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

Binocular Depth Estimation Using Convolutional Neural Network With Siamese Branches

Guodong Liu[1,2,3,4], Guolai Jiang[1,2,3,4], Rong Xiong[1,3,4,*] and Yongsheng Ou[1,3,4,*]

***ABSTRACT* - Binocular depth estimation is a hot research topic in computer vision. Traditional methods need high precision camera calibration and key point matching, but the results are not ideal. In this paper, we introduce an approach of binocular depth estimation method based on deep learning. A new convolutional neural network is designed, which consists of two sub-networks. The first sub-network is a deep network with siamese branches and 3D convolutional layer, it learns parallax and global information and generates a global depth estimation result in low resolution. The second is a fully convolutional deep network, which reconstructions the depth map to original resolution. The two sub-networks are connected by a pool pyramid. Experiments are taken on the Middlebury Stereo Dataset show that the proposed method can generate much more accurate depth image than traditional methods.**

Index Terms - Binocular depth estimation; Siamese branches; 3D convolution; pool pyramid.

I. INTRODUCTION

Accurate depth information is important for understanding the object geometric relations of the surrounding environment. In order to obtain depth information, depth estimation was proposed and image-based depth estimation became a hot research topic. And it has been widely applied in robotics [1], autopilot [2], 3D modeling [3] and other fields.

Depending on the number of images used, Image-based depth estimation methods can be divided into three types, i.e., monocular depth estimation [4], binocular depth estimation [5, 6, 7] and multi-ocular depth estimation.

Generally speaking, monocular depth estimation relies on uses experience, i.e. Determining the distance of an object by considering the known original size, the size of the object in the image, the viewing angle of the object in camera, as well as the reference target around the object, the occlusion and shadows, etc. And the known original size of the object is learned from data sets. However, monocular depth estimation has scale problem inherently. Estimating the depth only by experience does not work well in many cases, even humans will make mistakes when they are trying to estimate distance with one eye.

[1]The authors are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China. {gd.liu, gl.jiang, rong.xiong, ys.ou}@siat.ac.cn
[2]University of Chinese Academy of Sciences, Beijing 100049, China.
[3]Guangdong Provincial Key Lab of Robotics and Intelligent System,Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.
[4]CAS Key Laboratory of Human-Machine Intelligence-Synergic Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China.
Rong Xiong* and Yongsheng Ou* are the corresponding authors.
This work was jointly supported by National Natural Science Foundation of China (Grants No. U1613210), Guangdong Special Support Program (2017TX04X265), Science and Technology Planning Project of Guangdong Province (2019B090915002) and Shenzhen Fundamental Research Programs (JCYJ20170413165528221).

Whereas, Human can use two eyes to obtain the depth of an object directly in the brain. Scale problem does not exist in binocular depth estimation. Binocular depth estimation usually uses the parallax [5, 7] of two images to calculate the depth of each point.

In this paper, a binocular depth estimation method by using a deep neural network that receives two images to learn parallax, scale experience and other features. The main contributions of this paper are as follows: 1) A network architecture with siamese branches performs end-to-end binocular depth estimation, the siamese branches mean that the left view and right view use the same feature extraction method. 2) In order to learn parallax between left view and right view, a 3D convolutional layer is used to learn the features of sequence images from the 'frame' dimension. 3) A fully convolutional network and a pool pyramid are used to optimize the result.

II. RELATED WORKS

Image-based depth estimation has been wildly studied since the birth of computer vision. At the beginning, depth estimation relies on multi-views or dual-views, such as recover structure from motion [8], binocular semi-local matching [5], stereo matching [6]. Although these methods are effective in many cases, they depend on the performance of key point matching and do not perform well in texture less environment. Moreover, key points matching is very sensitive to image noise. In addition, such algorithms need high precision camera calibration before implementation.

There are also some studies focusing on depth estimation from a single image, such as calculating depth through light spot change [9], shape from light and shading [10], or Markov random field [11], etc. In recent years, with the rapid development of deep learning, it is also used for image-based depth estimation. David Eigen et al. first proposed utilizing the deep learning for depth estimation [4]. Iro Laina et al. used fully convolutional neural network to obtain better depth estimation results [12]. However, such single image methods are to learn the scale information of the object itself and the scale of the reference object in the image and some other information synthetically to judge the depth of each pixels, which requires a large training data set and the generalization ability of these methods is poor.

There are also some other ideas for depth estimation based on deep learning[13]. For example, [14] combined the task of semantic segmentation with depth estimation, where the predicted labels are used as additional constraints to facilitate the optimization task. [15] developed a novel classificational approach instead of regression, which combined the predict labels and depths.

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

Furthermore, many other methods can obtain the depth information directly through hardware, such as Microsoft Kinect series and Intel RealSense series. In these methods, the distance is measured by using the transmitting and receiving infrared light. Although such hardware-based methods have the drawback of missing data at the out of range areas or the edge areas, they are the most popular strategies of getting the ground truth.

III. METHOD

In our work, the depth estimation is regarded as a pixel-by-pixel image regression problem. And it is solved by deep neural network.

On one hand, pixels of the same color on the same object in the image can have different depth values, this requires that the network structure can learn the global information of the image. On the other hand, the average values of the depths of different objects may vary greatly, which makes the depth estimation task is similar to image semantic segmentation in some degree.

The network in this paper consists of two Sub-networks: siamese convolutional network and fully convolutional network. The siamese convolutional network is used to obtain the global depth image information, and the fully convolutional network is used to optimize the local depth image information.

A. Network I: Siamese Convolutional Network

The achievements of siamese networks in image similarity regression task and others are impressive [16]. Here, this paper improves the siamese networks for binocular depth estimation, left view and right view are input into two branches of siamese networks respectively, and the network can learn binocular matching and disparity information as far as possible on the 3D convolutional layer, and the fully-connected is used to predict the global depth in the end of the network.

As shown in Figure 1, the structure of network 1 can be divided into three parts. The first part is siamese part, which consists of two 2D convolutional layers. The input of the siamese part is left view image and right view image respectively, and the features of the two images are extracted separately in the same parameter value. Namely, the main point of siamese network is parameter value sharing, which means that the left view and right view use the same feature extraction method, it makes parallax learning part can learn effectively parallax information. In order to get better parallax learning results, maximization pooling is not set in the network layer of siamese part. This is because maximization pooling will make the feature position shift. In this part, we will get the feature maps of left view and right view after feature extraction, and combine them on the new dimension 'frame' dimension to generate a feature map of 'two frames' as the input of the second part.

The second part is parallax learning, which consists of a 3D convolutional layer and two 2D convolutional layers. Namely, we use a 3D convolutional kernel with a 'frame' dimension of 2 and a step size of 2 to extract features. The 3D convolution can learn the features of sequence images from the 'frame' dimension. The 3D convolution can learn the position changes between left feature maps and right feature maps, that is, the information related to parallax. At the same time, the 3D convolution also learns intra-frame 2D features. The number of feature maps output from this layer is 1 on the 'frame' dimension. We remove this dimension and input it into the next two layer of 2D convolutional layers to continue extracting features. Moreover, the two 2D convolutional layers contain a shortcut, which makes the two layers more effective in extracting features through learning residuals [17]. Then, flatten the feature map.

The third part is the fully-connected, which consists of a fully-connected hidden layer and a fully-connected output layer, it can learn global features, reshape output layer and a low resolution depth estimation image is obtained.

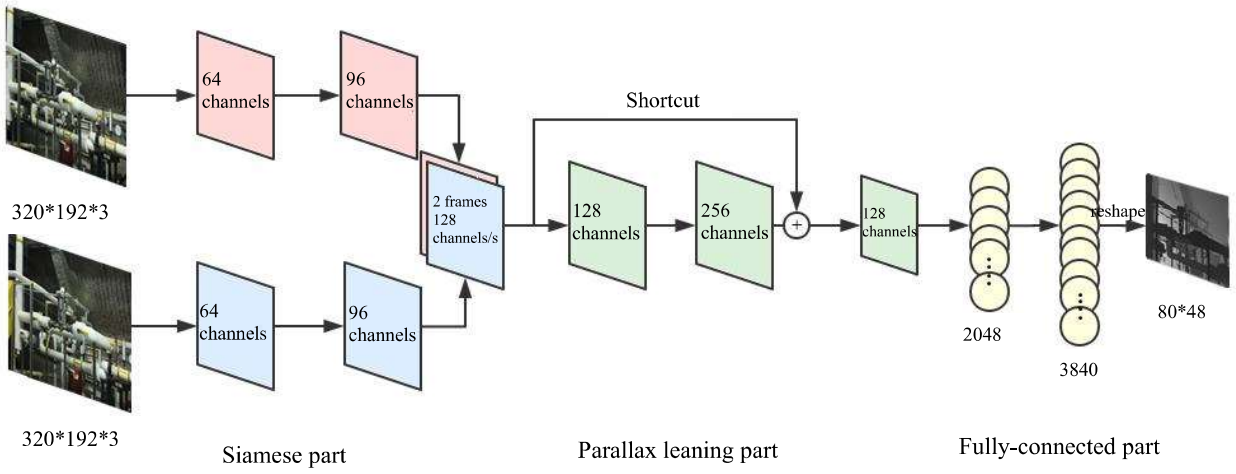

Figure 1 Sub-network 1: Siamese Convolutional Network, the operations and parameters are shown in the TABLE I

The all operations and parameters in the sub-network 1 is shown in TABLE I.

TABLE I
THE OPERATIONS AND PARAMETERS IN THE SUB-NETWORK 1

Parts	Layers	operations	parameters
Simese part	Hidden Layer 1	2D convolution	Kernel=7*7*64 Strides=2*2
		Activate function	Leaky ReLU
	Hidden Layer 2	2D convolution	Kernel=5*5*96 Strides=2*2
		Activate function	Leaky ReLU
Parallax learning part	Hidden Layer 3	3D convolution	Kernel=2*7*7*128 Strides=2*2*2
		Activate function	Leaky ReLU
	Hidden Layer 4	2D convolution	Kernel=5*5*256 Strides=1*1
		Activate function	Leaky ReLU
	Hidden Layer 5	2D convolution	Kernel=3*3*128 Strides=1*1
		Shortcut	Layer3+conv5
		Activate function	Leaky ReLU
		2*2 Max pooling	Strides=2*2
Fully-connected part	Hidden Layer 6	Activate function	Leaky ReLU
	Output Layer 7	Reshape	80 * 48

The sizes of the feature map in the sub-network 1 is shown in TABLE II.

TABLE II
THE SIZE OF THE FEATURE MAP IN THE SUB-NETWORK 1

Parts	Layers	Sizes (width * height * channels)
Input Layer		320*192*3
Simese part	Hidden Layer 1	160*96*64
	Hidden Layer 2	80*48*96
Parallax learning part	Hidden Layer 3	40*24*128
	Hidden Layer 4	40*24*256
	Hidden Layer 5	20*12*128
Fully-connected part	Hidden Layer 6	2048
	Output Layer 7	3840(80*48)

B. Network II: Fully Convolutional Network

Fully convolutional network is often used in image semantics segmentation task [18]. It is merely composed of convolutional layer or deconvolutional layer. Generally speaking, this network structure can keep the spatial location information and granularity, especially in the case of shallow network. The fully convolutional network in this paper consists of five layers of convolutional layer and one layer of deconvolutional output layer. The structure of the sub-network 2 is shown in Figure 2. It can be seen that the function of sub-network 2 is similar to super-resolution convolutional neural network[19]. The first layer is responsible for feature extraction from image. Then, the input of the second layer is composed of the output of the siamese convolutional network and the output of the first layer in this network. After all the convolutional layers, the deconvolutional layer reconstructs the resolution of the output image to the resolution of the input image and the high precision depth estimation is obtained. The idea of PSPNet's (Pyramid scene parsing network) pyramidal pooling [20] is applied in the step of combining the output of siamese convolutional network with the output of the first convolutional layer. Firstly, four new channels for combination are generated from the output of siamese convolutional network. The generation method of the four new channel is to linearly interpolate the global depth map of the output of siamese convolutional network to the required resolution as the first new channel, to pool the global depth map of the output of siamese convolutional network to 1*1 resolution and nearest neighbor interpolate to the required resolution as the second new channel. The same method is applied to the third new channel and the fourth new channel as used in the second new channel. The purpose of doing so is that depth estimation is sensitive to global depth scale and large local depth scale. And first new channel can provide depth reference of each pixel. The required resolution refers to the resolution of the feature map output from the first convolutional layer. In actual operations, it is necessary to increase the weight of the first new channel and decrease the weight of the other three new channels.

The all operations and parameters in the sub-network 2 is shown in TABLE III.

TABLE III
THE OPERATIONS AND PARAMETERS IN THE SUB-NETWORK 2

Layers	operations	parameters
Hidden Layer 1	Convolution	Kernel=7*7*60
	Batch normalization[21]	
	Activate function	Leaky ReLU
	2*2 Max pooling	Strides=2*2
Hidden Layer 2	Concatenate	pyramidal poolingt
	Convolution	Kernel=9*9*64
	Batch normalization[21]	
	Activate function	Leaky ReLU
Hidden Layer 3	Convolution	Kernel=7*7*96
	Batch normalization[21]	
	Activate function	Leaky ReLU
Hidden Layer 4	Convolution	Kernel=5*5*64
	Batch normalization[21]	
	Shortcut	Layer2+bn4
	Activate function	Leaky ReLU
Hidden Layer 5	Convolution	Kernel=5*5*96
	Batch normalization[21]	
	Activate function	Leaky ReLU
Hidden Layer 6	Convolution	Kernel=3*3*64
	Batch normalization[21]	
	Shortcut	Layer4+bn6
	Activate function	Leaky ReLU
Output Layer 7	Deconvolution	Kernel=5*5*1 Strides=2*2

The sizes of the feature map in the sub-network 2 is shown in TABLE IV.

TABLE IV
THE SIZE OF THE FEATURE MAP IN THE SUB-NETWORK 2

Layers	Sizes (width * height * channels)
Input Layer	320*192*3
Hidden Layer 1	160*96*60
Pyramidal Pooling	160*96*4
Hidden Layer 2	160*96*64
Hidden Layer 3	160*96*96
Hidden Layer 4	160*96*64
Hidden Layer 5	160*96*96
Hidden Layer 6	160*96*64
Output Layer 7	320*192*1

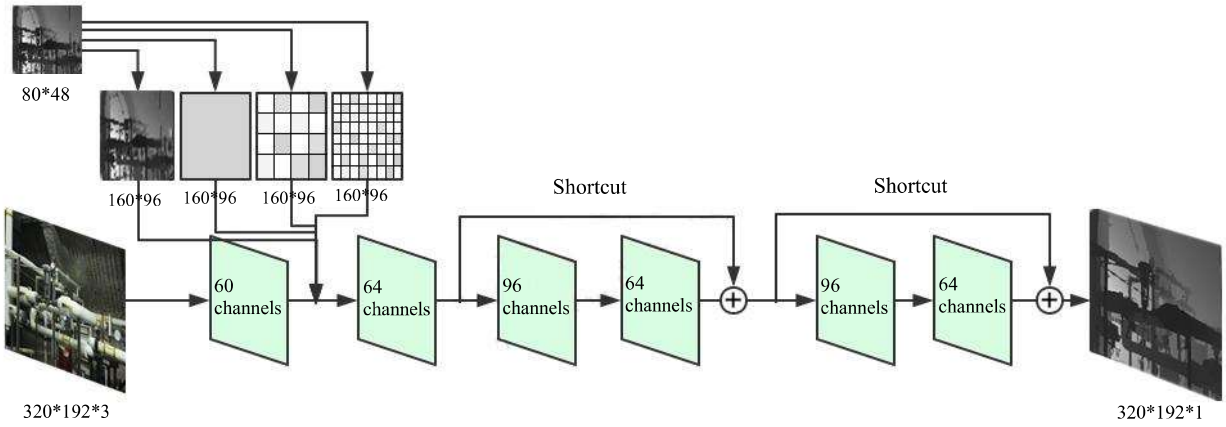

Figure 2 Sub-network 2: Fully Convolutional Network, the operations and parameters are shown in the TABLE III

In this paper, the two sub-networks need to be trained separately. Firstly, siamese convolutional network is trained, after a better result obtained trained fully convolutional networks to obtain the final high precision depth estimation results.

C. Loss Function

Loss function refer to L1 smooth loss, the loss of each pixel as follows:

$$\text{loss}_{\text{eachpix}} = \begin{cases} 0, & if\ y \le 0 \\ 0.5(y^*-y)^2, & if\ y>0\ and\ |y^*-y|<1 \\ |y^*-y|-0.5, & if\ y>0\ and\ |y^*-y|\ge 1 \end{cases} \quad (1)$$

Among them, y refers to ground truth, y* refers to the prediction results of the network. The ground truth of the depth estimation task generally has missing points and the value of missing points is 0. No loss calculation is carried out at these locations, which corresponds to the case of $Y \le 0$ of formula (1). In all cases of $Y > 0$, the number of all calculation points in each image is recorded as N and the average loss of the depth image is calculated by formula (2):

$$\text{loss} = \frac{\sum loss_{eachpix}}{N} \quad (2)$$

IV. EXPERIMENT

A. Data Augmentation

Five methods of data augmentation are applied in this paper:

1)The images is enlarged at random scale from 1.1 to 1.5, and then the original image size of image area is clipped at random position in the enlarged images. The same operations is done for the left eye and the corresponding right eye images and ground truth. Last but not least, the distance in ground truth divided by this random scale.

2)Random rotation transformation of - 7 ~ +7 was performed on the images, and the same operation was performed on the left eye images and the corresponding right eye images and ground truth.

3)Randomly change the whole brightness of the images by 0.77-1.2 times. Do the same operation for the left eye and the right eye images each time.

4)Gauss white noise with standard deviation of 0.001 was added to the left and right eye images.

5)Horizontal flips with 66.67% probability were used.

Except for the method 5), other methods are used many times in each experiment.

B. Experiments on MSD

This paper trains and evaluates this method on Middlebury Stereo Datasets [22, 23, 24], which is a high-precision binocular depth estimation data set. Each Ground Truth has been artificially corrected. This paper splits the data set into 85% training set and 15% test set, sets a small learning rate and batch size, and trains for 10 hours on a dual Intel Xeon E5 server to get the results. For sub-network 2, because it contains batch normalization, we set a larger batch size.

C. Results

Table V is the results of our method and traditional binocular stereo matching method (BM [6], SGBM [5]) and classical monocular depth Estimation method based on deep learning (E. David et al. [4]) on Middlebury Stereo Datasets [22]. In this paper, the resolution of the output depth image of Network 2 is 320*192. Because the resolution of the output depth image of Network 1 is 80*48, linear interpolation is used to enlarge it to 320*192. Similarly, because of the high resolution of Ground Truth, resize it to 320*192 for calculation.

We using four scale-invariant measurements evaluate each method in this paper, the four scale-invariant measurements are absolute relative difference, squared relative difference, root mean square error and root meansquare error log. The formulas of the four scale-invariant measurements are defined in reference [4]. It is worth at mentioning that we only evaluate the result of these method non-missing points of ground truth. We can see that the

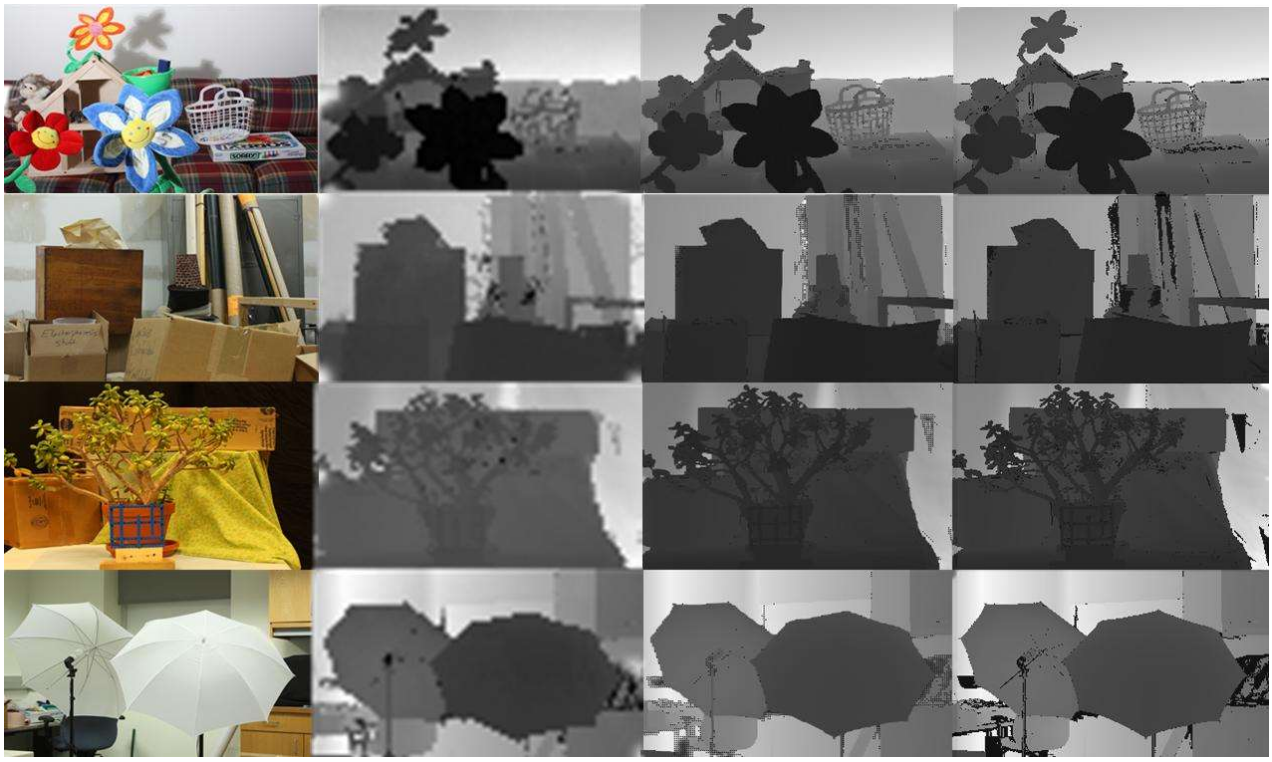

Figure 3 Results in Middlebury Stereo Dataset 2014 [22], the first column is the right view image, the second column is the output of our sub-network1, the third column is output of our entire network, the last column is the ground truth.

TABLE V
Comparing our method with the other methods

	BM([6])	SGBM([5])	E. David et al.[4]	Only sub-network 1	Our whole network
Abs Rel	2.63742	2.01571	0.17736	0.12709	**0.09736**
Sq Rel	11.84213	9.15700	1.47325	1.06545	**0.87335**
RMSE	149.6395	124.26890	6.67226	4.56353	**3.92954**
RMSE log	2.82405	2.2434	0.73527	0.43365	**0.21749**

propose method is much better than the traditional binocular stereo matching method BM and SGBM. The main shortcomings of traditional methods BM and SGBM lie in the key point matching : 1) The accuracy of key point matching is poor, which leads to the poor accuracy of depth estimation. 2) Mismatching of key points makes it impossible to estimate the depth of corresponding positions. However, the method of this paper relies on the powerful feature extraction ability of deep learning and uses a 3D convolutional layer to learn the parallax of the siamese branches feature images of left and right views. To some extent, this is equivalent to a dense matching with high robustness. Besides, Table V also shows that our method is better than the classical monocular depth estimation method based on deep learning (E. David et al.[4]), mainly because it is difficult for monocular to provide disparity information.

Figure 3 shows the results of our method on the Middlebury Stereo Dataset 2014 [22], Because the output depth image of sub-network 1 is enlarged, it is slightly blurred. The first column is the right view image, the second column is the output of our sub-network 1, the third column is output of our entire network, the last column is the ground truth.

V. CONCLUSION

This paper presented a binocular vision-based parallax learning deep neural network for binocular vision-based parallax learning deep neural network for binocular depth estimation. In order to retain the help of scale experience features and other features, this paper does not restrict the network to learn only parallax features. And experiments on Middlebury Stereo Dataset are carried out.

Based on the experiments, the proposed method achieves good results on Middlebury Stereo Datasets, which verified the effectiveness of our method. In addition, some further improvements can be made in the future. For example, better performance and generalization ability can be acquired by deepening the network layer of parallax learning part in the sub-network 1 or deepening the network layer of sub-network

2 [25], and utilizing the DenseNet idea [26] can also achieve better results and better generalization ability.

In the future, the multi-task learning combined with depth estimation and image semantics segmentation [27] will be considered. The visual features by these two tasks are similar, which may promote each other to make the deep network faster learn in the right direction.

ACKNOWLEDGMENT

I would like to express my gratitude to Sheng Xu, he helped me during the writing of this paper.

This work was jointly supported by National Natural Science Foundation of China (Grants No. U1613210), Guangdong Special Support Program (2017TX04X265), Science and Technology Planning Project of Guangdong Province (2019B090915002) and Shenzhen Fundamental Research Programs (JCYJ20170413165528221).

REFERENCES

[1] M. Chen, Z. Cai and Y. Wang, "A method for mobile robot obstacle avoidance based on stereo vision," *IEEE 10th International Conference on Industrial Informatics*, Beijing, 2012, pp. 94-98.

[2] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller and Yann LeCun, "Deep belief net learning in a long-range vision system for autonomous off-road driving," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, 2008, pp. 628-633.

[3] J. Lee, Y. Kim, S. Lee, B. Kim and J. Noh, "High-Quality Depth Estimation Using an Exemplar 3D Model for Stereo Conversion," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 7, pp. 835-847, 1 July 2015.

[4] E. David, P. Christian and F. Rob, "Depth map prediction from a single image using a multi-scale deep network," *2014 Neural Information Processing Systems (NIPS)*.

[5] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328-341, Feb. 2008.

[6] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Bombay, India, 1998, pp. 1073-1080.

[7] J. Shankar and M. Lenin, "Simulation of 3D cloud point from disparity map of stereo image," *2015 International Conference on Advanced Computing and Communication Systems*, Coimbatore, 2015, pp. 1-4.

[8] G. Liu and Q. Feng, "Recovering 3D Shape and Motion from Image Sequences Using Affine Approximation," *2009 Second International Conference on Information and Computing Science*, Manchester, 2009, pp. 349-352.

[9] V. Paramonov, I. Panchenko, V. Bucha, A. Drogolyub and S. Zagoruyko, "Depth Camera Based on Color-Coded Aperture," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, 2016, pp. 910-918.

[10] A. Ahmed and A. Farag, "Shape from Shading for Hybrid Surfaces," *2007 IEEE International Conference on Image Processing*, San Antonio, TX, 2007, pp. II - 525-II - 528.

[11] A. Saxena, M. Sun and A. Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824-840, May 2009.

[12] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks," *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, 2016, pp. 239-248.

[13] K. Yang, X. Hu, L. M. Bergasa, E. Romera and K. Wang, "PASS: Panoramic Annular Semantic Segmentation," in *IEEE Transactions on Intelligent Transportation Systems*.

[14] B. Liu, S. Gould and D. Koller, "Single image depth estimation from predicted semantic labels," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 1253-1260.

[15] L. Ladický, J. Shi and M. Pollefeys, "Pulling Things out of Perspective," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 89-96.

[16] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 4353-4361.

[17] 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)*, Las Vegas, NV, 2016, pp. 770-778.

[18] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 1 April 2017.

[19] S. Pillai, R. Ambruş and A. Gaidon, "SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation," *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, 2019, pp. 9250-9256.

[20] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 6230-6239.

[21] I. Sergey and C. Szegedy ,"Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." (2015).

[22] W. Porter and H. Hirschmüller, "High-resolution stereo datasets with subpixel-accurate ground truth," *2014 German Conference on Pattern Recognition (GCPR)*. volume 8753, pp. 31-42, Oct. 2014.

[23] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.

[24] H. Hirschmuller and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.

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

[26] G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 2261-2269.

[27] A. Mousavian, H. Pirsiavash and J. Košecká, "Joint Semantic Segmentation and Depth Estimation with Deep Convolutional Networks," *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, 2016, pp. 611-619.