Membrane Technology
ISSN (online): 1873-4049

Depth Detection Based on Multi-Scale Residual Multi-Layer Perceptron Iterative Network

Minghao Sun[1], Jifeng Sun [2,*]

[1]*Department of Electrical Engineering, University of Illinois at Urbana-Champaign, Champaign, USA*
[2]*Institute of Information and Electronic, South China University of Technology, Guangzhou 510641, Guangdong, China*
**Corresponding Author.*

Abstract:
Three-dimensional (3D) reconstruction can be used to satisfy the requirement of people on 3D cinema, 3D games, 3D medical imaging and 3D map. The traditional method on 3D reconstruction needs large computation and possesses low accuracy. With the development of AI, more attention to convolutional neural network for three-dimensional reconstruction is put. Current 3D reconstruction methods limit the robustness and integrity and reduce the accuracy of reconstructed models when dealing with occluded, texture-free, or low-textured local backgrounds. Aiming at the problems of low accuracy and cost volume construction in the current 3D reconstruction process, a 3D reconstruction method based on iterative refinement of multi-scale residuals was proposed. The basic idea is to obtain the initial depth map by feature extraction to generate point clouds, and then use multi-scale feature fusion pyramid multi-layer perceptron (MLP) network to obtain point features at different levels. This new multi-scale residual MLP iterative network is used to predict the depth value, and the residual between the depth prediction and the real value is used to estimate the loss. This method attains more precise depth information that enhances the ability of 3D reconstruction of objects, simplifies the network model, and reduces the computational burden. Investigational outcomes show that the suggested method can be used in processing the binocular vision with occluded, texture-free or low-textured local backgrounds and generating higher quality 3D reconstructed objects than the previous methods.

Keywords: Binocular vision, three-dimensional reconstruction, iterative refinement, multi-layer perceptron, cost volume.

INTRODUCTION

3D reconstruction in computer vision is key technique of the computer science. Its application is widely spread in industry, entertainment, military and transportation. Shape from shading (SFS), Structure from motion (SFM), laser radar, Synthetic Aperture Radar (SAR), millimeter radar and Positive electron tomography (PET) are typical 3D reconstruction methods.

Convolutional neural network (CNN) is extensively applied in various territories of the image processing field such as pattern recognition, 3D reconstruction and colorization.

Chang proposed an end-to end pyramid stereo matching network [1] which makes computation more efficient. Chen developed the over-smoothing issue of Convolutional neural network (CNN) in view of disparity estimation [2]. Wang studied the raw evaluation of stereo image depth on mobile devices at any time [3] that the single module regularization and the training are improved on the super-sliding and the whole structure. Voynov built multi-view large-scale dataset [4] among which the scenes are utilized in strengthening a diverse set of material properties which is a difficulty for previous algorithms. Around 1.4 million images of one hundred and seven different scenes derived from one hundred viewing paths depending under fourteen kinds of lighting are provided. Zhou proposed Sparse Fusion for 3D reconstruction [5], in which distilling view-conditioned diffusion was used. Singh developed multi-layer perceptron (MLP) network[6] which can be applied to pattern recognition and 3D reconstruction. The problems of 3D reconstruction on binocular vision are given as follows:

(1) Construction of the cost volume is difficulty;

(2) The precision of the 3D reconstruction is low;

(3) Large Computation is needed.

In this paper, so as to obtain HD quality 3D reconstruction on binocular vision, advanced convolution network that based on multi-scale residual MLP is utilized.

THE PRINCIPLE OF 3D RECONSTRUCTION BASED ON THE MULTI-SCALE RESIDUAL MLP ITERATIVE NETWORK

According the problem that construction of the cost volume is difficult and the precision of the 3D reconstruction is shallow, a 3D rebuilding technique based on multi-scale residual multi-layer perceptron (MLP) iterative network is offered. In this

method, a rough depth chart is built with smaller computation first. Then, by extracting the features from every point of the left picture and the right one, an iterative refined depth chart is got by a multi-scale residual MLP iterative network.

Structure of the Network

The structure of the multi-scale residual MLP iterative network for the depth detection is shown in Figure 1 in which two images from the binocular vision are reconstructed into 3D objects. There are two steps of this structure:

(1) The characteristics are moved out from the binocular vision images with a feature extraction network, the homograph of two feature diagrams employed forming the matching cost volume. The possibility volume got with the help of the possibility normalization along the direction of the depth. Therefore, the initial depth is achieved from the possibility volume. The initial depth is helpful to 3D reconstruction.

(2) A multi-scale residual MLP iterative network is used in refining the depth. A multi-scale pyramid network is put in order to get the multi-scale image features. The point features are got by the connection between the variance of the image feature and the normalized coordinates of the world coordinate system. Then, point features of different scales are achieved with three-layer residual MLP module. The exchange of the point features is implemented with a multi-layer perceptron (MLP). A possibility scalar SoftMax is outputted as the prediction of the residual depth. The view field is enlarged by multiple iteration in order to get more accuracy depth prediction. Loss is computed by the predicted feature points and the information of binocular vision images. A appreciate loss function is employed to do the iteration of the feature points & train the network.

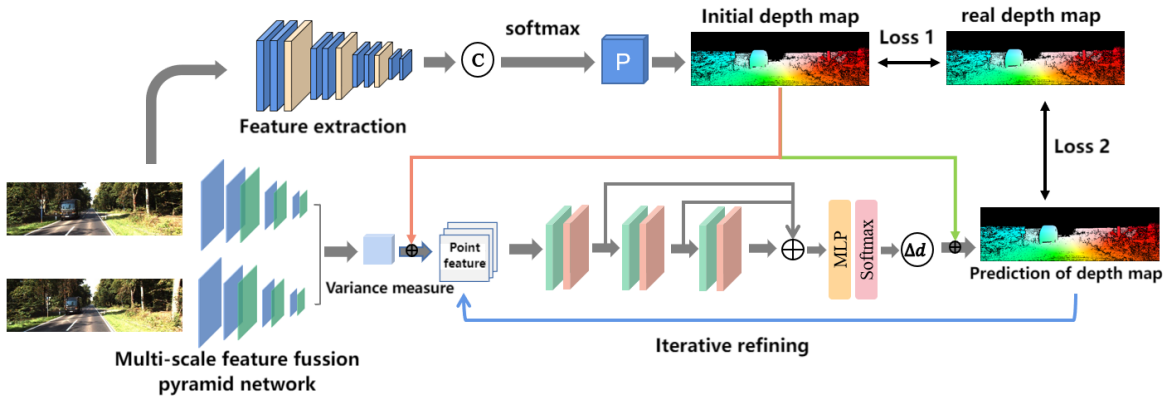

Figure 1. Structure of the multi-scale residual MLP iterative network for the depth detection

Achievement of the Depth Map

Achievement flow chart of the depth map shown in Figure 2 consists of feature extraction, construction of matching cost volume, cost polymerization, possibility normalization and computation of the depth. The feature extraction is the process to extract the feature matrix from the image information, and is important to 3D reconstruction, because the accuracy of the feature extraction gives effects to the accuracy of the follow-up depth. Although the multi-layer convolutional network of PSMNet[1] is good at the feature extraction, the gradient will disappear as the deep of the convolutional network increase, which affects the ability of the feature extraction.

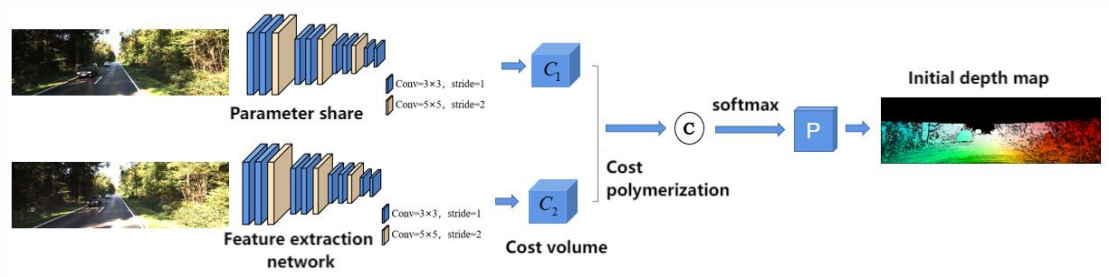

Figure 2. Achievement flow chart of the depth map

To achieve the goal of solving the issue that extensive computation is demanded in the feature extraction network or the network is just simply too complicated, a following four-parts convolutional network is used: three same subnets which are all composed of two 3× 3 convolution layers and a 5× 5 convolution layer; the fourth part consists of two 3× 3 convolution layers.

The output of the network is the feature map with two 32 channels, the size of the output is 1/8 of the input images. In the other word, the process of the above network is 1/8 down-sampling process. For the input image $I_i(i=1,2)$, the output feature map after the processing of the convolutional network is F_i.

The second step of the network is to construct the matching cost volume from the gotten image features and the basic parameters of the camera. Because the extracted features belong to different coordinate system, homography matrix must be used to transform all feature map to different planes in the camera coordinate system to form two feature volumes. Eq. 1 and Eq. 2 can be used to implement End-to-end depth map inference of the feature map to the reference coordinate system.

$$H_i(d) = K_i \cdot R_i \cdot \left(F_i - \frac{(t_1 - t_i) \cdot n_1^T}{d} \right) \cdot R_1^T \cdot K_1^T \tag{1}$$

$$V_i = H_i(d) \cdot F_i \tag{2}$$

$\{V_1, V_2\}$ got from Eq. 2 can be used as the input of the next part, i.e. the cost polymerization in which $\{V_1, V_2\}$ is polymerized as a cost volume C. In two input images of the binocular vision, W, H, D, and F are the width, height, the number of the sampling of the input images and the number of the feature map channels. A cost volume C with the size of $\frac{W}{8} \cdot \frac{H}{8} \cdot D \cdot F$ is got from the following Eq. 3:

$$C = \frac{\sum_{i=1}^{2} \left(V_i - \overline{V_i} \right)^2}{2} \tag{3}$$

Where $\overline{V_i}$ represents the average feature volume of two images of the binocular vision.

Substitute for the average used in Hartmann's cost volume computation[8], variance is applied to inference the similarity between the image blocks, the yielded cost volume can be a measure for difference between view features.

The cost volume got from the feature map may be affected by the lighting and object occlusion, therefore, Some sliding constraint is needed for more precision data. Softmax is used get cost volume C before possibility normalization refinement to produce the possibility volume P which can be applied to inference the depth of the objects more accurate. The prediction of the pixel depth and the determination of the confidence level of the measurement estimation can be done.

Depth D can be computed from possibility volume P with Eq. (4), which is Winner-Take-All of Collins.

$$D = \sum_{d=d_{\min}}^{d_{\max}} d \times P(d) \tag{4}$$

Where $P(d)$ is the pixel possibility estimation on depth d and is complete differentiable. The expectation along the direction of the depth is in the range of $[d_{\min}, d_{\max}]$. The output depth possesses the identical size as the 2d image feature map.

Afterwards, the initial depth map can be obtained. The next stage of work is to refine the depth information for getting more accurate 3D model of the objects. Two steps of the next step are the point feature extraction with a multi-scale feature fusion pyramid network and the prediction of the depth residual with a multi-scale residual MLP iterative network. The second step is different from the scheme based on the cost volume and the features are determined by the fixed space regions of the scene.

Multi-scale Feature Fusion Pyramid Network

The feature pyramid is widely applied in computer vision, in which low-resolution and high-resolution are respectively used to detect big targets and small targets. Pyramid network is commonly used to process the information with different scale or resolution. The pyramid structure used in this paper is shown in Figure 3. This network is the improved one of Lin's feature pyramid networks [8].

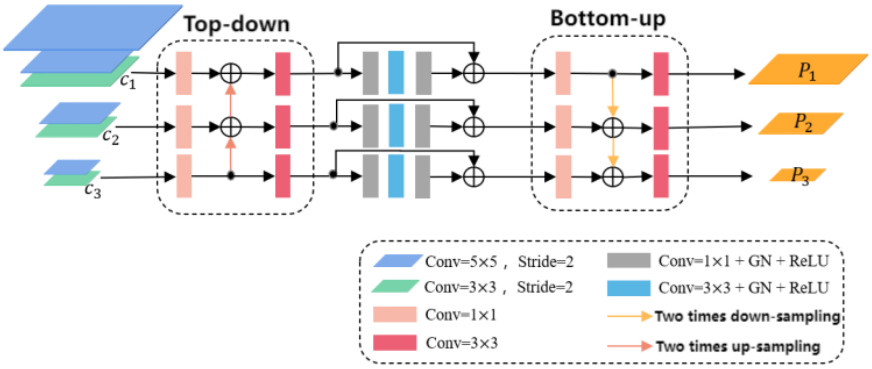

Figure 3. Multi-scale feature fusion pyramid network

As Figure 3, two input images are all put the feature fusion pyramid network and the image features of each layer is got. The residual block of different scale through each 3×3 convolutional layer are represented as c_1, c_2, c_3 .The signal will pass through the top-down process and the bottom-up process. In the case of top-down process, after passing through 1×1 convolutional layer, two times up-sampling with bilinear interpolation is done. The up-sampling feature map is linked with the feature block of next stage to form the input through 3×3 convolutional layer. In top-down process, strong semantics information is mapped from low resolution to features of high resolution. Therefore, the ability of extract the features of the small targets is improved. In the same way, the bottom-up process is similar with the top-down process, only the direction of propagation is contrary to that of the bilinear interpolation. In bottom-up process, the features are mapped from high resolution to features of low resolution. Therefore, the ability of extract the features of the large targets is improved.

For each image, $P_i \quad i=1,2$ is obtained with a multi-scale feature fusion pyramid network. The correct feature parameters can be obtained by scaling up or down each responding feature map with inner parameters of the camera. For i layer multi-scale feature fusion pyramid network, the variance measure of two views $\{N_1, N_2\}$ is computed with Eq. 5:

$$N_i = \frac{\left(P_i^1 - \overline{P}_i\right) + \left(P_i^2 - \overline{P}_i\right)}{2}, \quad i=1,2,3 \tag{5}$$

Where P_i^1 and P_i^2 represent the features of two different images after passing through the feature pyramid.

In order to obtain the feature of each 3D point, the feature point C_p can obtained the connection of N_i and the normalized point coordinate X_p as Eq. 6:

$$C_p = concat\left(N_i, X_p\right) \tag{6}$$

Where the initial value of X_p is the point coordinate of the initial depth, point feature information of next step multi-scale residual MLP network is used as X_p of new iteration.

Multi-Scale Residual MLP Iterative Network

After being processed by the multi-scale feature fusion pyramid network, the signal is connected with the normalized point coordinate X_p and the feature point C_p is obtained, as shown in Figure 4. However, the accuracy of C_p is low, the data structure is irregular and unordered. The refinement of the depth map is necessary. Local region [9] is important to the accuracy of the depth detection. The feature polymerization [10] is also necessary. In order to get the spatial and geometrical information, convolution, graph or attention mechanism [are used in the previous researches. Based in this research, a new network, called multi-scale residual MLP, is used to predict the depth residual. The multi-scale residual MLP is repeated on multi-stages to enlarge the receptive field and achieve complete geometrical information of the feature point. The method of the dynamic features extraction is applied to refine the depth of 3D reconstructed objects. Figure 4 shows a iterative process in the multi-scale residual MLP, in which three-layer residual module is used to extract the depth polymerization features, as shown in Figure 4(a). Each residual module is composed of two MLP layers, one batch normalization layer BN and one activation function ReLU, as shown in Figure 4(a).

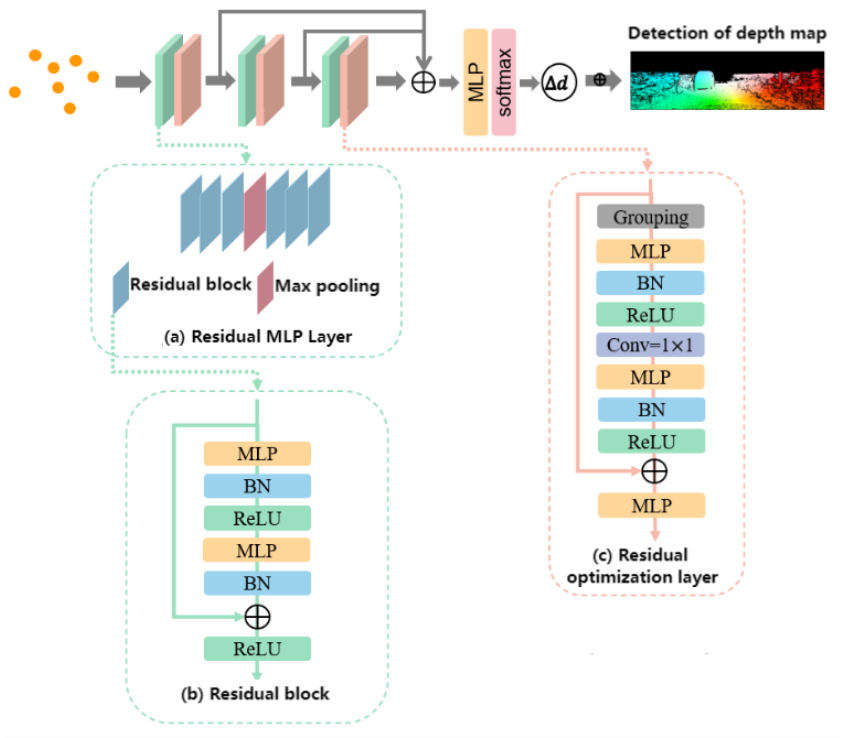

Figure 4. Multi-scale residual MLP iterative network

So as illustrated in the Figure, the residual optimization layer consists of Grouping, MLP and 1×1 convolutional layer. A residual connection is added to reduce the disappear of the gradient caused by the stacking of MLP.

For the input feature Cp, after the processing by three layer MLP iterative network, a possibility scalar softmax is obtained. Then, the depth prediction of the residual can be written as:

$$\text{d=E(ks)=}\sum_{k=-m}^{m} ks \times prob(\widetilde{p_k}) \tag{7}$$

Where $prob(\widetilde{p_k})$ is the weighted possibility sum of all imaginary point displacement. The depth refinement of the addition of the residual depth and the initial depth can be applied to get more accuracy 3D model.

Loss Function

Commonly speaking, the loss function for most 3D reconstruction of the binocular vision is related to the depth after the cost computation. Experimental research shows such loss function restrict the precision of 3D reconstruction. The loss function of this paper described in Eq. 8 consists of Loss 1 and Loss 2:

$$Loss=\sum_{p\in P_{valid}} \|D(p)-D_i(p)\| + \lambda \|D(p)-D_r(p)\| \tag{8}$$

Where P_{valid} is the real pixel set, $D(p)$ is real depth of the pixel p, $D_i(p)$ is predicted depth of pixel p, $D_r(p)$ is refined prediction of the depth on pixel p. λ is 1. $\|D(p)-D_i(p)\|$ is Loss 1, $\|D(p)-D_r(p)\|$ is Loss 2.

EXPERIMENT AND THE ANALYSIS

The experimental platform of our algorithm is Pytorch, the operating system is Ubuntu, GPU is NVIDA RTX 3080, RAM of the display-card is 12G.

In the case of the experiment in which KITTI is used as the training dataset, the size of the images is 512×256, The number of Views is 2, batch size is set as 2, maximum depth $D=192$. In the case of the experiment in which Adam is used as the training dataset, the initial learning rate is 1×10-3, $\beta_1=0.9$, $\beta_2=0.999$. In the case of the experiment in which Scene Flow is used as the training dataset, the learning rate is 1×10-3. Ten epochs are used.

Dataset

KITTI and Scene Flow are used as the training datasets and the test datasets proving the usefulness of the raised method in this research.

Comparative Experiment on KITTI

Eight previous networks, i. e. GC-Net**Error! Reference source not found.**, PSM-Net**Error! Reference source not found.**, HD3-Stereo**Error! Reference source not found.**, LEAStereo**Error! Reference source not found.**, AcfNet**Error! Reference source not found.**,HITNet**Error! Reference source not found.**, CFNet**Error! Reference source not found.** are used in the comparative experiment on KITTI 2012 and KITTI 2015.

Table 1. Comparative experiment result of the previous network and ours network with KITTI 2012 dataset

Method	2PE		3PE		4PE		5PE	
	Noc	All	Noc	All	Noc	All	Noc	All
GC-Net**Error! Reference source not found.**	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46
PSM-Net**Error! Reference source not found.**	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15
HD3-Stereo**Error! Reference source not found.**	2.00	2.56	1.40	1.80	1.12	1.43	0.94	1.19
LEAStereo[12]	1.90	2.43	1.13	**1.45**	0.83	1.08	0.67	0.88
AcfNet**Error! Reference source not found.**	1.83	**2.35**	1.17	1.54	0.92	1.21	0.77	1.01
HITNet**Error! Reference source not found.**	2.00	2.65	1.41	1.89	1.14	1.53	0.96	1.29
CFNet**Error! Reference source not found.**	1.90	2.43	1.23	1.58	0.92	1.18	0.74	0.94
Ours	**1.82**	2.39	**1.11**	1.56	**0.81**	**1.08**	**0.66**	**0.87**

Table 2. Comparative experiment result of the previous network and ours network with KITTI 2015

Method	Noc		All	
	D1-bg	D1-fg	D1-bg	D1-fg
GC-Net**Error! Reference source not found.**	2.02	5.58	2.21	6.16
PSM-Net**Error! Reference source not found.**	1.71	4.31	1.86	4.62
HD3-Stereo**Error! Reference source not found.**	1.56	3.43	1.70	3.63
LEAStereo**Error! Reference source not found.**	1.29	2.65	**1.40**	2.91
AcfNet**Error! Reference source not found.**	1.36	3.49	1.51	3.80
HITNet**Error! Reference source not found.**	1.54	2.72	1.74	3.20
CFNet**Error! Reference source not found.**	1.43	3.25	1.54	3.56
Ours	**1.29**	**2.63**	1.47	**2.89**

As exhibited in Table 1 and Table 2, most results of our method show it has superior performance over the traditional methods.

To obtain the visual effect of the above experiments, the visual result of disparity maps on two methods, as shown in Figure 5. Either on KITTI 2012 dataset or on KITTI 2012 dataset, Our method has better disparity map than CFNet[14].

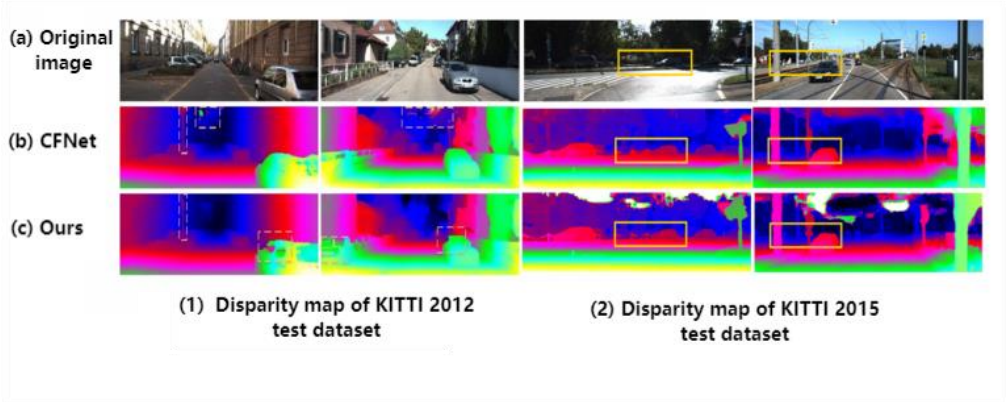

Figure 5. Comparative experiment results of disparity maps on KITTI dataset

Comparative Experiment on Scene Flow Dataset

To prove the effectiveness in the case that there is occlusion, Scene Flow dataset is utilized in completing the comparative experiment. Here the cost of EPE (End-Point-Error) is used to represent the average Euclidean distance between the predicted disparity map and real disparity map. Table 3 exhibits the result of such a comparative experiment among GC-Net**Error! Reference source not found.**, PSM-Net**Error! Reference source not found.**, HD3-Stereo**Error! Reference source not found.**, AcfNet**Error! Reference source not found.**, CFNet**Error! Reference source not found.** and our algorithm . The smaller EFE is, the more accurate the predicted disparity map. The result shows our algorithm possesses more accurate disparity map than the main previous algorithm.

Table 3. EPE of different algorithms on Scene Flow

algorithm	EPE[px]
GC-Net[9]	2.51
PSM-Net[10]	1.09
HD3-Stereo**Error! Reference source not found.**	1.08
AcfNet**Error! Reference source not found.**	0.87
CFNet**Error! Reference source not found.**	0.97
Ours	**0.81**

The visual result of disparity maps on two methods, as shown in Figure 6. In the case of occlusion image ans less-texture image, such as Scene Flow dataset, Our method has better disparity map than CFNet[11].

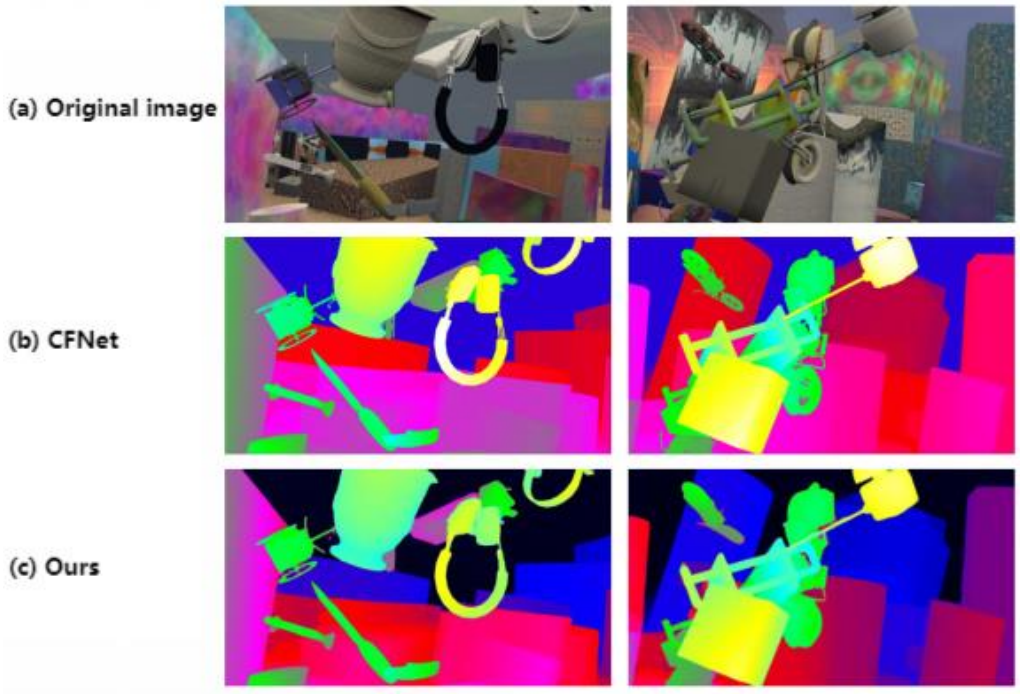

Figure 6. Comparative experiment results of disparity maps on Scene Flow dataset

Ablation Experimental Results and the Analysis

KITTI 2012 and KITTI 2015 are used as the dataset, two images of them are chosen as the test images. The size of the images is 512×256. The depth of the network is 192. t-PE, D1-bg and D1-fg are used as the cost to evaluate the effectiveness of the methods on the ablation experiment. Table 4 shows the result of such ablation experiment.

Table 4. Ablation experimental results

Method	Multi-scale feature fusion pyramid network	Multi-scale residual MLP iterative network	KITTI 2012				KITTI 2015			
			2PE		3PE		Noc		All	
			Noc	All	Noc	All	D1-bg	D1-fg	D1-bg	D1-fg
Method I	×	×	2.34	3.13	1.95	2.33	1.96	3.59	1.89	4.16
Method II	×	√	2.14	3.01	1.87	2.12	1.76	3.23	1.79	3.67
Ours	√	√	1.81	2.39	1.11	1.56	1.29	2.63	1.47	3.07

In method I, adaptive feature extraction network and multi-layer region matching weight network are removed. In method II, multi-layer region matching weight network is removed. When KITTI 2012 and KITTI 2015 are chosen as the dataset, the Ablation experimental results shows our method is better than method I and method II.

CONCLUSION

The multi-scale residual refined network is proposed for depth detection. There are three sub-networks or processing in this network, i. e. multi-scale feature fusion pyramid network, feature extraction network and iterative refining module. The experimental result shows the multi-scale residual refined network possesses better effect on 3D reconstruction than the previous methods and can produce more precision 3D object model with small computation and without constructing complicated cost volume.

The future work may be three parts as follows:

(1) It is to build a larger 3D image dataset.

(2) Lightweight networks is used to save the computation.

(3) 3D reconstruction in the complicated scene will be studied.

ACKNOWLEDGMENTS

We thank professor Lianwen Jin of SCUT for giving good advice. NSFC No. 62071183 provides the finance support for this research. We also thanks to National Science Foundation of China (NSFC).

REFERENCES

[1] Chang J R, Chen Y S. Pyramid stereo matching network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5410-5418.

[2] Chen C, Chen X, Cheng H. On the over-smoothing problem of cnn based disparity estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8997-9005.

[3] Wang Y, Lai Z, Huang G, et al. Anytime stereo image depth estimation on mobile devices. 2019 international conference on robotics and automation (ICRA). IEEE, 2019: 5893-5900.

[4] Voynov O, Bobrovskikh G, Karpyshev P, et al. Multi-sensor large-scale dataset for multi-view 3D reconstruction. Proceedings of CVPR, 2023: 21392-21403.

[5] Zhou Z, Tulsiani S, SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction, Proceedings of CVPR, 2023: 12588- 12597.

[6] Singh, Gurpreet, and M. Sachan. "Multi-layer perceptron (MLP) neural network technique for offline handwritten Gurmukhi character recognition." IEEE International Conference on Computational Intelligence & Computing Research IEEE, 2015.

[7] Hartmann W, Galliani S, Havlena M, et al. Learned multi-patch similarity. Proceedings of the IEEE international conference on computer vision. 2017: 1586-1594.

[8] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[9] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression. Proceedings of the IEEE International Conference on Computer Vision. 2017: 66-75.

[10] Kim K R, Koh Y J, Kim C S. Multiscale feature extractors for stereo matching cost computation. IEEE Access, 2018, 6: 27971-27983.

[11] Yin Z, Darrell T, Yu F. Hierarchical discrete distribution decomposition for match density estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6044-6053.

[12] Cheng X, Zhong Y, Harandi M, et al. Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems, 2020, 33: 22158-22169.

[13] Zhang Y, Chen Y, Bai X, et al. Adaptive unimodal cost volume filtering for deep stereo matching. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12926-12934.

[14] Tankovich V, Hane C, Zhang Y, et al. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14362-14372.

[15] Shen Z, Dai Y, Rao Z. Cfnet: Cascade and fused cost volume for robust stereo matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13906-13915.