IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received March 12, 2019, accepted April 8, 2019, date of publication April 19, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911618

Multi-Dimensional Residual Dense Attention Network for Stereo Matching

GUANGHUI ZHANG[1,2], DONGCHEN ZHU[1], WENJUN SHI[1,2], XIAOQING YE[1,2], JIAMAO LI[1,2], (Member, IEEE), AND XIAOLIN ZHANG[1,2,3], (Member, IEEE)

[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Information and Technology, ShanghaiTech University, Shanghai 200050, China

Corresponding author: Jiamao Li (jmli@mail.sim.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671014, in part by the Shanghai Municipal Science and Technology Major Project (Zhangjiang Lab) under Grant 2018SHZDZX01, and in part by the Shanghai Science and Technology Committee, China, under Grant 17511108202.

ABSTRACT Very deep convolutional neural networks (CNNs) have recently achieved great success in stereo matching. It is still highly desirable to learn a robust feature map to improve ill-posed regions, such as weakly textured regions, reflective surfaces, and repetitive patterns. Therefore, we propose an end-to-end multi-dimensional residual dense attention network (MRDA-Net) in this paper, focusing on more comprehensive pixel-wise feature extraction. Our proposed network consists of two parts: the 2D residual dense attention net for feature extraction and the 3D convolutional attention net for matching. Our designed 2D residual dense attention net uses a dense network structure to fully exploit the hierarchical features from preceding convolutional layers and uses residual network structure to fuse low-level structure information and high-level semantic information. The 2D attention module of the net aims to adaptively recalibrate channel-wise features to be more concerned about informative features. Our proposed 3D convolutional attention net further expands attention mechanism for matching. The stacked hourglass module of the net is exploited to extract multi-scale context information as well as geometry information. The novel 3D attention module of the net aggregates hierarchical sub-cost volumes adaptively instead of manually and then achieves a comprehensive recalibrated cost volume for more correct disparity computation. The experiments demonstrate that our approach achieves the state-of-the-art accuracy on Scene Flow dataset and KITTI 2012 and KITTI 2015 Stereo datasets.

INDEX TERMS Stereo matching, multi-dimensional, residual dense attention, hierarchical, 3D attention mechanism.

I. INTRODUCTION

Stereo matching is important for calculating disparity from stereo images, which plays a critical role in autonomous navigation, 3D reconstruction, and 3D tracking. Given a pair of rectified images, the task of stereo matching is to compute the disparity d for each pixel in reference image. Extensive studies have been carried out for this task, which includes conventional methods (including local and global methods) and popular deep learning methods. However, the disparity of ill-posed regions, such as weakly-textured regions, reflective surfaces, and repetitive patterns, still need to be calculated accurately.

Most conventional local methods are following the typical four-stage pipeline [3]: matching cost computation, cost aggregation, disparity computation, and disparity refinement. For matching cost computation, Census [34] alleviates the drawback of the sum of absolute differences method which is sensitive to light change, by introducing binary coding and hamming distance. Nevertheless, the light robustness of the Census transform is kept with a large kernel size. Zhan *et al.* [2] use the combination of the double-RGB gradient difference, lightweight Census, and color difference on guidance image to obtain a competitive matching cost. However, it requires heavy parameter-regulating. Global methods

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

such as Graph Cuts [4] and Belief Propagation [5] formulate stereo matching as an optimization problem by involving smoothness constraints for neighboring pixels, which require longer execution time. In brief, existing conventional methods are mainly limited by the feature extraction ability.

In recent years, some neural networks-based algorithms have shown their more powerful feature extraction ability than conventional methods. MC-CNN [1] first applies deep learning method to learn how to match corresponding points in matching cost computation. Next, inspired by MC-CNN, many methods put forth effort neural networks-based algorithms for stereo matching. For example, DRR [7] decomposes the disparity estimation task into multiple steps. CRL [9] cascades a residual learning network based on DispNetC [8] with an 2D encoder-decoder structure. However, they are based on 2D-only CNN, which cannot extract geometry information adequately. Therefore, GC-Net [10] introduces 3D CNN into stereo matching. Further, PSMNet [11] presents a pyramid stereo matching network inspired by PSPNet [12] and uses the stacked 3D CNN hourglass module. They can extract geometry information from stereo images well at the cost of computation burden. The fundamental thought of existing neural networks-based methods is to make efforts to fully exploit local and global context information as well as geometry information commonly. However, these networks are going deeper, which results in low-level structure information being missed, e.g. edges, corners. It is still difficult to find accurate corresponding points in inherently ill-posed regions.

Some recent studies attempt to incorporate complementary information, such as object, semantic, edge, or sparse depth information, to get superior disparity map [18]–[21]. Since these methods rely on the complementary information, such a category can be regarded as complementation-based methods. However, these methods usually require expensive corresponding ground truth data. In this paper, we make an effort to improve the above problems without adding complementary information. Inspired by PSMNet [11] and RDNet [25], we present an end-to-end multi-dimensional residual dense attention network with less network layers for stereo matching.

A 2D residual dense attention net is designed to extract hierarchical features of left and right images. The attention module of the net is used to focus on the informative channel features. A 3D convolutional attention net is designed to further extract multi-scale context information as well as geometry information by stacked hourglass architecture and adaptively recalibrate hierarchical sub-cost volumes by 3D attention module. To the best of our knowledge, this paper is the first one to propose 3D attention module to adaptively and effectively aggregate hierarchical sub-cost volumes, which boosts the performance through emphasizing informative aggregation features.

Our main contributions are listed below:

- An end-to-end complementation-free deep learning framework for stereo matching is proposed. The proposed MRDA-Net achieves state-of-the-art accuracy on the Scene Flow dataset, KITTI 2015, and KITTI 2012 dataset.
- The residual dense network structure is first introduced into stereo matching to extract hierarchical features, and attention mechanism is incorporated into it to pay more attention to informative features for improvement.
- A 3D attention module is proposed, which adaptively recalibrates and aggregates hierarchical sub-cost volumes obtained from different network depths. To the best of our knowledge, we are the first one to expand 2D attention mechanism into 3D attention for adaptive cost aggregation.

II. RELATED WORK

A. COMPLEMENTATION-FREE METHODS

The methods, which only take left and right images as input and ground truth of disparity as supervised information without any auxiliary information, are called complementation-free methods. According to the basic network structure, we roughly summarize the complementation-free methods into convolutional neural networks-based and recurrent neural networks-based.

1) CONVOLUTIONAL NEURAL NETWORKS-BASED

Mayer *et al.* [8] offer a large-scale synthetic dataset and propose an end-to-end network for disparity, optical flow, and scene flow estimation with an encoder-decoder structure. Based on [8], [9]–[11] predict dense disparity map in an end-to-end framework. Ye *et al.* [6] leverage ensemble learning in refining outliers within the hourglass architecture to improve the accuracy of stereo matching. In addition, 3D CNN has indeed contributed greatly to the accuracy improvement of stereo matching. GC-Net [10] first proposes 3D CNN architecture which explicitly incorporates context and geometry information over cost volume for stereo matching task. PSMNet [11] adopts 3D CNN architecture to regularize cost volume using stacked multiple hourglass networks with intermediate supervision. Similarly, Lu *et al.* [13] try to extract rich context and semantic information through multi-scale matching cost computation and multi-dimensional aggregation operation. PDSNet [14] exploits 3D CNN architecture to optimize cost volume, and presents an application-friendly stereo matching network, which does not need to set the maximum disparity value. Similar to abovementioned methods, our method makes use of a 3D CNN structure, and we incorporate 3D attention module for improvement.

2) RECURRENT NEURAL NETWORKS-BASED

Recurrent neural networks (RNNs) have been introduced into stereo matching task. Zhong *et al.* [23] take a continuous stereo video as input instead of stereo images, which exploit recurrent nature to memorize and learn from its past experiences to apperceive unseen environments. LRCR [24] proposes a left-right comparative recurrent model. The left-right

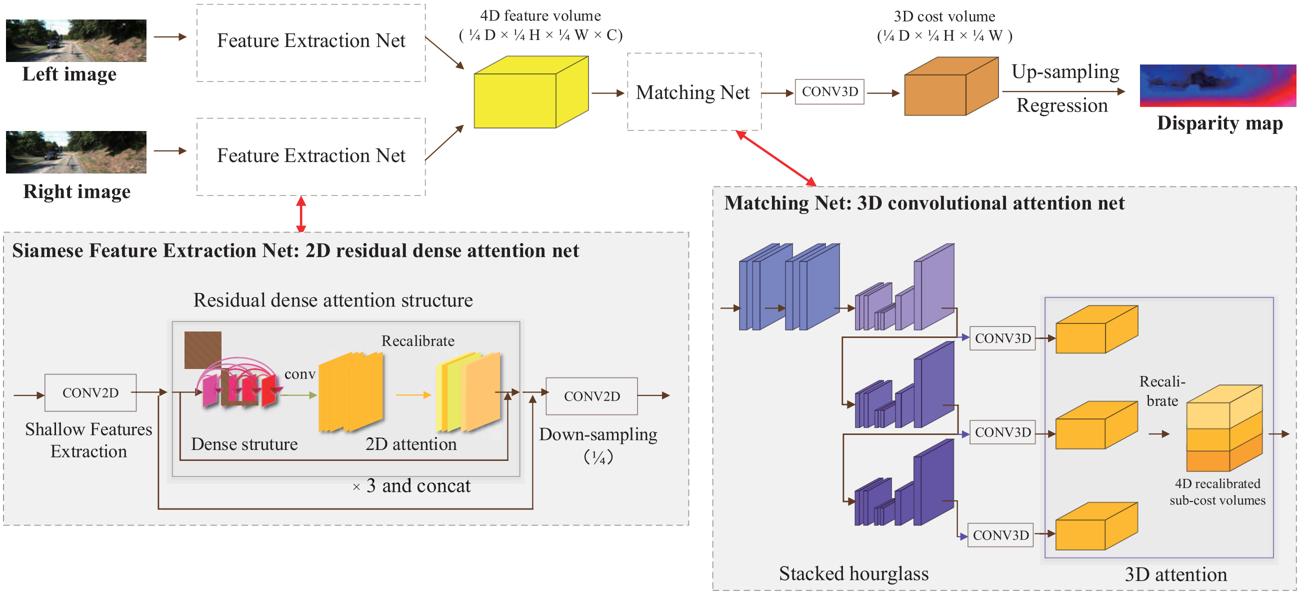


FIGURE 1. An overview of our MRDA-Net. The left and right images are fed to Siamese feature extraction consisting of shallow feature extraction module, stacked residual dense attention (RDA) module ("×3 and concat" means that 3 RDA submodules are stacked and then concatenated), and down-sampling module. The left and right feature volumes are used to obtain a 4D feature volume, which is fed into matching net consisting of the stacked hourglass module and the 3D attention module. The matching net is mainly performed by 3D CNN. For simplicity, the skip connections of the stacked hourglass structure drawing lessons from [11] are omitted.

consistency checking and disparity estimation are performed jointly rather than separately. However, the recurrent convolutional operation suffers from the burden on both computation and memory.

B. COMPLEMENTATION-BASED METHODS

Some studies attempt to incorporate object, semantic, edge, or sparse depth information to refine matching cost computation and cost aggregation [18]–[21]. Displets [18] uses object knowledge that obtained by modeling 3D vehicles to resolve the stereo ambiguities for disparity estimation. The SegStereo [19] and DispSegNet [20] incorporate prior semantic information into stereo matching. The EdgeStereo [21] borrows the network architecture of edge detector HED [22] and achieves the mutual promotion between the edge detection task and stereo matching task. Cheng *et al.* [15], [16] take extra sparse disparity data from LiDAR as complementary input, and learn the affinity matrix through a novel convolutional spatial propagation network following [17]. However, these complementation-based methods usually require corresponding ground truth of auxiliary data for training.

C. ATTENTION MECHANISM

Attention can be regarded as a guidance to assign larger weight to the most informative components of input. Tentative works have been proposed in recent years. Hu *et al.* [29] emphasize that the importance of different channel features is different, and propose a squeeze-and-excitation block to model channel-wise relationships, which has been verified to be effective in image classification task. Li *et al.* [30] successfully combine attention mechanism and spatial pyramid to extract precise dense features in semantic segmentation task. Jiao *et al.* [31] introduce the attention idea into the loss function of monocular depth estimation. To the best of our knowledge, few work has been proposed to investigate the effect of attention in stereo matching task. Most recently, Sang *et al.* [35] propose a MCANet, and introduce attention mechanism in stereo matching network, which is based on only-2D CNN. Its disparity accuracy is not good as ours. In this paper, we apply the 2D attention module into feature extraction net named residual dense attention net for obtaining feature volumes of left and right images. Moreover, we expand it into 3D attention module for adaptively aggregating hierarchical sub-cost volumes in the matching stage.

III. OUR APPROACH

In this paper, similar to [10], [11], [14], we decompose the stereo matching task into Siamese feature extraction, matching (including cost aggregation), and disparity computation. Siamese feature extraction net and matching net are two focal points of our work. An overview of our approach is illustrated in Fig. 1. On the whole, the first inputs are left and right images, the last output is a disparity map.

Specifically, the Siamese feature extraction net consists of three modules: shallow features extraction module, stacked residual dense attention module, and down-sampling module. The outputs of the Siamese feature extraction net are feature volumes of left and right images. The matching net is equivalent to matching cost computation and cost aggregation in

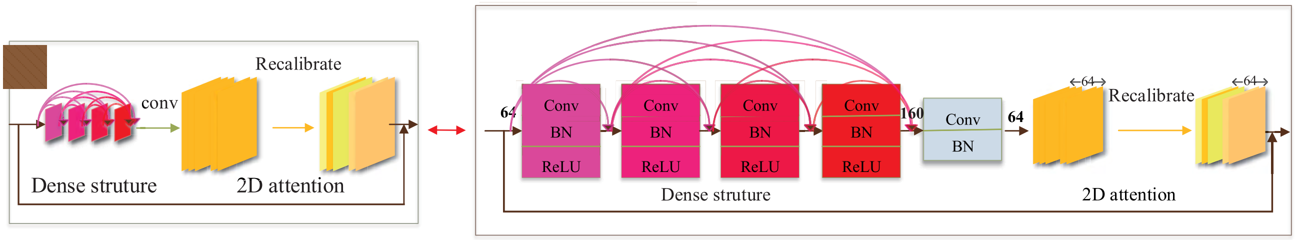


FIGURE 2. The detailed schematic diagram of the residual dense attention structure.

conventional stereo algorithms. We take experience from [26] to form a 4D feature volume that will be used as the input of the successive matching net, named 3D convolutional attention net. Last but not least, up-sampling is used to recover the size of cost volume to size of original image by linear interpolation. Then we use disparity regression as proposed in GC-Net [10] to estimate the continuous disparity map.

A. NETWORK ARCHITECTURE

1) SIAMESE FEATURE EXTRACTION NET: 2D RESIDUAL DENSE ATTENTION NET

Many studies have been carried out on feature extraction in stereo matching. It is difficult to know the context relationship only through the intensities of all pixels. Rich image features are required to understand the context information and geometry information, particularly for ill-posed regions. For example, a car running on the road has windows, tires, mirrors, and hoods, etc. To better compute the disparity of the car, we need to perceive not only the car, but also the above-mentioned parts of the car. In this paper, a residual dense attention net is proposed for incorporating richer hierarchical context information. Our concrete net architecture schematic is shown in the left box named Siamese Feature Extraction Net (FEN) in Fig. 1. I represents I_l and I_r for simplicity, due to the uniformity of operation on I_l and I_r. The input and output are denoted by I and FV_{out} respectively. The overall process can be formulated as

$$FV_{out} = F_{fen}(I) \tag{1}$$

where $F_{fen}(\cdot)$ denotes the FEN net operation. The FEN net includes the shallow feature extraction module, the stacked residual dense attention module, and the down-sampling module. Details of these modules will be elaborated in the following.

Shallow feature extraction module For the shallow feature extraction module, two convolutional layers are used to extract shallow features (SF) that can highlight low-level structure information. The module can be summarized as

$$SF = F_{sf}(I) \tag{2}$$

where F_{sf} denotes the convolution operation.

Stacked residual dense attention module For the stacked residual dense attention module, it can be regarded as a composite operation with stacked residual dense attention (RDA) submodules. In the composite operation, these RDA submodules are concatenated. Next three simple convolutional layers are successively conducted to fuse them to obtain deep features DF. Finally, we conduct the fusion of shallow features SF and deep features DF incorporating hierarchical context information with skip connection. The fused features (FF), which contain the low-level structure information and high-level semantic information, is achieved so far.

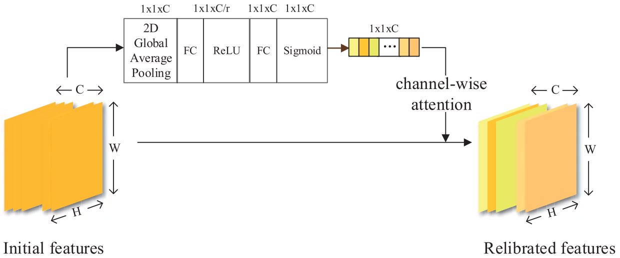


FIGURE 3. The detailed schematic diagram of the feature recalibration through 2D attention module. *r* denotes the compression ratio, which can be used to further fuse multi-channel features and reduce computation load. It is set to be 4 in our network.

Specifically, as shown in Fig. 2, the RDA submodule draws lessons from residual network structure [28] and dense network structure [27]. The nuclear primitive of the original residual structure, some convolutional layers, are replaced by a four-layer dense network structure and attention unit. Each layer of the dense network structure can be regarded as a composite function which consists of convolution, batch normalization, and ReLU. The attention unit in RDA submodule is shown in Fig. 3, which is responsible for recalibrating the features obtained from dense convolutional network through the relationship among channel features in detail. The attention unit can automatically learn and increase the weight of the informative features. Obviously, high-frequency channel-wise features are more informative for high-accuracy disparity computation. It should be promising to obtain improvements if the attention unit pays more attention to such channel-wise features. We will investigate the effect of the attention unit in Sec. IV-C.1.

Down-sampling module In order to reduce the computational complexity, we use naive convolution operation for down-sampling on feature volume ($H \times W \times 64$). The final size of feature volume is $H/4 \times W/4 \times 32$. The process can

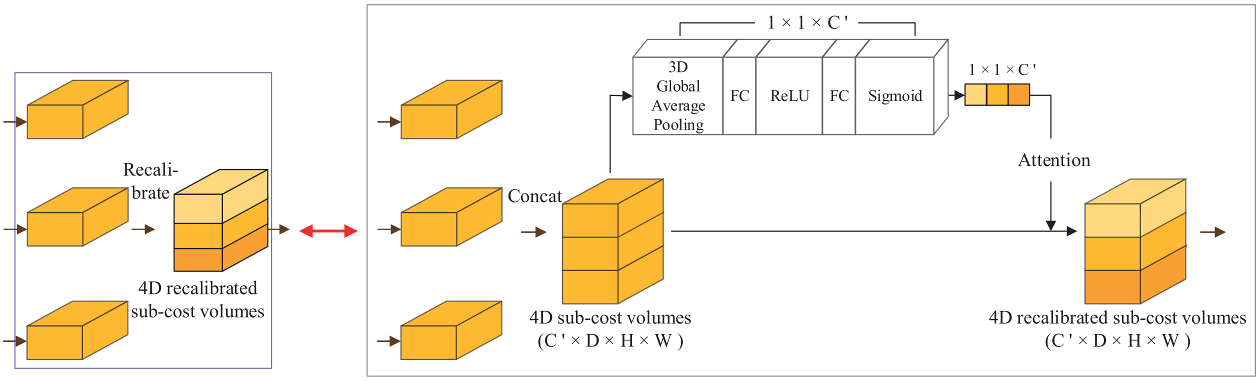


FIGURE 4. The detailed schematic diagram of the aggregated feature recalibration on multiple sub-cost volumes through 3D attention module. *C′* denotes the number of sub-cost volume, which is set to be 3 here. In the process of the recalibration, different colors represent different response values.

be modeled as

$$FV_{out} = F_{ds}(FF) \tag{3}$$

where F_{ds} is the down-sampling operation. The down-sampling operation is performed by two convolutional layers, in each of which 1/2 down-sampling is performed.

2) MATCHING NET: 3D CONVOLUTIONAL ATTENTION NET

As shown in Fig. 1, the matching net consists of the stacked hourglass module and the 3D attention module. The stacked hourglass module draws lessons from [11]. The stacked hourglass module is used to further extract multi-scale context information as well as geometry information. Similar to [11] and [15], we stack three hourglass architectures. We let each of them generate a cost volume, and then concatenate them to form 4D sub-cost volumes, which are fed into 3D attention module. Finally, the recalibrated sub-cost volumes are achieved by multiplying the corresponding response values obtained by 3D attention module.

In matching net, we present a 3D attention module for cost aggregation by analogizing the 2D attention mechanism [29], as shown in Fig. 4. The 3D attention module is used for adaptively recalibrating and aggregating multiple sub-cost volumes, which are from different network depths. The module indicates that the support regions of pixels in the cost aggregation of conventional method could be selected adaptively.

To be specific, we use the simplest aggregation strategy, 3D global average pooling, to obtain the initial response value of each sub-cost volume, noting that more sophisticated aggregation methods could be also considered. In addition, in order to selectively emphasize informative aggregation features of sub-cost volumes and suppress less useful ones, it must be able to learn a nonlinear interaction between sub-cost volumes, and it also must learn a non-mutually-exclusive relationship to ensure that multiple sub-cost volumes could be emphasized opposed to one-hot activation. So two fully connected (FC) layers and sigmoid activation function are adopted to obtain response values.

B. LOSS FUNCTION

We use disparity regression method *soft argmin* as proposed in GCNet [10] to compute disparity, which is different from winner-takes-all (WTA). The WTA directly chooses the optimal disparity value corresponding to the minimum cost at every pixel, while *soft argmin* converts the final cost volume into probability volume, and then multiplies each disparity with its corresponding probability to obtain the final estimated disparity at each pixel. In this way, the information is fully exploited to achieve a more reasonable result.

For regression problems, the most-used loss functions are mean square error (L2) and mean absolute error (L1). In CSPN [15], it is proved that L1 loss function performs better than L2. Therefore, we adopt the smooth L1 loss function to train our MRDA-Net. The loss function is defined as

$$L_{oss} = \frac{1}{N}\sum_{i=1}^{N} l_1\left(d_i - \hat{d}_i\right) \tag{4}$$

where N denotes the number of labeled pixels, d denotes the ground truth disparity, and $\hat{d}$ denotes the estimated disparity. In Equ. 4,

$$l_1(d_i - \hat{d}_i) = \begin{cases} 0.5\left(d_i - \hat{d}_i\right)^2 & if \ \left|d_i - \hat{d}_i\right| < 1 \\ \left|d_i - \hat{d}_i\right| - 0.5 & otherwise \end{cases} \tag{5}$$

where D denotes maximum disparity value.

IV. EXPERIMENT

In this section, some experiments are carried out to evaluate the proposed MRDA-Net. We evaluation our network from two perspectives: (1) the overall performance of our approach compared with other state-of-the-art methods. (2) The ablation study for our proposed 2D residual dense attention net and 3D convolutional attention net.

A. DATASETS AND IMPLEMENTATION DETAILS

We evaluate our MRDA-Net on Scene Flow dataset, KITTI 2015, and KITTI 2012 stereo dataset.

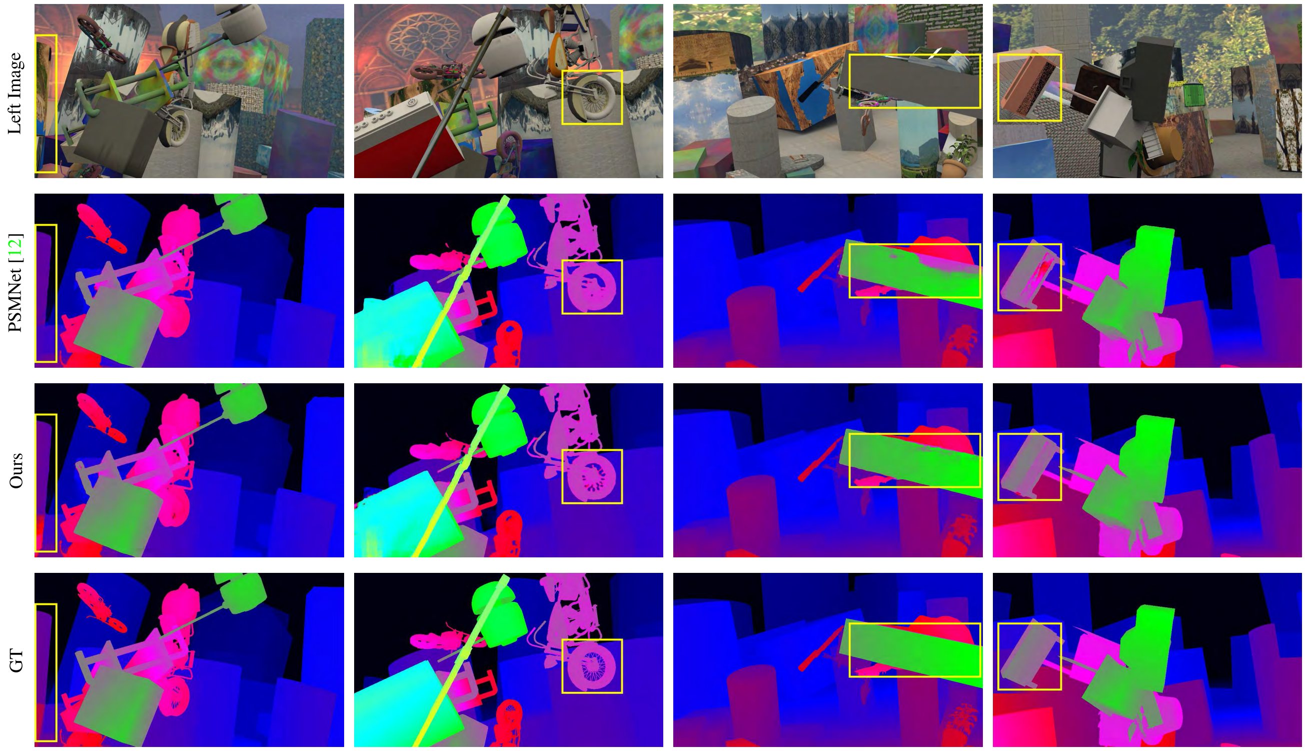


FIGURE 5. The visualization results of Scene Flow Dataset. The results show that it exactly improves challenging ill-posed regions, such as leftmost occlusions, detail regions, and plane regions.

1) SCENE FLOW DATASET

The Scene Flow dataset is a synthetic dataset, which includes not only outdoor scene but also indoor scene. It consists of three sub-datasets: Driving, Monkaa, and FlyingThing3d. The dataset contains 35454 training image pairs and 4370 testing image pairs, and provides elaborate as well as dense disparity maps. The resolution of the image is 540 × 960 ($H \times W$). Our MRDA-Net is trained on the complete dataset. Similar to PSMNet [11], if the disparity is larger than the maximum disparity value set in our experiment, the pixel corresponding to the disparity will be excluded in the loss computation.

2) KITTI 2015 DATASET

The dataset is collected using a driving car in real-world, which only has outdoor streetscape. The dataset contains 200 training image pairs and 200 testing image pairs, and provides sparse ground truth disparity maps of training image pairs that are obtained using LiDAR. For pixels that do not have ground truth, the disparity is set to 0. To test the generalization ability of our network, like [11], we randomly select 80% of training image pairs (160 image pairs) as training set to fine-tune our MRDA-Net, and then use the remaining 20% (40 image pairs) as validation set.

3) KITTI 2012 DATASET

Similar to KITTI 2015, KITTI 2012 has only outdoor streetscape. The dataset contains 194 training image pairs and 195 testing image pairs, and provides sparse ground truth disparity maps of training image pairs.

Our proposed MRDA-Net was implemented based on PyTorch platform using Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), The network is trained on two NVIDIA Titan-X GPUs with 12GB memory, the batch size is set to 4 for training, each GPU is allocated 2. The training process took about 30 hours for 10 epochs, the fine-tuning process took about 6 hours for 300 epochs. Following prior works [32], [33], we use a poly learning rate strategy on epoch. The specific learning rate is

$$l_r^{epoch} = l_0 \times (1 - epoch/nEpochs)^{0.9} \tag{6}$$

where base learning rate l_0 is set to 0.001 in both train and fine-tuning stages. *nEpochs* denotes the total number of epoch. For data augmentation, we only randomly crop the original input image pair into fixed size, which reduces the memory usage during training simultaneously.

B. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART

Comparison on Scene Flow Dataset To verify the effectiveness of the proposed network, we compared our method with our baseline and other state-of-the-art methods, including PSMNet [11], PDSNet [14], MCA-Net [35], CRL [9], DispNetC [8], GC-Net [10] on the Scene Flow test set. The qualitative evaluation results are illustrated in Tab. 1. It can be seen that our method outperforms the baseline network PSMNet by 13.7%, which proves our designed network

TABLE 1. The comparison results on Scene Flow test set. The "EPE" refers to End-point-error.

	MRDA-Net(our)	PSMNet [11]	PDSNet [14]	MCANet [35]	CRL [9]	DispNetC [8]	GC-Net [10]
EPE	**0.94**	1.09	1.12	1.30	1.32	1.68	2.51

TABLE 2. Network setting. The × means the module is not used, and the ✓ means the module is used.

Network setting				KITTI 2015	Scene Flow
	Residual Dense Architecture	2D attention	3D attention	Val 3-pixel-error(%)	EPE
Base(PSMNet [11])	×	×	×	1.98	1.09
MRDA-Net (ours)	×	×	✓	1.90	1.08
	✓	×	×	1.88	1.00
	✓	×	✓	1.82	0.97
	✓	✓	×	1.86	0.98
	✓	✓	✓	**1.78**	**0.94**

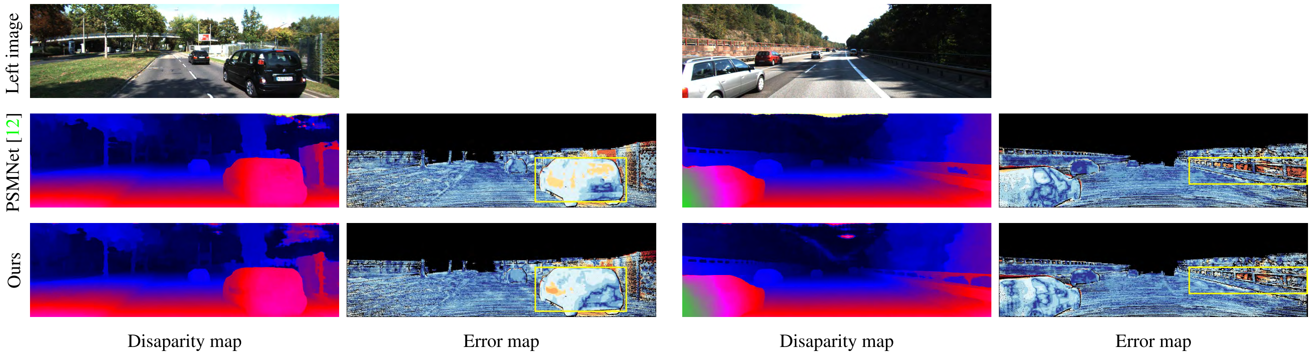


FIGURE 6. The visualization results of KITTI 2015 validation set.

is effective. We also show some disparity results on the dataset in Fig. 5. It is obvious that our network performs better in challenging ill-posed regions, such as leftmost occlusions, detail regions, and weakly-textured plane regions.

Comparison on KITTI 2015/2012 Dataset Further, we compare the results with PSMNet [11] under the same training conditions except batchsize. The batchsize of [11] is 12, while our batchsize is 4 during training because of the limitations of graphics card resources. The experimental results show that our MRDA-Net still performs better than PSMNet [11] on the same 40 validation sets of KITTI 2015 training set. The qualitative evaluation results are illustrated in Tab. 2. We compute the percentage of 3-pixel-error on KITTI 2015 validation set and end-point-error on Scene Flow test set. As listed in the last row with all modules, our designed MRDA-Net achieves the best result, a 1.78 3-pixel-error rate on KITTI 2015 validation set and a 0.94 end-point-error on Scene Flow test set. Our result outperforms the baseline PSMNet by 10.1% on KITTI 2015 validation set. Some disparity results are visualized in Fig. 6. The second row illustrates two examples of our baseline PSMNet, showing disparity maps and corresponding error maps. The third row illustrates our results and errors. As can be seen from Fig. 6, our results are more correct in reflective regions (ie. surface of vehicle) and repetitive patterns (i.e. fence regions), which are indicated by the yellow boxes.

TABLE 3. Result on KITTI 2015.

Methods	All pixels			Non-Occluded pixels		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
DispNetC [8]	4.32	4.41	4.34	4.11	3.72	4.05
CRL [9]	2.48	**3.59**	2.67	2.32	**3.12**	2.45
GC-Net [10]	2.21	6.16	2.87	2.02	5.58	2.61
PDSNet [14]	2.29	4.05	2.58	2.09	3.68	2.36
LRCR [24]	2.55	5.42	3.03	2.23	4.19	2.55
MRDA-Net	**1.93**	4.78	**2.41**	**1.71**	4.38	**2.15**

Comparison results on KITTI 2015 dataset. "D1-fg", "D1-bg", and "D1-all" refer that the error is evaluated over foreground regions, background regions, and all ground truth pixels, respectively. The comparisons are state-of-the-art stereo algorithms belonging to complementation-free methods. DispNetC and CRL belong to state-of-the-art 2D-only CNN-based methods. GC-Net and PDSNet belong to state-of-the-art 3D CNN-based methods. LRCR belongs to the state-of-the-art RNN-based methods.

Then we compare our MRDA-Net with other state-of-the-art methods. These results are obtained based on KITTI

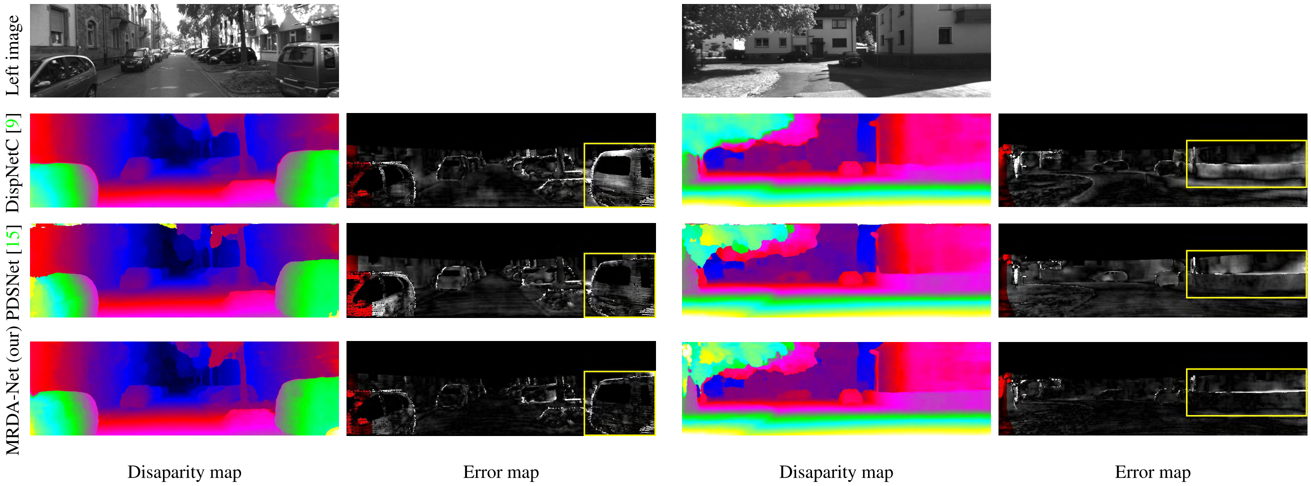


FIGURE 7. The visualization results of KITTI 2012 Dataset.

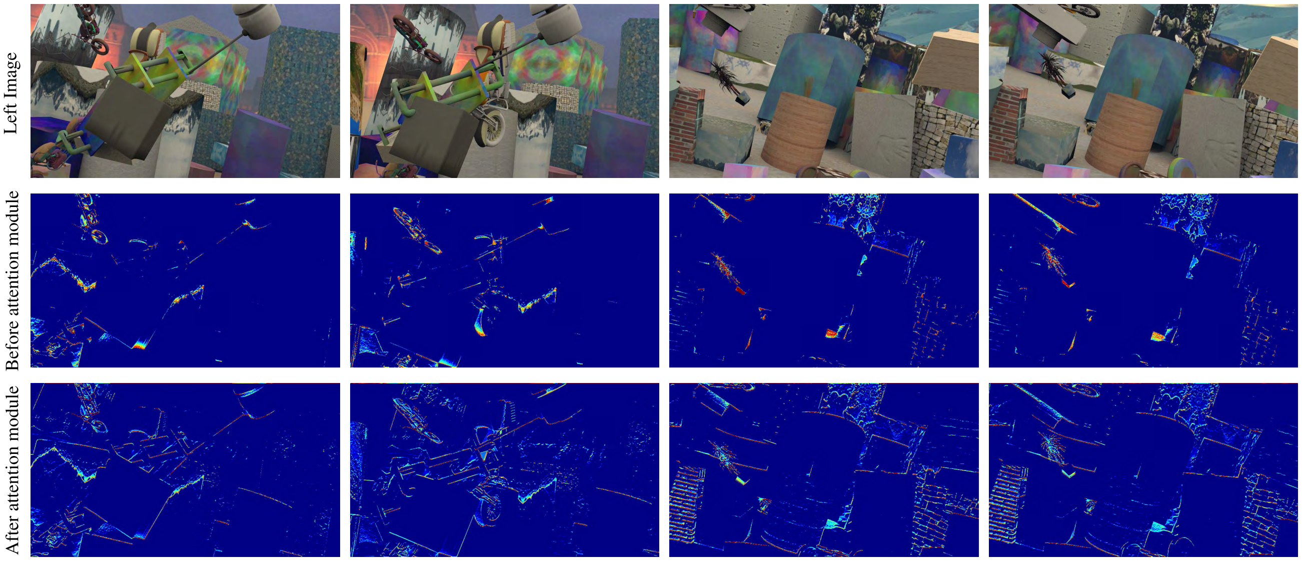


FIGURE 8. The visualization results of heatmaps in residual dense attention net, which are based on the model trained on Scene Flow dataset. The red means the pixels are assigned more response while the blue less response.

evaluation server. The qualitative evaluation results are illustrated in Tab. 3. In the table, we can see our network is superior to state-of-the-art methods in overall three-pixel-error (D1-all), especially background regions (D1-bg). For the error in the foreground regions (D1-fg), where the disparity values are usually larger than background, our results are not good as the methods [8], [9], [14]. It is probably because we set the maximum disparity value and down-sample the value for reducing computation of 3D CNN.

The evaluation on KITTI 2012 is similar to KITTI 2015. The qualitative evaluation results are illustrated in Tab. 4. Our experimental result performs better than these state-of-the-art algorithms. The two-pixel-error rate in non-occluded regions of the proposed MRDA-Net is low to 2.40%. Some disparity results are shown in Fig. 7. Our network more correctly predicts the disparities in reflective regions (i.e. surface of vehicle) and weakly-textured regions (ie. walls), as indicated by the yellow boxes.

TABLE 4. Result on KITTI 2012.

Methods	2px		3px	
	Out-Noc	Out-All	Out-Noc	Out-All
DispNetC [8]	7.38	8.11	4.11	4.65
GC-Net [10]	2.71	3.46	1.77	2.30
PDSNet [14]	3.82	4.65	1.92	2.53
MRDA-Net	**2.40**	**3.21**	**1.48**	**2.09**

"2px", "3px" means 2-pixel-error, 3-pixel-error, respectively. "Out-All" means that all pixels are taken into account in error estimation. "Out-Noc" means that only the pixels in non-occluded regions are considered.

C. ABLATION STUDY

1) ABLATION STUDY FOR 2D RESIDUAL DENSE ATTENTION NET

In this subsection, effectiveness of our design choice is empirically proved. For the ablation study of 2D residual dense attention net, we prove that our proposed net with attention mechanism is better than naive residual dense net. The comparison results are shown in Tab. 2. As can be seen, the sixth and eighth rows of the table indicate the 3-pixel-error rate on KITTI 2015 validation set reduces from 1.82 to 1.78, and the end-point-error on Scene Flow test set reduces from 0.97 to 0.94. Moreover, as illustrated by Fig. 8, the informative high-frequency features such as the edges and details are assigned higher response values after attention module. It shows that the net with the 2D attention module effectively emphasizes informative or challenging features to achieve better disparity map than without attention. In addition, according to the third and fifth rows of the table, we can see the residual dense architecture introduced by us is verified to be more effective than spatial pyramid pooling architecture used in [11].

2) ABLATION STUDY FOR 3D CONVOLUTIONAL ATTENTION NET

We compare the proposed adaptive cost aggregation by using 3D attention module with PSMNet [11] which needs to set parameters manually. The 3D convolutional attention net automatically aggregates hierarchical sub-cost volumes, which indicates that the support regions of pixels in the cost aggregation of conventional method are selected adaptively. As shown in Tab. 2, the contrast between the third and fourth rows, the contrast between the fifth and sixth rows, and the contrast between the seventh and eighth rows all prove the effectiveness of 3D convolutional attention module. Specifically, even in the case of adding only 3D attention module, as listed in the fourth row of the Tab. 2, the result is better than the best one in [11] obtained by tedious manual parameter adjustment, which is listed in the third row. When all modules are added, the result is significantly superior to [11]. As listed in the third and eighth rows of the Tab. 2, the 3-pixel-error rate on KITTI 2015 validation set reduces from 1.98 to 1.78, and the end-point-error on Scene Flow test set reduces from 1.09 to 0.94.

V. CONCLUSION

We propose an end-to-end complementation-free stereo matching network, named Multi-dimensional Residual Dense Attention Network (MRDA-Net). Our proposed network is based on two key parts. One is 2D residual dense attention net, which is mainly used to extract delicate and robust features of input images with 2D attention mechanism paying attention to informative features. The net can be applied into other pixel-wise tasks, e.g. semantic segmentation. The other is 3D convolutional attention net. A novel 3D attention module is proposed to adaptively recalibrate and aggregate hierarchical sub-cost volumes to obtain robust cost volume for more correct disparity computation. Extensive experiments demonstrate that our network effectively improves the overall disparity accuracy and disparities of challenging regions such as detail regions, weakly-textured plane regions. In the future work, we will consider to incorporate the idea of spatial attention into stereo matching network for further research.

REFERENCES

[1] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1592–1599.

[2] Y. Zhan, Y. Gu, K. Huang, C. Zhang, and K. Hu, "Accurate image-guided stereo matching with efficient matching cost and disparity refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1632–1645, Sep. 2016.

[3] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proc. IEEE Workshop Stereo Multi-Baseline Vis.*, Dec. 2001, pp. 131–140.

[4] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[6] X. Ye, J. Li, H. Wang, H. Huang, and X. Zhang, "Efficient stereo matching leveraging deep local and context information," *IEEE Access*, vol. 5, pp. 18745–18755, 2017.

[7] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7187–7196.

[8] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.

[9] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 878–886.

[10] A. Kendall *et al.*, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75.

[11] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[13] H. Lu, H. Xu, L. Zhang, and Y. Zhao. (2018). "Cascaded multi-scale and multi-dimension convolutional neural network for stereo matching." [Online]. Available: https://arxiv.org/abs/1803.09437

[14] S. Tulyakov, A. Ivanov, and F. Fleuret. (2018). "Practical deep stereo (PDS): Toward applications-friendly deep stereo matching." [Online]. Available: https://arxiv.org/abs/1806.01677

[15] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 108–125.

[16] X. Cheng, P. Wang, and R. Yang. (2018). "Learning depth with convolutional spatial propagation network." [Online]. Available: https://arxiv.org/abs/1810.02695

[17] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 1519–1529.

[18] F. Güney and A. Geiger, "Displets: Resolving stereo ambiguities using object knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4165–4175.

[19] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia. (2018). "Segstereo: Exploiting semantic information for disparity estimation." [Online]. Available: https://arxiv.org/abs/1807.11699

[20] J. Zhang, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson. (2018). "Dispsegnet: Leveraging semantics for End-to-End learning of disparity estimation from stereo imagery." [Online]. Available: https://arxiv.org/abs/1809.04734

[21] X. Song, X. Zhao, H. Hu, and L. Fang. (2018). "Edgestereo: A context integrated residual pyramid network for stereo matching." [Online]. Available: https://arxiv.org/abs/1803.05196

[22] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.

[23] Y. Zhong, H. Li, and Y. Dai. "Open-world stereo video matching with deep RNN." [Online]. Available: https://arxiv.org/abs/1808.03959

[24] Z. Jie *et al.*, "Left-right comparative recurrent model for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3838–3846.

[25] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[26] L. Yu, Y. Wang, Y. Wu, and Y. Jia. (2018). "Deep stereo matching with explicit cost aggregation sub-architecture." [Online]. Available: https://arxiv.org/abs/1801.04065

[27] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[29] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. (2017). "Squeeze-and-excitation networks." [Online]. Available: https://arxiv.org/abs/1709.01507

[30] H. Li, P. Xiong, J. An, and L. Wang. (2018). "Pyramid attention network for semantic segmentation." [Online]. Available: https://arxiv.org/abs/1805.10180

[31] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 55–71.

[32] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587

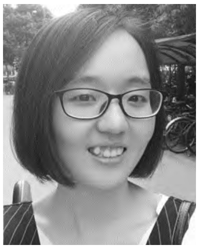
[33] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[34] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 151–158.

[35] H. Sang, Q. Wang, and Y. Zhao, "Multi-scale context attention network for stereo matching," *IEEE Access*, vol. 7, pp. 15152–15161, 2019.

GUANGHUI ZHANG received the B.S. degree from the South China University of Technology, China, in 2016. She is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences. Her current research interests include stereo vision, 3D reconstruction, and 3D scene understanding.

DONGCHEN ZHU received the B.S. degree from Wuhan University, China, in 2013, and the Ph.D. degree in information and communication engineering from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2018, where she is currently an Assistant Professor. Her current research interests include computer vision, stereo vision, 3D reconstruction, and artificial intelligence.



WENJUN SHI received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, China, in 2015. She is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences. She is also an intern student with Shanghai Eyevolution Technology Co., Ltd., Shanghai, China. Her current research interests include 3D reconstruction, 3D scene understanding, and visual odometry.

XIAOQING YE received the B.S. degree from Wuhan University, China, in 2014. She is currently pursuing the Ph.D. degree in information and communication engineering with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences. Her current research interests include stereo vision, 3D reconstruction, and autonomous driving.

JIAMAO LI received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2012. He is currently a Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences. His current research interests include computer vision, machine vision, 3D micro-imaging, and artificial intelligence.

XIAOLIN ZHANG received the Ph.D. degree from Yokohama National University, in 1995. He was a Professor with the Tokyo Institute of Technology, Japan, from 2012 to 2013. He is currently a Professor with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences; and the University of Chinese Academy of Sciences. His research interests include bionics, brain science, computer vision, and artificial intelligence.