedure and dataset content, and a comprehensive comparison towards other datasets. In addition, we present additional qualitatively and quantitatively results and further experiments that analyze the performance of current state-of-the-art methods for specific situations, such as the presence of textured regions and variations in the scene illumination.

The remainder of the paper is structured as follows: Section 2 starts with a comprehensive presentation of the current state-of-the-art in deriving depth maps from single and stereo images and reviews existing RGB-D datasets that are used for training and benchmarking purposes. A thorough description of the proposed geometric quality metrics is provided in Section 3. Section 4 is devoted to a detailed description of the new `IBims-1` RGB-D indoor dataset and a quantitative and qualitative comparison towards the related `NYU-v2` (Silberman et al., 2012) dataset. In Section 5, an assessment of the performance of several current SIDE methods regarding both established and proposed error metrics on `IBims-1` is outlined, which reveals novel insights into the performance and differences among the methodologies. Besides the benchmarking protocol, Section 6 presents additional experiments that aimed to highlight specific properties of the methods, such as robustness towards image augmentations and the influence of texture and illumination cues on the depth estimation. Section 7 concludes the paper with a concise summary and shares the newly gained insights for further improvements in the field of SIDE.

2. Related work

The task of image-based depth perception is a long-standing and active research field, which has been already addressed by a variety of different techniques. The following section provides an overview of both established and novel algorithms with a focus on most recent learning-based methods. Since those data demanding methods rely on a multitude of aligned RGB and depth image pairs for training, the availability of RGB-D datasets has recently increased significantly. Therefore, we introduce and discuss existing datasets used in the field of SIDE in the second part of this section.

2.1. Methodologies

Recovering depth information from images can be addressed by *single-view* or *multi-view* approaches. The following sections provides an overview of different groups for deriving image-based depth maps. A summary including relevant literatures is listed in Table 1.

Multi-view. Traditionally, depth information is derived by geometric constraints from multiple observations of a scene using stereo camera setups or leveraging camera motion. The former rely on a prior calibration of the stereo setup and dense point correspondences across the stereo images to estimate depth via geometric triangulation. The task of optimal pixel-wise disparity estimation is usually addressed by local, semi-global, or global optimization methods (Szeliski, 2010). While local methods (Yoon and Kweon, 2006) evaluate pixel correspondences in a point-wise approach, yielding fast, but often inaccurate correspondences due to their sensitivity towards appearance changes and occlusions, global (Kolmogorov and Zabih, 2001; Felzenszwalb and Huttenlocher, 2006) and semi-global methods (Hirschmuller, 2005), on the other hand, make explicit smoothness assumptions and solve for a global optimization problem formulated as energy minimization frameworks, resulting into accurate and less noisy depth maps, but requiring significantly increased computation times. A prominent representative for semi-global methods constitutes the well-known *semi-global-matching (SGM)* algorithm (Hirschmuller, 2005). Methods that leverage monocular camera motion are utilizing *Structure-from-Motion (SfM)* or *Simultaneous Localization and Mapping (SLAM)* methods to transform multiple single-view images to a stereo problem, which can be addressed by multi-view stereo (MVS) methods subsequently (Szeliski, 2010). Extensive studies in the field of two or more frame stereo correspondence algorithms can be found in Seitz et al. (2006), Scharstein and Szeliski (2002), Hartley and Zisserman (2003). A further line of approaches was developed with the emergence of light field cameras using an array of micro-lenses placed in front of the image sensor (Heber and Pock, 2016; van Doorn et al., 2011).

Single-view active methods. Another research direction endeavors to ease the multi-view requirement by addressing the task of depth estimation by a sequence of images from the same perspective. Depth information is obtained either by variations of the camera parameters (shape from focus/defocus Favaro and Soatto, 2005; Suwajanakorn et al., 2015), by different lighting conditions of the scene (photometric stereo Ackermann et al., 2015) or by utilizing polarization cues (Ngo et al., 2015; Kadambi et al., 2015).

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

Table 1
Overview of different approaches for image-based depth estimation.

Group	Approach	Method	Literature
Multi-view	Calibrated stereo setup	Local, semi-global, global	Yoon and Kweon (2006), Hirschmuller (2005), Kolmogorov and Zabih (2001) and Felzenszwalb and Huttenlocher (2006)
	Unordered image stacks	SfM + MVS	Hartley and Zisserman (2003), Seitz et al. (2006) and Szeliski (2010)
	Light-field cameras		Heber and Pock (2016) and van Doorn et al. (2011)
Single-view	Active	Shape from focus/defocus	Favaro and Soatto (2005) and Suwajanakorn et al. (2015)
		Lightning conditions	Ackermann et al. (2015)
		Polarization cues	Ngo et al. (2015) and Kadambi et al. (2015)
	Passive	Shape from shading	Horn (1970) and Zhang et al. (1999)
		Atmospheric optics	Nayar and Narasimhan (1999)
	Learning-based	Parametric	Saxena et al. (2006, 2008, 2009), Hoiem et al. (2007), Liu et al. (2010), Ladicky et al. (2014), You et al. (2014), Li et al. (2014), Shi et al. (2015), Hane et al. (2015), Ranftl et al. (2016a), Baig and Torresani (2016) and Furukawa et al. (2017)
		Non-parametric	Konrad et al. (2012, 2013), Karsch et al. (2014), Liu et al. (2014), Choi et al. (2015) and Kong and Black (2015)
	Deep learning-based	Supervised	Eigen et al. (2014), Eigen and Fergus (2015), Li et al. (2015), Liu et al. (2015), Wang et al. (2015), Zoran et al. (2015), Zhuo et al. (2015), Liu et al. (2016), Laina et al. (2016), Kim et al. (2016), Roy and Todorovic (2016), Wang et al. (2016), Chakrabarti et al. (2016), Li et al. (2017), Liu et al. (2018), Fu et al. (2018), Xu et al. (2018), Lee et al. (2018), Hu et al. (2019), Hao et al. (2018), Heo et al. (2018), Yang and Zhou (2018) and Ramamonjisoa and Lepetit (2019)
		Unsupervised	Ummenhofer et al. (2017), Garg et al. (2016), Godard et al. (2017), Zhuo et al. (2015), Kuznietsov et al. (2017), Zhan et al. (2018) and Yin and Shi (2018)

Single-view passive methods. Most prominently, *shape from shading (SfS)* methods (Horn, 1970) exploit intensity or color gradients of a single image under the assumption of homogeneous lighting and Lambertian surface properties. Although these methods work on single-shots, they only perform well for largely known environments or synthetic data but rather poor on real images in unconstrained environments (Zhang et al., 1999). Another early approach aimed at exploiting light sources and illumination conditions, such as haze and fog in an image to recover the relative scene depth by relying on atmospheric optical models (Nayar and Narasimhan, 1999).

Single-view learning-based methods. As one of the first learning-based approaches, Torralba and Oliva (2002) focused on absolute depth estimation for a query image by incorporating the size of known objects depicted in the image. Instead of decomposing the image into its constituent elements, the absolute scene depth of the image is derived from the global image structure represented as a set of features from Fourier and wavelet transforms. The features of the query image were finally compared towards a model trained with 4000 images and corresponding scene depths in a cluster-weighted modeling approach. With the release of first RGB-D datasets (Saxena et al., 2009; Geiger et al., 2012; Silberman et al., 2012), data-driven approaches became feasible and rapidly began to outperform established model-based methods. A pioneer work of a supervised learning-based approach was firstly proposed by Saxena et al. (2006) by training a discriminatively-trained *Markov random field (MRF)* incorporating multi-scale local and global-image features to infer depth. An extension of this work to 3D scene reconstruction was proposed later (Saxena et al., 2009). Since then, a variety of approaches have been proposed to exploit the monocular cues using hand-crafted features together with graphical models (Hoiem et al., 2007; Saxena et al., 2008; You et al., 2014; Li et al., 2014; Shi et al., 2015; Hane et al., 2015; Ranftl et al., 2016b; Baig and Torresani, 2016; Furukawa et al., 2017). Better depth estimates have been achieved by incorporating semantic labels (Liu et al., 2010; Ladicky et al., 2014).

Single-view non-parametric learning-based methods. Another cluster of work estimate depth using non-parametric learning-based methods (Konrad et al., 2012, 2013; Karsch et al., 2014; Liu et al., 2014; Choi et al., 2015; Kong and Black, 2015). These methods assume similarities between RGB values and depth cues across a large set of images. First, similar images of the input image are retrieved from a RGB-D database by feature-based matching. The depth complements of the nearest neighbors are combined and cross-bilateral filtered for smoothing the final depth map (Konrad et al., 2013), warped towards the input image using SIFT flow (Liu et al., 2011; Karsch et al., 2014), or optimized via a conditional random field (CRF) (Liu et al., 2014).

Single-view deep learning-based methods. In conjunction with the undeniable influence of deep learning within the field of computer vision, the research was driven towards the use of CNNs for depth estimation. Since 2014, some works have significantly improved SIDE performance with the use of deep models, demonstrating the superiority of deep features over hand-crafted features (Eigen et al., 2014; Eigen and Fergus, 2015; Li et al., 2015; Liu et al., 2015; Wang et al., 2015; Zoran et al., 2015; Zhuo et al., 2015; Liu et al., 2016; Laina et al., 2016; Kim et al., 2016; Roy and Todorovic, 2016; Wang et al., 2016; Chakrabarti et al., 2016; Li et al., 2017; Liu et al., 2018; Fu et al., 2018; Xu et al., 2018; Lee et al., 2018). These methods pursue the problem of SIDE as a regression problem by building upon successful architectures and learning a deep CNN to estimate the continuous depth map. The first work using deep models was proposed by Eigen et al. (2014) in a two-scale architecture. A coarse global prediction is performed with one network in a first stage, while another network locally refines the prediction in a successive second stage. An extension to this approach uses deeper models and additionally predicts normals and semantic labels (Eigen and Fergus, 2015).

Some works have harnessed the power of pre-trained CNNs in the form of fully convolutional networks (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Laina et al., 2016; Li et al., 2017). The convolutional layers from networks such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) are fine-tuned, while the fully connected layers are re-learned from scratch to encode a spatial feature mapping of the scene. One main limitation using CNNs for depth prediction is decrease of resolution of the output map due to repeated pooling operations in the deep feature extractors. In order to preserve the local structures of output depth maps, several authors have attempted to cope with this problem by up-sampling (Chakrabarti et al., 2016; Eigen and Fergus, 2015; Li et al., 2017), up-convolution blocks (Laina et al., 2016), skip connections between the up-sampling blocks (Li et al., 2017) and space-increasing discretization (Fu et al., 2018).

Improving the quality of predicted depth maps was also addressed by combining CNNs and graphical models, such as *conditional random*

fields (CRFs) (Liu et al., 2015; Li et al., 2015; Wang et al., 2015; Liu et al., 2016; Kim et al., 2016; Xu et al., 2017, 2018). A *deep convolutional neural field (DCNF)* combining CNNs and CRFs in a unified framework for estimating depth on each superpixel while enforcing smoothness within a CRF was proposed by Liu et al. (2015, 2016). Li et al. (2015) and Wang et al. (2015) use hierarchical CRFs to refine their patch-wise CNN predictions from superpixel down to pixel level. CRFs can be exploited to fuse the multi-scale information derived from inner layers of a CNN (Xu et al., 2017, 2018). A combination of CNNs and regression forests with very shallow architectures at each tree node reduces the need for big data (Roy and Todorovic, 2016). Exploiting the Fourier frequency domain in a deep learning algorithm was proposed by Lee et al. (2018).

After the first success of applying deep architectures for SIDE, authors began to focus on tackling major challenges, such as distorted depth discontinuities (Hu et al., 2019; Hao et al., 2018; Ramamonjisoa and Lepetit, 2019) or planar regions (Wang et al., 2016; Heo et al., 2018; Liu et al., 2018; Yang and Zhou, 2018).

Unsupervised deep learning-based. Recently, unsupervised or semi-supervised learning is introduced to learn depth estimation (Ummenhofer et al., 2017; Garg et al., 2016; Godard et al., 2017; Zhou et al., 2017; Kuznietsov et al., 2017; Zhan et al., 2018; Yin and Shi, 2018). This is accomplished by an intermediate task of a view synthesis, and allows training by only using stereo pairs as input with known baselines. These methods design reconstruction losses to estimate the disparity map by recovering a right view with a left view.

Use of synthetic data. With the emergence of synthetic datasets, first work was done to exhibit the possibility to render noise-free and dense depth maps in a very large scale. However, the large domain gaps between synthetic data and real data is still a very challenging task. First works in this field are trying to handle this gap (Guo et al., 2018; Zheng et al., 2018).

Ordinal depth prediction. Some applications only require relative or ordinal depth, such as 2D-to −3D conversion (Karsch et al., 2014), image refocusing (Anwar et al., 2017), or foreground–background segmentation (Camplani and Salgado, 2014). Methods in this field predict dense relative depths from pairwise relationships (closer-than and further-than relationships) estimates for rare points in the input image (Zoran et al., 2015; Chen et al., 2016).

2.2. Existing RGB-D datasets

In order to train supervised SIDE methods as well as to evaluate and compare them with other approaches, any dataset containing corresponding RGB and depth image pairs can be considered, which also comprises, *e.g.*, benchmarks originally designed for the evaluation of MVS approaches.

This variety of freely available datasets can be categorized according to different criteria (*cf.* Table 2). Some of them exhibit an adequate number of samples for training deep models, others concentrate on few, but highly accurate, RGB-D image pairs allowing for exhaustive analysis and comparison of different methodologies. The amount and quality of depth maps also depends on the choice of the sensor used for the acquisition campaign. In general, RGB-D image pairs are commonly generated either by active sensors, such as RGB-D cameras or laser scanners, or passively by the use of stereo images. While active RGB-D sensors, such as the Microsoft Kinect version 1 and 2, the Occipital Structure Sensor, and the Intel RealSense are pre-calibrated setups, ready to produce aligned depth maps in a large quantity without manual effort, LiDAR sensors are slow and usually need an additional camera and registration technique. However, the quality of generated depth maps from LiDAR are superior to RGB-D sensors in terms of resolution, completeness, range, and accuracy. More recently, researches started to make use of the big amount of image data from freely available image databases, such as Flickr, to generate RGB-D image pairs utilizing stereo vision algorithms. With the generation of synthetic data, data-depending deep learning methods can be fed with innumerable training data.

The following datasets can currently be considered for the task of SIDE. Among the datasets that rely on precise laser scan data, Strecha et al. (2008) 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` (Schöps et al., 2017) and the `Tanks & Temples` (Knapitsch et al., 2017) datasets, have been released, which stand out due to their high resolution indoor and outdoor images and accurate ground-truth point clouds acquired from a laser scanner. 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 are acquired in a sequential matter. The scans can be used to generate depth maps aligned with the captured RGB images, but, however, it cannot be guaranteed that 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` (Saxena et al., 2009), 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 (Geiger et al., 2012) 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 for training and evaluating SIDE in indoor scenarios is the `NYU depth v2` (Silberman et al., 2012) dataset 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` (Chang et al., 2017), `ScanNet` (Dai et al., 2017), and `2D-3D-S` (Armeni et al., 2017) datasets provide even larger amounts of indoor scenes collected from RGB-D cameras, such as the Matterport Camera or the Structure sensor (Occipital, 2016). These datasets are valuable additions to the `NYU-v2` dataset but also suffer from the same weaknesses, as the used sensors have a similar design to the Kinect v1 sensor.

Recently, RGB-D datasets have been published using solely RGB images, such as `DIW` (Chen et al., 2016), `MegaDepth` (Li and Snavely, 2018), and `ReDWeb` (Xian et al., 2018). These datasets provide depths maps generated from stereo images utilizing freely available large-scale data platforms (*e.g.*, Flickr). They offer a huge variety of different scenes containing both indoor and outdoor scenes and can be easily computed using established MVS methods. However, the scale is unknown and the provided depth maps are therefore only relatively scaled, which only allows for ordinal depth estimation. Nevertheless, first investigations on training deep networks on these images reveal better generalization capabilities, but, however, they are ineligible when a metric scale is needed.

The LiDAR-based `LIVE Color+3D Database` (Su et al., 2017) offers highly-accurate registered RGB-D image pairs for 98 outdoor scenes similar to `Make3D`, but with an increased resolution and dense depth maps. The large range of scene depths and the high quality of

Table 2
Comparison of existing datasets related to SIDE evaluation with respect to different dataset characteristics. Interval distinguishes between still image acquisition (still) and continuous image acquisition (cont). Density specifies the completeness of provided depth maps. Higher resolutions are specified in brackets, if available in the datasets.

Benchmark	Setting	Sensor	Scenes	Images	Interval	Range (in m)	Density	Resolution (in Mpx)
MegaDepth (Li and Snavely, 2018)	Outdoor	RGB	196	130k	Still	Relative	Dense	1.9
DIW (Chen et al., 2016)	Various	RGB	–	470k	Still	Relative	2 points	0.15
ReDWeb (Xian et al., 2018)	Various	RGB	–	3.6k	Still	Relative	Dense	0.19
SceneNet RGB-D (McCormac et al., 2017)	Indoor	Synthetic	57	5m	Cont.	1–5 m	Dense	0.08
SUNCG (Song et al., 2017)	Indoor	Synthetic	45k	130k	Cont.	1–8 m	Dense	0.31
360-D (Zioulis et al., 2018)	Indoor	Various	–	22k	Still	1–10 m	Dense	0.13
NYU-v2 (Silberman et al., 2012)	Indoor	RGB-D	464	654	Still	1–10 m	Gaps	0.31
Matterport3D (Chang et al., 2017)	Indoor	RGB-D	90	200k	Cont.	1–10 m	Gaps	0.8
ScanNet (Dai et al., 2017)	Indoor	RGB-D	707	2.5m	Cont.	0.4–3.5 m	Gaps	0.31
2D-3D-S (Armeni et al., 2017)	Indoor	RGB-D	6	25k	Cont	1–10 m	Gaps	1.3
ETH3D (Schöps et al., 2017)	Various	LiDAR	25	898	Cont.	1–20 m	Gaps	0.4 (24)
Tanks & Temples (Knapitsch et al., 2017)	Various	LiDAR	14	150k	Cont.	1–20 m	Dense	2
Kitti (Geiger et al., 2012)	Street	LiDAR	–	697	Cont.	1–80 m	Sparse	0.5
Strecha (Strecha et al., 2008)	Outdoor	LiDAR	6	30	Still	1–10 m	Dense	6
Make3D (Saxena et al., 2009)	Outdoor	LiDAR	–	534	Still	1–80 m	Sparse	0.017
LIVE Color+3D Database (Su et al., 2017)	Outdoor	LiDAR	–	98	Still	2–100 m	Dense	2.07
IBims-1 (Koch et al., 2018)	Indoor	LiDAR	70	100	Still	0.3–25 m	Dense	0.31 (1.5)

the depth maps allow for detailed investigations of SIDE methods in outdoor scenarios, however, the scene variety is rather limited.

With the appearance of synthetic datasets, such as SceneNet RGB-D (McCormac et al., 2017), SUNCG (Song et al., 2017), and 360-D (Zioulis et al., 2018), first attempts were made to train deep models with rendered RGB-D image pairs of this multitude of synthetically generated indoor scenes. However, the rendered RGB images are still far from realistic shots and are therefore not suited for testing the applicability of SIDE methods in real world environments.

3. Novel evaluation metrics for depth estimation

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 (Eigen et al., 2014; Eigen and Fergus, 2015; Laina et al., 2016; Li et al., 2017; Xu et al., 2017):

Absolute relative difference: $\text{rel}\left(\boldsymbol{Y},\boldsymbol{Y}^*\right) = \frac{1}{T}\sum_{i,j}\left|y_{i,j}-y^*_{i,j}\right|/y^*_{i,j}$

Squared relative difference: $\text{srel}\left(\boldsymbol{Y},\boldsymbol{Y}^*\right) = \frac{1}{T}\sum_{i,j}\left|y_{i,j}-y^*_{i,j}\right|^2/y^*_{i,j}$

RMS (linear): $\text{RMS}\left(\boldsymbol{Y},\boldsymbol{Y}^*\right) = \sqrt{\frac{1}{T}\sum_{i,j}\left|y_{i,j}-y^*_{i,j}\right|^2}$

RMS (log): $\log\left(\boldsymbol{Y},\boldsymbol{Y}^*\right) = \sqrt{\frac{1}{T}\sum_{i,j}\left|\log y_{i,j}-\log y^*_{i,j}\right|^2}$

Threshold: percentage of $\boldsymbol{Y}$ such that $\max(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}) = \sigma < thr$

The absolute relative difference error measures the relative per-pixel error linear to the absolute distance. In other words, an error of 0.1 m at a depth of 1 m is penalized equally to an error of 1 m at a depth of 10 m. An alternative with a squared influence of the relative per-pixel error is given by the squared relative difference. In contrast, the RMS error equally penalizes an error of 0.1 m at both depths. The threshold error on the other hand considers per-pixel proportions rather than per-pixel differences and measures the ratio of pixels, for which the relative difference between prediction and ground truth depths is below a threshold (thresholds are usually set to 1.25, 1.25^2, and 1.25^3).

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. A general weakness of most current state-of-the-art SIDE methods is that the outputs tend to have spatially distorted or blurry object edges. While these local structures only affect a rather small part of the entire image, missing or blurry depth discontinuities have only a minor effect on the global error metrics, impeding a fair comparison of different methods.

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 Section 3.1. Visual illustrations of our metrics explained below are depicted in Fig. 2. 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 entire image area?

3.2.1. 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.

3.2.2. Planarity Error (PE)

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

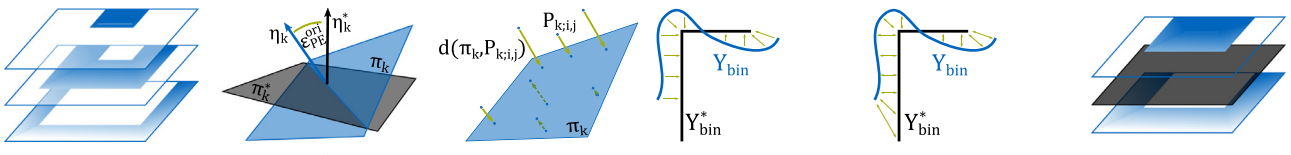


Fig. 2. Visualizations of our proposed error metrics. The *distance-related assessment* (a) applies standard metrics for different depth range intervals. The flatness and orientations of predicted planar regions can be evaluated with our *planarity errors* (b and c). The location accuracy and completeness of depth discontinuities is rated by the *depth boundary errors* (d and e), while the consistency of depth predictions with respect to a virtual depth plane can be assessed with our *directed depth errors* (f).

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 origin o. In detail, a masked depth map $\boldsymbol{Y}_k$ of a particular planar surface and an intrinsic matrix is used together in order to project the masked depth map 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 $\mathcal{P}_k = \left\{\boldsymbol{P}_{k;i,j}\right\}_{i,j}$, respectively. The planarity error

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

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

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

is defined as the 3D angle difference between the normal vectors of predicted and ground truth 3D planes. Figs. 2b and 2c illustrate the proposed planarity errors. Note that for each individual planar mask the predicted depth maps are median scaled w.r.t. the ground truth depth map. This eliminates scaling differences of compared methods, which would influence the planarity error by favoring underestimated depth predictions.

3.2.3. Location Accuracy of Depth Boundaries (DBE)

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. In order to evaluate predicted depth maps, edges $\boldsymbol{Y}_{\text{bin}}$ are extracted and compared to a set of ground truth edges $\boldsymbol{Y}_{\text{bin}}^*$ via *truncated chamfer distance* of the binary edge images. Specifically, a *Euclidean distance transform* is applied to the ground truth edge image $\boldsymbol{E}^* = DT\left(\boldsymbol{Y}_{\text{bin}}^*\right)$, while distances exceeding a given threshold θ are truncated to a maximum distance θ. We define the depth boundary errors (DBEs), comprised of an accuracy measure

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

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

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

by accumulating both ground truth and predicted edges multiplied with their corresponding distance maps of ground truth and predicted edges $\boldsymbol{E}^*$ and $\boldsymbol{E} = DT\left(\boldsymbol{Y}_{\text{bin}}\right)$. Therefore, the completeness error penalizes both missing and extra edges in the predictions in an equal manner. A visual explanation of the DBEs are illustrated in Figs. 2d and 2c.

3.2.4. Directed Depth Error (DDE)

For many applications, it is of high interest that depth images are consistent over the whole image area. Although the absolute depth error, the squared depth error and the RMS errors 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)

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

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

$$\varepsilon_{\text{DDE}}^{-}\left(\boldsymbol{Y}, \boldsymbol{Y}^*, \boldsymbol{\pi}^*\right) = \frac{\left|\left\{y_{i,j} \,|\, d_{\text{sgn}}(\boldsymbol{\pi}^*, \boldsymbol{P}_{i,j}) < 0 \wedge d_{\text{sgn}}(\boldsymbol{\pi}^*, \boldsymbol{P}_{i,j}^*) > 0\right\}\right|}{T} \tag{7}$$

as the proportions of correct, too far and too close predicted depth pixels $\varepsilon_{\text{DDE}}^0$, $\varepsilon_{\text{DDE}}^+$ and $\varepsilon_{\text{DDE}}^-$. In practice, a reference depth plane $\boldsymbol{\pi}^*$ is defined at a certain distance (*e.g.*, at 3 m) orthogonal to the camera view 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 maps.

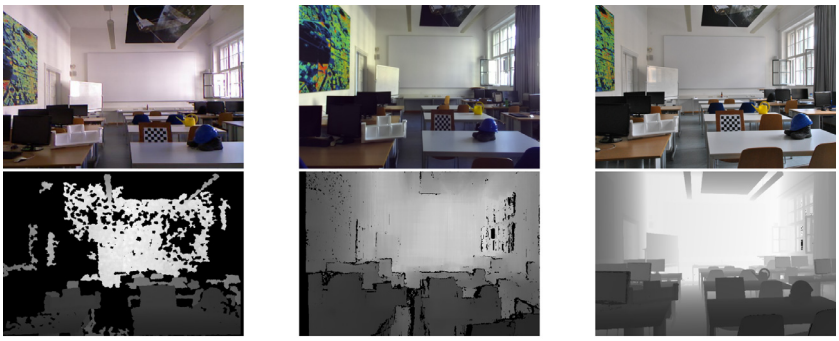
4. The IBims-1 dataset

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. Sensor comparison

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.

For the creation of depth maps, we considered various sensors and instruments. Common mass market RGB-D products, such as Microsoft Kinect, not only allow for fast and convenient capturing of scenes, but



(a) Kinect v1 (b) ZED Cam (c) Laser scan + DSLR

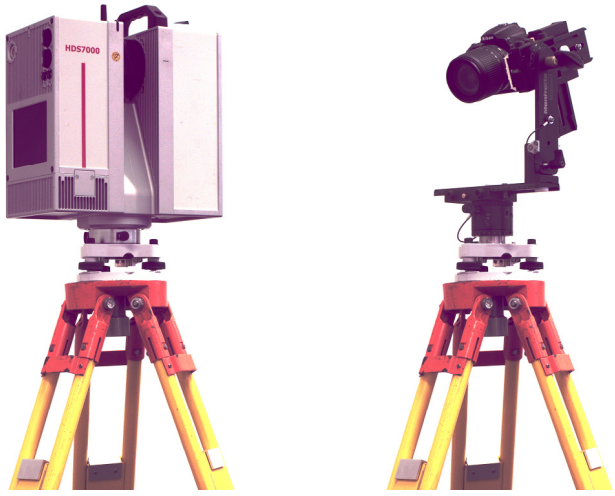
Fig. 3. Comparison of the depth map quality of different sensors.

also provide registered images and depth maps at the same time. However, the overall quality – especially in terms of resolution, accuracy and depth range – of the resulting depth maps and images turn out to be insufficient for the intended usage as reference data. Stereo rigs, such as the Stereolabs ZED camera, outperform RGB-D products in several crucial areas, such as outdoor scenes. They are equally easy to use but also show deficits in certain areas. As the stereo reconstruction only produces results for textured surfaces, the produced depth maps are often incomplete and suffer from noise. Precise geodetic instruments, such as tacheometers, laser trackers, or laser scanners, can provide highly accurate distance measurements. Among them, laser scanners excel in recording highly accurate dense point clouds in 360°. Fig. 3 shows a comparison of depth maps acquired from different sensors capturing the same scene. Beside differences in the image quality and intrinsics of the RGB images, the depth map generated with the Kinect v1 lack from numerous areal gaps as well as distorted object boundaries. The depth map provided by the ZED Cam on the other hand features almost dense depth estimates due to an internal interpolation of texture-less regions but, however, show the same deficits around object boundaries caused by the parallax effect. In addition, the overall noise level – especially for high distances – is relatively large compared to the Kinect v1. The high density and extremely high accuracy of the laser scanner allows for generating accurate, dense and detailed depth maps of superior quality compared to the other sensors.

As we want to generate highly accurate depth maps for high-resolution images, we finally chose a laser scanner as our sensor of choice. They do, however, fall short of expectations regarding provided imagery. As only a few instruments can capture RGB images at all, this is, in practice, most commonly done using an auxiliary camera. For this reason, we decided to design our own acquisition setup, as it is explained in the following section.

4.2. Acquisition process

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. Dependent on the scene depth of the individual images in our dataset we varied the point spacing of the acquired scans to ensure at least one depth value for each pixel in a down-sampled version of the RGB image of 640×480 px. However, for most scenes we exceeded the required point density by a multiple in order to provide nearly dense depth maps in a higher resolution as well (1500×1000 px). 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 our data acquisition, we chose two calibrated DSLR cameras: one Nikon D5500 DSLR camera equipped with a Nikon AF-S Nikkor 18–105 mm lens, mechanically fixed to a focal length of 18 mm and a Nikon D3000 DSLR camera equipped with the same lens, mechanically fixed to focal lengths of 18 mm and 21 mm.



(a) Laserscanner (b) Camera

Fig. 4. Our hardware setup used for the acquisition of `IBims-1` with a laser scanner (a) and a DSLR camera (b) mounted on a survey tripod. A custom panoramic tripod is used in order to achieve a coincidence of the optical center of the camera and the origin of the laser scanner coordinate system to avoid occlusions in the resulting depth maps.

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. An illustration of our setup is depicted in Fig. 4. 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.

4.3. 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 (Moreno-Noguer et al., 2007). This registration of the camera relative to the point cloud

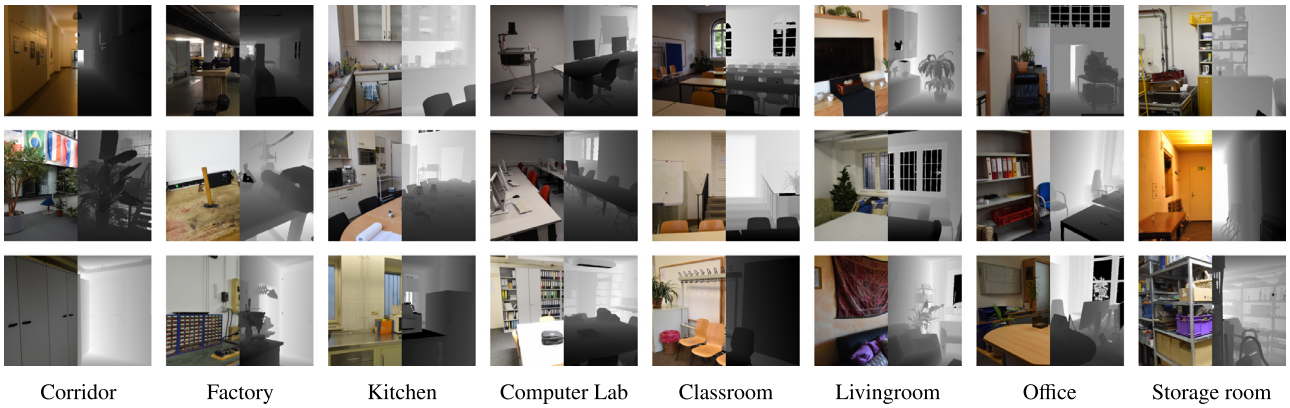


Fig. 5. Sample RGD-D image pairs of our IBims-1 dataset covering different scenes. Illustrations are composed of the RGB image (left) and the corresponding depth map (right).

Table 3
Statistics of plane annotations in NYU-v2 and IBims-1. Number of instances (Inst.) of a specific plane type (Type) occurred in the dataset (Images), the average size of each object mask (Avg. Size), and accuracy of fitted 3D reference planes. The larger deviations in planes fitted to the images of NYU-v2 can be attributed to the inaccurate and noisy measurements of the utilized RGB-D sensor. A reliable assessment of planarity errors based on NYU-v2 is therefore only possible to a limited extent.

Dataset	Type	Images	Inst.	Avg. Size (in px)	Mean Dev. (in mm)	Std. Dev. (in mm)
NYU-v2	Floor	132	132	29 389	17.42	14.25
NYU-v2	Table	44	44	27 989	17.80	17.19
NYU-v2	Wall	168	168	34 975	28.17	22.66
IBims-1	Floor	47	51	22 813	1.57	1.85
IBims-1	Table	46	54	15 704	1.18	1.50
IBims-1	Wall	82	140	46 744	1.79	2.38

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. After the pose estimation of the camera, 3D points were projected to the 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 of the depth maps.

4.4. Registration accuracy

In order to present a high-quality RGB-D reference dataset, it is crucial that RGB images and depth images are aligned properly. The reprojection errors of the 2D-3D correspondences used for the camera pose estimations provide a first evidence of the registration accuracy of our dataset. For each of the 100 RGB-D image pairs of our dataset we manually selected 8–10 point correspondences. The mean reprojection error for all 2D-3D correspondences is 0.81 px with respect to the NYU-like resolution.

Since the reprojection error is only calculated on the basis of a single points, it is difficult to make a general statement about the overall registration accuracy. For this reason we also investigate the alignment on the basis of edges with the assumption that most depth discontinuities in a edge map correspond to intensity changes in the RGB image. We therefore compute dominant edges in depth maps and RGB images respectively using a Sobel operator and compare them using a *directed chamfer distance*. Note, that we only consider edges in the RGB image which are located in the local neighborhood of extracted depth edges (*e.g.*, within 10 px) for excluding gradients caused by texture or illumination changes. In average, around 450 edge pixels were extracted and compared for each RGB-D image pair. The averaged chamfer distance considering all images is 1.20 px. Since some depth edges do not correspond to intensity changes in the RGB image and vice versa, this metric serves only as a vague proof of the registration accuracy, but, however, yields an overall quality measure showing how accurate our RGB and depth maps are aligned.

4.5. Contents

Following the described procedure in Sections 4.2 and 4.3, 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. This main part of the dataset contains 100 RGB-D image pairs in total. 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 Fig. 5.

Additionally, several manually created masks are provided. Unreliable or invalid pixels in the depth map are labeled by two different sets of binary masks. One of which flags transparent objects, mainly windows, which could be assigned with an ambiguous depth. While the laser scanner captured points behind those objects, it may be intended to obtain the distance of the transparent object for certain applications. The other mask for invalid pixels indicates faulty values in the 3D point cloud. Those mainly originate from scanner-related errors, such as reflecting surfaces, as well as regions out of range. Three further sets of masks label planar surfaces of three different types, *i.e.*, tables, floors, and walls. Each instance is contained in a separate mask. Examples for planar masks are shown in Fig. 6, while statistics of the plane annotations are listed in Table 3. It is worth mentioning that the plane masks do not coincide with the object boundaries, but rather keeping a buffer area of several pixel towards the object boundaries. The reason for this is that these masks are used for investigating the capability of predicting planar regions. Object boundaries often cause distortions in the predicted depth map which is target of our DBE but should not influence the PE.

In order to allow for evaluation following the proposed DBE metric, we provide distinct edges for all images. 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 applying a Canny edge detector on the depth maps. Since the scenes in our dataset exhibit various depth ranges, the selection of dominant

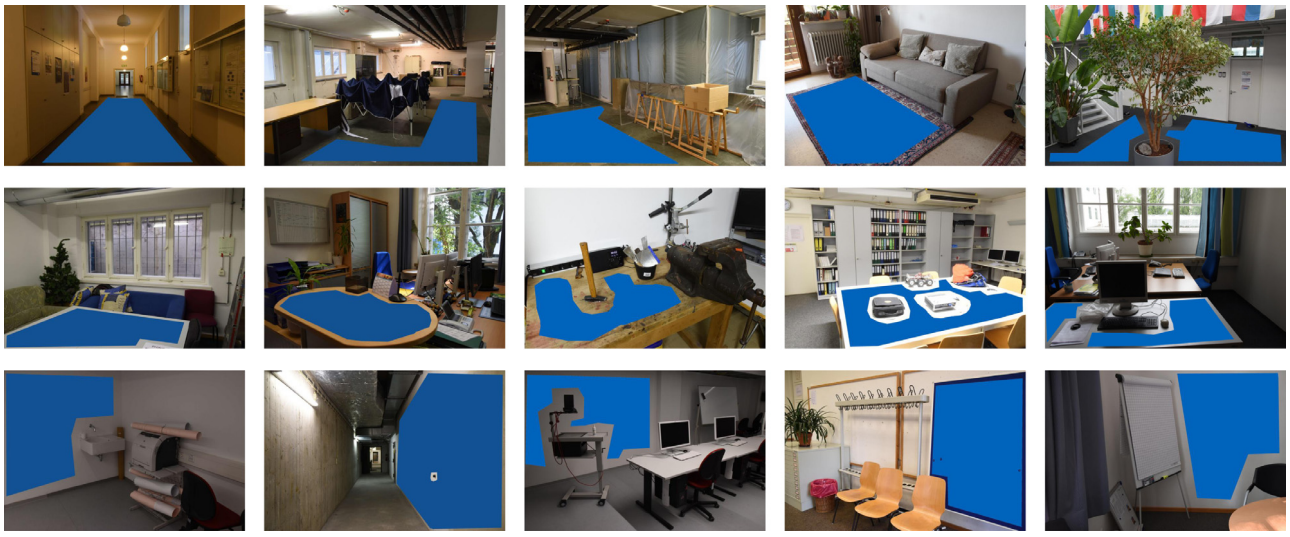


Fig. 6. Annotation samples showing provided plane masks (■) for *floors* (top), *table tops* (mid) and *walls* (bottom).

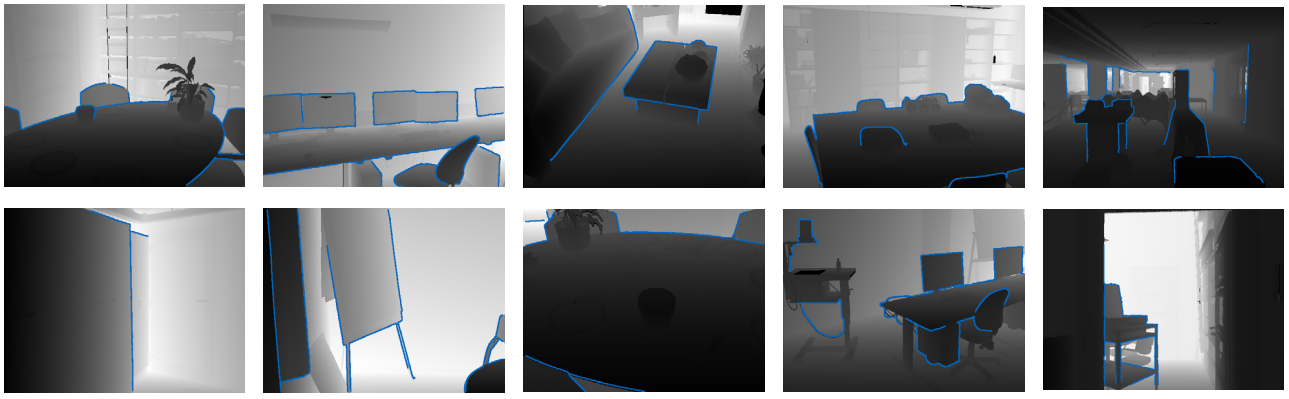


Fig. 7. Annotation samples showing provided edge masks (▬) for distinct depth discontinuities.

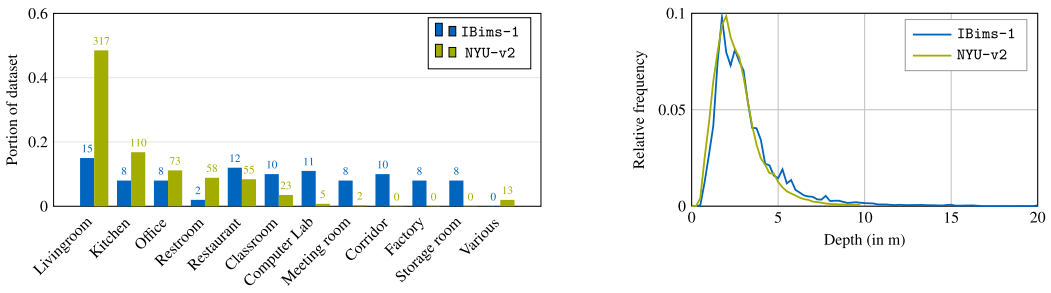
edges vary with the depth range of the individual RGB-D image pairs. For this reason, we only consider distinct depth edges that exceed a depth change of at least 15 % of the overall depth range in the individual image. Fig. 7 shows examples of the ground truth edges for different scenes from IBims-1.

Additionally, we provide an *auxiliary dataset* which consists of four parts: (1) Four outdoor RGB-D image pairs, containing vegetation, building, cars and 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 (Asuni and Giachetti, 2014) hung on a wall. Those could potentially give valuable insights, as they reveal what kind of image features SIDE methods exploit. 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) 56 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 many 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.

4.6. Comparison of IBims-1 and NYU-v2

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 offices, lecture, and living rooms, computer labs, as well as more challenging ones, such as long corridors, potted plants and factory rooms. A comparison regarding the scene variety between NYU-v2 and IBims-1 can be seen in Fig. 8a. Furthermore, IBims-1 features statistics comparable to NYU-v2, such as the distribution of depth values, shown in Fig. 8b, and a comparable field of view.

However, comparing the depth map quality of both datasets, raw depth maps of NYU-v2 show a large amount of missing and erroneous depth values due to parallax effect, limited range (up to 10 m), and relatively high noise level, as this is already investigated in Zennaro et al. (2015). In total, 36 % of all depth values in the raw depth maps in NYU-v2 are missing. Missing values were interpolated using the colorization method of Levin et al. (2004), which results in erroneous measurements and artifacts, such as flying pixels. Moreover, transparent and specular surfaces are not masked in NYU-v2 resulting in distorted depth values in the dataset. Due to the high point density and accuracy of the scans in IBims-1, no interpolation is needed for NYU-like resolution in IBims-1, resulting in dense and valid depth values. Fig. 9 visualizes the quality of NYU-v2 and compares it towards IBims-1. In contrast to the imprecise and incomplete depth maps in NYU-v2, the seamless depth maps in IBims-1 facilitate the extraction of accurate and complete depth discontinuities. The high accuracy of these depth maps also guarantees the extraction of accurate 3D planes in the range of a few millimeters, while deviations of more



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

(b) Distribution of depth values

Fig. 8. IBims-1 dataset statistics compared to the NYU-v2 dataset. Scene variety (a) and distribution of depth values (b)

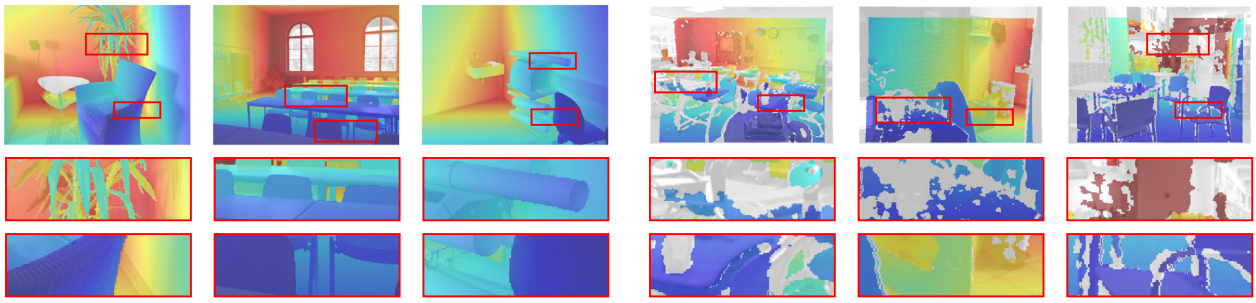


Fig. 9. Visualization of registration accuracy and depth completeness of IBims-1 (left) and NYU-v2 (right). Overlay of grayscale RGB images and colored depth maps for various samples (invalid or missing depth values are depicted in gray). Top: full image. Middle and bottom row: detailed views.

than 2 cm were noted when making use of the NYU-v2 dataset,[2] as shown in Table 3. Although in principle NYU-v2 allows to generally assess the planarity of depth predictions, the results would not satisfy our accuracy requirements for providing reliable conclusions about the performance of the methods.

5. Evaluation of SIDE methods

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. As outlined in our review of the state-of-the-art in Section 2.1, a multitude of different deep learning-based SIDE approaches have been developed over the past few years. Naturally, not all of them can be subjected to detailed investigation. We chose an exemplary subset of the available approaches for the evaluation experiments, that either represent a milestone in the development of SIDE, or constitute current approaches that address specific aspects of particular interest, which have been identified in the development of our geometrically interpretable error metrics. In order to allow a fair comparison, only methods that were trained on indoor scenes, namely the NYU-v2 dataset, were examined. This preliminary selection was further narrowed down to accessible methods for which we received either source code or predictions for our dataset, which ultimately led to a comparison of eight methods, namely those proposed by Eigen et al. (2014), Eigen and Fergus (2015), Liu et al. (2015), Laina et al. (2016), Li et al. (2017), PlaneNet (Liu et al., 2018) and Sharpnet (Ramamonjisoa and Lepetit, 2019). Since all of these methods were solely trained on the NYU-v2 dataset, differences in the results are expected to arise from the developed methodology rather than the training data. For the evaluation using our dataset, only valid depth areas were considered by applying the provided corresponding masks to the raw depth maps. The quantitative results on both datasets with all error metrics are listed in Table 4. A detailed analysis of the individual metrics is given in the following sections. Although a runtime evaluation would be of great interest for many application fields, the realization of a revealing comparison was infeasible, since runtime is highly dependent on implementation details and utilized frameworks, which varied between the examined methods. Furthermore, the lack of available source code for some methods prevents a comparison on the same hardware setup.

5.1. Established global error metrics

The results of evaluation using commonly used global metrics on IBims-1 and NYU-v2 listed in Table 4 by computing the statistical error metrics on the complete images. This is the standard evaluation procedure in all recent publications. The revealed lower overall scores for our dataset are expected since the dataset is previously unseen by these methods. As the methods are trained to predict depths in the range of the NYU-v2 dataset (*i.e.*, 1–10 m), they are not able to estimate depths beyond this range which are also encompassed in our dataset. This highly affects the RMS error, which turned out to be almost three times as large as in NYU-v2. Moreover, our dataset uncovers different generalization capabilities of the methods, as the order of the rankings has changed between NYU-v2 and IBims-1. However, the ranking according to different standard metrics did not change substantially among the methods, as most metrics are highly correlated to each other. This proves our claim for further sophisticated evaluation criteria, which are analyzed in the following sections.

5.2. Distance-related assessment

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. Fig. 10 shows the error band of the relative and RMS errors of the method proposed by Li et al. (2017) 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

[2] Plane annotations for NYU-v2 were also made available on our webpage.

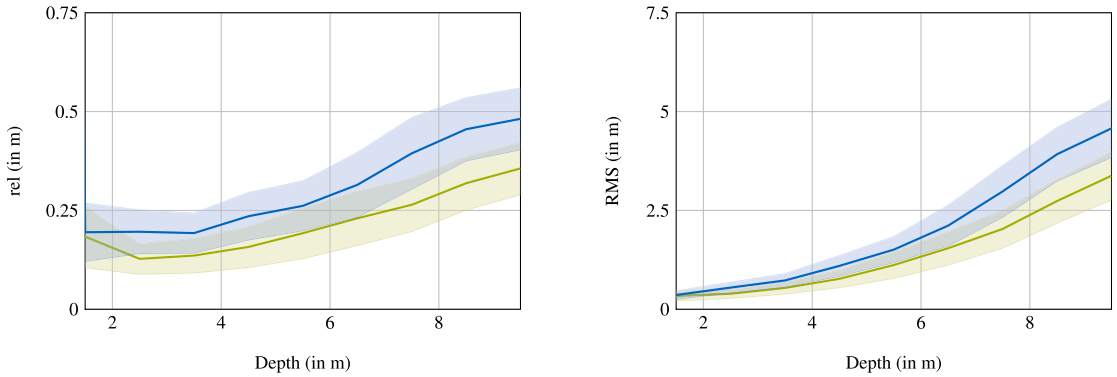


Fig. 10. Distance-related global errors (left: relative error and right: RMS) for the shared depth range of NYU-v2 (mean: ——, ±0.5 std: ■) and IBims-1 (mean: ——, ±0.5 std: ■) using the method of Li et al. (2017).

Table 4
Quantitative results for standard metrics on NYU-v2 and standard metrics, proposed PE, DBE, and DDE metrics on IBims-1 applying different SIDE methods (**best**, **second best**). Higher the better for ↑ and lower the better for ↓.

Method	Dataset	Standard metrics ($\sigma_i = 1.25^i$)						PE (cm/°)		DBE (px)		DDE (%) for d =3 m		
		rel ↓	$\log_{10}$ ↓	RMS ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑	$\epsilon_{\text{PE}}^{\text{plan}}$ ↓	$\epsilon_{\text{PE}}^{\text{orie}}$ ↓	$\epsilon_{\text{DBE}}^{\text{acc}}$ ↓	$\epsilon_{\text{DBE}}^{\text{comp}}$ ↓	$\epsilon_{\text{DDE}}^{0}$ ↑	$\epsilon_{\text{DDE}}^{-}$ ↓	$\epsilon_{\text{DDE}}^{+}$ ↓
Eigen et al. (2014)	NYU-v2	0.22	0.09	0.76	0.61	0.89	0.97	–	–	–	–	–	–	–
Eigen and Fergus (2015) (AlexNet)	NYU-v2	0.19	0.08	0.67	0.69	0.91	**0.98**	–	–	–	–	–	–	–
Eigen and Fergus (2015) (VGG)	NYU-v2	0.16	**0.07**	0.58	0.75	0.95	**0.99**	–	–	–	–	–	–	–
Laina et al. (2016)	NYU-v2	**0.14**	**0.06**	**0.51**	**0.82**	0.95	**0.99**	–	–	–	–	–	–	–
Liu et al. (2015)	NYU-v2	0.21	0.09	0.68	0.66	0.91	**0.98**	–	–	–	–	–	–	–
Li et al. (2017)	NYU-v2	**0.15**	**0.06**	0.53	0.79	**0.96**	**0.99**	–	–	–	–	–	–	–
Liu et al. (2018)	NYU-v2	**0.14**	**0.06**	**0.51**	0.81	**0.96**	**0.99**	–	–	–	–	–	–	–
Ramamonjisoa and Lepetit (2019)	NYU-v2	**0.14**	**0.06**	**0.46**	**0.84**	**0.97**	**0.99**	–	–	–	–	–	–	–
Eigen et al. (2014)	IBims-1	0.32	0.17	1.55	0.36	0.65	0.84	7.70	24.91	9.97	9.99	70.37	27.42	2.22
Eigen and Fergus (2015) (AlexNet)	IBims-1	0.30	0.15	1.38	0.40	0.73	0.88	7.52	21.50	4.66	8.68	77.48	18.93	3.59
Eigen and Fergus (2015) (VGG)	IBims-1	**0.25**	**0.13**	1.26	0.47	0.78	**0.93**	**5.97**	**17.65**	4.05	8.01	79.88	18.72	**1.41**
Laina et al. (2016)	IBims-1	0.26	**0.13**	1.20	0.50	0.78	0.91	**6.46**	19.13	6.19	9.17	81.02	17.01	1.97
Liu et al. (2015)	IBims-1	0.30	**0.13**	1.26	0.48	0.78	0.91	8.45	28.69	**2.42**	**7.11**	79.70	14.16	6.14
Li et al. (2017)	IBims-1	**0.22**	**0.11**	**1.09**	**0.58**	**0.85**	**0.94**	7.82	22.20	3.90	8.17	**83.71**	**13.20**	3.09
Liu et al. (2018)	IBims-1	0.29	0.17	1.45	0.41	0.70	0.86	7.26	**17.24**	4.84	8.86	71.24	28.36	**0.40**
Ramamonjisoa and Lepetit (2019)	IBims-1	0.26	**0.11**	**1.07**	**0.59**	**0.84**	**0.94**	9.95	25.67	**3.52**	**7.61**	**84.03**	**9.48**	6.49

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 camera with slightly different intrinsics and image quality, as well as for unseen scenarios. A comparison of the performance on a larger depth range for different methods and error metrics, as shown in Fig. 11, clearly shows a trend of decreasing accuracy over an increasing distance. Best results are achieved in a very close range up to 4 m, which corresponds to the maximum of the depth distribution of the NYU-v2 dataset on which the methods were trained (*cf.* Fig. 8b). Training on this highly imbalanced dataset with current state-of-the-art methods results in predicting depths below a RMS error of 1 m for distances up to 5 m, but linearly increases together with the scene depth for distances greater than 5 m. While most methods do not differ significantly in predicting depth values at various ranges and correspond to the ranking in Table 4, the method of Ramamonjisoa and Lepetit (2019) performs notably better at larger distances. However, since the results exhibit a deficiency in close ranges up to 2–3 m, which corresponds to the peak of the depth distribution in IBims-1, errors in this range decisively contribute to the global errors listed in Table 4. Such enhanced distinction and assessment of the performance would not have been feasible by solely relying on established global error metrics.

5.3. Planarity

To investigate the quality of reconstructed planar structures, we evaluated the different methods with the planarity and orientation errors $\epsilon_{\text{PE}}^{\text{plan}}$ and $\epsilon_{\text{PE}}^{\text{orie}}$, respectively, as defined in Section 3.2.2, for different planar objects. In particular, we distinguished between horizontal and vertical planes and used masks from our dataset. Beside a combined error, including all planar labels, we separately computed the errors for the individual objects as well. Results for averaged errors among all types of planar regions are listed in Table 4, while results for individual plane types are shown in Fig. 12. The results reveal different performances for individual classes, especially orientations of floors and table tops were predicted in a significantly higher accuracy, while the absolute orientation error for walls is surprisingly high. Considering the flatness of the predictions, tables can be reconstructed more reliable than floors or walls. 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. (2017) achieved much better results in predicting orientations of horizontal

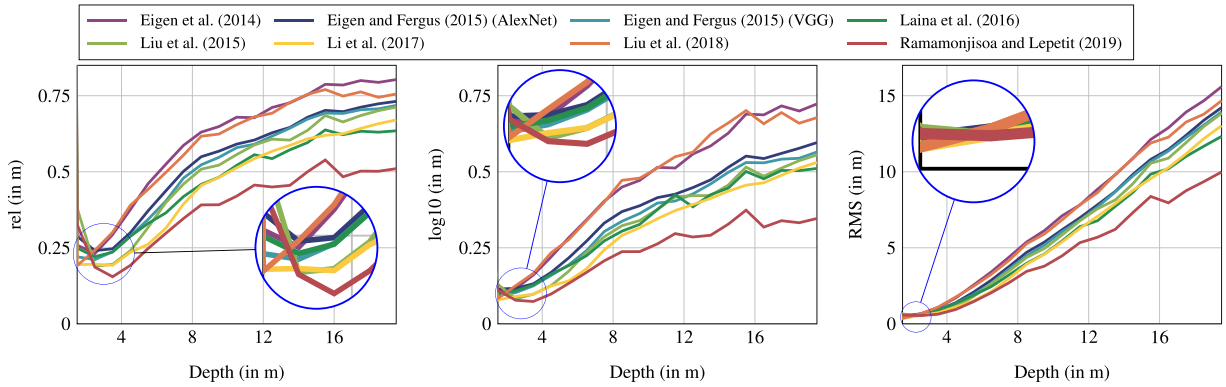


Fig. 11. Comparing distance-related global errors up to 20 m on `IBims-1` for the examined methods. From left to right: relative error, log10 error and RMS error.

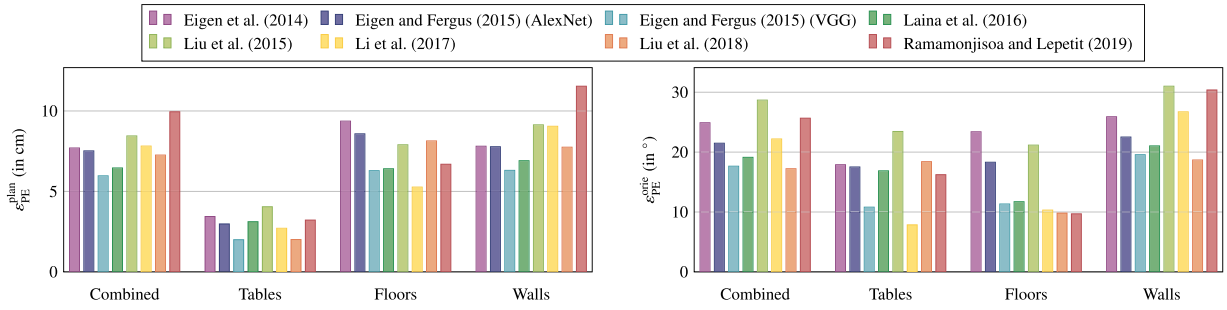


Fig. 12. Results for the planarity metrics $\epsilon_{\text{PE}}^{\text{plan}}$ (left) and $\epsilon_{\text{PE}}^{\text{orie}}$ (right) on `IBims-1` for individual plane types and a combination of all (Combined).

planes but also performed rather bad on vertical surfaces. In contrast, orientation results for (Liu et al., 2015) exhibit large errors for all types of planes. Reason for this could lie in problems of smooth depth transitions for adjacent superpixels representing flat, but textured or differently illuminated areas. This oversegmentation results in strong depth changes in planar regions. The method of Ramamonjisoa and Lepetit (2019) revealed large differences in the accuracy of the reconstruction of planar objects, notably for floors and walls. In striving at preserving accurate and sharp depth transitions, this network tends to be more sensitive to texture changes and high frequencies, yielding fragmented and falsely determined planes. The performance of PlaneNet (Liu et al., 2018), which focuses on the preservation of planar regions, strongly depends on a prior semantic segmentation of the input image. For each detected planar region in the segmentation step, the method estimates reasonable 3D plane parameters, but, however, pixel-accurate segmentations of planar regions often fails, which results in imprecise and fragmented 3D planes. Visual results showing residuals of projected depth maps and ground truth 3D planes are depicted in Fig. 13, which reveals different depth map characteristics based on the used methodology. 3D illustrations, displaying projected 3D points, fitted 3D plane and ground truth 3D plane for the scenes in Fig. 13 are shown in Fig. 14. Despite the considerably lower accuracy of fitted ground truth 3D planes in `NYU-v2`, planarity errors can principally be determined in the same manner, although, as already outlined in Section 4.6 inaccurate ground truth 3D planes limit the reliability of the derived results. The evaluations have shown that, similar to the global metrics, better overall results can be achieved, which is partly attributed to the slight domain shift between both dataset. However, similar to the results on `IBims-1`, a difference in the performance regarding the reconstruction of planar regions could be observed which results in a similar ranking of the investigated methods.

5.4. 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 $\epsilon_{\text{DBE}}^{\text{acc}}$ and $\epsilon_{\text{DBE}}^{\text{comp}}$, respectively, introduced in Section 3.2.3. We set the distance threshold of the *truncated chamfer distance* to $\theta = 10$ px, which also defines the upper bound of the accuracy and completeness errors. Quantitative results for all methods are listed in Table 4. Comparing the accuracy error of all methods, Liu et al. (2015) and Ramamonjisoa and Lepetit (2019) achieved best results in preserving actual depth boundaries, while other methods tended to produce smooth edges, and thus failed to reconstruct precise and complete depth transitions. This smoothing property and the small output resolution of some methods also affected the completeness error, resulting in missing edges expressed by larger values for $\epsilon_{\text{DBE}}^{\text{comp}}$. A comparison of depth boundaries from different methods can be seen in Fig. 15. Preserving sharp depth discontinuities is a main challenge using CNN-based methods, due to the intensive number of strided convolutions and spatial poolings, which reduce the output resolution, and, thus, local details of the image. However, methods that explicitly address this aspect have proven to enhance the reconstruction of object contours, which is also evident in the proposed DBE metrics.

5.5. 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, such as image refocusing and 3D cinematography. For the quantitative results listed in Table 4 we defined a reference plane at 3 m distance and computed the proportions of correct $\epsilon_{\text{DDE}}^{0}$, overestimated $\epsilon_{\text{DDE}}^{+}$, and underestimated $\epsilon_{\text{DDE}}^{-}$ depth values towards this plane according to the error definitions in Section 3.2.4. A visual illustration of correctly and falsely predicted depths is depicted in Fig. 16 and a comparison of different thresholds of d is shown in Fig. 17. The results show that, apart from the approach of Ramamonjisoa and Lepetit (2019), the methods tended to underestimate depth, although the amount of correctly estimated depth values almost

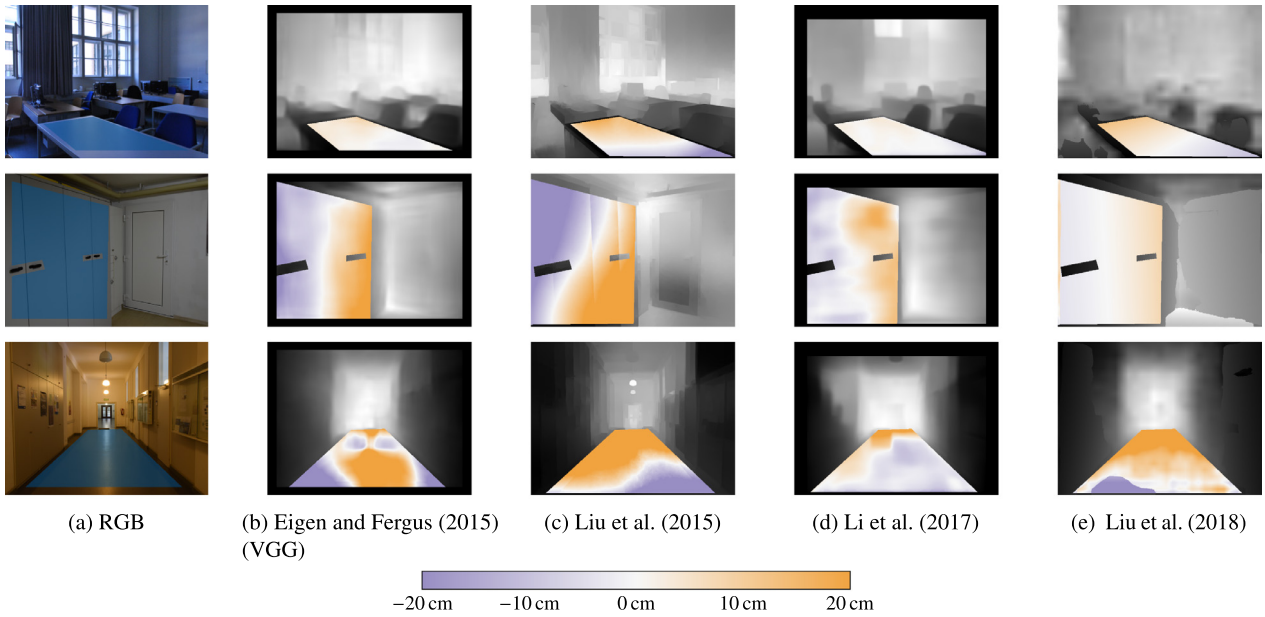


Fig. 13. Visual results after applying *planarity errors* (PEs) on different planar regions (top: table, middle: wall, bottom: floor). RGB with corresponding plane masks (■) (a). Predictions using different methodologies (b-e). Colors in the predictions correspond to orthogonal differences of projected depths towards the reference plane.

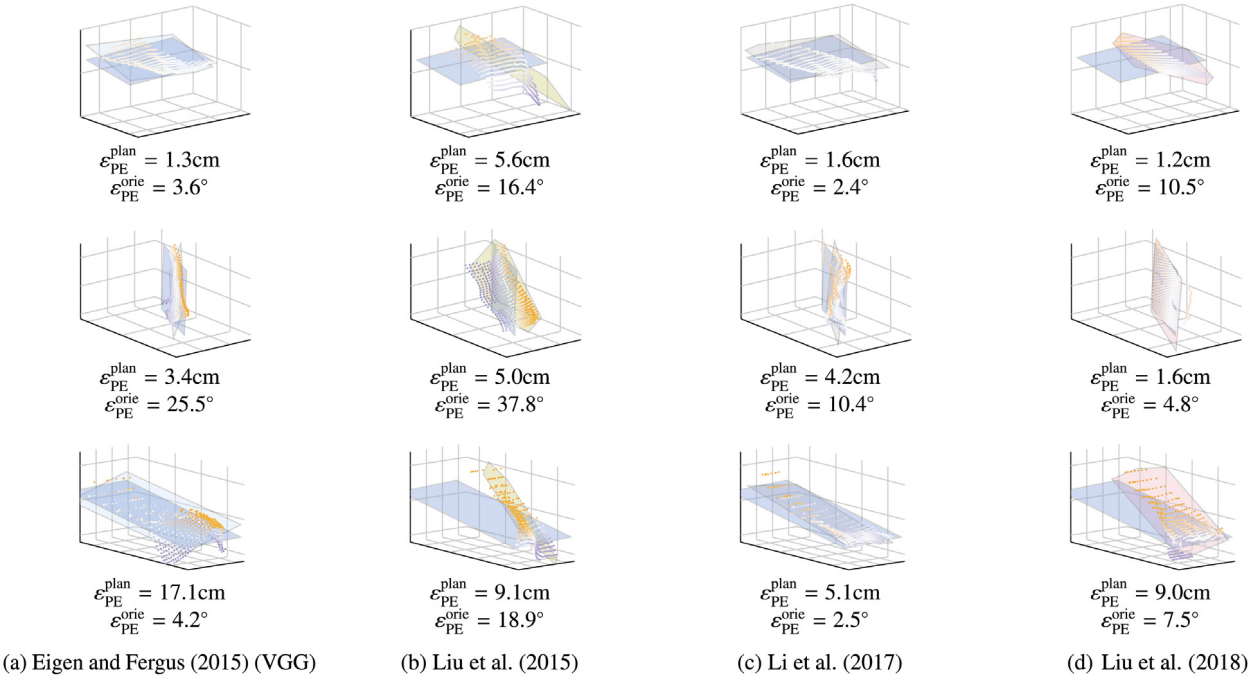


Fig. 14. 3D visualizations of predicted 3D planes from Fig. 13. Ground truth 3D planes (■), projected 3D points from predictions and fitted 3D planes. Color coding of the 3D points is similar to definitions in Fig. 13.

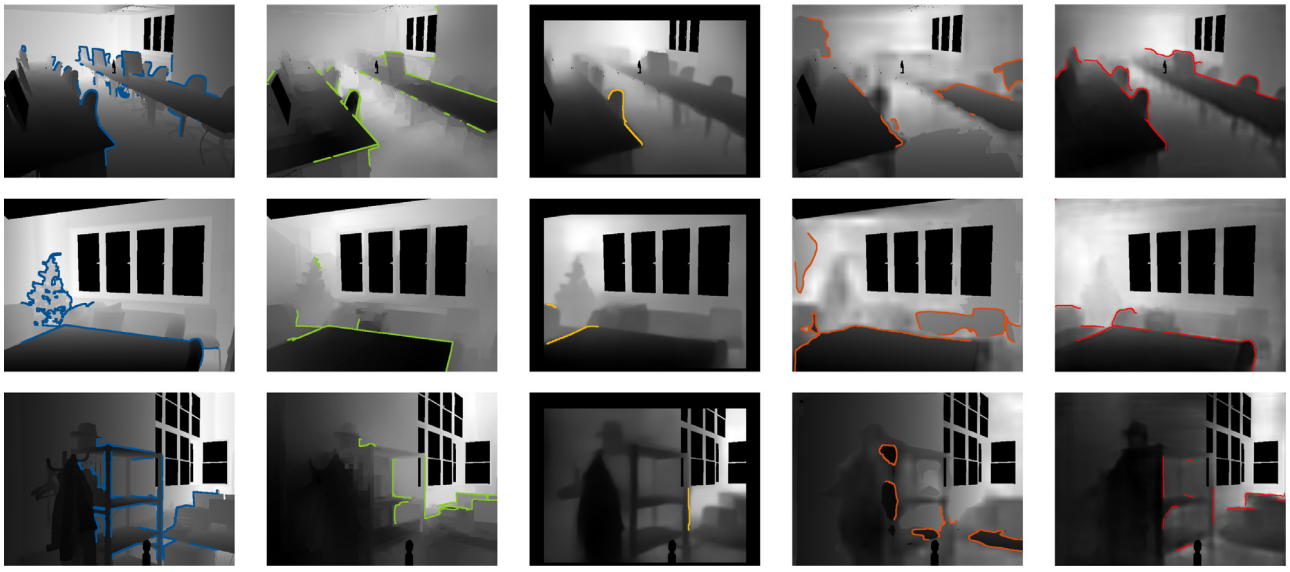
reaches 85 % for the methods of Li et al. (2015) and Ramamonjisoa and Lepetit (2019). For shorter distances up to 3 m, the methods of Eigen et al. (2014) and PlaneNet (Liu et al., 2018) tend to underestimate to a larger extend compared to other methods, while the method of Liu et al. (2015) rather overestimated short distances. It is worth noting that the method of Ramamonjisoa and Lepetit (2019) exhibited a largely well-balanced distribution of over- and underestimated depths.

6. Influence on the performance of SIDE methods

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, the planarity of predicted textured vertical surfaces, and the influence of different illumination in the scene.

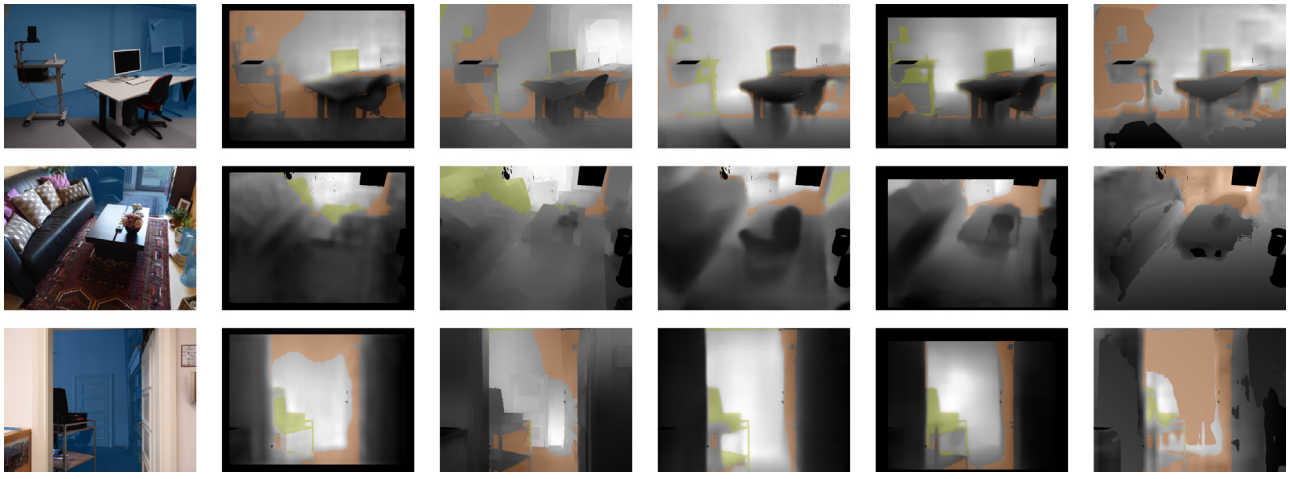
6.1. Augmentation

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



(a) Ground truth (b) Liu et al. (2015) (c) Li et al. (2017) (d) Liu et al. (2018) (e) Ramamonjisoa and Lepetit (2019)

Fig. 15. Visual results after applying *depth boundary errors* (DBEs) on `IBims-1`. Overlay of ground truth depth map with ground truth edge (—) (a) and depth map predictions with extracted edges (colored) using different methods (b-e).



(a) RGB (b) Eigen et al. (2014) (c) Liu et al. (2015) (d) Laina et al. (2016) (e) Li et al. (2017) (f) Liu et al. (2018)

Fig. 16. Visual results after applying *directed depth errors* (DDEs) on `IBims-1`. Ground truth depth plane at $d = 3\,\text{m}$ separating foreground from background (■) (a). Differences between ground truth and predictions (b-f). Color coded are depth values that are either estimated too short (■) or too far (■).

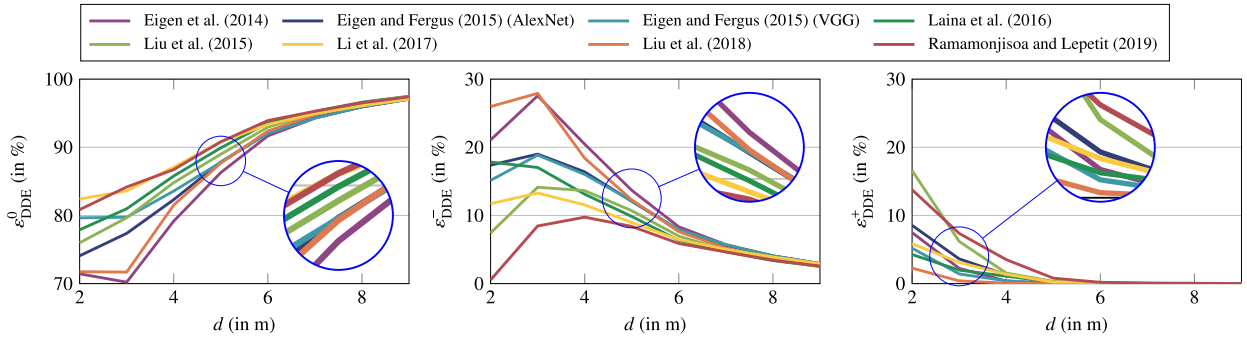


Fig. 17. *Directed depth errors* (DDEs) for different distances d of the virtual plane separating foreground and background. From left to right: Proportions of correct ($\varepsilon^0_{\text{DDE}}$), too-close ($\varepsilon^-_{\text{DDE}}$) and too-far ($\varepsilon^+_{\text{DDE}}$) predicted pixels for different methods.

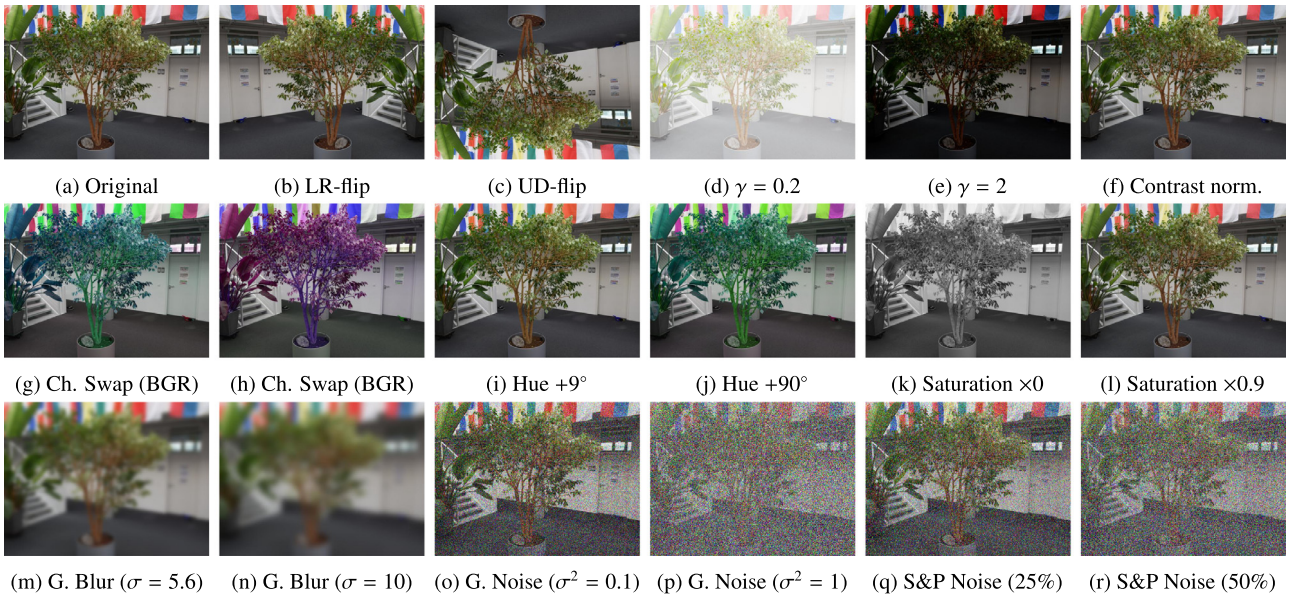


Fig. 18. Different geometric and radiometric augmentation samples applied to `IBims-1`.

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. Examples from this auxiliary dataset are shown in Fig. 18.

Table 5 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.*, grayscale input images which yielded an equal error to the reference. The results for blurring the input images with a Gaussian kernel of various standard deviations, as well as adding a different amount of Gaussian and salt and pepper noise to the input images are depicted in Fig. 19. Minor blurring did not change the results, as the examined methods considerably down-sample the input images and are thus robust to blurring up to a certain standard deviation. However, the performance of all methods starts to linearly decrease for blurring the image with $\sigma > 2$, whereby the methods of Eigen et al. (2014) and Liu et al. (2015) are more robust for larger blurring than the other methods. PlaneNet (Liu et al., 2018) could not handle blurring the image for standard deviations of the Gaussian distribution $\sigma > 2$ due to a failed vanishing point estimation.

The results for adding noise to the images, shown in Fig. Figs. 19b and 19c, give certain thresholds for the maximum tolerable amount of noise. All of the considered methods were able to cope with up to 10% of Salt and Pepper noise and Gaussian noise with variance of 0.01 until the quality of results decreased notably. The AlexNet version of Eigen and Fergus (2015) seems to be more robust to noise as opposed to the VGG version, which is, however, less sensitive to blurred input images. Again, the method of Liu et al. (2015) performed best on large noise levels, while PlaneNet (Liu et al., 2018) could not cope with a large amount of noise.

6.2. 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 Fig. 20, gradients seem 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. All of the examined networks respond to these patterns. However, this effect is less severe for Laina et al. (2016), which respond with only a constant offset to the alternating gradients in the pattern. Edges in the input also seem to influence the result as to be seen in Stripes and Boxes. Again, Laina et al. (2016) gave a constant offset, while the result of Liu et al. (2015) clearly contained artifacts of the superpixel approach, which is even more evident in Curves. Although `NYU-v2`, which served as training data for all methods, also contains such textured surfaces in terms of paintings and drawings on walls, the networks are unable to distinguish between intensity changes due to real depth discontinuities and solely texture. Further research in this field is needed in order to improve the applicability of SIDE in the fields of 3D room modeling or robot navigation.

6.3. Illumination

Illumination plays a significant role in recovering the 3D structure of a scene, especially for indoor scenarios where different types of natural and artificial illumination come together. This can be considered as a combination of under- and overexposure and intensity-based gradients on planar regions. As both effects were already discussed in sections Sections 6.1 and 6.2 separately, this experiment represents a real world scenario revealing these effects for current state-of-the-art methods. For this experiment we captured a static scene containing a table covering small objects in the foreground, as well as a white wall in the background separated by a floor lamp. We generated one ground truth depth map using a Kinect v1 and changed the scene illumination by various artificial lights, such as diffuse lighting from a floor lamp, and directional lighting from a spot appended on the floor lamp and a flashlight illuminates the scene from different viewpoints outside of the image. Depending on the illumination type, shadows cause strong gradients especially on the background wall. RGB images, predictions and quantitative results of the examined methods are visualized in Fig. 21. The results clearly show the impact of directional lighting of the spot creating depth changes according to the strong gradients

Table 5
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 (Reference).

Method	Reference	Geometric		Contrast			Ch. Swap		Hue		Saturation	
		LR	UD	$\gamma = 0.2$	$\gamma = 2$	Norm.	BGR	BRG	+9°	+90°	×0	×0.9
Eigen et al. (2014)	0.322	−0.003	0.087	0.056	0.015	0.000	0.017	0.018	0.001	0.021	0.003	−0.001
Eigen and Fergus (2015) (AlexNet)	0.301	0.006	0.147	0.105	0.023	−0.002	0.017	0.008	0.002	0.017	0.007	−0.001
Eigen and Fergus (2015) (VGG)	0.254	0.003	0.150	0.109	0.008	0.000	0.010	0.013	0.000	0.012	0.009	−0.001
Laina et al. (2016)	0.255	−0.004	0.161	0.078	0.022	−0.001	0.007	0.009	0.000	0.007	0.003	−0.001
Liu et al. (2015)	0.301	−0.004	0.079	0.021	0.011	−0.001	0.006	0.004	0.000	0.009	0.004	0.001
Li et al. (2017)	0.222	0.001	0.152	0.024	0.004	0.001	0.016	0.014	0.003	0.019	0.015	0.001
Liu et al. (2018)	0.287	0.003	0.204	0.069	0.025	−0.001	0.009	0.027	0.000	0.010	0.027	0.002
Ramamonjisoa and Lepetit (2019)	0.257	0.008	0.156	0.003	−0.003	0.000	0.012	0.010	−0.002	0.012	0.004	0.001

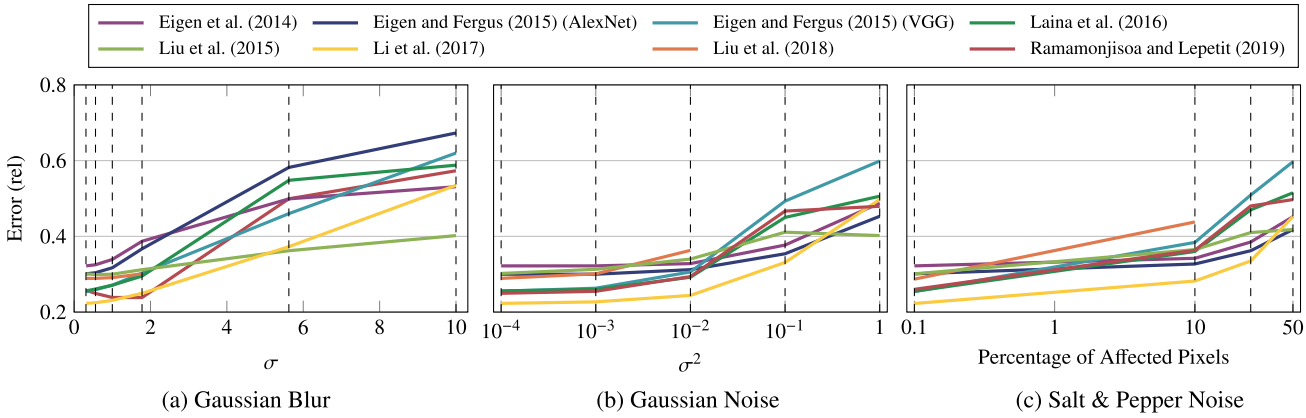


Fig. 19. Quality of SIDE results for different methods after applying different augmentations with increasing intensity on `IBims-1`. Vertical lines (-) correspond to discrete augmentation intensities.

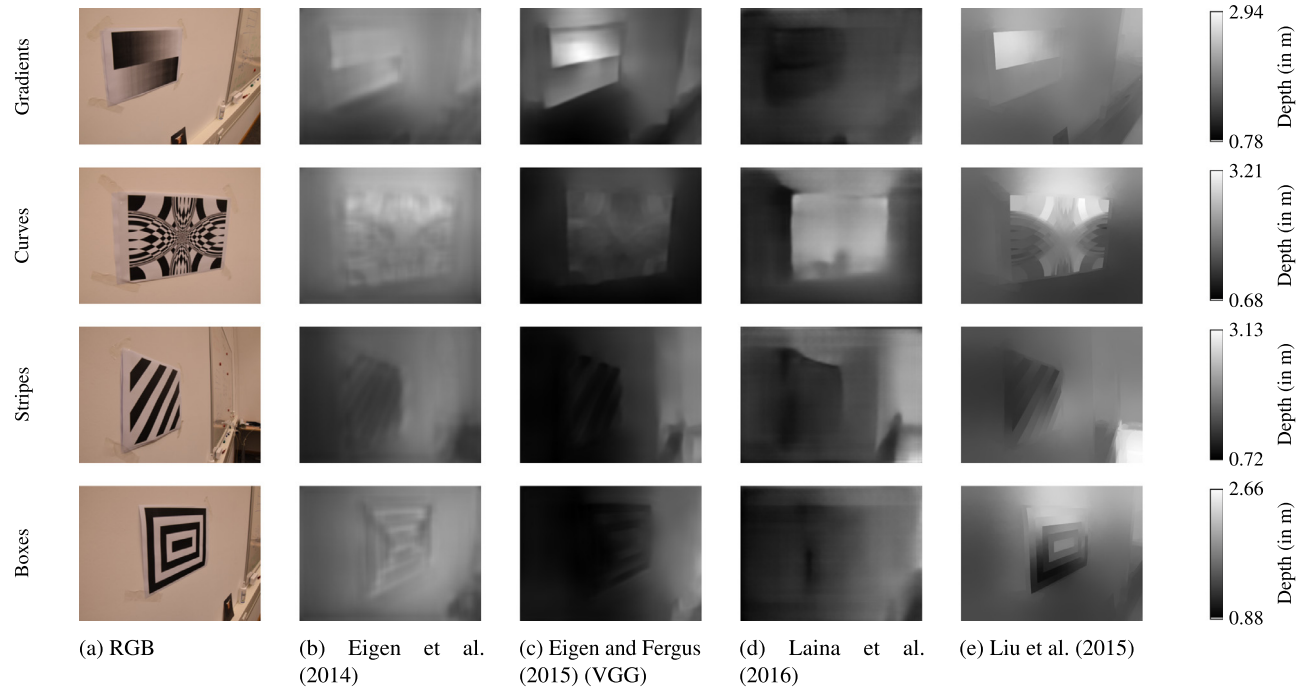


Fig. 20. Predictions for different printed samples from the Pattern dataset (Asuni and Giachetti, 2014) on a planar surface (rows). Predictions using different methods (b-e) of the input images (a). Predicted depth maps are color-coded according to the colormaps shown in the last column.

on the right side of the wall, while diffuse lighting did not influence the results notably. While comparable performances of the different methods – especially for diffuse lighting – can be observed when using the global error metric, more distinguishable results can be noted applying the *planarity errors*. As in the evaluation in Section 5, Liu et al. (2015) experiences difficulties in estimating the correct plane, while PlaneNet (Liu et al., 2018) successfully segmented the wall in each image and produces accurate 3D planes, although problems in the

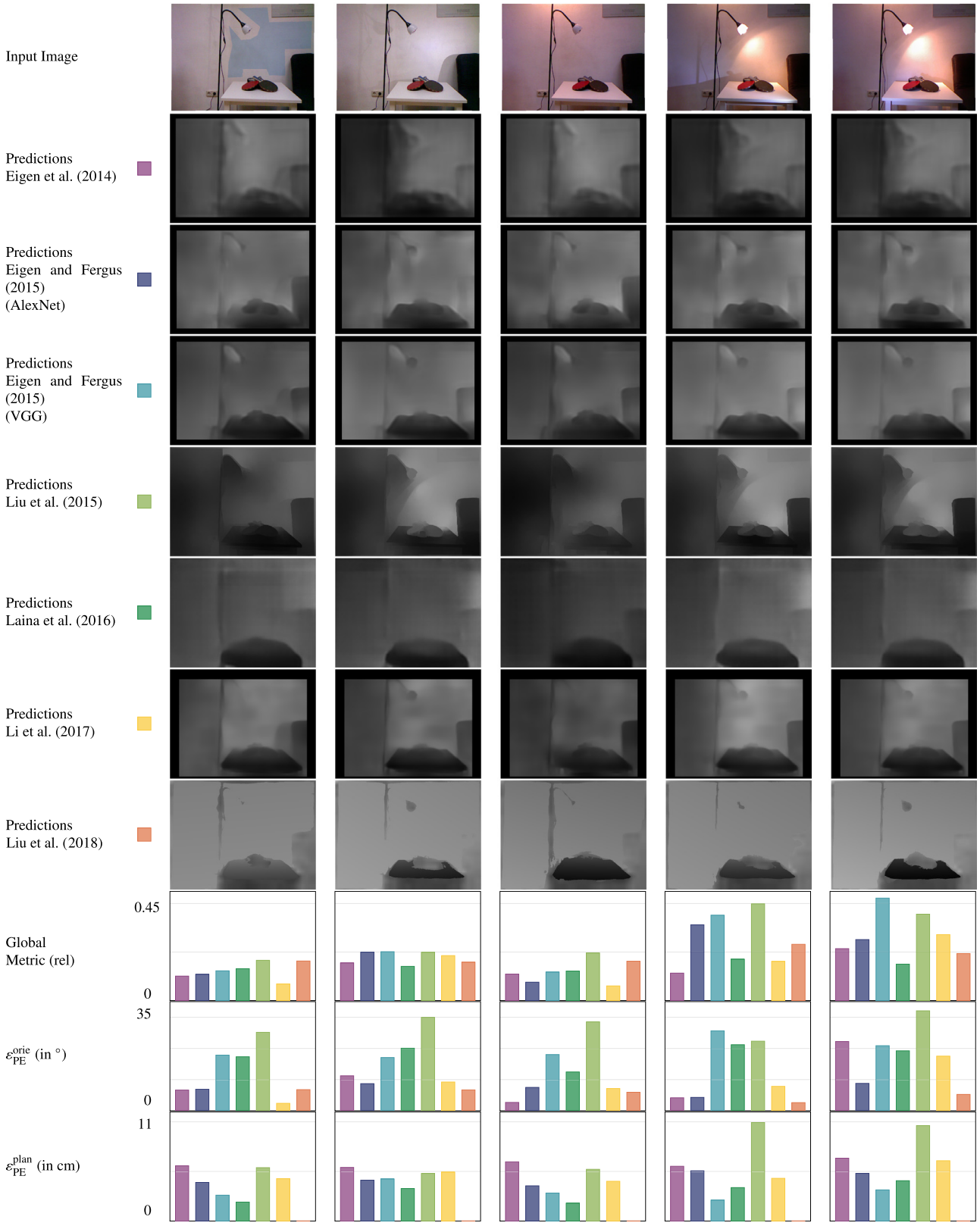


Fig. 21. Influence of different illumination on SIDE methods for a static scene. From top to bottom: Input RGB images, predictions using different SIDE methods, and errobars for global relative distance error and *planarity error* for annotated wall in the top-left image ( ).

predictions of objects on the tables can be noticed, resulting in larger errors for the global metric.

7. Conclusions

We presented a novel set of quality criteria and a new high-quality RGB-D dataset for the evaluation of SIDE methods. We pointed out, that established error metrics which are used to assess the quality of predicted depth maps do not consider meaningful geometric properties, such as the preservation of depth boundaries and planar regions, the depth consistency across the image, and the depth range in the image. In order to gradually establish SIDE methods in industrial applications, different properties of the derived depth maps are decisive which highly depend on the application field. For instance, 3D indoor room modeling emphasizes accurate and correct plane estimations rather the reconstruction of detailed and small-scaled furnishings. Developing realistic occlusion-aware augmented reality applications, on the other hand, requires the reconstruction of precise and sharp depth discontinuities in occluded contours (Ramamonjisoa and Lepetit, 2019). With the growing popularity of 3D movies, SIDE techniques are partially used to substitute the costly and time-consuming stereoscopic video recording process or the manual 2D-to −3D conversion of single RGB images to arrive stereo pairs (Xie et al., 2016). Since most 3D animations mainly consist of a few discrete depth layers, the consistency of depth estimates for certain depth ranges becomes an important issue. In the field of autonomous driving, the accuracy assessment of distance estimates is often considered as non-linear, since higher accuracies are required for objects close to the camera than for faraway objects (Liebel and Körner, 2019). Therefore, a distance-related assessment of the depth maps would provide valuable insights into the performance of the methods for different depth ranges.

As all of these application samples focus on different geometric properties of SIDE, measurable evaluation metrics are needed to compare and understand the performance of both existing and novel methodologies in this field. We elaborated simple, but geometrically interpretable error metrics for the mentioned properties above. Particularly, these are *distance-related error metrics*, *planarity errors*, *depth boundary errors*, and *directed depth errors*. Since these metrics require precise, dense, and noise-free RGB-D image pairs, existing RGB-D datasets cannot fully satisfy these high demands. For this reason, we introduced a new high-quality indoor RGB-D dataset, recorded with a custom acquisition setup combining a *laser scanner* and a *DSLR* camera to capture accurately aligned RGB-D image pairs. In our experiments, we were able to assess the quality of current state-of-the-art SIDE approaches w.r.t. to above mentioned properties, and unlike commonly used global metrics, our proposed set of quality criteria enabled us to unveil even subtle differences between the considered methods. In particular, our experiments have shown that the prediction of planar surfaces, which is crucial for many reconstruction applications, is lacking accuracy and CNN-based methods tend to produce smooth predictions resulting in blurry or vanishing depth boundaries. Although new methods that tackle specific aspects of the analyzed properties have been proposed recently, they still struggle to find a good trade-off between these aspects. Intuitively, a method that is designed and trained to predict sharp edges at depth discontinuities based on a single image, such as Sharpnet (Ramamonjisoa and Lepetit, 2019), tends to be sensitive to texture changes. Hence, a drop in the planarity metrics could be observed. Detecting planar regions in images and accurately predicting continuous depth values for such areas, as proposed in the PlaneNet approach of Liu et al. (2018), on the other hand, comes at the cost of disregarding finer details in favor of dominant planes. Our experiments showed that, again, the increased performance with respect to the targeted property is opposed by notable shortcomings in other aspects, most prominently the detection of edges. Additional experiments were conducted to test the robustness of the methods in terms of geometrical and radiometrical distortions, in the presence of textured planar surfaces and under varying lighting conditions. The results have proven a high robustness to minor blurring or noising of the input image, as well as to radiometrical changes. On the other hand, gradients and sharp intensity changes of planar objects, either caused by texture or illumination, can easily jar the methods in producing large depth changes. 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. Together with our new proposed error metrics, it serves as an independent evaluation protocol for indoor depth prediction and helps to improve future developments in this field.

Acknowledgments

This research was funded by the German Research Foundation (DFG), Germany for Tobias Koch and the Federal Ministry of Transport and Digital Infrastructure (BMVI), Germany 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

Ackermann, J., Goesele, M., et al., 2015. A survey of photometric stereo techniques. Found. Trends Comput. Graph. Vis. 9 (3–4), 149–254.

Anwar, S., Hayder, Z., Porikli, F., 2017. Depth estimation and blur removal from a single out-of-focus image. In: BMVC.

Armeni, I., Sax, S., Zamir, A.R., Savarese, S., 2017. Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105.

Asuni, N., Giachetti, A., 2014. 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, pp. 63–70. http://dx.doi.org/10.2312/stag.20141242.

Baig, M.H., Torresani, L., 2016. Coupled depth learning. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, pp. 1–10.

Camplani, M., Salgado, L., 2014. Background foreground segmentation with rgb-d kinect data: An efficient combination of classifiers. J. Vis. Commun. Image Represent. 25 (1), 122–136.

Chakrabarti, A., Shao, J., Shakhnarovich, G., 2016. Depth from a single image by harmonizing overcomplete local network predictions, in: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 2658–2666.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y., 2017. Matterport3d: Learning from rgb-d data in indoor environments. In: International Conference on 3D Vision (3DV). pp. 667–676. http://dx.doi.org/10.1109/3DV.2017.00081.

Chen, W., Fu, Z., Yang, D., Deng, J., 2016. Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems. pp. 730–738.

Choi, S., Min, D., Ham, B., Kim, Y., Oh, C., Sohn, K., 2015. Depth analogy: Data-driven approach for single image depth estimation using gradient samples. IEEE Trans. Image Process. 24 (12), 5953–5966.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2.

Devernay, F., Beardsley, P., 2010. Stereoscopic cinema. In: Image and Geometry Processing for 3-D Cinematography. Springer, pp. 11–51.

Dhamo, H., Tateno, K., Laina, I., Navab, N., Tombari, F., 2019. Peeking behind objects: Layered depth prediction from a single image. Pattern Recognit. Lett. 125, 333–340.

van Doorn, A.J., Koenderink, J.J., Wagemans, J., 2011. Light fields and shape from shading. J. Vis. 11 (3), 21.1–21.21. http://dx.doi.org/10.1167/11.3.21, arXiv:/data/journals/jov/933483/jov-11-3-21.pdf.

Eigen, D., Fergus, R., 2015. 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, http://dx.doi.org/10.1109/ICCV.2015.304.

Eigen, D., Puhrsch, C., Fergus, R., 2014. 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,

Favaro, P., Soatto, S., 2005. A geometric approach to shape from defocus. IEEE Trans. Pattern Anal. Mach. Intell. 27 (3), 406–417. http://dx.doi.org/10.1109/TPAMI.2005.43.

Felzenszwalb, P.F., Huttenlocher, D.P., 2006. Efficient belief propagation for early vision. Int. J. Comput. Vis. 70 (1), 41–54.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2002–2011.

Furukawa, R., Sagawa, R., Kawasaki, H., 2017. Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface–. In: The IEEE International Conference on Computer Vision (ICCV). pp. 4640–4648.

Garg, R., Carneiro, G., Reid, I., 2016. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Proceedings of European Conference on Computer Vision (ECCV). Springer, pp. 740–756. http://dx.doi.org/10.1007/978-3-319-46484-8_45.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3354–3361.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, p. 7.

Guo, X., Li, H., Yi, S., Ren, J., Wang, X., 2018. Learning monocular depth by distilling cross-domain stereo networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 484–500.

Hane, C., Ladicky, L., Pollefeys, M., 2015. Direction matters: Depth estimation with a surface normal classifier, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 381–389.

Hao, Z., Li, Y., You, S., Lu, F., 2018. Detail preserving depth estimation from a single image using attention guided networks. In: 2018 International Conference on 3D Vision (3DV). IEEE, pp. 304–313.

Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press.

Hassner, T., Basri, R., 2006. Example based 3d reconstruction from single 2d images. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). IEEE, p. 15.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heber, S., Pock, T., 2016. Convolutional networks for shape from light field, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3746–3754, http://dx.doi.org/10.1109/CVPR.2016.407.

Heo, M., Lee, J., Kim, K.R., Kim, H.U., Kim, C.S., 2018. Monocular depth estimation using whole strip masking and reliability-based refinement, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 36–51.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2. IEEE, pp. 807–814.

Hoiem, D., Efros, A.A., Hebert, M., 2007. Recovering surface layout from an image. Int. J. Comput. Vis. 75 (1), 151–172.

Horn, B.K.P., 1970. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical Report, MIT, Cambridge, MA, USA, URL https://dspace.mit.edu/handle/1721.1/6885.

Hu, J., Ozay, M., Zhang, Y., Okatani, T., 2019. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1043–1051.

Izadinia, H., Shan, Q., Seitz, S.M., 2017. Im2cad, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5134–5143.

Kadambi, A., Taamazyan, V., Shi, B., Raskar, R., 2015. Polarized 3D: High-quality depth sensing with polarization cues, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3370–3378, http://dx.doi.org/10.1109/ICCV.2015.385.

Karsch, K., Liu, C., Kang, S.B., 2014. Depth transfer: Depth extraction from video using non-parametric sampling. IEEE Trans. Pattern Anal. Mach. Intell. 36 (11), 2144–2158. http://dx.doi.org/10.1109/tpami.2014.2316835.

Kim, S., Park, K., Sohn, K., Lin, S., 2016. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: Proceedings of European Conference on Computer Vision (ECCV). Springer, pp. 143–159.

Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. Graph. 36 (4).

Koch, T., Liebel, L., Fraundorfer, F., Körner, M., 2018. Evaluation of CNN-based single-image depth estimation methods, in: Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS), pp. 331–348, https://doi.org/10.1007/978-3-030-11015-4_25.

Kolmogorov, V., Zabih, R., 2001. Computing visual correspondence with occlusions using graph cuts, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 2, pp. 508–515, http://dx.doi.org/10.1109/ICCV.2001.937668.

Kong, N., Black, M.J., 2015. Intrinsic depth: Improving depth transfer with intrinsic images, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3514–3522.

Konrad, J., Brown, G., Wang, M., Ishwar, P., Wu, C., Mukherjee, D., 2012. Automatic 2d-to-3d image conversion using 3d examples from the internet. In: Stereoscopic Displays and Applications XXIII, Vol. 8288. International Society for Optics and Photonics, p. 82880F.

Konrad, J., Wang, M., Ishwar, P., Wu, C., Mukherjee, D., 2013. Learning-based, automatic 2d-to-3d image and video conversion. IEEE Trans. Image Process. 22 (9), 3485–3496.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

Kuznietsov, Y., Stückler, J., Leibe, B., 2017. Semi-supervised deep learning for monocular depth map prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6647–6655.

Ladicky, L., Shi, J., Pollefeys, M., 2014. Pulling things out of perspective, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 89–96.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In: Fourth International Conference on 3D Vision (3DV). IEEE, pp. 239–248.

Lee, J.H., Heo, M., Kim, K.R., Kim, C.S., 2018. Single-image depth estimation based on fourier domain analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 330–339.

Levin, A., Lischinski, D., Weiss, Y., 2004. Colorization using optimization. In: ACM Transactions on Graphics (Tog). ACM, pp. 689–694.

Li, J., Klein, R., Yao, A., 2017. A two-streamed network for estimating fine-scaled depth maps from single rgb images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3372–3380.

Li, X., Qin, H., Wang, Y., Zhang, Y., Dai, Q., 2014. Dept: depth estimation by parameter transfer for single still images. In: Asian Conference on Computer Vision. Springer, pp. 45–58.

Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M., 2018. 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.

Li, Z., Snavely, N., 2018. MegaDepth: Learning single-view depth prediction from internet photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050.

Liebel, L., Körner, M., 2019. Multidepth: Single-image depth estimation via multi-task regression and classification. arXiv preprint arXiv:1907.11111.

Liu, B., Gould, S., Koller, D., 2010. Single image depth estimation from predicted semantic labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1253–1260.

Liu, M., Salzmann, M., He, X., 2014. Discrete-continuous depth estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 716–723.

Liu, F., Shen, C., Lin, G., 2015. 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.

Liu, F., Shen, C., Lin, G., Reid, I., 2016. Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Anal. Mach. Intell. 38 (10), 2024–2039.

Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y., 2018. PlaneNet: Piece-wise planar reconstruction from a single RGB image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2579–2588.

Liu, C., Yuen, J., Torralba, A., 2011. Sift flow: Dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 978–994.

Mancini, M., Costante, G., Valigi, P., Ciarfuglia, T.A., 2018. J-MOD 2: joint monocular obstacle detection and depth estimation. IEEE Robot. Lett. 3 (3), 1490–1497.

McCormac, J., Handa, A., Leutenegger, S., Davison, A.J., 2017. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation. In: Proceedings of the International Conference on Computer Vision (ICCV), vol. 4, pp. 2697–2706.

Moreno-Noguer, F., Lepetit, V., Fua, P., 2007. Accurate non-iterative o (n) solution to the pnp problem, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–8.

Nayar, S.K., Narasimhan, S.G., 1999. Vision in bad weather. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR), vol. 2, IEEE, pp. 820–827.

Ngo, T.T., Nagahara, H., Taniguchi, R.i., 2015. Shape and light directions from shading and polarization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2310–2318, http://dx.doi.org/10.1109/CVPR.2015.7298844.

Occipital, I., 2016. Structure sensor-3d scanning, augmented reality, and more for mobile devices.

Phan, R., Androutsos, D., 2013. Robust semi-automatic depth map generation in unconstrained images and video sequences for 2d to stereoscopic 3d conversion. IEEE Trans. Multimed. 16 (1), 122–136.

Ramamonjisoa, M., Lepetit, V., 2019. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. arXiv preprint arXiv:1905.08598.

Ranftl, R., Vineet, V., Chen, Q., Koltun, V., 2016. Dense monocular depth estimation in complex dynamic scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4058–4066.

Ranftl, R., Vineet, V., Chen, Q., Koltun, V., 2016. Dense monocular depth estimation in complex dynamic scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4058–4066.

Roy, A., Todorovic, S., 2016. Monocular depth estimation using neural regression forest, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5506–5514, http://dx.doi.org/10.1109/cvpr.2016.594.

Saxena, A., Chung, S.H., Ng, A.Y., 2006. Learning depth from single monocular images. In: Advances in Neural Information Processing Systems. pp. 1161–1168.

Saxena, A., Chung, S.H., Ng, A.Y., 2008. 3-d depth reconstruction from a single still image. Int. J. Comput. Vis. 76 (1), 53–69.

Saxena, A., Sun, M., Ng, A.Y., 2009. Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 31 (5), 824–840.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. 47 (1–3), 7–42.

Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. 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), pp. 2538–2547.

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 519–528, http://dx.doi.org/10.1109/CVPR.2006.19.

Shi, J., Tao, X., Xu, L., Jia, J., 2015. Break ames room illusion: depth from general single images. ACM Trans. Graph. 34 (6), 225.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. In: Proceedings of European Conference on Computer Vision (ECCV). Springer, pp. 746–760.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition.

Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T., 2017. Semantic scene completion from a single depth image. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, pp. 190–198.

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. 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.

Su, C.C., Cormack, L.K., Bovik, A.C., 2017. Bayesian depth estimation from monocular natural images. J. Vis. 17 (5), 22.

Suwajanakorn, S., Hernandez, C., Seitz, S.M., 2015. Depth from Focus with Your Mobile Phone, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3497–3506, http://dx.doi.org/10.1109/CVPR.2015.7298972.

Szeliski, R., 2010. Computer vision: algorithms and applications. Springer Science & Business Media.

Torralba, A., Oliva, A., 2002. Depth estimation from image structure. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 24 (9), 1226–1238.

Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T., 2017. Demon: Depth and motion network for learning monocular stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 5, pp. 5038–5047.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L., 2015. 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.

Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., Yuille, A.L., 2016. Surge: Surface regularized geometry estimation from a single image, in: Advances in Neural Information Processing Systems, pp. 172–180.

Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z., 2018. Monocular relative depth perception with web stereo data supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 311–320.

Xie, J., Girshick, R., Farhadi, A., 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In: Proceedings of European Conference on Computer Vision (ECCV). Springer, pp. 842–857.

Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N., 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 161–169.

Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E., 2018. Structured attention guided convolutional neural fields for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3917–3925.

Yang, F., Zhou, Z., 2018. Recovering 3d planes from a single image via convolutional neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 85–100.

Yin, Z., Shi, J., 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2.

Yoon, K.J., Kweon, I.S., 2006. Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal. Mach. Intell. (4), 650–656.

You, X., Li, Q., Tao, D., Ou, W., Gong, M., 2014. Local metric learning for exemplar-based object detection. IEEE Trans. Circuits Syst. Video Technol. 24 (8), 1265–1276.

Zennaro, S., Munaro, M., Milani, S., Zanuttigh, P., Bernardi, A., Ghidoni, S., Menegatti, E., 2015. Performance evaluation of the 1st and 2nd generation kinect for multimedia applications. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on. IEEE, pp. 1–6.

Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I., 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 340–349.

Zhang, R., Tsai, P.-S., Cryer, J.E., Shah, M., 1999. Shape from shading: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 21 (8), 690–706. http://dx.doi.org/10.1109/34.784284.

Zheng, C., Cham, T.J., Cai, J., 2018. T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 798–814.

Zhou, T., Brown, M., Snavely, N., Lowe, D.G., 2017. Unsupervised learning of depth and ego-motion from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6612–6619.

Zhuo, W., Salzmann, M., He, X., Liu, M., 2015. Indoor scene structure analysis for single image depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 614–622, http://dx.doi.org/10.1109/cvpr.2015.7298660.

Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P., 2018. Omnidepth: Dense depth estimation for indoors spherical panoramas.. In: The European Conference on Computer Vision (ECCV). pp. 453–471.

Zoran, D., Isola, P., Krishnan, D., Freeman, W.T., 2015. Learning ordinal relationships for mid-level vision, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 388–396.