Open access • Proceedings Article • DOI:10.1109/ICRA.2017.7989590

# Parse geometry from a line: Monocular depth estimation with partial laser observation — Source link

Yiyi Liao, Lichao Huang, Yue Wang, Sarath Kodagoda ...+2 more authors

**Institutions:** Zhejiang University, University of Technology, Sydney

**Published on:** 01 May 2017 - International Conference on Robotics and Automation

**Topics:** Monocular and Obstacle avoidance

# Parse Geometry from a Line: Monocular Depth Estimation with Partial Laser Observation

Yiyi Liao[1], Lichao Huang[2], Yue Wang[1], Sarath Kodagoda[3], Yinan Yu[2], Yong Liu[1]

***Abstract*— Many standard robotic platforms are equipped with at least a fixed 2D laser range finder and a monocular camera. Although those platforms do not have sensors for 3D depth sensing capability, knowledge of full geometry is an essential part in many robotics activities. Therefore, recently, there is an increasing interest in depth estimation using monocular images, of which the estimated depth might be unreliable in robotics applications as this task is inherently ambiguous. In this paper, we have attempted to improve the precision of monocular depth estimation by introducing 2D planar observation from the remaining laser range finder without extra cost. Specifically, we construct a dense reference map from the sparse laser range data, redefining the depth estimation task as estimating the distance between the real and the reference depth. To solve the problem, we construct a novel residual of residual neural network, and tightly combine the classification and regression losses for continuous depth estimation. Experimental results suggest that our method achieves considerable promotion compared to the state-of-the-art methods on both NYUD2 and KITTI, validating the effectiveness of our method on leveraging the additional sensory information. We further demonstrate the potential usage of our method in obstacle avoidance where our methodology provides comprehensive depth information compared to the solution using monocular camera or 2D laser range finder alone.**

## I. INTRODUCTION

Depth information plays an important role in daily lives of human, and is also a valuable cue in computer vision and robotics tasks. Many research works have demonstrated the benefit of introducing the depth information for tasks such as object recognition and scene understanding [1]–[4]. Recently, some researchers have opted to use monocular cameras to estimate the depths because of its inherent practical value. Monocular depth estimation is particularly challenging as it is a well known ill-posed problem. It is non trivial due to the vast amount of monocular depth cues, such as object sizes, texture gradients and overlaps needed for such depth estimations, in addition with the global scale of the scene. Thanks to the development of the deep convolutional neural networks over the recent years, remarkable advances are achieved on the task of monocular depth estimation [5]–[9]. A possible reason is that the monocular cues can be better modeled with the larger capacity of the deep network. However, the global scale of the scene remains a major ambiguity in monocular depth estimation [5]. With this ambiguity, the depth estimation result might be unreliable for robotics applications such as obstacle avoidance.

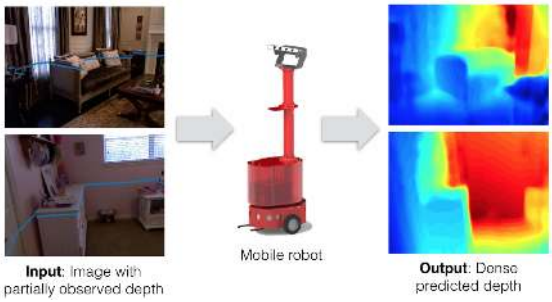



Fig. 1. Illustration of our proposed method. The input of our method is a single image and a planar of 2D laser range data, obtained from a monocular camera and a 2D laser range finder. The aim of this paper is to precisely estimate the dense depth of the full scene.

In this regard, a natural option to consider is to see whether the global ambiguity can be resolved using complementary sensory information. We want to exploit this idea by introducing limited direct depth observations to the monocular depth estimation task using a planar 2D laser range finder. With the availability of cheap 2D laser range finders, this option can be attractive in robotics applications in terms of accuracy and cost of sensing. Illustrative examples can be found in Figure 1. Ideally, the partially observed depth information can be employed to better estimate the global scale while the monocular image can be exploited for the relative depth estimation. To achieve this goal, we construct a novel convolutional neural network architecture to solve the partially observed depth estimation task.

For mobile robots, the proposed configuration for partial depth observation is very common. Many well-known robots such as Pioneer[1] and K5 Security Robot[2] are equipped with a camera and a 2D laser range finder. The 2D laser range finder is indispensable for navigation and obstacle avoidance on the mobile robot [10], [11]. We demonstrate that our method facilitates greater perception for obstacle avoidance compared to that of a single 2D laser range finder, as the latter has a very limited vertical field of view which might be insufficient to completely reflect the surrounding environments especially with voids. Therefore, our method is a painless extension for most mobile robots with no requirement of additional cost.

[1]Yiyi Liao, Yue Wang and Yong Liu are with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, 310027, China. Yong Liu is the corresponding author of this paper, e-mail: yongliu@iipc.zju.edu.cn.

[2]Lichao Huang and Yinan Yu are with the Horizon Robotics, China.

[3]Sarath Kodagoda is with the Centre for Autonomous Systems (CAS), The University of Technology, Sydney, Australia.

[1]http://www.mobilerobots.com/ResearchRobots/PioneerP3DX.aspx

[2]http://knightscope.com

The key task of this paper is to effectively leverage the limited and sparse 2D laser data for precisely estimation of the completed and dense depth. We formulate this problem as an end-to-end learning task based on a novel fully convolutional neural network. The contribution of this paper can be summarized as follows:

- We introduce the 2D laser range data to the task of monocular depth estimation by constructing a dense reference depth map from the sparse observation. With the dense reference map, the task of estimating the depth is redefined as estimating the residual depth between the real depth and the reference depth.
- To explicitly estimate the residual depth, we construct a novel network architecture named residual of residual neural network. Besides, the network combines both classification and regression losses for effectively estimating the continuous depth value.
- We conduct experiments on both indoor and outdoor environments and gain considerable promotion compared to the state-of-the-art monocular depth estimation methods, as well as another partially observed depth estimation method. We further demonstrate its potential usage in obstacle avoidance for mobile robots.

The remainder of the paper is organized as follow: Section II gives a review of related works. The methodology for solving the partially observed depth estimation task is given in Section III, and the experimental results are presented in Section IV. Finally, we conclude the paper in Section V.

## II. RELATED WORKS

In recent years, deep learning methods are intensively exploited in depth estimation using single images [5]–[9]. Deep networks are validated to be more effective compared to the conventional methods based on hand-crafted features and graphical models [12], [13]. Eigen et al. [5] firstly proposed to regress the depth value in the end-to-end framework using deep neural network. That work was extended to simultaneously estimate the depths, surface normals and semantic labels in their latter work [6]. Liu et al. [7] proposed to combine Conditional Random Field (CRF) and the deep neural network for depth estimation. The deep neural network learned the unary and pairwise potentials and the CRF was jointly optimized with the network. More recently, Laina et al. [8] and Cao et al. [9] tackled the depth estimation task based on the Residual Neural Network (ResNet) [14], which won the first place on the classification and detection competition tasks at ILSVRC and COCO in 2015. Specifically, Laina et al. [8] regressed the depth value using the fully convolutional ResNet, and a novel up-convolutional scheme was developed for fine-grained estimation. Instead of regression, Cao et al. [9] regarded the depth estimation as a classification task, where the estimation probability could be obtained for refinement using CRF. Taking the advantage of the deep ResNet, Laina et al. [8] and Cao et al. [9] set a new baseline in the depth estimation task. As can be seen, previous works indicate a trend of learning all the variations in depth estimation using one deeper model, which might be difficult for disambiguating the global scale. Differently, in this paper, we propose to take the advantage of an external but common sensor on mobile robots to get an relatively reliable estimation of the global scale, upon which the variation that need to be modeled by the network could be reduced. To solve this partially observed depth estimation task, we propose a novel architecture of fully convolutional network.

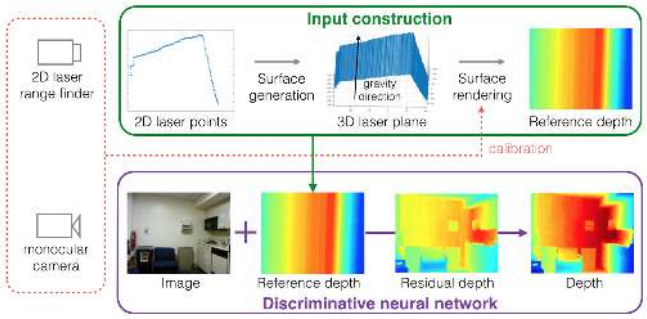



Fig. 2. Pipeline of the proposed method. A 3D surface is generated from the 2D laser scan along the gravity directly, which is then rendered to the image plane to generate a dense reference depth map. By combining the image and the reference depth, we estimate the depth based on a discriminative neural network, where the residual between the reference depth and the actual depth is explicitly estimated within the network.

In robotics scenario, there are also attempts for depth estimation with partially observed depths. A popular topic is the dense depth reconstruction with fusion of sparse 3D laser range data and the monocular image [15], [16]. As the 3D laser range data is obtained, this problem is usually formulated as an inpainting problem considering the measurement compatibility and the smoothness regularization. Both [15] and [16] made efforts on designing the regularization term and presented outstanding performances, where the former introduced a second order smoothness term and the latter searched optimal regularizer for different scenes. It is to be noted that the inpainting method requires the laser range data to cover most of the scene, which means it can only work with 3D laser range finder. With only a single planar view of 2D laser range data, the inpainting formulation is intractable due to severely insufficient information. In this paper, we state the partially observed task as a discriminative learning task, which can estimate the depth even with a planar view. A relatively similar work to this paper is [17]. The authors proposed to use a Multi-modal Auto-Encoder to impute the missing depth in the sparse depth map estimated by structure from motion. In this paper, we alternatively estimate the residual between the real depth and the reference depth, which is shown to be more effective in the experiments.

## III. METHOD

In this section, we present our novel methods for the partial observation task. The pipeline of our method is illustrated in Figure 2, which is composed of the input construction and the discriminative neural network.

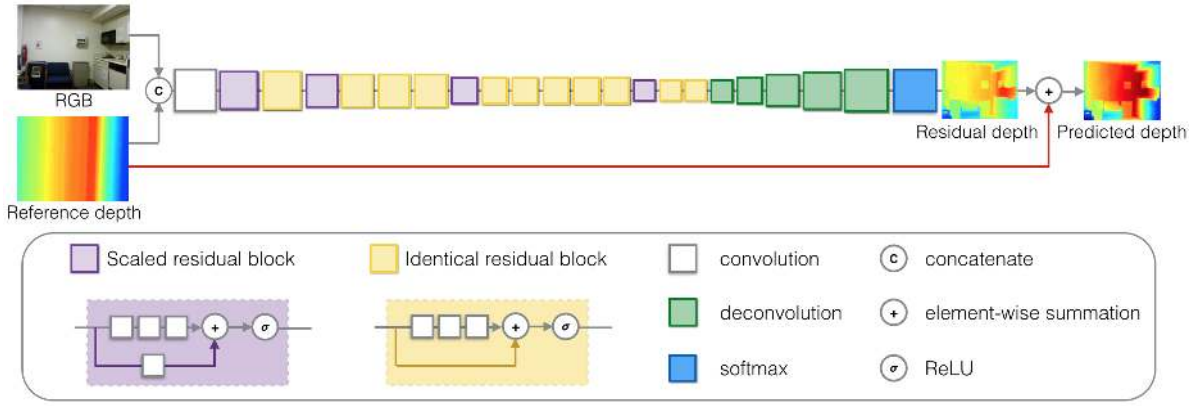



Fig. 3. Our residual of residual network architecture. The network is designed based on the ResNet-50, with 50 convolutional layers and 5 deconvolutional layers. We add a global identity skip to send the reference depth to the last feature layer, which is denoted as the red line in the figure. The global identity skip encourages the network to explicitly learn the residual depth. Best viewed in color.

### A. Construction of reference depth

Given the partially observed depth data with a 2D laser range finder, a naive idea of generating a sparse depth map is by associating the available laser range depths to pixels and rest of all the pixels are padded with zeros [17]. Intuitively, the network can hardly learn useful information from the sparse depth map generated using 2D laser range data due to the extremely sparse distribution. On the other hand, such sparse depth map is a mixture of two different categories of values. The zero value filled in the unknown region is the logical code to denote whether there is a valid depth value, while the depth value is a real number for the distance. Mixing the logical code and the depth value in the same map might confuse the network.

To avoid the problems mentioned above, we construct a dense reference map where every pixel is assigned with a depth value, which we name as "*reference depth*" map. The generation process is visualized in Figure 2. Firstly, median filtering is applied to the laser scan readings for smoothness, followed by interpolation between adjacent laser points. The linear interpolation is employed for imputing the missing depth values. Secondly, at each point in the imputed laser scan, we generate a line along the gravity direction in 3D space, resulting in a family of lines composing a surface vertical to the ground plane. Finally, by rendering this virtual surface to the image plane of the corresponding monocular camera, we can obtain the dense reference depth map. A related work to our reference depth map is Stixel [18], which compactly represents the scene as many vertical rectangular sticks. Their work validates the feasibility of our depth map construction pipeline in real situation.

After constructing the reference depth map, it is concatenated with the corresponding image as the input of the subsequent neural network for learning. Note that with a dense reference depth map as the input, the task of the depth estimator is transformed to *sculpting a depth value from the reference*, while it is originally formulated as *creating a depth value from the unknown*. This transformation significantly changes the declaration of the depth estimation problem, which we consider as one crucial reason for boosting the performance.

### B. Residual of residual network

In order to sculpt the depth from the reference map, we want to estimate the difference between the real depth and the reference map, which we formally denote as "*residual depth*". Here we use the residual neural network (ResNet) [14] as our network backbone because of its inherent design to learn the residual, as well as its superior performance on a wide range of computer vision tasks recently. It should be noted that, the residual in this paper, is assigned with exact physical concept, which reveals whether the actual depth is closer or further than the reference depth at each pixel. However, in the original ResNet, the network is actually learning a transformed residual since the input experiences nonlinear transformations during the feed-forward propagation. Thus we propose the "*residual of residual network*" to explicitly estimate the residual depth, which is particularly suited to the branch of partially observed estimation tasks.

Our network architecture is illustrated in Figure 3, where the main branch is a fully convolutional network extended from the standard ResNet-50. ResNet is composed of two kinds of residual blocks, i.e. the scaled residual blocks and the identical residual blocks. In both residual blocks, there are two branches, the lower one is an identity skip connection aiming at preserving the information of the block input, and the upper one consists of three convolutional layers which are encouraged to model the residual with respect to the input. Both kinds of blocks can be represented as

$$x_{l+1} = \sigma(F(x_l, W_l) + h(x_l)) \tag{1}$$

where $x_l$ is the input of the $l$th block, $W_l$ is the weights of the three convolutional layers of the $l$th block, $\sigma(x) = max(0, x)$ is the nonlinear ReLU layer. For the scaled residual blocks, $h(\cdot)$ is a convolutional layer to scale the feature maps, while in identical residual blocks $h(x)$ is an identity mapping. Considering the nonlinear ReLU layer, as $F(x_l, W_l)$ and $h(x_l)$ are not ensured to be positive, we have

the inequality

$$\sigma(F(x_l, W_l) + h(x_l)) \neq \sigma(F(x_l, W_l)) + \sigma(h(x_l)) \quad (2)$$

Due to the existence of the nonlinear mapping, the block does not exactly learning the residual $x_{l+1} - x_l$ between the input and the output, even though $h(\cdot)$ is the identity mapping, i.e. identical residual block. Consequently, the full network learns a transformed residual between the network input $x_0$ and the final network output $x_L$, rather than the residual depth $x_L - x_0$, where $x_0$ denotes the reference depth and $x_L$ denotes the actual depth in our model. In addition, the deconvolutional layers located between the residual blocks and the network output further transform the feature maps and interrupt our initial attempt for learning the residual depth.

To encourage the network to explicitly learn to sculpt the depth from the reference depth, we add a global identity skip as the red line in Figure 3 to directly send the reference depth to the last feature map before the output, which can be presented as $x_L = x_{L-1} + x_0$. Thus $x_{L-1}$ is enforced to explicitly learn the residual depth map $x_L - x_0$. Hence there are two categories of identity skips in our network, i.e. the local identity skips connecting each residual block and the global identity skip connecting the full network, that is the reason it is called the residual of residual network.

### C. Combination of classification and regression

Previous works usually formulate the depth estimation problem as a classification task or a regression task [5]–[9]. In this paper, we construct a loss function tightly combining the classification loss and the regression loss. Specifically, a softmax layer is added on top of the final deconvolution layer as visualized in Figure 3. For classification, all depth values are discretized into $K$ bins, where the center depth value of each bin is denoted as $k$. We set $K = 101$ in our experiments. Let us denote the input feature vector of the softmax layer as $f_i$, then the probability of the corresponding sample $i$ being assigned to the discretized depth $k$ is computed as

$$p_i^k = \frac{exp(f_i^T \theta_k)}{\sum_{k=1}^{K} exp(f_i^T \theta_k)} \quad (3)$$

Then the predicted depth is given as

$$\hat{y}_i = \arg\max_k p_i^k \quad (4)$$

As (4) can only provide a discretized depth estimation, we propose to calculate the predicted depth using the expected value as

$$\bar{y}_i = \sum_{k=1}^{K} p_i^k k \quad (5)$$

By calculating the expected value, we obtain a continuous depth estimation. The expectation is also more robust than the discretized value with the maximal probability. Furthermore, it is also more convenient for calculating the gradients with respect to the expected value (5) compared to the argmax value (4).

With the predicted probability, the softmax classification loss is given as

$$L_c = \sum_{i=1}^{M} \sum_{k=1}^{K} \delta([y_i] - k) \log(p_i^k) \quad (6)$$

where $M$ is the number of all samples, $y_i$ is the ground truth depth and $[y_i]$ is the center depth value of the discretized bin that $y_i$ falls in. $\delta(x) = 1$ when $x = 0$, otherwise $\delta(x) = 0$. For regression, we use L1 loss to generate a constant gradient even when the difference is small, which is formulated as

$$L_r = \sum_{i=1}^{M} |y_i - \bar{y}_i| \quad (7)$$

Then we combine the classification loss and the regression loss to train the network as

$$L = L_c + \alpha L_r \quad (8)$$

where $\alpha$ is the constant weight term. We set $\alpha = 1$ in our experiments.

When compared with individual classification or regression losses, our combination of both these two losses brings the following remarks:

- The classification loss alone cannot distinguish the difference across discretized bins while regression is able to provide larger penalty to the larger predictive errors.
- If the estimated depth falls into the correct bin, then the classification loss would vanish. The regression loss still works to eliminate the small loss within the bin, leading to a finer estimation.
- Compared to the solution with regression loss alone, our method can provide a probabilistic distribution. Furthermore, as the depth is computed as the expected value, it has a fixed range and thus is more robust compared to the direct regression.

### D. Estimation refinement

As shown in Figure 3, the network outputs the predicted depth by summing the reference depth and the residual depth, without additional trainable layers. We further refine the predicted depth by applying the median filtering. It can reduce the noises generated from the summation and slightly promote the estimation performance.

## IV. EXPERIMENTS

The experiments were preferably carried out on publicly available data sets for benchmarking and easier comparison. In this paper, we evaluate our method on the indoor dataset NYUD2 [1] and the outdoor dataset KITTI [19].

### A. Experimental setup

NYUD2 [1] is an indoor dataset collected using the Microsoft Kinect. It covers 464 scenes with 4 million raw image and depth pairs. We follow the official split to use 249 scenes for training and 215 scenes for testing. We sampled 50,000 images from the raw training data for training, where

the missing depth values were masked out during the training process. Test set includes 654 images with filled-in depth values, which is the same as the other monocular depth estimation methods [5]–[9]. We simulated a laser scan that was perpendicular to the gravity direction, with a fixed height above the ground plane. In our experiment, the height was set as 80cm. Since NYUD2 is collected using a hand-held Kinect, the camera pose varies a lot between different frames, leading to an uncertain gravity direction. Thus we follow Gupta. et al [20] to estimate the gravity direction, of which we observe the accuracy is acceptable in our experiments.

For the outdoor dataset KITTI [19], we use three scene categories ("City", "Residential" and "Road") in the raw data for training and testing, the same as Eigen et al. [5]. We sampled 5,000 images captured by the left color camera from 30 scenes for training, and 632 images from other 29 scenes for testing. With the relative small training set, the network is initialized with the weights learned from NYUD2. The ground truth depth was obtained with a Velodyne HDL-64E 3D laser scanner, where the missing depth was masked out for both training and testing. As the Velodyne laser scanner observed 64 laser scans in each frame, we simulated a 2D laser range finder by taking one of the laser scans as our partially observed data. The laser scan was selected to be within a fixed range of polar angle in the spherical coordinate, which was set as $88^\circ \pm 2^\circ$ in our experiment. As the sensors were fixed on a mobile car in the KITTI, the gravity direction was fixed for all frames and could be obtained from the offline calibration. It is to be noted that the gravity direction is also fixed in practical applications of both indoor and outdoor robots, and there is no requirement for the additional estimation of the gravity direction.

For the network configuration, both image and reference depth in NYUD2 were reshaped as $320 \times 256$, and the predicted size was $160 \times 128$. The input size of KITTI was set as $320 \times 96$, with output size $160 \times 48$. Though the depth is only available at the half bottom of the image in KITTI, we input the full image into the network for learning the context. Note that the predicted result was up-scaled to the original size for evaluation on both NYUD2 and KITTI.

We implemented our residual of residual network based on Caffe [21]. Following ResNet, we also used batch normalization for efficient convergence and the batchsize was set as 16. The loss was summed over all valid pixels and the learning rate $\eta = 10^{-6} \times 0.98^{\lfloor n/1000 \rfloor}$, $n$ denotes the iteration number. The model was trained for 80,000 iterations, which took about 33 hours on a Nvidia Titan X. Following [5], we used online data augmentation to avoid over-fitting. Specifically, the data augmentation steps include rotation, scaling, color transformation and flips.

We used the following standard metrics to evaluate our performance, where $y_i$ is the ground truth depth, $\bar{y}_i$ is the estimated depth value and $N$ is the number of total pixels:

- Root Mean Squared Error (rms): $\sqrt{\frac{1}{N}\sum_i (\bar{y}_i - y_i)^2}$
- Mean Absolute Relative Error (rel): $\frac{1}{N}\sum_i \frac{\bar{y}_i - y_i}{y_i}$
- Mean log10 Error (log10): $\frac{1}{N}|\sum_i \log_{10} \bar{y}_i - \log_{10} y_i|$
- Threshold $\delta_k$: percentage of $y_i$, s.t. $\max(\frac{\bar{y}_i}{y_i}, \frac{y_i}{\bar{y}_i}) < \delta^k$, $\delta = 1.25$ and $k = 1, 2, 3$.

### B. Model evaluation

To demonstrate the benefits gained from the input construction, the network design and the refinement process, we conducted comparison experiments on NYUD2 against some variants of our proposed method. Table I illustrates the comparison results. Note that all variants listed in the table were implemented based on ResNet-50 with the same network capacity. Specifically, we first performed the monocular depth estimation using only RGB image as a baseline. Then the laser information was added to the input as the reference depth map ("Ref."), without the residual of residual structure ("Res. of Res."). Furthermore, we added the global identity skip to explicitly estimate the residual depth. Finally, the refinement was performed to refine the predicted depth of our residual of residual network. All network architectures were trained in two different loss function settings, the classification loss alone ("C") or the combination of classification and regression ("C.+R."). As can be seen from the table,

- By comparing the results in the first two rows to the following rows, it can be seen that the performance is substantially promoted with our reference depth map as the additional input. It validates the effectiveness of constructing a dense reference map from the sparse partial observation as described in Section III-A, which redefines the depth estimation task as sculpting the depth from the reference.
- Comparison between the results with and without the "Res. of Res." structure demonstrates the superiority by adding the global identity skip. As we explained in Section III-B, the ResNet is inherently suited to our partially observed task as we want to learn the residual depth, thus the standard ResNet also achieves a relative good performance without the global identity skip. By adding the global identity skip to preserve the exact physical concept of the residual, our residual of residual network further boosts the estimation performance.
- By comparing the performances with classification loss alone and with combination of classification and regression, it can be seen that the combination raises the performances in all model variants. It demonstrates the benefit of our loss design as introduced in Section III-C.
- The refinement introduced in Section III-D further brings slight improvement to the estimation accuracy.

### C. Comparison with the state-of-the-art

In Table II, we compared with the state-of-the-art depth estimation methods using our best model suggested in Table I. We conducted experiments on both NYUD2 and KITTI to validate the generalization ability of the proposed method. Quantitative results in Table II illustrates that the depth

TABLE I

MODEL EVALUATION ON NYUD2.

| Input | Res. of Res. | Loss | Refined | Error (*lower is better*) | | | Accuracy (*higher is better*) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | rms | rel | log10 | $\delta_1$ | $\delta_2$ | $\delta_3$ |
| RGB | – | C. | – | 0.642 | 0.184 | 0.071 | 76.2 | 92.7 | 97.4 |
| RGB | – | C.+R. | – | 0.617 | 0.173 | 0.068 | 77.2 | 93.8 | 97.8 |
| RGB + Ref. | No | C. | – | 0.537 | 0.124 | 0.051 | 86.2 | 95.1 | 97.9 |
| RGB + Ref. | No | C.+R. | – | 0.507 | 0.126 | 0.050 | 86.3 | 95.7 | 98.4 |
| RGB + Ref. | Yes | C. | No | 0.480 | 0.108 | 0.045 | 87.0 | 95.8 | 98.5 |
| RGB + Ref. | Yes | C.+R. | No | 0.451 | 0.106 | 0.044 | 87.4 | 96.2 | 98.8 |
| RGB + Ref. | Yes | C.+R. | Yes | 0.442 | 0.104 | 0.043 | 87.8 | 96.4 | 98.9 |

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART ON NYUD2 AND KITTI.

| Dataset | Method | Error (*lower is better*) | | | Accuracy (*higher is better*) | | |
|---|---|---|---|---|---|---|---|
| | | rms | rel | log10 | $\delta_1$ | $\delta_2$ | $\delta_3$ |
| NYUD2 | Liu et al. [7] | 0.824 | 0.230 | 0.095 | 61.4 | 88.3 | 97.1 |
| | Eigen et al. [5] | 0.907 | 0.215 | – | 61.1 | 88.7 | 97.1 |
| | Eigen et al. [6] | 0.641 | 0.158 | – | 76.9 | 95.0 | 98.8 |
| | Cao et al. [9] | 0.645 | 0.150 | 0.065 | 79.1 | 95.2 | 98.6 |
| | Laina et al. [8] | 0.583 | 0.129 | 0.056 | 80.1 | 95.0 | 98.6 |
| | Ours | **0.442** | **0.104** | **0.043** | **87.8** | **96.4** | **98.9** |
| KITTI | Saxena et al. [13] | 8.734 | 0.280 | – | 60.1 | 82.0 | 92.6 |
| | Eigen et al. [5] | 7.156 | 0.190 | – | 69.2 | 89.9 | 96.7 |
| | Mancini et al. [22] | 7.508 | – | – | 31.8 | 61.7 | 81.3 |
| | Cadena et al. [17] | 6.960 | 0.251 | – | 61.0 | 83.8 | 93.0 |
| | Ours | **4.500** | **0.113** | **0.049** | **87.4** | **96.0** | **98.4** |

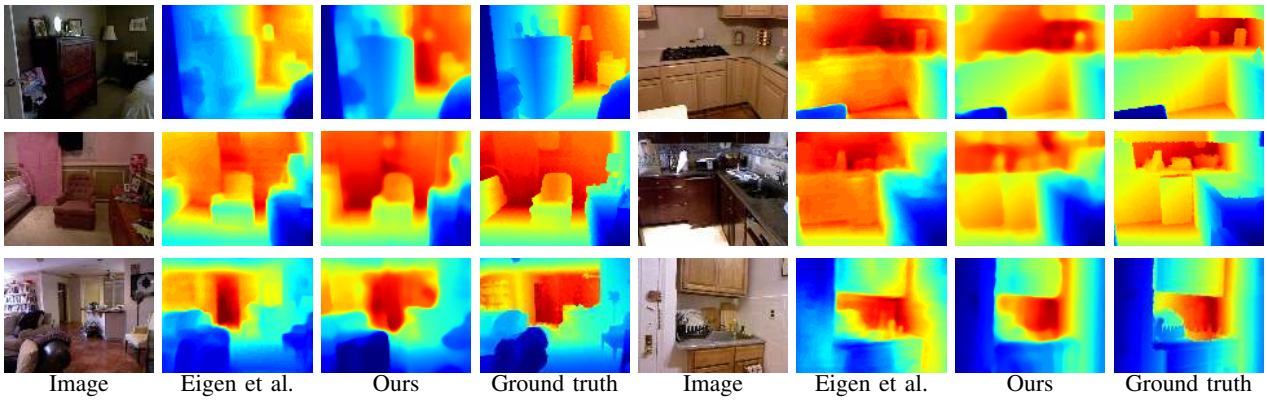



Fig. 4. Comparison on NYUD2. For each test image, the corresponding depth estimated by Eigen et al. [6], our method and the ground truth are given from left to right. Red denotes far and blue denotes close in the depth maps. It can be seen that the estimation of Eigen et al. [6] usually has a bias due to the ambiguity of the global scale. Best viewed in color.

estimation accuracy is substantially promoted by adding a single planar view of laser range data, validating the effectiveness of resolving the scale ambiguity with additional sensory information. It is to be noted that Cadena et al. [17] also tackled the depth estimation task with partially observed depth. They proposed to reconstruct the depth depth from a sparse depth map obtained based on structure from motion, while we formulate the partially observed estimation task as a residual learning task and achieves a considerable promotion.

Corresponding qualitative comparisons are given in Figure 4 and Figure 5. For NYUD2, we compared with Eigen et al. [6] and the ground truth in Figure 4. As can be seen, taking the advantage of a single planar of laser data, our methods parse the scenes better with a more accurate estimation of the global scale. Figure 5 illustrates the comparison between our method and the ground truth obtained from the Velodyne laser range finder. Images and depth maps are cropped to show the region with laser observations. It is

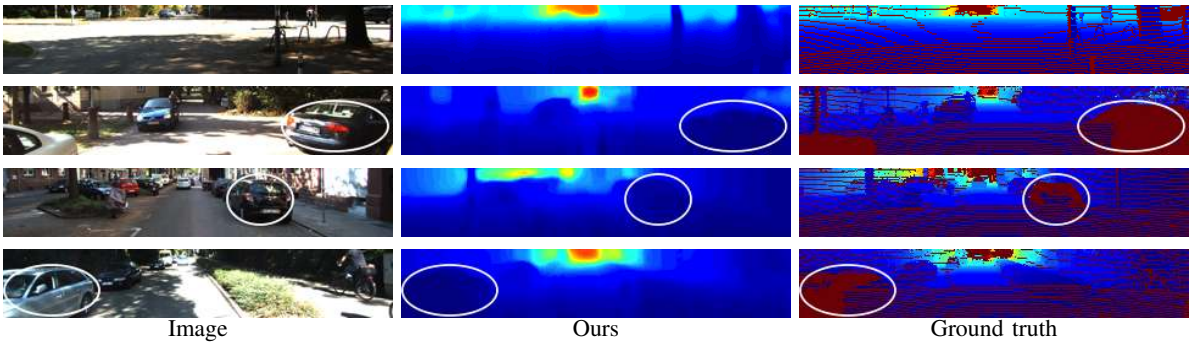



Fig. 5. Comparison on KITTI. For each test image, the corresponding depth estimated our method and the ground truth are given from left to right. Red denotes far and blue denotes close in the depth images, and missing values are denoted as dark red. The white circles demonstrate that missing depth value usually occurs due to the reflection when using the 3D laser range finder, while our method gives a reliable depth estimation. Best viewed in color.

to be noted that the 3D laser range finder usually cannot observe valid depth values when scanning the windows of the cars, which might lead to unsafe predictions in the high-level decision tasks. On the contrary, our method provides a relatively stable depth estimation regardless of the reflection.

### *D. Analysis and potential usage*

To reveal the intrinsic impact of introducing the limited and sparse partial observation, we analyzed the depth estimation performances at different heights of the scene on NYUD2. Specifically, we generated a set of scans that are perpendicular to the gravity direction and sampled above the ground from 10cm to 210cm at equal distances. For all test images on NYUD2, the same evaluation metrics mentioned above were applied to evaluate on those individual scans. Figure 6 reveals the comparison between our monocular depth estimation results (second row in Table I) and our refined partially observed result (last row in Table I). It is to be noted that there is a minimum point in each error metric in Figure 6(a), 6(b), 6(c), and a corresponding maximum point in each accuracy metric in Figure 6(d), 6(e), 6(f). This is corresponding to the height of our laser scan (80cm), which demonstrates that the observed laser information is effectively preserved based on our residual of residual neural network. Furthermore, the partial observation not only increases the performance at the height of the 2D laser range finder, but also gives a considerable promotion to the performances of the overall scene.

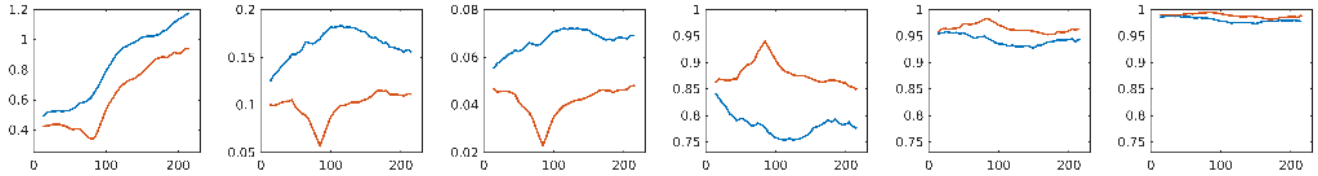
Figure 6 also demonstrates our method has a reliable root mean squared error below the heights of 80cm. This is particularly suited to the robotics applications. We demonstrate that our method has the potential for obstacle avoidance in Figure 7. Specifically, we parsed the scene based on different methods for comparison. This includes the simulated 2D laser range finder, the estimated dense depth and the ground truth depth provided by the Kinect. Following the general method for obstacle avoidance based on point clouds, we projected each dense depth map to the 3D space, and then down-projected all the 3D points within the height $(0, M]$ to the 2D plane to obtain the nearest obstacle in the scene. Here 0 is the height of the ground and $M$ is a safe range that is usually set to be higher than the robot. We set $M = 100$cm in this example. The simulated laser range scan was also set to be perpendicular to the gravity direction, which can be directly presented in the 2D plane. For thorough comparison, we simulated two laser range finders at 20cm and 80cm above the ground plane respectively. Figure 7 demonstrates the images and the corresponding obstacle maps generated from different methods. As can be seen from Figure 7, the 2D laser scan with a single planar view is too limited to correctly parse the full scene. For example, the laser scanner set at 20cm fails to detect some parts of the obstacles in Figure 7(a), whereas the laser scanner set at 80cm fails in parts in Figure 7(b), and both of them provide some inaccurate predictions in Figure 7(c) and 7(d). For depth estimation with only monocular images, there is usually a bias between the real distance and the estimated distance. By learning the depth taking the advantage of the 2D laser scan, our method is capable to provide a comprehensive depth estimation with relatively high accuracy, leading to a reliable obstacle map.

## V. Conclusion

This paper explores the monocular depth estimation task. By introducing sparse 2D laser range data into the depth estimation task, our method effectively alleviates the global scale ambiguity and produces a more reliable estimation result. We formulated the partially observed depth estimation task as a residual learning problem, which was implemented based on our residual of residual network with the aim of explicitly learning the residual depth. In addition, a novel loss function was proposed to combine classification and regression for the task of depth estimation. We conducted experiments on both indoor and outdoor datasets including NYUD2 and KITTI. The performance was compared with state-of-the-art techniques and our method shows superior results on both datasets validating the effectiveness of the proposed method. Furthermore, it suggests a promising direction to use sparse laser data to guide dense depth estimation using learning methods.

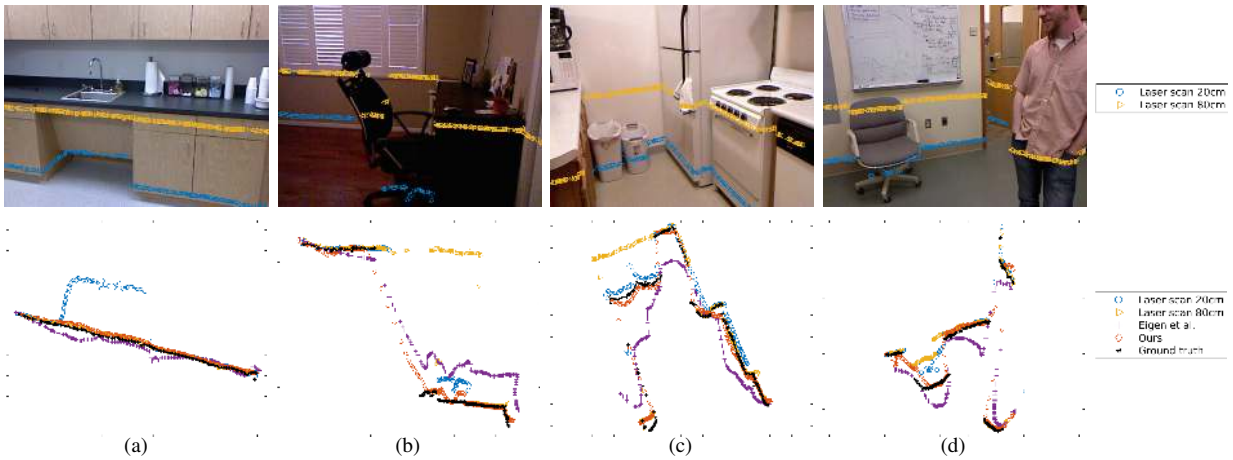
## References

[1] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012.



(a) rms (b) rel (c) log10 (d) $\delta_1$ (e) $\delta_2$ (f) $\delta_3$

Fig. 6. Evaluation results at different heights. In each figure, the horizontal axis is the height, and the vertical axis is the evaluation metric. The blue line denotes the performance of the depth estimation with only RGB image, and the red line denotes the performance of the depth estimation with combination of the RGB image and a single line of laser range scan with height at 80cm. Best viewed in color.




(a) (b) (c) (d)

Fig. 7. Illustration of obstacle avoidance. The first row shows the image with the projected laser scan. The height of the laser is set at 20cm to simulate the real situation. The second row demonstrates the corresponding obstacle maps in the two-dimensional space, with the gravity direction eliminated. Best viewed in color.


[2] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*, pp. 345–360, Springer, 2014.

[3] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2318–2325, IEEE, 2016.

[4] Y. Wang, S. Huang, R. Xiong, and J. Wu, "A framework for multi-session rgbd slam in low dynamic workspace environment," *CAAI Transactions on Intelligence Technology*, 2016.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, pp. 2366–2374, 2014.

[6] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[7] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields.," *IEEE transactions on pattern analysis and machine intelligence*, 2015.

[8] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," *arXiv preprint arXiv:1606.00373*, 2016.

[9] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *arXiv preprint arXiv:1605.02305*, 2016.

[10] A. Cherubini, F. Spindler, and F. Chaumette, "Autonomous visual navigation and laser-based moving obstacle avoidance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2101–2110, 2014.

[11] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Place classification with a graph regularized deep neural network," *IEEE Transactions on Cognitive and Developmental Systems*, 2016.

[12] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems*, pp. 1161–1168, 2005.

[13] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] A. Harrison and P. Newman, "Image and sparse laser fusion for dense scene reconstruction," in *Field and Service Robotics*, pp. 219–228, Springer, 2010.

[16] P. Piniés, L. M. Paz, and P. Newman, "Too much tv is bad: Dense reconstruction from sparse laser with non-convex regularisation," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 135–142, IEEE, 2015.

[17] C. Cadena, A. Dick, and I. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding," in *Proc. Robotics: Science and Systems*, 2016.

[18] H. Badino, U. Franke, and D. Pfeiffer, "The stixel world-a compact medium level representation of the 3d-world," in *Joint Pattern Recognition Symposium*, pp. 51–60, Springer, 2009.

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

[20] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 564–571, 2013.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[22] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," *arXiv preprint arXiv:1607.06349*, 2016.