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

Unsupervised Monocular Depth Estimation with Encoder-decoder Network*

Defang Chen, Xin Ma* and Yibin Li
School of Control Science and Engineering
Center for Robotics
Shandong University
Jinan, 250061, China
maxin@sdu.edu.cn

***Abstract*— Although fully supervised learning methods could achieve good results for monocular depth estimation, it is very expensive to obtain ground truth depth data for training. In this paper, an encoder-decoder network is proposed to estimate monocular depth with unsupervised learning. During training, this network is able to enhance the weight of the effective features in the image by learning. Therefore, the encoder is able to extract more effective features and the decoder generates more targeted results. Compared to supervised learning, the proposed network is trained by stereo pairs and image reconstruction error, rather than ground truth depth. The proposed network achieves competitive results for monocular depth estimation on the KITTI driving dataset.**

***Index Terms*— CNN, Depth estimation, Unsupervised learning.**

I. Introduction

In applications such as robotics and autonomous driving, depth data is usually obtained by Lidar. However, the cost of obtaining depth data via Lidar is very high. In the past few decades, depth estimation from a single image has attracted increasing attentions. Depth estimation from a single image can not only replace Lidar in some applications, but also can be used for pose estimation, 2D to 3D transformation, and so on. Although various methods that depth estimation from a single image have been proposed, they still cannot achieve the desired results.

Traditional depth calculation methods based on binocular images or multi-view images have been widely studied. These methods estimate the depth of each pixel by simultaneously acquiring the images of the left and right cameras, calculating the disparity and stereo matching. For monocular images, it is necessary to obtain observations of multiple angles of the estimated scene, so this method cannot be applied to fields such as robots or autonomous driving. In recent years, convolutional neural network (CNN) are very popular in image processing applications, and some methods based on CNN have been proposed to solve scene depth estimation of monocular images [1]–[3]. These methods use the ground truth depth as the supervisor training CNN prediction model, and predict the scene depth through the trained model. The training of this model requires a large number of datasets with the truth value of the scene depth. The acquisition of these datasets requires the Lidar or the depth camera, so the cost of acquiring the dataset is very expensive and only suitable for predicting the depth of scenarios similar to dataset. In order to overcome these problems, some unsupervised methods have become a hot topic of research [4]–[6]. Garg et al. [4] use an unsupervised CNN model to estimate the depth of single image, which is equivalent to the fully supervised method. The method does not rely on ground truth depth but uses photometric warp error as a supervision of the CNN model.

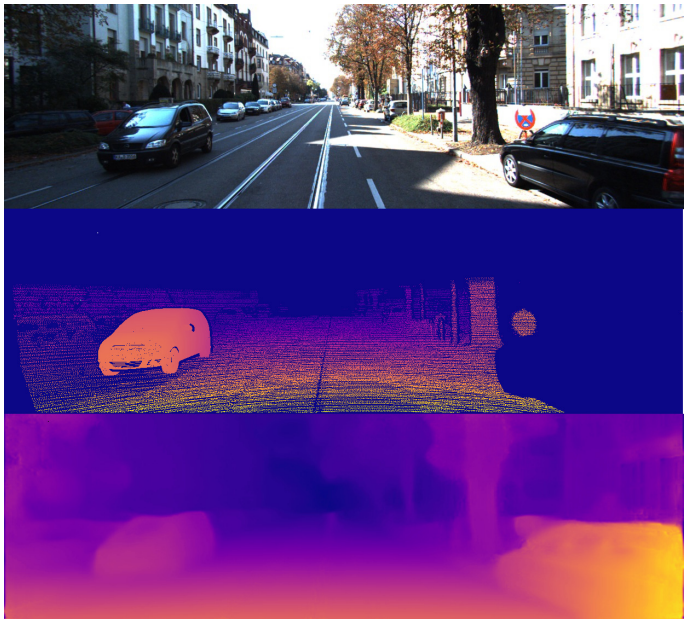

Fig. 1. The result of the depth estimation proposed in this article. Top to bottom: Input monocular image, ground truth disparity, and our result. The ground truth disparity is spare and local, but our result is dense and full-scale.

*This work was supported by the National Key Research and Development Program of China (No. 2018YFB1305803), National 863 High-Tech Program of China (No. 2015AA042307) and the Fund of National Nature Science Foundation of China under Grant No. 61673245.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

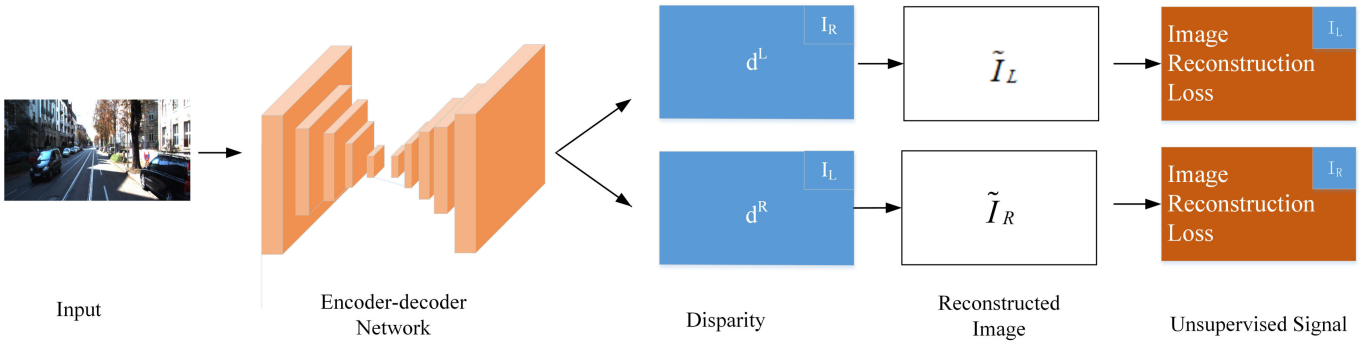

Fig. 2. The proposed monocular depth estimation framework.

On this basis, an unsupervised monocular depth prediction framework is proposed by Godard et al. [5]. The framework reconstructs the left and right images by photometric warp, and proposes the left and right consistency loss of disparity to train the model. This method achieves better results than the fully supervised methods on the KITTI dataset [7], [8] and achieves the best results on the Cityscapes dataset. However, the feature extraction capability of this method is weak, and it is impossible to distinguish unhelpful features in the image, such as the sky.

To overcome the problems mentioned, we propose an unsupervised monocular depth estimation model based on a powerful encoder decoder network. Fig. 1 shows an example of our results. The model uses image reconstruction errors as a supervisor, instead of ground truth depth. The proposed network uses a full convolution structure in the form of encoder-decoder, and adds the "Squeeze-and-Excitation" (SE) block to the encoder-decoder structure [9] innovatively. During training, the SE block structure can increase the weight of effective image features and reduce the invalid or small image feature mapping weight by learning feature weight. At the time of model training, our method still turns the depth estimation into an image reconstruction problem.

II. Related Works

There are many methods for predicting depth based on images. These methods train the CNN model to predict the depth of the image in a large number of monocular images, binocular image pairs or video sequences. Among the methods for predicting depth of monocular images, there are mainly methods based on supervised learning and methods based on unsupervised learning.

A. Supervised monocular depth estimation

Liu et al. [1] present a deep convolutional neural field model for estimating depth from single monocular images. Eigen et al. [2] performs depth estimation using rough global prediction for the entire image and local refinement prediction. Li et al. [3] uses conditional random field (CRF) to improve the accuracy of depth estimation. Kendall et al. [10] present an end-to-end deep learning architecture to predicts disparity from rectified image pairs of stereo. These methods require a large number of high quality images and their corresponding truth depth values when training the network, but the cost of obtaining ground truth depth for training is very difficult. At the same time, predicting depth from a single image is an uncertain problem, since an image can correspond to multiple depths, and these depth values may also be reasonable. For these problems, we prefer to trains the model through geometric relationship between images, instead of using ground truth depth.

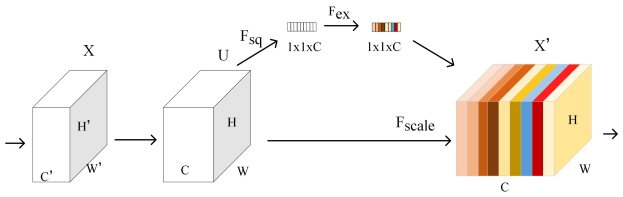

Fig. 3. SE Block.

B. Unsupervised monocular depth estimation

Garg et al. [4] reconstructs the source image using the predicted depth and the known inter-view displacement, and the photometric error is used as a supervisor for the network in the reconstruction. Godard et al. [5] present unsupervised monocular depth estimation with left-right consistency. Zhan et al. [6] use deep feature reconstruction to predict monocular depth.

III. Methods

The proposed monocular depth estimation framework is shown in Fig. 2. An encoder-decoder network that combines SE blocks to extract features and generate disparity value, image reconstruction and three loss functions for model training are introduced in this section.

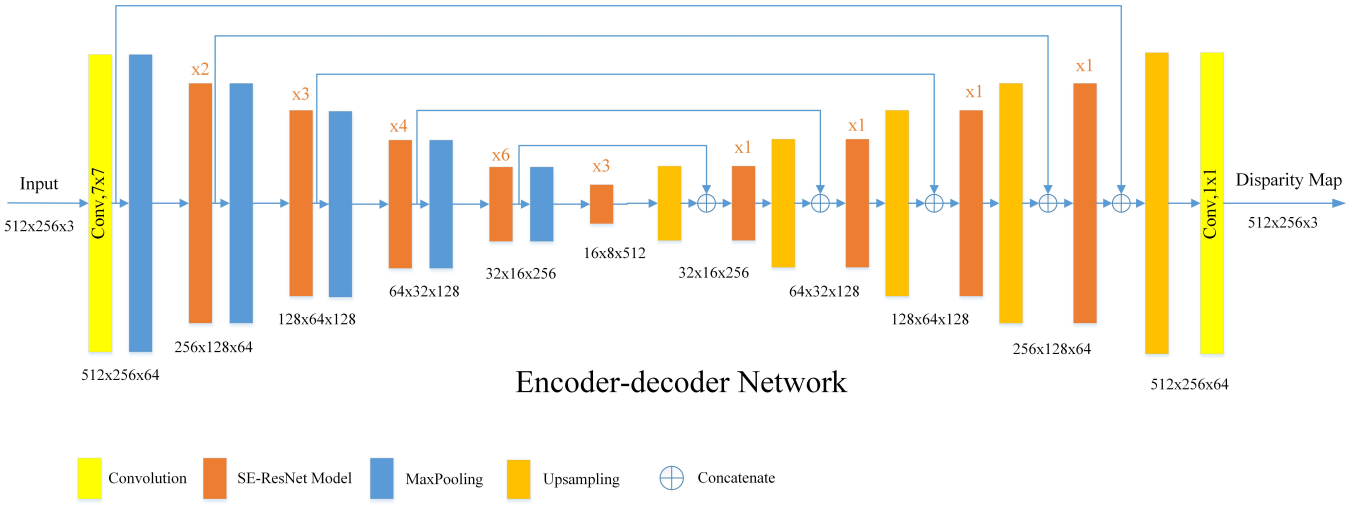

Fig. 4. The proposed Encoder-decoder network.

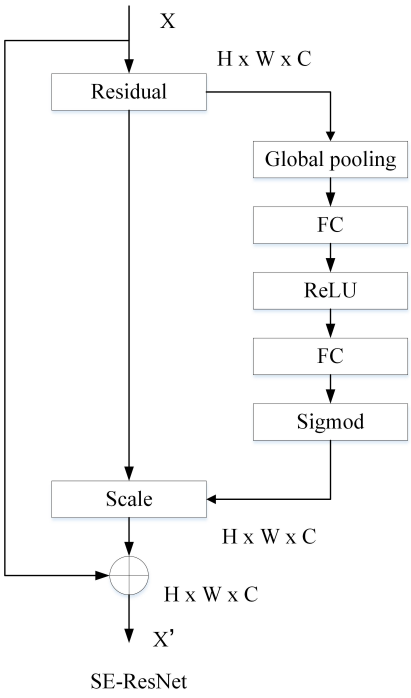

Fig. 5. SE-ResNet Block.

A. *Encoder-decoder network*

The success of CNN in image processing has been fully reflected in previous work. The powerful feature extraction capabilities of CNN can extract a lot of useful information from images. However, the feature extraction capabilities of different CNN frameworks are different. From VGG [11], Inception models [12] to ResNet [13], the feature extraction capabilities of CNN have been improved significantly. In order to extract more features, a lot of works continue to deepen the number of network layers, but deeper CNN is not necessarily useful. Because there are a lot of useless features in the image for depth estimation, such as sky. To overcome this problem, we add SE block to encoder-decoder network. The SE block is shown in Fig. 3. F_{sq} denote global average pooling, as in

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(\mathrm{i}, \mathrm{j}). \tag{1}$$

F_{ex} denote two fully connected (FC) layers, as in

$$F_{ex}(z, W) = \sigma\left(W_2 \delta\left(W_1 z\right)\right), \tag{2}$$

where σ refers to the Sigmod function and δ refers to the ReLU [14] function. The final output of SE block is the adjusted U, as in

$$X_c' = F_{scale}(u_c, s_c). \tag{3}$$

The principle of SE block is to adjust the size of the scale to enhance the important features and weaken the unimportant features, so that the extracted features are more directional. The SE-ResNet obtained by combining SE block and ResNet, as shown in Fig. 5. The SE-ResNet not only can the feature extraction ability of the deep network be enhanced, but also the influence of useless features in the image can be reduced.

The proposed encoder-decoder network, as shown in Fig. 4, consists of an encoder and a decoder. The encoder consists of SE-ResNet blocks for extracting monocular image features. The decoder uses a combination of upsampling and SE-ResNet block to improve resolution. High resolution features are obtained by up-sampling low-resolution features, and then features are learned by convolution. Furthermore, the decoder resolve higher resolution details by uses skip connections to

concatenate the features of encoder with the features obtained by upsampling. Finally, the decoder generates left and right disparity map with the same resolution as input. Therefore, depth (z) can be obtained from known baseline distance (b) of the binocular camera, camera focal length (f) and predicted disparity (d), as in

$$z = \frac{bf}{d}. \tag{4}$$

B. Image reconstruction

During training, A large number of image pairs I_l and I_r that captured from the same binocular camera as input to network, and the binocular camera was already calibrated. In Section III(A), the encoder-decoder network has generated left-view disparity (d_l) and right-view disparity (d_r) from a single image. The geometric relationship between the binocular image pair and the disparity, as in

$$d_l = u_l - u_r, \tag{5}$$

where u_l denote the left coordinate of the spatial point P on the left-view, u_r denote the right coordinate of the spatial point P on the right-view. According to geometric relationship between the binocular image pair and the disparity, the right image can be reconstructed by left image and right-view disparity, as in

$$\tilde{I}_r = I_l\left(d_r\right). \tag{6}$$

In the same way, the left image can be reconstructed by right image and left-view disparity, as in

$$\tilde{I}_l = I_r\left(d_l\right). \tag{7}$$

Therefore, the proposed unsupervised method does not require ground truth depth as the supervision, but instead uses the error between the reconstructed images ($\tilde{I}_l, \tilde{I}_r$) and the original images (I_l, I_r).

C. Training loss

The main supervised signal is image reconstruction error, as shown in Section III(B), instead of ground truth depth. Furthermore, similar to [6], disparity smoothing loss and left-right consistency are used as auxiliary supervision. The loss function is shown in

$$L = \alpha L_{ir} + \beta L_{ds} + \gamma L_{dc}, \tag{8}$$

where L_{ir} denote image reconstruction loss, L_{ds} denote disparity smoothness loss, L_{dc} denote disparity consistency loss, α, β, γ denote weights.

1) Image reconstruction loss: During training, the network learning generates disparity value, and reconstructs new left ($\tilde{I}_l$) and right ($\tilde{I}_r$) image from the disparity value and left (I_l) and right (I_r) image. The error between the reconstructed image and the original image supervise the accuracy of the network to generate the disparity map, instead of ground truth depth. Similar to [15], image reconstruction loss is calculated according to SSIM [16], as in

$$L_{ir} = \frac{1}{N}\sum_{i,j}\alpha\frac{1 - SSIM\left(\mathrm{I}_{ij}, \tilde{\mathrm{I}}_{ij}\right)}{2} + (1-\alpha)\left\|\mathrm{I}_{ij} - \tilde{\mathrm{I}}_{ij}\right\|. \tag{9}$$

2) Disparity smoothness loss: Similar to [17], we use disparity smoothness loss to encourages the predicted disparity to be smooth, as in

$$L_{ds} = \frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}\right| e^{-\|\partial_x I_{ij}\|} + \left|\partial_y d_{ij}\right| \mathrm{e}^{-\|\partial_y I_{ij}\|}. \tag{10}$$

3) Disparity consistency loss: Inspired by [5], the left-right disparity consistency loss makes the left disparity value and the mapped right disparity value equal, as in

$$L_{lr}^{l} = \frac{1}{N}\sum_{i,j}\left|\mathrm{d}_{ij}^{l} - \mathrm{d}_{ij+d_{ij}^{l}}^{r}\right|. \tag{11}$$

IV. Experiment

A. Experimental Details

The unsupervised monocular depth estimation with encoder-decoder network in this paper is realized by deep learning open-source framework TensorFlow (1.12.0) [18], and using two RTX2080 GPU. During the training, network spends 22 hours to training 26 million trainable parameters for 50 epochs. The training set has a total of 30 thousand pairs of binocular images, and the image size is 512×256. During the training, the batch size is set to 8. Through some experimental comparisons, α, β and γ of loss function are 1, 1 and 0.025, respectively, the result is the best. Inspired by [19], nearest neighbour sampling and convolution are used instead of deconvolution in decoder network. This replacement overcomes the checkerboard effect of deconvolution and improves the accuracy of generating disparity. During optimization, the Adam optimizer [20] are used to optimize parameters of model for 50 epochs, where $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^{-8}$. Early, we use a variable learning rate, the initial value is $\lambda = 10^{-3}$. Learning rate is reduced by an order of magnitude every 10 epochs. We have found through experiments that it is very effective to use cosine annealing to reduce the learning rate. Cosine annealing reduces the learning rate as the cosine function decreases. Using cosine annealing, the learning rate drops to a small value when the network loss approaches the global minimum. In addition, Data augmentation works well in

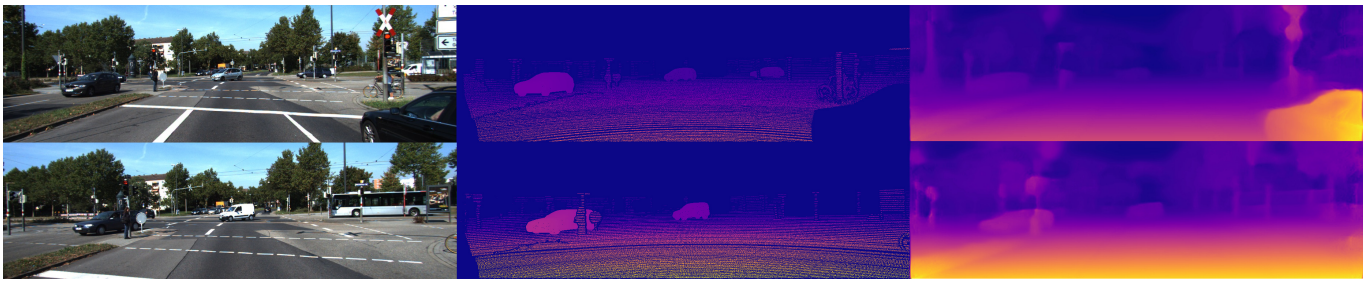

Fig. 6. Results of our depth estimation. Left to right: Inputs, ground truth, our results.

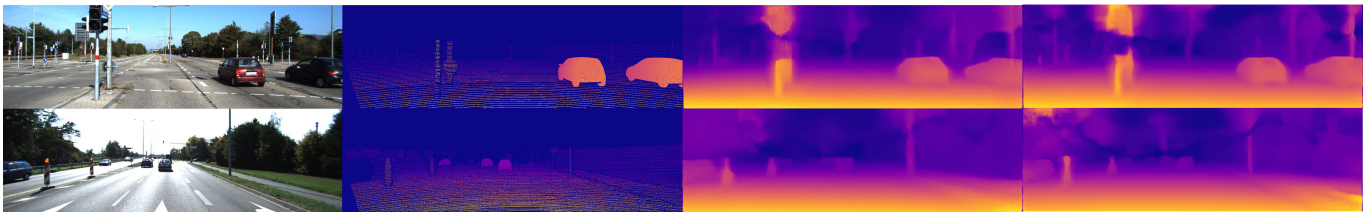

Fig. 7. Our results and results from existing mainstream unsupervised method. Left to right: Inputs, ground truth, Godard et al. [5], our results.

TABLE I

COMPARISON OF PERFORMANCE BETWEEN OUR METHOD AND EXISTING APPROACHES. K IS KITTI DATASET. THE RESULTS OF OTHER METHODS COME FROM THEIR PAPERS. FOR ABS REL, SQ REL AND RMSE, LOWER IS BETTER. FOR $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, HIGHER IS BETTER.

Method	Supervised	Dataset	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	No	K	0.361	4.826	5.102	0.638	0.804	0.894
Eigen et al. [2] Coarse	Yes	K	0.214	1.605	6.563	0.673	0.884	0.957
Eigen et al. [2] Fine	Yes	K	0.203	1.548	6.307	0.702	0.890	0.958
Liu et al. [1]	Yes	K	0.201	1.584	6.471	0.68	0.898	0.967
Garg et al. [4]	No	K	0.169	1.08	5.104	0.273	0.74	0.904
Godard et al. [5]	No	K	0.148	1.344	5.927	0.247	0.922	0.964
Godard et al. [5] pp	No	CS+K	0.118	0.923	5.015	0.854	0.947	0.976
Ours	No	K	0.115	0.912	4.975	0.86	0.945	0.976

CNN. Therefore, we randomly flip some of the images in the dataset and add noise to the image.

We use rectified binocular image pairs in the KITTI dataset to train encoder-decoder network. KITTI dataset contains total of 42 thousand rectified binocular image pairs from 61 scenes. During the training, 30 thousand rectified binocular image pairs in the dataset were selected as the training dataset. Furthermore, KITTI dataset provides 200 high quality disparity images and depth values from Lidar to validate predicted results.

B. Results

We evaluate results of the proposed method in this paper on KITTI dataset and compare with other depth estimation methods. Some of our results are shown in Fig. 6. We use color map to visually show different depths. Warm colors indicate a closer distance, and dark colors indicate a farther distance. Similar to [2], averaged relative error (Rel) and root mean squared error (RMSE) are main evaluation metrics between our results and ground truth. The Rel is shown in

$$Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|}{D_i^*}, \tag{12}$$

where N is the total number of pixels, D_i is the i-th pixel value of predicted depth image, and D_i^* is the i-th pixel value of ground truth depth image. The RMSE is shown in

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{1}(D_i - D_i^*)^2}, \tag{13}$$

where N is the total number of pixels, D_i is the i-th pixel value of predicted depth image, and D_i^* is the i-th pixel value of ground truth depth image.

In Fig. 6, our results clearly show the depth of most objects. Fig. 7 shows our results and results from existing mainstream unsupervised methods. In Table I, we use the above evaluation metrics to compare several depth estimation methods. From the table, the proposed model of unsupervised

monocular depth estimation not only does not rely on ground truth depth, but preforms better results.

V. Conclusion

An unsupervised monocular depth estimation model based on powerful encoder-decoder network is presented in this paper. The model is trained by rectified binocular image pairs in KITTI dataset, instead of ground truth depth. Our encoder-decoder network based on SE block extracted features are more directional, so that our method has achieved competitive results. In this paper, we illustrate proposed approach, network architecture, training, and results. In addition, we also compare with mainstream methods for monocular depth estimation.

Monocular depth estimation still has some challenges. The results based CNNs are far less accurate than Lidar. Light has a greater impact on depth estimation. In future work, models with higher accuracy still need to be studied.

References

[1] F. Liu, C. Shen, G. Lin and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 2024-2039, 2016.

[2] Eigen D, Puhrsch C, Fergus R, "Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network," in Conference on Neural Information Processing Systems, Montreal,2014, pp.2366-2374.

[3] Bo Li, Chunhua Shen, Yuchao Dai, A. van den Hengel and Mingyi He, "Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFs," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1119-1127.

[4] Garg R, BG V K, Carneiro G, et al, "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," in European Conference on Computer Vision, Amsterdam, 2016, pp. 740-756.

[5] C. Godard, O. M. Aodha and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6602-6611.

[6] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal and I. M. Reid, "Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 340-349.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset," International Journal of Robotics Research (IJRR), vol.32, no.11, pp. 1231-1237, 2013.

[8] A. Geiger, P. Lenz and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3354-3361.

[9] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7132-7141.

[10] A. Kendall, H. Martirosyan, S. Dasgupta and P. Henry, "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 66-75.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint, arXiv:1409.1556, 2015.

[12] C. Szegedy et al., "Going Deeper with Convolutions," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[14] Nair V, Hinton G E, "Rectified Linear Units Improve Restricted Boltzmann Machines," in Proceedings of the 27th international conference on machine learning (ICML-10), Haifa, 2010, pp. 807-814.

[15] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Is L2 a Good Loss Function for Neural Networks for Image Processing?" arXiv preprint, arXiv:1511.08861, 2015.

[16] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image Quality Assessment: from Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.

[17] P. Heise, S. Klose, B. Jensen and A. Knoll, "PM-Huber: PatchMatch with Huber Regularization for Stereo Matching," in IEEE International Conference on Computer Vision, Sydney, NSW, 2013, pp. 2360-2367.

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint, arXiv:1603.04467, 2016.

[19] A. Odena, V. Dumoulin, and C. Olah. Odena A, Dumoulin V, Olah C, "Deconvolution and Checkerboard Artifacts," Distill, 2016, doi.10.23915/distill.00003.

[20] Kingma, Diederik P., and Jimmy Ba., "Adam: A Method for Stochastic Optimization," arXiv preprint, arXiv:1412.6980, 2014.