Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Full Resolution Dense Depth Recovery by Fusing RGB Images and Sparse Depth*

Yong Luo, Guoliang Liu, Hanjie Liu, Tiantian Liu and Guohui Tian
School of Control Science and Engineering
Shandong University
Jinan, Shandong, China
{*liuguoliang*}*@sdu.edu.cn*

Shiqing Xin
School of Computer Science and Technology
Shandong University
Qingdao, Shandong, China
xinshiqing@sdu.edu.cn

***Abstract*— Full resolution depth is required in many real world engineering applications. However, exist depth sensors only offer sparse depth sample points with limited resolution and noise, e.g., LiDARs. We here propose a deep learning based full resolution depth recovery method from monocular images and corresponding sparse depth measurements of target environment. The novelty of our idea is that the structure similar information between the RGB image and depth image is used to refine the dense depth estimation result. This important similar structure information can be found using a correlation layer in the regression neural network. We show that the proposed method can achieve higher estimation accuracy compared to the state of the art methods. The experiments conducted on the NYU Depth V2 prove the novelty of our idea.**

***Index Terms*— Visual odometry, depth prediction, deep learning, view synthesis**

I. INTRODUCTION

Perceiving and interpreting the 3D structure of the surrounding environment is critical for a wide range of engineering applications, such as advanced driver assistance system, autonomous driving, intelligent vehicle, robot, etc. However, existing depth sensors have their own limitations and are difficult to meet the requirements of engineering applications. For instance, multi-line lidars are sophisticated and complex in structure, cost prohibitive and provide only sparse measurements of distance objects. The quality of the depth information obtained by Kinect developed by Microsoft is affected by the illumination conditions, which can generate a lot of noisy points under strong light. These sparse depth points can degrade the algorithm performance since many useful structure information is missed.

With the improvement of computer performance and the emergence of large-scale datasets, depth estimation from a single monocular image based on deep learning shows great potential, which has great robustness to handle the problems due to light condition changes, image noise and image motion blurring. However, depth estimation from a single RGB image alone is inherently unreliable and ambiguous: a single image on its own cannot explicitly provide accurate depth cue, i.e., a color image for a given scene, there are countless 3D scene structures to generate this 2D image. To address this inherently fundamental limitation of depth estimation from monocular images, Ma et al. [1] used uniformly sample sparse depth measurements to generate sparse depth map and construct a deep regression model that takes both sparse depth map and RGB image as input to predict a full resolution depth image. Their results show that the additional sparse depth samples can improve the accuracy and robustness of depth estimation, and their method achieves the state-of-the-art performance. Sparse depth measurements are easily obtained from 3D sensors, e.g., LiDARs, Kinect, and stereo camera, etc. Combining with the RGB image, a full resolution depth image can be estimated using the regression network [2][3].

Based on the idea of leaning dense depth from sparse depth and RGB image, we here propose a correlation layer based end-to-end depth estimation neural network that can further improve the performance of deep reconstruction. The novel idea is that the depth image and the RGB image have similar structure information, e.g., edge information and corner information, which means there are some key common information can be found by matching these two images. The correlation layer does that job for matching high level features of depth image and RGB image. In this way, the detail information of estimated dense depth can be refined by our idea. We evaluate the depth reconstruction performance of our method quantitatively and qualitatively on the NYU-Depth-V2 [2] dataset. The experimental results show that our learning network can achieve higher accuracy than the work of Ma et al. [1].

II. RELATED WORK

A. Supervised Single Image Depth Estimation

Monocular depth estimation refers to reconstruct the 3D structure of a scene from monocular images. Saxen et al.

* This work is partially supported by National Natural Science Foundation of China (#61603213, #91748115), the National Key R&D Program of China (#2018YFB1306504), Young Scholars Program of Shandong University (#2018WLJH71), the Fundamental Research Funds of Shandong University, and the Taishan Scholars Program of Shandong Province. (Corresponding author: Guoliang Liu).

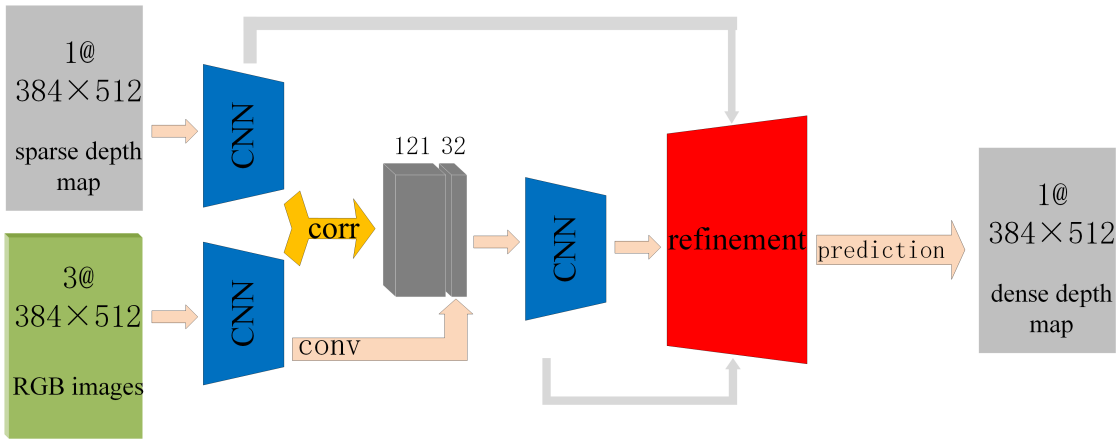

Fig. 1: The architecture of our neural network includes two CNN (Convolutional Neural Network) modules, the correlation layer and the refinement module. Two CNN modules have nine convolutional layers (three for the first CNN module) with stride of 2 in six of them,i.e., the stride of the odd layer from the fourth layer is set to 1, and one ReLU nonlinearity after each layer. The CNN layers size are defined as 7×7 for the first layer, 5×5 for the following two layers and 3×3 starting from the fourth layer. The correlation layer is a core module of our network that can find similar structure information between the RGB image and corresponding depth image. The refinement module combines the shrank CNN feature map and the dense depth prediction after upsampled to achieve a more refined depth prediction. We use the same network architecture for NYU-Depth-v2 dataset.

[4] proposed a patch-based model that estimates the absolute scale of each patch, and uses a Markov Random Field model to combine patch predictions to infer depth map. However, this approach is difficult to estimate thin structures, and such local prediction lacks the global context required to generate realistic outputs. Instead of hand-crafting features, many recent methods use a convolutional neural networks (CNN) to learn depth information. Eigen et al.[5] [6] proposed a two-stacked convolutional neural network to predict global feature of image, and refine local detail of depth prediction by combining coarse-scale depth prediction with fine-scale prediction. Li et al. [7] combined learning-based image features with hierarchical CRFs to refine the depth prediction. Liu et al. [8] proposed a deep structured learning method based on continuous depth and Gaussian assumptions on the pairwise potentials to avoid the hand-crafted features. Laina et al. [9] used RGB-D images to train the ResNet-based encoder-decoder architecture for depth prediction of indoor scenes with higher accuracy.

B. *Semi-supervised and Unsupervised Depth Estimation*

Inspired by the image wrap technique "spatial transformer" [10], the semi-supervised and unsupervised learning methods attract plenty of intention in the field of depth estimation. Kuznietsov et al. [11] proposed a semi-supervised training network for depth prediction using stereo images and sparse ground-truth depth readings from a supplementary sensing cue, e.g., 3D laser. Garg et al.[12] propose a deep convolutional neural network for single-view depth prediction, which is trained in an unsupervised manner from a stereo pair analogous to an autoencoder. Zhou et al. [13] jointly unsupervised trained the depth estimator and the pose estimator based on temporal geometric consistency, and achieved remarkable results. However, the result of the monocular depth estimation lacks global scale information compared to the real scene. Godard et al. [14] proposed an unsupervised monocular depth estimation method using left-right consistency, which can reconstruct the scaled dense depth. Li et al. [15] combined the ideas of [13] and [14], and proposed a scaled deep learning network for monocular depth estimation and pose estimation. However, the construction of the left-right consistency loss is computationally intensive and complex. Compared to work presented in [15], our method is not only fast, but also simple to implement, and can realize scaled scene reconstruction.

C. *Learning-Based Sensor Fusion*

Another line of related work is depth reconstruction from sensor fusion information. Mancini et al.[16] proposed a convolutional neural network that took both RGB images and optical flow images as input to reconstruct the depth of the scene. Cadena et al.[17] developed auto-encoders to simultaneously generate complete scene segmentations and depth maps from RGB images, depth images and semantic label information. Kuznietsov et al. [11] proposed a semi-supervised training network for depth prediction using stereo images and (sparse) ground-truth depth readings from a supplementary sensing cue, e.g., 3D laser. Liao et al.[18] combined the 2D laser measurement of the mobile ground robot and RGB images information to reconstruct the depth

of the scene, which can achieve higher precision than using RGB images alone. Ma et al. [1] proposed an end-to-end learning scheme to predict dense depth map from sparse depth maps and RGB images. Compared to work presented in [18], Ma et al. [1] does not have assumptions about the direction or position of the sensor, and the spatial distribution of the input depth samples. Their results show that the additional sparse depth samples can improve the accuracy and robustness of depth estimation, and their method achieves the state-of-the-art performance. In contrast to the work in [1], our method can achieve higher accuracy on the public datasets by designing a correlation layer between depth and RGB image, since they share similar structure features.

III. METHODOLOGY

We describe the network structure of proposed deep learning based dense depth estimation system in this section. In addition, we discuss the loss functions used in this paper, and training process in detail.

A. Network Structure

Our network consists of four part, i.e. two CNN (Convolutional Neural Networks) modules, a correlation layer and a refinement module, as shown in Fig. 1. The first step is to create two separate, yet identical processing streams for sparse depth map and RGB image. With this architecture, the network first extracts multi-scale local and global meaningful high-level representations of sparse depth map and RGB image respectively. The first CNN module has three convolutional layers (conv1, conv2, con3) with a step size of 2, followed by a ReLU activation function on each layer. The convolution kernel size of the first layer is 7×7, and the convolution kernel size of the latter two layers is 5×5.

We combine the high-level feature representations of both the sparse depth map and the RGB image in the correlation layer, which plays the matching process for finding similar structure features between depth and RGB image. The correlation layer was first proposed in FlowNetCorr for optical flow estimation [19]. Here, we use the correlation layer to match high-level structure features of depth image and RGB image. The sparse depth map and RGB image are input into the first CNN module, and two feature maps of P_1 and P_2 are obtained with the size $w \times h \times c$ (width $\times$ height $\times$ number of channels). The correlation $cor(x_1, x_2)$ of two feature map patches p_1 and p_2 centered on x_1 and x_2 respectively is defined as

$$cor(x_1, x_2) = \sum_{a \in [-k,k] \times [-k,k]} \langle p_1(x_1 + a), p_2(x_2 + a) \rangle \quad (1)$$

for a squared patch of size $K = 2k+1$. This operation is similar to convolution, and the two feature patches are multiplied and then added, but the operation does not require training weights. It can be known from (1) that the computational complexity of the cross-correlation operation of two feature patches is $c \times K \times K \times (w \times h) \times (w \times h)$. In order to reduce the computational complexity, we limit the matching range of the feature patches. Each feature patch x_1 performs a cross-correlation operation with a neighboring block of the corresponding point on feature map P_2, and the neighboring block length is $D = 2d + 1$. Therefore, we can obtain D^2 feature maps by performing cross-correlation operations. In this paper, we set $d = 5$, $D^2 = 121$, i.e., the number of output channels is 121. We also perform a 1×1 convolution operation on the advanced features of the RGB image, and the output is 32 channels, which then is combined with the output of the correlation layer.

The second CNN module takes the above combined features as input, and includes six convolution layers (conv3_1, conv4, conv4_1, conv5, conv5_1, conv6), where the odd layer has a step size of 2, the even layer has a step size of 1, with convolution kernel that has size 3×3. Each layer is followed by a ReLU activation function.

To achieve pixel-level dense depth prediction, we use a refinement module which employs an upconvolutional layer for unpooling and convolution. The idea of refinement module is very similar to the fully convolutional networks proposed in [20].The network predicts dense depth map on small-sized feature maps while backward deconvolution operations. The bilinear interpolation is used to resize the dense depth map, which is then combined with the features obtained by the deconvolution operation. The deconvolution operation is performed five times, and the resolution of the predicted dense depth map is one-half of the input image. A full resolution of dense depth map can be obtained by a bilinear interpolation. This refinement module preserves the prediction information of scene depth, and absorbs the multi-level feature information of the network. Therefore, the refinement module can correct the dense depth prediction multiple times.

B. Loss Function

The depth estimation network outputs five dense depth maps with different size (s = 5). If $\hat{D}_s$ represents the predicted dense depth map, we adjust the input sparse depth map d_s to be the same size as $\hat{D}_s$. Let p denote a pixel coordinate in the sparse depth map d_s, we train our network by minimizing the loss function L_{depth}:

$$L_{depth} = \sum_s \sum_p \left| \hat{D}_s(p) - d_s(p) \right|, \quad (2)$$

IV. EXPERIMENTS

We implement the proposed neural network using PyTorch. Our network model was trained on NYU-Depth-v2 dataset using TITAN X (Pascal) GPUs with 12GB of memory. Weight decay is applied to all weights and biases of the convolution layers. We evaluate our method using the similar metrics in [1]. During training, the training data is augmented

TABLE I: Depth estimation comparison on the NYU-Depth-v2 dataset.

Input	Sparse Depth Samples	Method	REL	RMSE	δ_1	δ_2	δ_3
RGB	0	Eigen et al.[6]	0.158	0.641	76.9	95.0	98.8
	0	Laina et al.[9]	0.127	0.573	81.1	95.3	98.8
RGBd	225	Liao et al.[18]	0.104	0.442	87.8	96.4	98.9
	100	Ma et at.[1]	0.056	0.278	95.8	99.0	99.7
	200	Ma et at.[1]	0.050	0.245	97.0	99.4	99.8
	500	Ma et at.[1]	0.044	0.238	97.4	99.4	99.8
	100	Ours	0.055	0.255	96.1	99.2	99.8
	200	Ours	0.046	0.219	97.3	99.5	99.9
	500	Ours	**0.040**	**0.183**	**98.1**	**99.7**	**99.9**

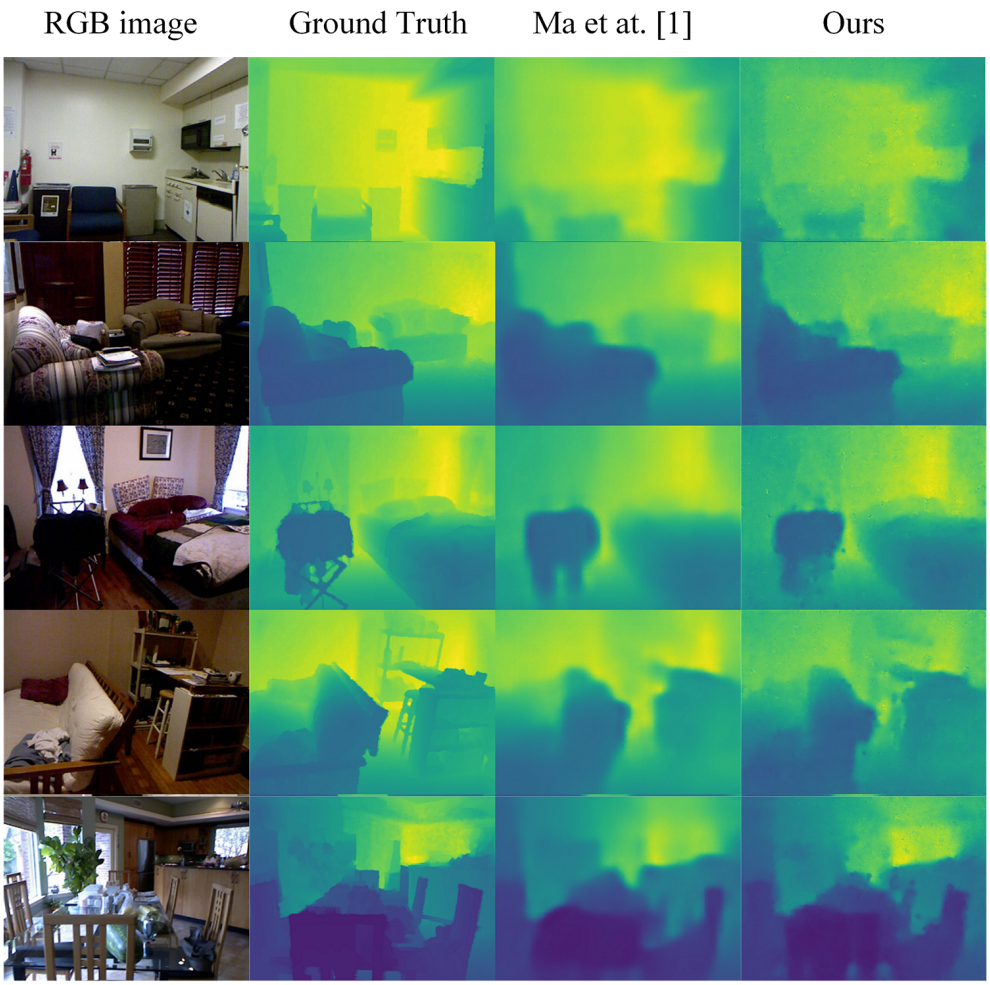

Fig. 2: Depth estimation on NYU-Depth-v2 dataset with 1000 sparse depth samples and RGB images. From left to right: RGB image, ground truth depth map, depth estimation result in [1], our depth estimation result.

in an online manner with random transformations, i.e., scaling and Random cropping.

For the NYU-Depth-v2 dataset, we use the official toolbox to project the depth values onto the image plane, and draw out using the cross-bilateral filter, then sparsely sample the depth image to generate the sparse depth map. The original frame which has size 480×640 is then down-sampled to a size of 256×512. To demonstrate the performance of our method, we compare our dense depth estimator with RGB-based approaches [9][6], i.e., the fusion approach [18]

that uses an 2D laser scanner mounted on a mobile robot, and recent published sparse-to-dense approach [1] on the NYU-Depth-v2 dataset. The quantitative comparison of depth estimation results are summarized in Table I, which clear show that the additional sparse depth image can improve the prediction accuracy compared to the RGB-based approaches, i.e., first three rows compared to last six rows. Moreover, our method outperforms the fusion approach and the sparse-to dense approach for the same number of sparse depth samples. Therefore, we conclude that improved performance can be achieved using the proposed idea, i.e., correlation layer based regression neural network for structure feature matching. The qualitative comparison results between the sparse-to-dense method and our method are shown in Fig. 2, where we can see a refined dense depth map is achieved using the proposed method, e.g., the structure of the environment is more clear.

V. Conclusion

In this paper, we propose an end-to-end learning scheme for estimating dense depth images from RGB images and sparse depth maps. The proposed idea is to use a correlation layer in the regression neural network to find the similar structure information between the RGB image and depth image, which can help to refine the structure information of the estimated depth image. The experiment results on the public datasets prove the novelty of our idea, i.e., the NYU-Depth-V2 dataset. The performance of our method is superior to the state of the art methods due to the refined depth structure.

References

[1] Ma, F. and Karaman, S.: 'Sparse-to-dense: Depth prediction from sparse depth samples and a single image', *IEEE International Conference on Robotics and Automation*, 2018, pp. 1-8.

[2] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R., 'Indoor segmentation and support inference from rgbd images', *European Conference on Computer Vision*, 2012, pp. 746-760.

[3] Geiger, A., Lenz, P., and Urtasun, R., 'Are we ready for autonomous driving? the kitti vision benchmark suite', *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[4] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-d scene structure from a single still image," 11 2007, pp. 1–8.

[5] Eigen, D., Puhrsch, C. and Fergus, R., 'Depth map prediction from a single image using a multi-scale deep network', *NIPS*, 2014, **2**, pp. 2366-2374.

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

[7] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2015, pp. 1119–1127.

[8] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[9] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N., 'Deeper depth prediction with fully convolutional residual networks', *Fourth International Conference on 3D Vision*, 2016, pp. 239-248.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *CoRR*, vol. abs/1506.02025, 2015. [Online]. Available: http://arxiv.org/abs/1506.02025

[11] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 2215–2223.

[12] R. Garg, V. K. B. G, and I. D. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," *CoRR*, vol. abs/1603.04992, 2016. [Online]. Available: http://arxiv.org/abs/1603.04992

[13] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," 04 2017.

[14] Godard, C. O., Aodha, M. and Brostow, G. J., 'Unsupervised monocular depth estimation with left-right consistency', *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6602-6611.

[15] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *CoRR*, vol. abs/1709.06841, 2017. [Online]. Available: http://arxiv.org/abs/1709.06841

[16] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," *CoRR*, vol. abs/1607.06349, 2016. [Online]. Available: http://arxiv.org/abs/1607.06349

[17] C. Cadena, A. Dick, and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding," 06 2016.

[18] Liao, Y., Huang, L., Wang, Y., Kodagoda, S., Yu, Y., and Liu, Y., 'Parse geometry from a line: Monocular depth estimation with partial laser observation', *arXiv:1611.02174*, 2016.

[19] Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., Smagt, P. v. d., Cremers, D., and Brox, T., 'Flownet: Learning optical flow with convolutional networks', *IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[20] Long, J., Shelhamer, E. and Darrell, T., 'Fully convolutional networks for semantic segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39: 4**, pp. 640-651.