2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 24–28, 2017, Vancouver, BC, Canada

# GeoCueDepth: Exploiting Geometric Structure Cues to Estimate Depth from a Single Image

Yiming Zeng, *Student Member*, *IEEE*, Yu Hu, *Member*, *IEEE*, Shice Liu, Qiankun Tang, Jing Ye, *Member*, *IEEE*, and Xiaowei Li, *Senior Member*, *IEEE*

***Depth estimation from a single image is very challenging due to the inherent ambiguity of mapping a color image to a depth map. Previous work tackles this problem by exploiting various levels of features with multi-scale deep convolutional neural networks. However, most of the local geometric structure related monocular depth cues are lost when being propagated through convolutional neural network. Moreover, the error of depth cues related to local geometric structures is not considered in the loss function. In this work, we propose the GeoCueDepth convolutional neural network to exploit local geometric structure cues and propose a training loss that takes the geometric error into consideration, which significantly improve the performance of depth prediction in both accuracy and sharpness. Experiments show that the proposed method achieves 0.122 average relative error and 0.078 square relative error on the NYU Depth v2 data set, which outperforms state-of-the-art monocular depth estimation approaches.***

## I. Introduction

Depth estimation plays an important role in autonomous robot navigation, grasping, human-computer interaction, 3D modeling and augmented reality. While dedicated depth sensors like 3D laser scanner or RGB-D camera can give accurate depth measurement, depth estimation from color images is still fundamental when the depth sensor is not equipped or is not able to generate clean data.

Previous depth estimation methods mainly focus on stereo [1,2] and motion (Structure-from-Motion, SfM) [3]. When stereo views are provided, the depth map can be estimated from accurate image correspondence that is described conventionally by hand-crafted features. In the motion case, based on the point correspondences between relevant frames, the 3D scene can be reconstructed through triangulation. On the other hand, depth estimation from a single image is expected in many applications such as virtual shopping, real estate, object recognition [4], human detection [5] and automatic 2D-to-3D video conversion. However, unlike depth estimation from stereo and motion, monocular depth estimation is very challenging due to the inherent ambiguity of mapping a color image to a depth map, as a given color image can be mapped to numerous possible world scenes.

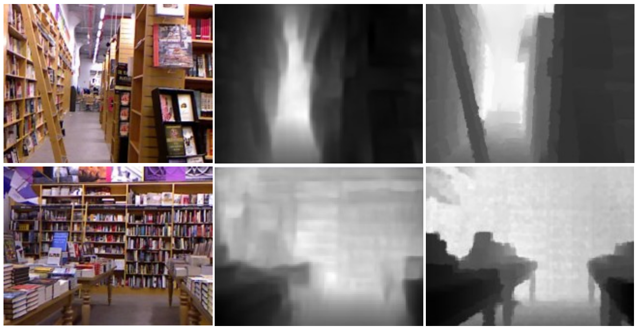
Figure 1. Examples of depth estimation results using conventional CNN method. The accuracy is high (i.e. square relative error between the pre-dicted depth image and the groudtruth are 0.259 in the first row and 0.270 in the second row, respectively) but many geometric structures such as strong edges and corners are lost. From left to right: color image, predicted depth image [11], and ground truth.

A color image consists of many depth cues which can be classified to semantical cues, such as object location, object size, occultation, and geometric structure cues, e.g. edge, corner, vanishing point, perspective, texture gradient. Visual system of human beings can inference 3D structure information from 2D image according to monocular depth cues [2]. It still remains a difficult challenge in robotic vision to predict depth map from these cues.

Previous work tackles this problem by exploiting various levels of depth cues. At an early stage, efforts of exploring monocular depth cues related to geometric structure are primarily based on hand-crafted features [6~10]. Later on, depth features were learned from deep convolutional neural net-works (CNNs) [11~18].

However, we observe that the monocular depth cues related to local geometric structures are mostly lost during being propagated through the CNN layers. In addition, quantification metrics such as square relative error and root mean square relative error cannot represent the underlying geometric structure of the scene. Therefore, prior work sometimes will be trapped in a dilemma that the accuracy of the predicted depth map looks high but geometric details such as strong edges and corners are lost, as shown in Figure 1.

Our objective is to build a CNN that can well exploit the geometric cues to enhance the depth estimation performance. We demonstrate the effectiveness of the proposed approach in terms of both accuracy and sharpness and compare it with the other state-of-the-art monocular depth estimation approaches on the widely used NYU Depth v2 dataset [30].

The main contributions of this work are:

**a depth cue expressway architecture**, in which the monocular depth cues that have been extracted in low layers are forwarded to the last summation layer to provide

* This work is supported in part by National Natural Science Foundation of China under Grant No. 61274030, 61532017, and in part by National Basic Research Program of China (973 Program) under Grant No.2011CB302503.

The authors are with the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, P.R. China, and are also with the Graduate University of Chinese Academy of Sciences, Beijing, 101407, P.R. China (e-mail: {zengyiming, huyu, liushice, tangqiankun, lxw}@ict.ac.cn. Yu Hu is the corresponding author: 86-10-62600632; fax: 86-10-62600600; e-mail: huyu@ict.ac.cn).

978-1-5386-2682-5/17/$31.00 ©2017 IEEE

geometric structure details of the scene, thus to improve the accuracy and sharpness of the predicted depth map.

**a local geometric structure error,** which considers both accuracy and local geometric structure details. In addition, a loss function derived from this metric encourages the predicted depth map to retain more geometric features such as strong edges and corners.

## II. Related Work

Depth estimation from a single image is an ill-posed problem because a captured color image scene can be mapped to numerous real world scenarios [11]. Previous work addresses this issue by exploiting various levels of depth cues. Saxena et al. [6] used linear regression and Markov Random Field (MRF) with multi-scale hand-crafted features to predict depth map. Then they developed their theory to Make3D [7] for 3D model generation. Despites of low-level and mid-level cues, high-level cues such as user annotations [8], semantic object labels [9] were also used to predict depth map.

In recent years, various deep CNNs were proposed, e.g. AlexNet [20], VGG [21], ResNet [22], which were applied to object classification, object detection and semantic segmentation tasks and achieved very good performance. The growing interest for CNNs has inspired ideas of end-to-end learning depth map from CNNs [11~18].

The CNN-based monocular depth estimation approaches have two ways to process depth cues. One way is to transform monocular depth cues into intermediate products by CNN at first, then post-process these intermediate products by Conditional Random Fields (CRFs) [15, 16] or random forest [18], and harmonize the information to form a depth map at last. Liu et al. [15] assumed an image can be over-segmented into image patches (super-pixels) and defined these super-pixels as nodes of CRF. For a super-pixel, its unary potential and pairwise potential were learned by CNN and were fed to the CRF structured loss layer. Predicting the depth map is to maximize the conditional probability. Afterwards, Li et al. [17] combined the approaches of Liu et al. [15] and Eigen et al. [11], and used a hierarchical CRF to integrate global context CNN results and regional CNN results. In addition to CRFs, random forest is also introduced to accomplish this task [18]. Chakrabarti et al. [19] solved the problem of monocular depth estimation by using a neural network to produce a mid-level representation that summarized depth cues.

The other way is to directly regress a color image into the depth map in the pixel level. The pioneer work was proposed by Eigen et al. [11]. They proposed a two-scale network architecture: the coarse scale makes a global prediction based on the entire image, and the fine scale refines this prediction locally. Later on, they extended the work to a three-scale network architecture for pixel level tasks including depth estimation, surface normal prediction and semantic labeling [12]. More recently, Mancini et al. [13] proposed an encoder –decoder architecture and Laina et al. [14] introduced a single-scale CNN architecture that consisted of the fully convolutional architecture and the residual learning [22] to estimate depth.

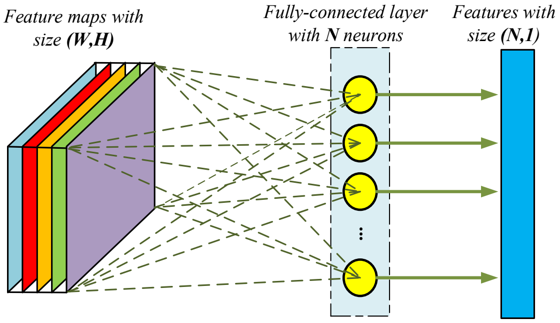
Feature maps with size (W,H)
Fully-connected layer with N neurons
Features with size (N,1)
⋮

Figure 2. Fully-connected layer. The input of the fc layer is ***C*** channel of feature maps with witdth ***W*** and height ***H***, the output of the fc layer is a vector of size ***(N,1)***, where ***N*** is the number of neurons in the fc layer. The two-dimensional feature maps are reduced to a one-dimensional vector.

Moreover, relative depth annotations were used by Zoran et al. [23] and Chen et al. [24] to train their depth prediction CNN. In addition to the supervised methods, semi-supervised [28] and unsupervised [29] depth estimation methods were proposed recently.

## III. Motivation

While the prior work on monocular depth estimation has achieved good performance, we find that there are still two issues that limit the accuracy and sharpness of the estimated depth map.

### *A. The problem of depth cue vanishing*

To better understand the performance contribution from different layers, Yosinski et al. [25] introduced a visualization tool to give insights into the function of intermediate feature layers. We utilized this tool to visualize feature maps generated by each layer and observed that some monocular depth cues were lost step by step in the CNN pipeline. One of the root causes is that the fully-connected layer flattens the two-dimensional feature maps to a one-dimensional vector, resulting in vanishing of local geometric information such as edges, corners, and perspective, as shown in Figure 2.

The other root cause is the pooling operation. Convolution layers usually contain a pooling operation. No matter what pooling operation is chosen, e.g. the max-pooling, the average-pooling or the stochastic-pooling, the pooling operation actually emphasizes the relative relationship among neighboring features rather than the absolute value of a feature. Therefore, pooling has a negative effect on retaining geometric structure related depth cues.

Motivated by this observation, we propose a “depth cue expressway” architecture to retain the geometric structure information. It forwards geometric structure related monocular depth cues to the last layer where the depth map is predicted. Since these cues distributed in feature maps are directly utilized to estimate depth map, our model not only improves sharpness, but also benefits accuracy. We will elaborate our model with depth cue expressway architecture in Section IV.B.

### *B. The problem of neglecting geometric error*

The most commonly used loss function for depth regression task is the L2 loss, which minimizes the Euclidean distance between the predicted depth $d_i$ and the ground truth depth $d_i^*$ for a pixel *i*.

$$L_2 = \frac{1}{2n}\sum_{i \in P}(d_i - d_i^*)^2$$

where *P* represents the entire image of the scene, *n* is the pixel number of *P*. The L2 loss only measures the numerical difference of $d_i$ and $d_i^*$, having no description about the correlation between the neighborhoods $Ng(d_i)$ and $Ng(d_i^*)$ of pixel *i*.

Prior work [12, 15] has shown that if the geometry related features could be considered in some way, then the depth estimation performance could be improved. Eigen et al. [12] considered image gradients of the prediction with the ground truth in the loss function to encourage predictions to have not only close-by values, but also similar local structures. Liu et al. [15] used a pairwise potential item to measure the similarity of adjacent super-pixels. Different with these work, we use three principle geometric feature metrics, i.e. curvature, gradient, and contrast, to quantify the correlation of $Ng(d_i)$ and $Ng(d_i^*)$, and derive a loss function to train our convolutional neural network. This loss function encourages small pieces of the depth map to have similar geometric structure. We will present our loss function in detail in Section IV.C.

## IV. PROPOSED APPROACH

In the section, we first define the local geometric structure error to measure the difference of two images in terms of geometric structures. Then we present the GeoCueDepth approach. The proposed GeoCueDepth approach consists of a novel neural network architecture that has dedicated depth cue expressways to forward feature maps with rich geometric cues and a novel training loss function that specifically considers geometric structures.

### A. Relative Local geometric structure error

To solve the aforementioned problems of neglecting the geometric structure error, we consider curvature, gradient, and contrast of Tamura's texture features [26] to quantify the geometric structure information.

The relative error in curvature and the relative error in gradient are given by,

$$T_1\left(d_i, d_i^*\right) = \left|\kappa\left(d_i\right) - \kappa\left(d_i^*\right)\right| / \left|\kappa\left(d_i^*\right)\right|$$
$$T_2\left(d_i, d_i^*\right) = \left|\nabla\left(d_i\right) - \nabla\left(d_i^*\right)\right| / \left|\nabla\left(d_i^*\right)\right|$$

where $\kappa$ represents curvature [31] and $\nabla$ represents the sum of horizontal and vertical gradients. Note that for the corner case that the denominator is 0, a mask should be set to exclude this pixel.

Moreover, contrast implies the depth hierarchy, for example, high contrast means abundant depth hierarchy. The original contrast of Tamura's texture features is calculated on the entire image, as follows,

$$\xi\left(P\right) = \sigma^2 / \left(\mu_4\right)^{\frac{1}{4}}$$

where $\mu_4\left(x\right) = E\left[\left(x - E\left(x\right)\right)^4\right]$ and $\sigma^2(x) = E\left[\left(x - E\left(x\right)\right)^2\right]$

We modify the original contrast metric by taking $Ng(d_i)$ into consideration to measure local structure,

$$T_3\left(d_i, d_i^*\right) = \left|\xi\left(Ng(d_i)\right) - \xi\left(Ng(d_i^*)\right)\right| / \left|\xi\left(Ng(d_i^*)\right)\right|$$

Based on the abovementioned geometric errors, we define the Relative Local Geometric Structure Error (RLGSE) as follows:

$$RLGSE\left(d_i, d_i^*\right) = \frac{1}{n}\sum_j \sum_{i \in P} \beta_j T_j\left(d_i, d_i^*\right)$$

where $\beta_j, j = 1, 2, 3$ is the empirical coefficient for each error.

In the following, we use RLGSE to guide the selection of feature maps that should be forwarded to the last layer to produce the depth map, as well as to guide the modification of the training loss function

### B. Depth cue expressway architecture

As we have mentioned, conventional CNN-based monocular depth estimation approaches have the depth cue vanishing problem. To deal with this problem, intermediate results which are rich in monocular depth cues should be retained and be utilized to enhance the depth prediction performance.

To give an insight to monocular depth cues distributed in the primitive model, we visualize this primitive model with the tool proposed by Yosinski et al. [25]. By observing the visualized feature maps, we find monocular depth cues such as strong edges, corners, and texture gradient are obvious in the feature maps generated by the low level layers, and then gradually vanishes along with a sequential convolution and pooling operations. Therefore, we quantitatively analyze this phenomenon. We calculate the average RLGSE of the feature maps in multi-channels,

$$e_i = RLGSE(F(x, \{W_i\}), D^*) / C_i,$$

where $F(x, \{W_i\})$ is the feature map generated by the *ith* layer, $D^*$ is the ground truth of scene depth, and $C_i$ is the number of channels.

We plot $e_i$ in Figure 3. The RLGSE value gradually decreases along with the CNN layers. We can see there are several abrupt drops in the curve. The points marked in red represent local minimums, which means the feature maps

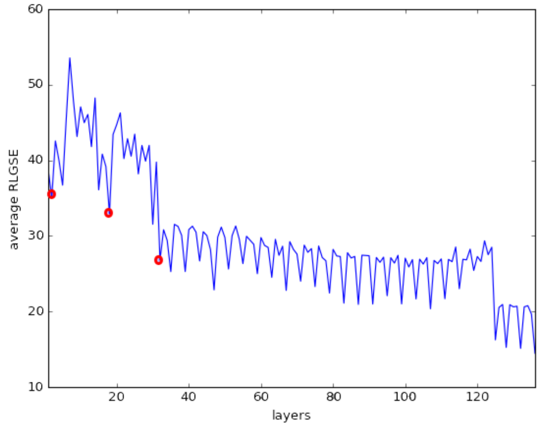

Figure 3 The RLGSE of feature maps. The marked points correspond to three layers: pool1, res3a_branch2c and res3b3.

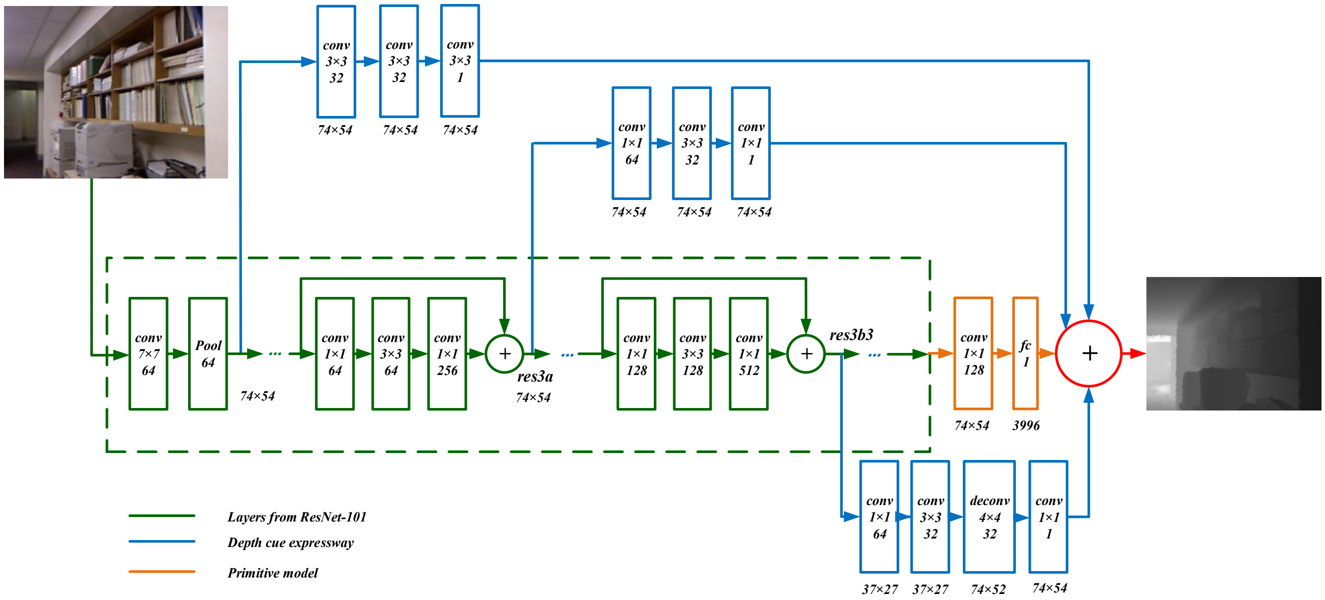

Figure 4 Our GeoCueDepth convolutional neural network. Our model consists of a main part derived from ResNet-101, and a depth cue expressway architecture.The convolutional kernel size and channels of feature maps are listed in the rectangles and the size of feature maps size labeled below. The depth cue expressway is a fully convolutional network layer. There are three depth cue expressways inserted after the pool1 layer, the res3a layer and the res3b3 layer respectively.

contain abundant geometric structures that are similar to the ground truth scene. Please note the RLGSE rises again after the drops, indicating the geometric structure related features diminish in the subsequent layers. That is also consistent with our observation on the visualized feature map. We use the local minimums in low level layers to guide the CNN design because the size of feature maps in high level layers is too small to give meaningful geometric cues.

To utilize these depth cues that will disappear, we forward the intermediate feature maps generated by the pool1, res3a_branch2c and res3b3 layers via three dedicated depth cue expressways to the summation layer which predicts depth map. Depth cue expressway is a fully convolutional network which consists of three convolutional layers and has no pooling layer to avoid reducing the size of features and filtering out structure details. When the 3×3 kernel is used, padding is needed to maintain the size of feature maps.

We designed a depth prediction convolutional neural network, which is derived from the well-known ResNet of 101 layers (ResNet-101). We removed the global pooling layer with all the layers behind it in ResNet-101 and appended a convolutional layer and a fully-connected layer.

As shown in Figure 4, the assembly of the layers in green and in orange forms a primitive model. The GeoCueDepth model further consists of three depth cue expressways. Please note that the res3a_branch2c layer is within the residual building block. In order to preserve the completeness of the residual building block, the depth cue expressway is attached to the output of the residual block. The depth cue expressway inserted after the res3b3 layer gets the input feature map with size of 37×27, therefore, an extra deconvolution layer is inserted to upsample the feature map. Finally, depth maps forwarded by these four branches are merged at the summation layer. The ResNet-101 part in the GeoCueDepth network can utilize high-level depth cues and the depth cue expressway part pays more attention to the geometric structure related monocular depth cues, which benefits the depth estimation performance a lot. These skip connections were used in FCN [32] to combine coarse feature maps with fine feature maps. Different with the skip connections in FCN, our expressway is used to preserve local geometric structure.

### *C. Training loss function*

There are various types of loss used in training convolutional neural networks to estimate depth map. However, existing loss functions have not taken the geometric structure into consideration. We propose a training loss function which is derived from the RLSGE metric, thereto consider the geometric structure errors,

$$Loss=\frac{1}{2n}\left[\sum_{i\in P}\left(d_i-d_i^*\right)^2+\sum_j\sum_{i\in P}\lambda_j S_j^2\left(d_i,d_i^*\right)\right]$$

where

$$S_1\left(d_i,d_i^*\right)=\kappa\left(d_i\right)-\kappa\left(d_i^*\right)$$

$$S_2\left(d_i,d_i^*\right)=\nabla\left(d_i\right)-\nabla\left(d_i^*\right)$$

$$S_3\left(d_i,d_i^*\right)=\xi\left(Ng(d_i)\right)-\xi\left(Ng(d_i^*)\right)$$

In this training loss function, we replace the relative component $T_j\left(d_i,d_i^*\right)$ to quadratic component $S_j^2\left(d_i,d_i^*\right)$ to make it derivable. To achieve small loss, not only the predicted depth should be close to the ground truth depth value, but also the curvature, gradient and contrast of Tamura's texture features should be similar with the ground truth. Thus, our loss function encourages the GeoCueDepth model to generate similar geometric structure with the color image of the scene. This loss function is flexible and easy to be extended if more structure features need to be considered, as

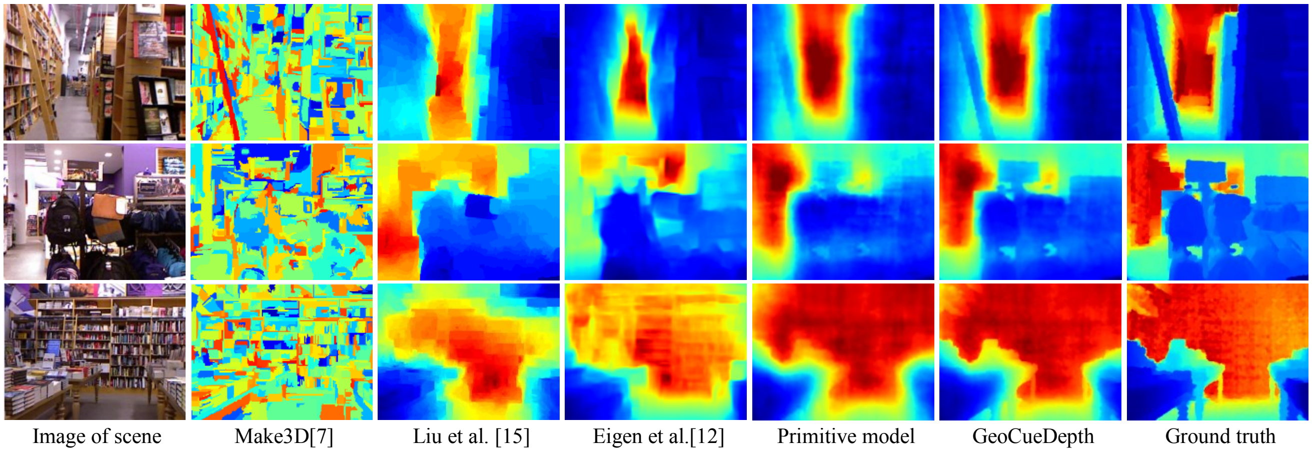

Figure 5 Examples of the predicted depth maps on the NYU Depth v2 dataset. As we can see, with effective utilization of geometric structure related monocular depth cues, our approach yields the depth map retainig more geometric structures such as strong edges and corners.

long as the error metric is derivable. Otherwise, appropriate transformation is needed for getting derivable items.

## V. Experiments

In this section, we describe the evaluation results on the NYU depth v2 dataset which contains images taken by Kinect camera in 464 indoor scenes, with the official split consisting in 249 training and 215 test scenes.

### *A. Experimental framework*

We train the primitive model based on the pre-trained RetNet-101 and fine-tune the whole model by stochastic gradient descent (SGD). In the GeoCueDepth architecture, xavier initialization is used in the convolutional layers, and rectified linear units are used in the whole model. Dropout layers are inserted before the fully-connected layer to avoid overfitting. We initialize the learning rate as 1e-5, and use the learning rate policy of dropping the learning rate in "steps" #20000 by a factor for gamma 0.1 every stepsize.

To improve diversity and variability of dataset and get more training example, we use data argument techniques including scale, rotation, crop, HSL shift and flip. Note that when using scale, we assume depth is simply proportional to the field of view, ignoring other image-forming condition. So when image of the scene is zoomed in by a factor ω, the corresponding depth should be divided by the factor ω. When the image is rotated, cropping should within the boundaries of original image. We use HSL shift instead of RGB shift, as the random shift in Hue, Saturation and Lightness are more similar to the variant light conditions in the natural environment.

TABLE I DEPTH ESTIMATION MEASUREMNENTS

| Metric | Expression |
|---|---|
| Average relative error<br>(rel) | $\frac{1}{N}\sum_{i\in P}\frac{\lvert d_i - d_i^*\rvert}{d_i^*}$ |
| Square relative error<br>(sqr-rel) | $\frac{1}{N}\sum_{i\in P}\frac{\left(d_i - d_i^*\right)^2}{d_i^*}$ |
| Root mean squared error<br>(rms) | $\sqrt{\frac{1}{N}\sum_{i\in P}\left(d_i - d_i^*\right)^2}$ |
| Root mean squared error log<br>(rms-log) | $\sqrt{\frac{1}{N}\sum_{i\in P}\left[\ln\left(d_i\right) - \ln\left(d_i^*\right)\right]^2}$ |
| Average log10 error<br>(log10) | $\frac{1}{N}\sum_{i\in P}\left\lvert\log_{10}\left(d_i\right) - \log_{10}\left(d_i^*\right)\right\rvert$ |
| Threshold $\delta$ | $\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right) < \delta, \delta \in \left\{1.25, 1.25^2, 1.25^3\right\}$ |

We trained and tested our model with the Caffe [27] framework on a NVIDIA GeForce GTX 1080Ti with 11GB frame buffer. It took about two days to train. Another advantage of directly regressing depth map is that the prediction time is 31ms per image, which almost equals to the forward time of the CNN, therefore, our model can meet the real time requirement for many applications.

### *B. Evaluation*

We evaluate our approach with measurements commonly used in previous work, as shown in Table 1. Table 2 reports the performance of our GeoCueDepth method, along with other popular state-of-the-art methods that are also trained on the NYU Depth v2 dataset. As can be observed, our approach outperforms the conventional methods [7,12,15] with large margins. In particular, the accuracy of  increases significantly from 0.715 to 0.865, and the error of log10 reduces by half. Therefore, the joint effect of forwarding monocular depth cues and loss function considering geometric structure benefits both accuracy and sharpness.

The predicted depth maps are illustrated in Figure 5. Firstly, we compare the depth map predicted by the primitive model with the GeoCueDepth model. For both models, many geometric structure details are retained, for instance, the edges of desk legs and the ladder. But with more geometric structure cues, the depth map predicted by our GeoCueDepth is closer to the ground truth and is sharper.

Then we compare the GeoCueDepth results against other CNN-based methods that regress the depth map directly. Our results are significantly better than that of [7, 15]. Comparing with the recent work of Eigen et al. [12], our depth predictions have retained more geometric structures and exhibit noteworthy visual quality, for example, in the first row, the edge of the ladder is sharper, in the second row, the edge of

TABLE 2 DEPTH ESTIMATION PERFORMANCE ON NYU DEPTH V2

| Method | Accuracy<br>(higher is better) | | | Error<br>(lower is better) | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | rel | sqr-rel | rms | rms-log | log10 |
| Make3D [7] | 0.447 | 0.745 | 0.897 | 0.349 | - | 1.214 | - | - |
| Eigen et al.[12] | 0.769 | 0.950 | 0.988 | 0.158 | 0.121 | 0.641 | 0.214 | - |
| Liu et al. [15] | 0.614 | 0.883 | 0.971 | 0.230 | - | 0.824 | - | 0.095 |
| Wang et al. [16] | 0.605 | 0.890 | 0.970 | 0.220 | 0.210 | 0.745 | 0.262 | 0.094 |
| Chakrabarti et al. [19] | 0.650 | 0.895 | 0.968 | 0.208 | 0.195 | 0.770 | 0.270 | - |
| Laina et al. [14] | 0.811 | 0.953 | 0.988 | 0.127 | - | 0.573 | **0.195** | 0.055 |
| Our primitive model | 0.715 | 0.936 | 0.985 | 0.179 | 0.150 | 0.619 | 0.305 | 0.107 |
| Our GeoCueDepth model | **0.865** | **0.963** | **0.992** | **0.122** | **0.078** | **0.430** | 0.260 | **0.050** |

the TV set is sharper, and in the third row, the legs of the desk and chairs are more evident.

## VI. CONCLUSION

We have proposed a new model based on deep learning for the task of monocular depth estimation. The GeoCueDepth model employs a depth cue expressway architecture and enhances the training loss function to consider local geometric structure errors. The curvature, gradient, contrast cues and the expressway are able to successfully highlight geometric structures that are in the ground truth scene depth. This enables the proposed model to outcome a more detailed depth representation of a scene. Our extensive evaluation on the NYU v2 dataset demonstrates the effectiveness of the proposed model by showing better overall accuracy and sharpness. This work shows that features of local geometric structures are helpful for depth estimation. In the future, we plan to conduct experiments on more datasets like KITTI and try more geometric structures.

## REFERENCES

[1] A. Torralba and A. Oliva, "Depth estimation from image structure", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, pp. 1226-1238, 2002.

[2] I. Howard, Perceiving in depth, 1st ed. Oxford: Oxford University Press, 2012.

[3] A. Cretual, F. Chaumette, and G. Sandini "Image-based positioning with respect to a non-structured scene using 2D image motion" in IROS, 2000.

[4] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in IROS, 2015.

[5] B. Choi, C. Mericli, J. Biswas, and M. Veloso, "Fast human detection for indoor mobile robots using depth images," in ICRA, 2013.

[6] A. Saxena, S. Chung and A. Ng, "Learning Depth from Single Monocular Images", in NIPS, 2005, pp. 1161-1168.

[7] A. Saxena, Min Sun and A. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 31, no. 5, pp. 824-840, 2009.

[8] B. Russell and A. Torralba, "Building a database of 3d scenes from user annotations", in CVPR, 2009, pp. 2711-2718.

[9] B. Liu, S. Gould and D. Koller, "Single image depth estimation from predicted semantic labels", in CVPR, 2010, pp. 1253-1260.

[10] K. Karsch, C. Liu and S. Kang, "Depth Extraction from Video Using Non-parametric Sampling", in ECCV, 2012, pp. 775-788.

[11] D. Eigen, C. Puhrsch and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network", in NIPS, 2014, pp. 2366-2374.

[12] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture", in ICCV, 2015, pp. 2650-2658.

[13] M. Mancini, G. Costante, P. Valigi and T. A. Ciarfuglia, "Fast robust monocular depth estimation for Obstacle Detection with fully convolutional networks," in IROS, 2016, pp. 4296-4303.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab, "Deeper depth prediction with fully convolutional residual networks", in 3DV, 2016, pp. 239-248.

[15] F. Liu, C. Shen and G. Lin, "Deep convolutional neural fields for depth estimation from a single image", in CVPR, 2015, pp. 5162-5170.

[16] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price and A. Yuille, "Towards unified depth and semantic prediction from a single image", CVPR, 2015, pp. 2800-2809.

[17] B. Li, C. Shen, Y. Dai, A. Hengel and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs", in CVPR, 2015, pp. 1119-1127.

[18] A. Roy and S. Todorovic, "Monocular Depth Estimation Using Neural Regression Forest", in CVPR, 2016, pp. 5506-5514.

[19] A. Chakrabarti, J. Shao and G. Shakhnarovich, "Depth from a Single Image by Harmonizing Overcomplete Local Network Predictions", in NIPS, 2016, pp. 2800 - 2809.

[20] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", in NIPS, 2012, pp. 1097-1105.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv: 1409.1556, 2014.

[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", in CVPR, 2016, pp. 770-778.

[23] D. Zoran, P. Isola, D. Krishnan and W. Freeman, "Learning Ordinal Relationships for Mid-Level Vision", in ICCV, 2015, pp. 388-396.

[24] W. Chen, Z. Fu, D. Yang and J. Deng, "Single-Image Depth Perception in the Wild", in NIPS, 2016.

[25] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization", In ICML, Workshop on Deep Learning, 2015.

[26] H. Tamura, S. Mori and T. Yamawaki, "Textural Features Corresponding to Visual Perception", IEEE Transactions on Systems, Man, and Cybernetics, vol. 8, no. 6, pp. 460-473, 1978.

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding", in MM, 2014, pp. 675-678.

[28] Y. Kuznietsov, J. Stuckler and B. Leibe, "Semi-Supervised Deep Learning for Monocular Depth Map Prediction", arXiv preprint arXiv:1702.02706, 2017.

[29] C. Godard, O. Aodha and G. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", arXiv preprint arXiv: 1609.03677, 2016.

[30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images", In ECCV, 2012, pp. 746-760.

[31] Y. Gong, "Spectrally regularized surfaces," 2015.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015, pp. 3431–3440.