IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 27, 2020, accepted June 10, 2020, date of publication June 18, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003375

Multi-Attention Network for Stereo Matching

XIAOWEI YANG[1,2], LIN HE[1,3], YONG ZHAO[4,5], (Member, IEEE), HAIWEI SANG[5], ZU LIU YANG[4], AND XIAN JING CHENG[6]

[1]School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
[2]Guizhou Tea Research Institute, Guizhou Academy of Agricultural Sciences, Guiyang 550006, China
[3]Liupanshui Normal College, Liupanshui 553004, China
[4]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China
[5]School of Mathematics and Big Data, Guizhou Education University, Guiyang 550018, China
[6]Research Institute of Qianbei Information Technology, Zunyi Normal University, Zunyi 563006, China

Corresponding author: Lin He (helin6568@163.com)

This work was supported in part by the Science and Technology Program of Shenzhen under Grant JCYJ20180503182133411, in part by the Project of Science and Technology Department of Guizhou Province under Grant QiankeheZhicheng [2019]239, in part by the Project of Science and Technology Department of Guizhou Province under Grant QiankeheJichu [2019]1250, in part by the Guizhou Provincial Science and Technology Cooperation Project under Grant QiankeheLHZi [2017]7072, and in part by the Guizhou Provincial Department of Education Youth Science and Technology Talents Growth Project under Grant QiaojiaoheKYZi [2017]251.

ABSTRACT In recent years, convolutional neural network (CNN) algorithms promote the development of stereo matching and make great progress, but some mismatches still occur in textureless, occluded and reflective regions. In feature extraction and cost aggregation, CNNs will greatly improve the accuracy of stereo matching by utilizing global context information and high-quality feature representations. In this paper, we design a novel end-to-end stereo matching algorithm named Multi-Attention Network (MAN). To obtain the global context information in detail at the pixel-level, we propose a Multi-Scale Attention Module (MSAM), combining a spatial pyramid module with an attention mechanism, when we extract the image features. In addition, we introduce a feature refinement module (FRM) and a 3D attention aggregation module (3D AAM) during cost aggregation so that the network can extract informative features with high representational ability and high-quality channel attention vectors. Finally, we obtain the final disparity through bilinear interpolation and disparity regression. We evaluate our method on the Scene Flow, KITTI 2012 and KITTI 2015 stereo datasets. The experimental results show that our method achieves state-of-the-art performance and that every component of our network is effective.

INDEX TERMS Neural network, stereo matching, multi-scale attention module, feature refinement module, 3D attention aggregation module.

I. INTRODUCTION

Binocular stereo vision simulates the operating principle of biological vision systems. It applies two cameras to acquire two digital images of the same three-dimensional scene from different angles and even from a different time and space. By using stereo matching algorithms, it can calculate the dispaity of two images. This technology has been widely used in many real-world scenarios, such as automatic driving [1], [2], 3D reconstruction [3], [4], pose estimation [5], and robot positioning and ranging [6].

Traditional stereo-matching algorithms have four primary steps: matching cost computation, cost aggregation, disparity calculation and disparity refinement [7], [8]. Traditional stereo matching methods are roughly divided into global matching [9]–[11], local matching [12], and semi-global matching [13]. It is difficult to estimate disparity accurately because these methods generally use artificially designed features to describe matching cost and cost aggregation; consequently, these methods can only learn a few linear combinations of data features.

Over the past few years, CNN methods have achieved unprecedented performance in many computer vision tasks, such as detection [14]–[16] semantic segmentation [17]–[19], image denoising [20], [21] and classification [22], [23]. Convolutional networks have also been used learn how to estimate disparity. The MC-CNN scene disparity estimation method proposed by Zbontar *et al.* [24] pioneered a Siamese network to compute the similarity between two image patches for stereo matching. Based on MC-CNN, Embedding-CNN proposed by Chen *et al.* [25], innovatively uses a full convolutional network to extract features from the entire image.

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

The network utilizes the dot product operation via a sliding window and obtains the matching score of each pixel within the search range of disparity. Mayer *et al.* created a large synthetic dataset to train an end-to-end network called DispNet [26] to estimate disparity; DispNet consists of a set of convolution layers to extract features, a cost volume formed by patch-wise correlation, an encoder-decoder structure for the second-stage process, and a classification layer to estimate disparity. Pang *et al.* [27] proposed a two-stage architecture called cascade residual learning (CRL). In the first stage, the network introduces a nontrivial upconvolution module to produce fine-grained disparities. In the second stage, the final disparity is rectified with the residual signals by the difference between the initial disparity and the ground-truth disparity. Khamis *et al.* [28] proposed StereoNet, which is the first end-to-end deep architecture for real-time stereo matching. The network achieves high disparity precision by using a very low-resolution cost volume that encodes all of the information. Song *et al.* [29] proposed an effective multitask learning network called EdgeStereo to improve the quality of disparity estimates. EdgeStereo consists of an edge detection subnetwork and a disparity estimation subnetwork; EdgeStereo utilizes geometric clues, such as edge contours and corresponding constraints, to obtain better generalization capability than other state-of-the-art disparity estimation networks for stereo matching. Guo *et al.* [30] proposed a new cost volume by group-wise correlation. Along the channel dimension, the right features and the left features are divided into groups and to obtain multiple matching cost proposals, correlation maps are computed among each group. Zhang *et al.* [31] proposed a deep guided aggregation network (GA-Net) that has two novel neural net layers, aimed at capturing the whole-image and local guide cost aggregation dependencies respectively.

To make better use the global context information for stereo matching, we propose a novel convolutional neural network. The network combines a spatial pyramid module with an attention mechanism to extract global context information at the pixel-level and the multi-scale module combines high-level feature information at four different scales without consuming too many computing resources. In the cost aggregation, we introduce an image feature refinement module to enhance the representation of feature maps at each stage. In addition, we design a 3D aggregate attention module to obtain high-quality channel attention vectors. Due to the fusion of high-level and low-level features information, the network can use high-level semantic information to guide low-level texture information and reduce the loss of information. The aggregate attention module can also suppress features that have lower discrimination ability. In addition, we simplify the hourglass structure proposed by PSMNet [32] and improve the performance of the network.

In summary, the main contributions of our work are as follows:

- We propose a novel convolutional neural network for stereo matching without any postprocessing.
- We present a context multi-scale attention module that can use a channel-wise attention vector to effectively select multi-scale information features at the pixel-level.
- We design an image feature refinement module to refine the feature map and strengthen the representational ability of the feature map of each stage during the cost aggregation.
- We introduce a 3D aggregation attention module, which can use high-level information to guide low-level texture information and identify high-quality channel attention vector features.
- Our MAN achieves state-of-the-art performance on the Scene Flow dataset, KITTI stereo 2012 and KITTI stereo 2015 benchmarks.

II. RELATED WORK

There are many studies on stereo matching. We learn from methods employing convolutional neural networks. Deep-learning models can learn more robust and discriminative features that help improve the accuracy of disparity estimation. In this paper, we review only the works most relevant to our own.

Brandao *et al.* [33] proposed a Siamese network with deconvolution and pooling operations for similarity computation in a wider receptive field. Unlike [33], who used direct upsampling, Lu *et al.* [34] concatenated deconvolution features with corresponding feature maps from the encoder structure, which can preserve both low-level fine information and high-level coarse information. Zhu *et al.* [35] designed a deep-learning network called CFPNet, which consists of a multiscale 2D local feature extraction module, a cross-form spatial pyramid module and a multiscale 3D feature matching and fusion module. The network utilizes the multiscale local feature extraction module to extract multiscale features, and the cross-form spatial pyramid module was designed to aggregate global context information with different locations and scales. Moreover, the network uses the multiscale 3D feature matching and fusion module to regularize the cost volume by using two parallel 3D deconvolution structures of two different receptive fields. Kendall *et al.* [36] proposed GC-Net, which uses 3D cost filtering and the soft argmax to incorporate global context information. Inspired by GC-Net, Chang and Chen [32] developed a network called PSMNet, which uses a pyramid spatial pooling module to enrich features with better global context and a stacked hourglass 3D residual network to extend the global context information for cost volume regularization. Xie *et al.* [37] introduced vortex pooling, which improves upon the atrous spatial pooling approach used in DeepLab. Vortex pooling uses average pooling in grids of varying dimensions before using dilated convolutions to utilize information from the pixels. Chabra *et al.* [38] proposed StereoDRNet, which uses the feature extraction described in vortex pooling and improves cost filtering. Rao *et al.* [39] proposed MSDC-Net, which consists of two modules: a multiscale fusion 2D convolution module and multiscale residual 3D convolution module.

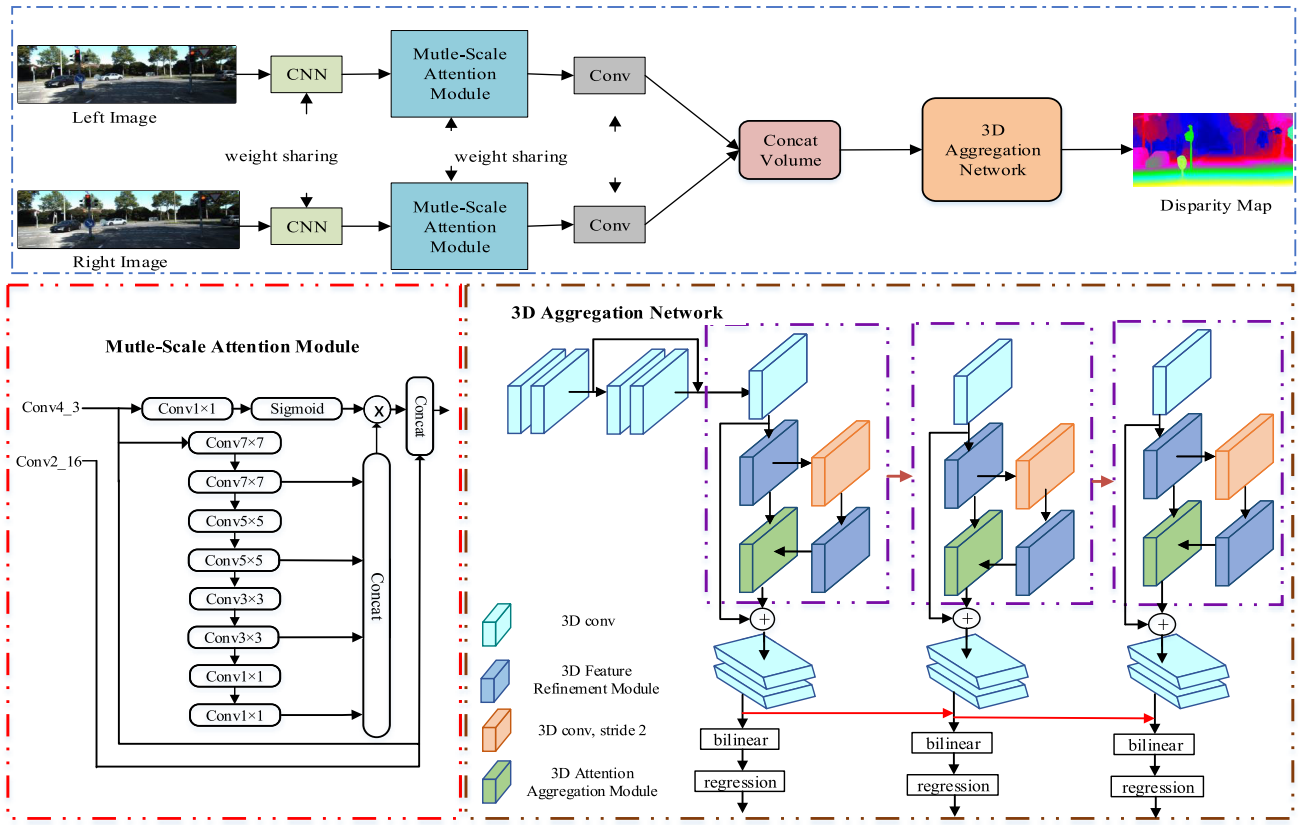


FIGURE 1. Overview of our multi-attention neural network. The pipeline of our module consists of the following steps: (a) Input images. (b) Feature extraction. (c) Cost volume. (d) Disparity estimation.

The network introduces the multiscale fusion 2D convolution module to extract cross-scale features to improve the ability to understand context. Because the encoder-decoder network often requires a vast amount of calculation and is difficult to train, the network utilizes the multiscale residual 3D convolution module to learn the regional support of global information from the cost volume.

Compared with traditional methods, convolutional neural networks have greatly improved the accuracy of stereo matching [27]–[36]. However, there are still some technical difficulties in estimating disparity for CNNs in ill-posed regions, such as reflective surfaces, repetitive patterns, thin structures and textureless areas. Therefore, regional support from global context information must be incorporated into stereo matching. Since overlooking the global context information will lose high-frequency information that helps generate fine details in disparity maps. Focusing on above problems, we learn from the experience of semantic segmentation research [40]–[42] and propose a stereo-matching method based on a multi-attention neural network. During feature extraction, we combine a spatial pyramid pooling module with an attention mechanism to design the multi-scale attention module that replaces the operation of the stack of convolutional layers and spatial pooling module. The multi-scale attention module can make use of the channel-wise attention vector to extract the global context information at the pixel-level so that we can utilize more valid global context information in the ill-posed regions to estimate disparity. Simultaneously, in the cost aggregation, we use the feature refinement module to refine the feature map and strengthen the representational ability of feature map and we also design the 3D attention aggregation module to identify the high-quality channel feature information. The multi-attention neural network will be described in detail below.

III. MULTI-ATTENTION NEURAL NETWORK

In this section, we will give a detailed description of our method Multi-Attention Network (MAN). The framework of MAN is shown in Figure 1.

A. BASIC NETWORK ARCHITECTURE

The network processes the left and right stereo images using CNN with shared weights for feature extraction respectively. The multi-scale attention module connects the subareas of different scales to use the global information. The size of an output feature maps is 1/4 the size of an original stereo images. Exactly like the method in GC-Net [36], we concatenate the output feature maps along every disparity level to form a 4D cost volume, which is then delivered to the 3D

aggregation network for cost volume regularization. Finally, we estimate the disparity map based on bilinear interpolation and disparity regression [36]. In the remainder of this section, we will discuss each component in detail.

B. FEATURE EXTRACTION

In current stereo-matching researches [32], [34], [35], [39], the accuracy of neural network algorithms which are exploiting more global context information will be greatly improved. In the original research work, the pyramid pooling module and the ASPP module can effectively extract feature information at different scales. However, the structure of the pyramid module is short of the global context prior to select the features channel-wise. Furthermore, the application of channel-wise attention vector is not enough to extract multi-scale features effectively. Therefore, we combine the space pyramid pooling module with the attention mechanism to design a multi-scale attention module (MSAM) that can integrate different scales information step-by-step, which is shown in Figure 1.

Specifically, the MSAM fuses features from four different pyramid scales by a U-shaped structure. Each scale has two consecutive convolution layers. The convolution operation of each scale is based on the previous one. Then we connect four different scales respectively to better extract global information. The network uses a large convolution kernel size, which does not bring too much computational burden because the resolution of the high-level feature map is small. The U-shaped structure of the module increases the network depth to improve the matching performance of the network and fuses global context information features more accurately. Unlike PSPNet [43] and ASPP concatenates different pyramid scale feature maps, the input of pyramid module passing through a 1×1 convolution and sigmoid operations is multiplied by the pyramid features so that we can use channel-wise attention vector to select the pixel-level global information effectively.

C. COST VOLUME

Traditional stereo methods use a winner-takes-all (WTA) approach, which chooses the disparity of the lowest distance between two stereo feature maps. Instead, in the range of disparity D_{max}, we concatenate the left feature map f_l and the corresponding right feature map f_r of each disparity d to form a 4D cost volume [34] (height $\times$ width $\times$ disparity $\times$ feature size), which is defined mathematically in Equation 1:

$$C_{\text{concat}}(d, x, y) = Concat\{f_l(x, y), f_r(x - d, y)\} \quad (1)$$

D. 3D AGGREGATION NETWORK

In PSMNet, a stacked hourglass structure is proposed to better obtain context characteristics. We design a 3D aggregation network, which simplifies the hourglass structure and improves the performance of the network. Additionally, we introduce an image feature refinement module and a 3D attention aggregation module to obtain high-quality feature information.

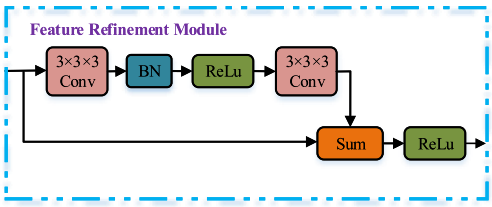


FIGURE 2. The schematic diagram of the Feature Refinement Module.

In the 3D aggregation network, the feature maps of each stage will be processed by FRM, as illustrated in Figure 2. The upward side is a basic residual block, which can refine the feature map and add to the original feature map to enhance the representational ability of each stage, inspired from the architecture of ResNet [44]. In other words, the original input without any changes is directly added to the output. Our network in Figure 2 has two layers. The shortcut connections in Equation 2 introduce neither extra parameter nor computation complexity. Equation 2 is as follows:

$$H(x) = F(x) + x \quad (2)$$

where x and $H(x)$ are the input and output vectors respectively of the layers. The function $F(x)$ represents the residual mapping to be learned.

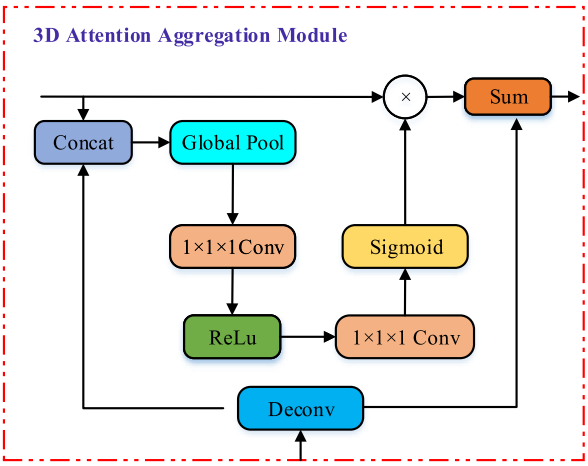


FIGURE 3. The 3D Attention Aggregation Module structure.

We introduce the 3D AAM to change the weights of the features at each stage and obtain a high-quality channel attention vector. The 3D AAM structure is shown in Figure 3. In our network architecture, the convolution operator outputs a probability distribution map, which gives the probability of each disparity at each pixel. The final score on the probability distribution map is summed over all channels of the features maps in Equation 3.

$$s_c = F(y; k) = \sum_{i=1,j=1}^{H} k_{i,j} y_{i,j} \quad (3)$$

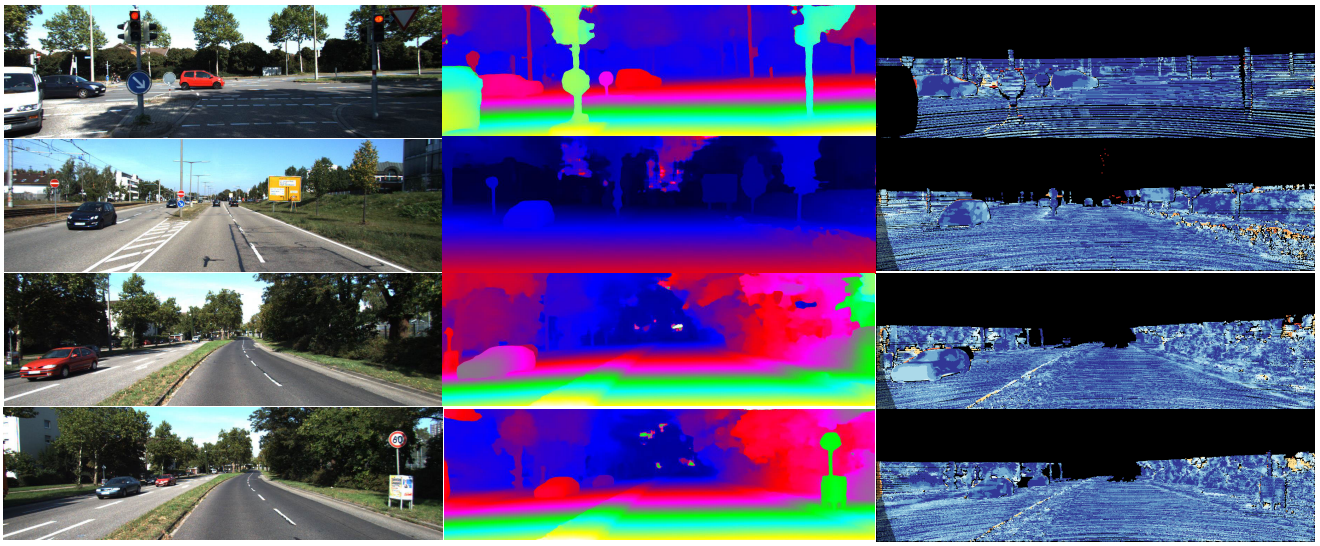


FIGURE 4. KITTI 2015 test data qualitative results. From left: left stereo input image, our disparity prediction, error map.

in which y is the output feature of the network. k represents the convolution kernel, and $c \in \{1, 2, \ldots, C\}$. C is the number of channels. H is the set of pixel positions. Finally, we use the softmax function to obtain the probability of the disparity d:

$$\delta_i(s_c) = \frac{\exp(s_c)}{\sum_{j=1}^{C} \exp(s_j)} \tag{4}$$

where δ is the disparity prediction probability and s is the output of the network.

As shown in Equations 3 and 4, the final output is the highest probability pixel offset. If the prediction result of a certain pixel is s_0, but its true label is s_1, then we can introduce a parameter β to change the highest probability value from s_0 to s_1, as Equation 5 shows.

$$\bar{s} = \beta s = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_C \end{bmatrix} \cdot \begin{bmatrix} s_1 \\ \vdots \\ s_C \end{bmatrix} = \begin{bmatrix} \beta_1 k_1 \\ \vdots \\ \beta_C k_C \end{bmatrix} \times \begin{bmatrix} y_1 \\ \vdots \\ y_C \end{bmatrix} \tag{5}$$

where $\bar{s}$ is the new prediction of network and $\beta = sigmoid(y; k)$.

In Equation 3, it shows that the weights of different channels are equal. However, the features in different stages have different degrees of discrimination, thereby resulting in an inconsistent prediction. We should inhibit the indiscriminative features and extract the discriminative features to obtain more valid image features matching for disparity estimation. In Equation 5, the β value applies on the feature maps y, which represents the feature selection with the 3D attention aggregation module. Therefore, we can make our network to obtain discriminative features to obtain the prediction that calculate a high precision disparity estimation.

In detail, the size of high-level features based on deconvolution is the same as the size of low-level features; thus we can concatenate these features to better use the high-level semantic information and the low-level texture information. Simultaneously, the operation of a continuous 1×1 convolution can adjust the number of feature channels to extract high-quality channel attention vectors. These vectors can guide the selection of the low-level features. To reduce the loss of information, the deconvolved high-level features are added to the guided low-level features, which can obtain the final output features. In the next section, we will verify the modules via experiments.

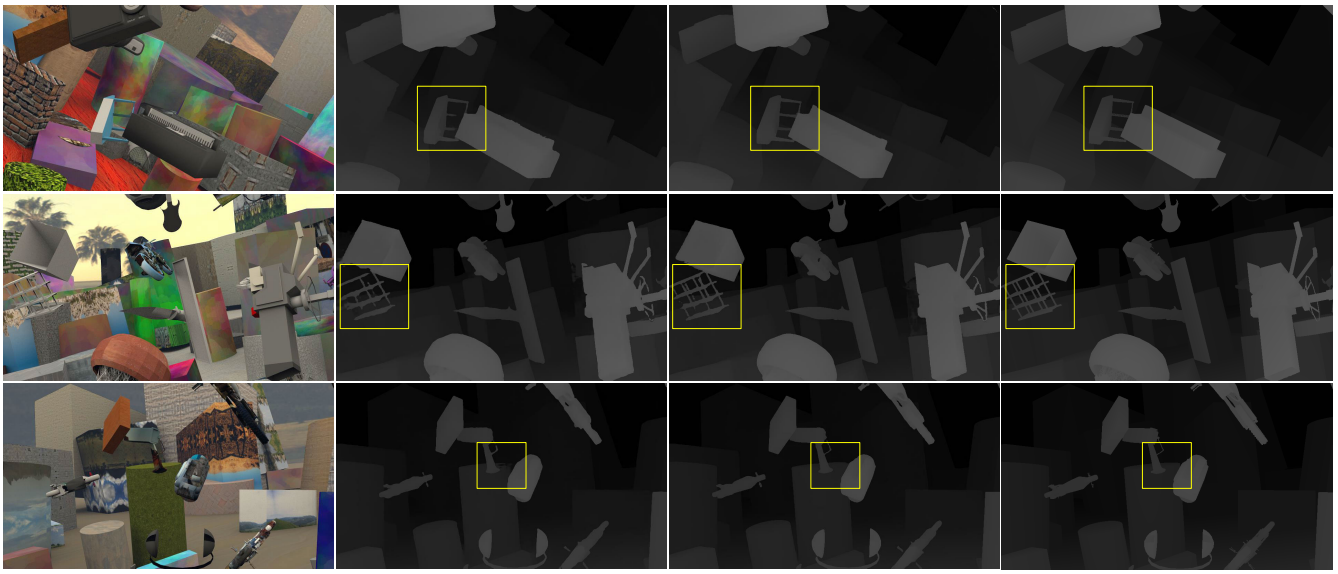
E. DISPARITY REGRESSION AND LOSS FUNCTION

We introduce the disparity regression proposed in [36] to predict the disparity map. The output feature map size is $(H, W, D+1)$ and D represents the maximum disparity. With the softmax operation $\sigma(\cdot)$, we can calculate the probability of each disparity d from the predicted cost C_d. Additionally, we use the sum of each disparity d probability weighted to calculate the predicted disparity. The disparity regression is defined as follows:

$$\text{soft argmin} = \sum_{d=0}^{D} d \times \sigma(-C_d) \tag{6}$$

We train our module using ground truth depth data from a random initialization. Because the ground truth value labels may be sparse, we average the loss over the labeled pixels N. We adopt the *Smooth* $L1$ loss to train our method MAN because of its robustness and low sensitivity to outliers. Therefore, we set the loss function as follows:

$$L(d, \hat{d}) = \frac{1}{N} \sum_{i=1}^{N} \text{smooth}_{L_1}\left(d_i - \hat{d}_i\right) \tag{7}$$



(a) Left image (b) PSMNet (c) Ours (d) Ground truth

FIGURE 5. Results of our model and PSMNet on Sceneflow test dataset.

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & if\,|x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

where d is the ground-truth disparity, $\hat{d}$ is the predicted disparity, and N is the total number of all labeled pixels.

IV. EXPERIMENTS AND DISCUSSION

We visualize the disparity maps generated by MAN and compare our method with others on Scene Flow [26], the KITTI stereo 2015 [45] and KITTI sterso 2012 [2] datasets. In addition, we will conduct multiple comparative experiments for the modules we proposed. In this section, we will present the experimental datasets, details, and results.

A. DATASETS AND EVALUATION METRIC

In this work, we use three public datasets to train and test our network.

1) SCENE FLOW

This is a synthetic stereo dataset that has 35454 training stereo image pairs and 4370 testing stereo image pairs. The image dimensions are $H = 540$ and $W = 960$. The dataset provides an elaborate and dense ground-truth disparity map. If a disparity is larger than the limits set in the experiment, we will abandon the pixels that have large disparities in the loss computation. For the Scene Flow dataset, we evaluate our network with the metric End-Point-Error (EPE), which is the average disparity error in pixels. We can also define the EPE as the average disparity error between the predicted disparity and the ground-truth disparity. The mathematical formula is designed as follows:

$$EPE(y, \hat{y}) = \frac{1}{NM} \sum_{j=1}^{N} \sum_{i=1}^{M} |y_i - \hat{y}_i| \tag{9}$$

where N is the number of the stereo image pairs for calculating end-point-error, M is the the number of pixels in the stereo image, y_i represents the disparity value of the i-th pixel in the predicted disparity map, and $\widehat{y_i}$ represents the disparity value of the i-th pixel in the ground-truth disparity map.

2) KITTI 2015

this is a real-word dataset from the perspective of a car with dynamic views, including urban, country and highway. It contains 200 training stereo pairs with sparse ground-truth disparities obtained by LiDAR and another 200 testing stereo pairs without ground-truth disparities. Both the stereo image pairs and ground-truth disparity have a size of (376 x 1280 pixels). To make the network have better performance, we will use all of the training data as a training set. For the KITTI 2015 dataset, we evaluate our network with the metrics: the percentage of disparity prediction outlier $D1$ (we consider a pixel to be correctly estimated if the disparity or flow end-point error is $< 3\,px$ or $< 5\%$) in the background ($D1$-bg), foreground ($D1$-fg) and all of the pixels ($D1$-all). The mathematical formula is designed as follows:

$$PE(y, \hat{y}) = \frac{1}{NM} \sum_{j=1}^{N} \sum_{i=1}^{M} \begin{cases} 1, & \text{if } |y_i - \hat{y}_i| \geq t \\ 0, & \text{if } |y_i - \hat{y}_i| < t \end{cases} \tag{10}$$

where PE is the pixel error rate, which refers to the proportion of pixels whose endpoint error exceed the threshold t, N is the number of the stereo image pairs for calculating end-point-error, M is the number of pixels in the stereo image,

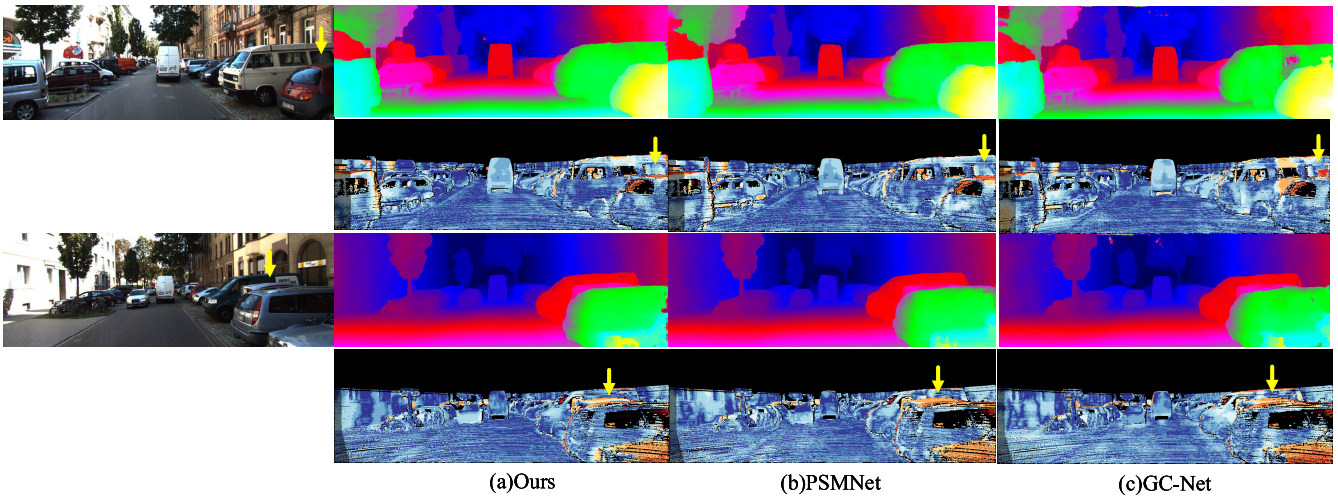


FIGURE 6. Comparison with other methods on the KITTI 2015 test dataset. The left panel shows the left input image of stereo image pair. For each input image, the disparity obtained by (a) MAN, (b) PSMNet, (c) GC-Net, is illustrated above the corresponding error maps.

y_i represents the disparity value of the *i-th* pixel in the predicted disparity map, and $\widehat{y_i}$ represents the disparity value of the *i-th* pixel in the ground-truth disparity map; when the threshold t is 1, 2, 3, and 5, the corresponding error rates are 1 pixel error rate ($1PE$), 2 pixel error rate ($2PE$), 3 pixel error rate ($3PE$) and 5 pixel error rate ($5PE$) respectively.

3) KITTI 2012

this is a real-word dataset from the perspective of a car with dynamic views, including urban, country and highway. It contains 194 training stereo pairs with sparse ground-truth disparities obtained by LiDAR and another 195 testing stereo pairs without ground-truth disparities. Both the stereo image pairs and ground-truth disparity have a size of (376 x 1280 pixels). To improve the performance of the network, we will set all of the training data as the training set. For the KITTI 2012 dataset, we use the percentage of bad pixels whose disparity errors are greater than a threshold ($> t$px); this percentage is denoted as t-pixel error.

B. EXPERIMENTAL DETAILS

Our module is implemented based on Pytorch. The model is optimized using the Adam method [48] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train our method with a batch size of 8 and images are randomly cropped from the input images to size $H = 256$ and $W = 512$. The maximum disparity size of the cropped images is 192. Our training is divided into two stages. In the first training stage, we fuse the Scene Flow dataset to pretrain the module. The initial learning rate is set to 0.0005. We also set the momentum to 0.9 and weight decay to 0.0001. We run SGD for 70K iterations in total; this operation takes 59 hours on four NVIDIA 1080 Ti Graphics Processing Units (GPUs).

In the second stage, we use the KITTI stereo 2015 dataset and the KITTI stereo 2012 dataset to fine-tune the module pretrained on Scene Flow. We set the maximum iteration to 25K (1000 epochs). The learning rate of the fine-tuning is set to 0.001 for the first 5K iterations, and reduced by 10^{-4} every 100 epochs. In the last 200 epochs, we train the module with a learning rate of 10^{-4}. The fine-tuning process takes approximately 17 hours and can obtain the final module.

To prove the effectiveness of the modules we proposed in this paper, we use the same data augmentation strategy as PSMNet, such as chromatic transformations (brightness, contrast and color) and spatial transformations (translation, rotation and scaling cropping).

C. EXPERIMENTAL RESULTS

1) KITTI 2015 BENCHMARK RESULTS

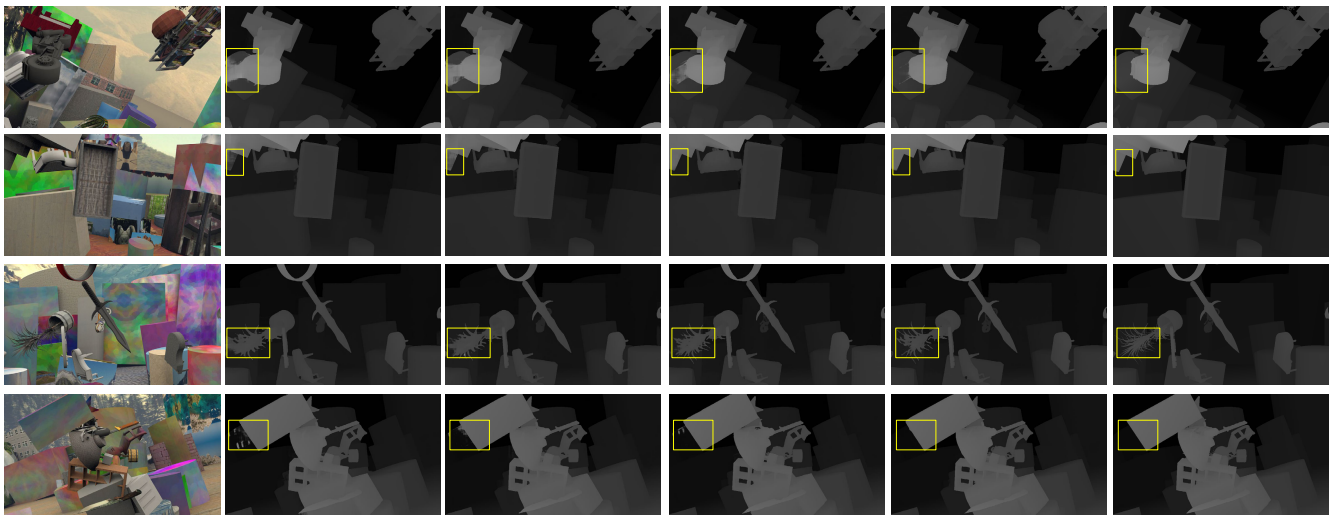
Since the ground truth disparity of the 200 testing images were not given, we calculated the disparity maps for the 200 testing images in the KITTI 2015 dataset and submitted the results to the KITTI evaluation server for the performance evaluation. Some disparity maps of the Multi-Attention Network on the KITTI 2015 test set are shown in Figure 4. We also compare our method with iResNet [46], GC-Net [36], DispNetC [26], CRL [27], MC-CNN-acrt [24], PSMNet [32], CFPNet [35] and SegStereo [47]. The comparison results are shown in Table 1. The results show, for all 200 test images, the percentage of pixels with a disparity error greater than three pixels or 5%. The qualifier '*bg*' refers to background pixels that contain static elements, '*fg*' refers to dynamic object pixels, and '*all*' refers to all pixels ($fg + bg$). We analyze the source of errors in other methods, and we find that most of the incorrect estimates are for regions with low texture quality, occlusions and reflections. The Table 1 shows that our method is effective, and our method outperforms the PSMNet algorithm and other methods in all performance evaluation indicators except speed. The all error rate is 2.10 on the KITTI stereo 2015 test datasets. Compared to that of PSMNet, it reductes 0.22.

TABLE 1. KITTI 2015 test set results. The online leaderboard ranks all methods according to the D1-all error of "All Pixels". The qualifier '*bg*' refers to background pixels that contain static elements, '*fg*' refers to dynamic object pixels, and '*all*' refers to all pixels (*fg* + *bg*).

Method	All PIxels			Non-Occluded			Runtime(s)
	D1-bg	*D1-fg*	*D1-all*	*D1-bg*	*D1-fg*	*D1-all*	
iResNet [46]	2.35	**3.23**	2.50	2.15	**2.55**	2.22	0.12
GC-Net [36]	2.21	6.16	2.87	2.02	5.58	2.61	0.9
DispNetC [26]	4.32	4.41	4.34	4.11	3.72	4.05	**0.06**
CRL [27]	2.48	3.59	2.67	2.32	3.12	2.45	0.47
MC-CNN-acrt [24]	2.89	8.88	3.89	2.48	7.64	3.33	67
PSMNet [32]	1.86	4.62	2.32	1.71	4.31	2.14	0.41
CFPNet [35]	1.90	4.39	2.31	1.73	3.92	2.09	0.95
SegStereo [47]	1.88	4.07	2.25	1.76	3.70	2.08	0.6
MAN	**1.71**	4.03	**2.10**	**1.57**	3.66	**1.92**	1.65

TABLE 2. Performance comparison of our proposed network with different settings. S, M and L represent different convulution kernel size. We computed the end-point-error and pixel percentages with errors larger than 1 pixel, 2 pixels and 3 pixels on the Scene Flow test set.

Network Settings								Experimental results on Scene Flow datasets			
Basic network	Multi-Scale Attention Module				3D Aggregation Network						
	S	*M*	*L*	*Attention mechanism*	*basic*	*FRM*	*3D AAM*	>1px	>2px	>3 px	End Point Error
✓					✓			9.896	5.972	4.406	1.039
✓			✓	✓	✓			10.165	6.289	4.769	1.109
✓		✓		✓	✓			9.831	5.730	4.380	1.026
✓	✓			✓	✓			9.657	5.612	4.275	0.924
✓	✓			✓	✓	✓		8.694	4.801	3.530	0.849
✓	✓			✓	✓	✓	✓	**8.488**	**4.647**	**3.419**	**0.832**



(a) Left image (b) Basic network (c) Basic network + MSAM (d) Basic network + MSAM + FRM (e) Basic network + MSAM + FRM+ 3D AMM (f) Ground truth

FIGURE 7. The qualitative comparison of the ablation experiments on the Scene Flow test dataset. Compared with the yellow boxes, the end-point-error rate gradually decreases, which can prove that our proposed modules are effective.

The left images in Figure 4 and Figure 6 are from the KITTI stereo 2015 test set, and the corresponding disparity maps and error maps are from the results submitted on the KITTI stereo 2015 benchmark. Our method can obtain more accurate disparity estimating by comparing the corresponding error maps, particularly in regions of car windows, as indicated blow the yellow arrows in Figure 6.

2) ABLATION EXPERIMENTS ON SCENE FLOW DATASETS

In this subsection, we analyze the effectiveness of each module of our network in details. We perform the ablation experiments on Scene Flow test datasets with the same configuration of the experimental environment. The relevant experimental results are shown in Table 2. In addition, we visualize some results of each ablation experiment and compare the corresponding disparity maps by the yellow boxes in Figure 7. The end-point-error rate gradually decreases, which can prove that our proposed modules are effective. Figure 5 illustrates some examples of the disparity maps estimated by the proposed PSMNet and ours. Our method obtains more robust results than PSMNet, as indicated by the yellow boxes in Figure 5.

TABLE 3. Performance comparison of our proposed network with different methods on KITTI 2012 datasets. We computed the pixel percentages with errors larger than 2, 3 and 5 pixels on KITTI 2012 test sets.

Method	> 2px		> 3px		> 5px		Mean Error	
	Noc	*All*	*Noc*	*All*	*Noc*	*All*	*Noc*	*All*
PSMNet [32]	2.44	3.01	1.49	1.89	0.90	1.15	0.5	0.6
PDSNet [49]	3.82	4.65	1.92	2.53	1.12	1.51	0.9	1.0
GC-Net [36]	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7
L-ResMatch [50]	3.64	5.06	2.27	3.40	1.50	2.26	0.7	1.0
MSDC-Net [39]	2.71	3.37	1.63	2.09	0.98	1.26	0.5	0.6
SegStereo [47]	2.66	3.19	1.68	2.03	1.00	1.21	0.5	0.6
PBCP [51]	3.62	5.01	2.36	3.45	1.62	2.32	0.7	0.9
SGM-Net [52]	3.60	5.15	2.29	3.50	1.60	2.36	0.7	0.9
MAN	**2.12**	**2.75**	**1.35**	**1.81**	**0.86**	**1.15**	**0.5**	**0.5**

TABLE 4. Results on the KITTI 2015 testing dataset with different weight values(β_1, β_2, β_3) for Loss 1, Loss 2, and Loss 3.

β_1	β_2	β_3	KITTI 2015 val error (%)			
			>1px	>2px	>3px	EPE
0	0	1	12.93	3.34	1.88	6.46
0.1	0.3	1	12.18	3.04	1.68	6.27
0.3	0.5	1	12.35	3.06	1.72	6.27
0.5	0.7	1	**12.13**	**3.01**	**1.67**	**6.21**
0.7	0.9	1	12.47	3.06	1.70	6.28
1	1	1	12.52	3.07	1.71	6.29

TABLE 5. Ablation experiments on the KITTI 2015 test datasets. We compute the percentages with errors larger than 3 pixels on the KITTI 2015 test sets.

Network Settings					Pixel error
Feature extraction		3D Aggregation Network			
basic network	MSAM	basic	FRM	3D AMM	>3px (%)
✓		✓			2.56
✓	✓	✓			2.49
✓	✓	✓	✓		2.16
✓	✓	✓	✓	✓	**2.10**

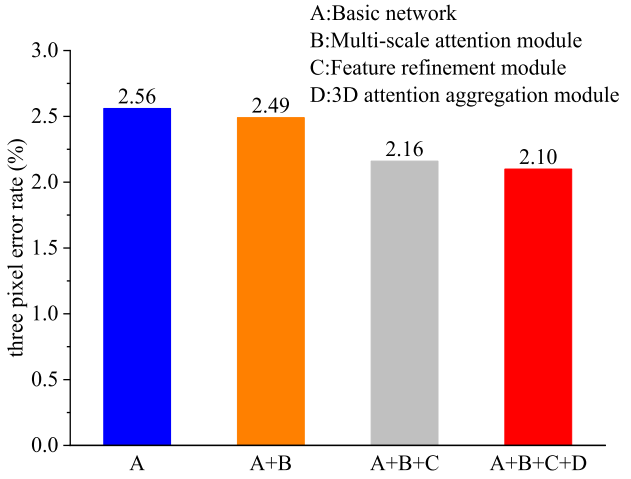


FIGURE 8. Ablation experiments on the KITTI 2015 test datasets. The blue histogram represents the three pixel error of basic network. The orange histogram, gray histogram, and red histogram are the results of the modules we proposed gradually added to the basic network. The three pixels error rate gradually decreases, which can prove that our proposed modules are effective.

Ablation for the multi-scale attention module: to make full use of the global context information, the encoder-decoder architecture uses stacked deconvolutional layers to recover the high-resolution prediction in the unary feature extraction, thereby creating too many parameters. In contrast, we combine the pyramid module with the attention mechanism to design a multi-scale attention module to effectively extract channel-wise information features at the pixel-level. Our experiments provide three specifications for multi-scale attention modules S(7-7-5-5-3-3-1-1), M(15-15-11-11-7-7-3-3) and L(21-21-15-15-9-9-3-3). The experimental results are shown in Table 2. When the size of the convolution kernel is too large, the receptive field will also increase, while this will lose a lot of texture details. When we use the convolution kernel size of the S specification, the neural network has the smaller end-point-error rate in the Scene Flow test set. The multi-scale attention module can increase the depth of the network and improve the matching performance of the network. The end-point-error on Scene Flow is reduced from 1.039 to 0.924. The experiments prove that the MSAM can obtain more valid global information to estimate disparity.

Ablation for feature refinement module: in the cost aggregation, we design a feature refinement module to enhance the representational ability of features in each stage. We perform the ablation experiment on the Scene Flow test dataset. The results of this ablation experiment are shown in Table 2. When the neural network introduces the feature refinement module, the end-point-error on Scene Flow is significantly reduced from 0.924 to 0.849. The experiments show that the feature refinement module we designed is effective.

Ablation for 3D attention aggregation module: we also introduce a 3D attention aggregation module to learn the high-level semantic information and low-level texture information. The features in different stages have different degrees of discrimination. We should extract the discriminative features and inhibit the indiscriminative features to obtain more valid image features matching for disparity estimation. The experimental results show that, when the network adopts the 3D attention aggregation module, the end-point-error is reduced by 0.017 on the Scene Flow test set.

3) KITTI 2012 BENCHMARK RESULTS

Similarly, we evaluate the Multi Attention Network on the KITTI 2012 stereo dataset and submit the testing results to the KITTI 2012 online leaderboard for the evaluation result. Figure 9 shows some qualitative results of our method on the KITTI 2012 benchmark. The performance of the proposed MNA, along with those competing methods, is presented in Table 3. Figure 10 illustrates some examples of the disparity maps together with the corresponding error maps estimated by the proposed MAN, PSMNet [32] and GC-Net [36]. Among these three methods, the Multi-Attention Network effectively yields precise estimations, as indicated by the black boxes in Figure 10.

4) ABLATION EXPERIMENT FOR LOSS WEIGHT

The 3D aggregation network also has three outputs for training. We divide the KITTI 2015 training data into a validation set (20%) and a training set (80%) and use the model trained with Scene Flow data to fine-tune on the KITTI training set for 300 epochs. We use the average end-point-error of the last

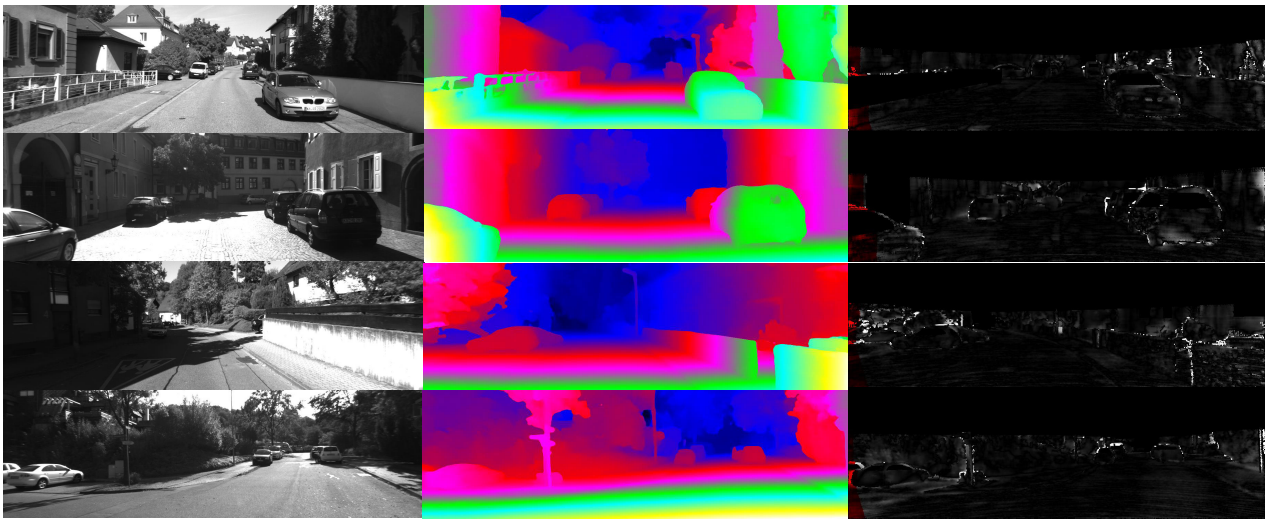


FIGURE 9. KITTI 2012 test data qualitative results. From the left: left stereo input image, our disparity prediction, error map.

TABLE 6. Ablation experiments on the KITTI 2012 test datasets.

Network Settings					KITTI 2012 val error (%)							
Feature extraction		3D Aggregation Network			Pixel error							
basic network	MSAM	basic	FRM	3D AMM	>2px		>3px		>4px		>5px	
					Noc	All	Noc	All	Noc	All	Noc	All
✓		✓			2.50	3.29	1.64	2.26	1.26	1.76	1.05	1.16
✓	✓	✓			2.49	3.27	1.62	2.23	1.26	1.74	1.04	1.45
✓	✓	✓	✓		2.22	2.83	1.43	1.86	0.91	1.18	0.91	1.18
✓	✓	✓	✓	✓	**2.12**	**2.75**	**1.35**	**1.81**	**1.04**	**1.40**	**0.86**	**1.15**

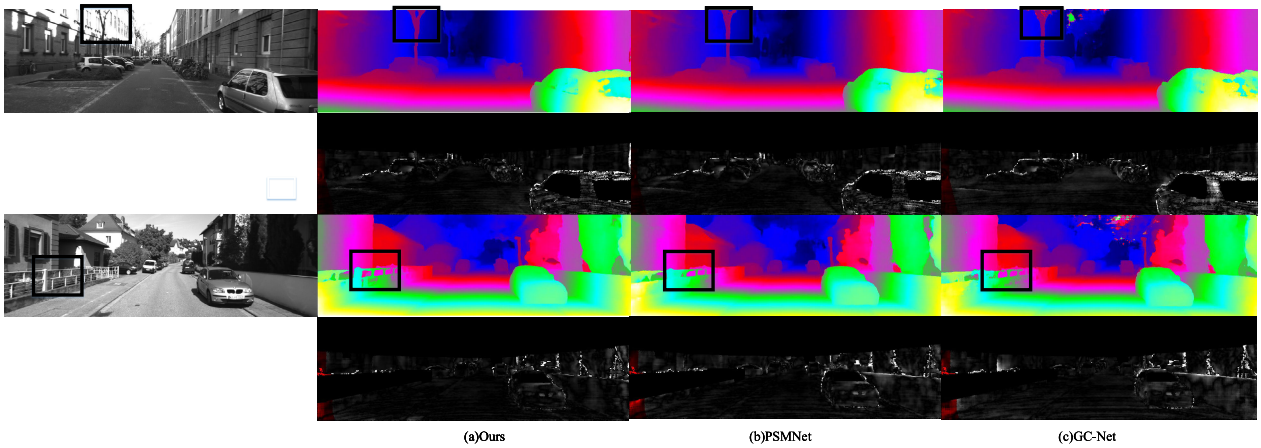


FIGURE 10. Results of disparity estimation for KITTI 2012 test sets. The left panel shows the left input image of the stereo image pair. For each input image, the disparity obtained by (a) Ours, (b) PSMNet [32], and (c) GC-Net [36], is illustrated above its error map. From the block boxes, we find that our method can achieve higher-precision matching compared with other methods.

100 epochs on the validation dataset as the evaluation index. The experimental results which applies various combinations of loss weights are shown in Table 4. When the weights of Loss 1, Loss 2 and Loss 3 are set as 0.5, 0.7 and 1.0, respectively, the network achieves the best performance, which is a 6.21% end-point-error on the KITTI 2015 validation dataset.

5) ABLATION EXPERIMENTS ON KITTI DATASETS

Furthermore, we perform the ablation experiments on the KITTI 2015 and KITTI 2012 training set to prove the effectiveness of our modules respectively. We set all the KITTI training set as a training set. Each ablation experiment finetunes 1000 epochs. We submit the results to the KITTI evaluation server for the performance evaluation. According to the online leader board, the corresponding results are shown in Table 5 and Table 6.

V. CONCLUSION

In this work, we propose a new efficient network (the Multi-Attention Network) for disparity matching. In the unary

feature extraction stage, we combine the spatial pyramid module with the attention mechanism to design the multi-scale attention module, which can obtain rich global context information. The multi-scale attention module has a good robustness and generalization ability; therefore, this module is also suitable for other computer vision tasks. In the cost aggregation, we introduce the feature refinement module, which can enhance the image feature representational ability of each stage. Additionally, to identify the high-quality channel attention vector, we design a 3D attention aggregation module which uses high-level semantic information to guide the low-level texture information, and then we combine them to reduce the loss both of high-level information and low-level information. We also simplify the stacked hourglass architecture and improve the performance of the neural network. The experiments demonstrate that our proposed modules are helpful for stereo matching. Our network achieves state-of-the-art performance on the Scene Flow, KITTI stereo 2015 and KITTI stereo 2012 benchmarks without any additional postprocessing or regularization.

In future work, we will learn how to combine edge detection algorithms and stereo-matching algorithms into a new multitask neural network that utilizes the edge information generated by the edge detection subnetwork to better guide and assist the stereo matching algorithm to obtain more precise disparity maps.

REFERENCES

[1] K. Schmid, T. Tomic, F. Ruess, H. Hirschmuller, and M. Suppa, "Stereo vision based indoor/outdoor navigation for flying robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3955–3962.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[3] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, Jun. 2018.

[4] C. Bourdy, "3D reconstruction and interpretation in human binocular vision by processing of disparity information," *J. Opt.*, vol. 20, no. 6, p. 243, 1989.

[5] B. Musleh, D. Martin, J. M. Armingol, and A. de la Escalera, "Continuous pose estimation for stereo vision based on UV disparity applied to visual odometry in urban environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 3983–3988.

[6] J. P. Hespanha, Z. Dodds, G. D. Hager, and A. S. Morse, "Decidability of robot positioning tasks using stereo vision systems," in *Proc. 37th IEEE Conf. Decis. Control*, Dec. 1998, pp. 3736–3741.

[7] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[8] D. Kong and H. Tao, "A method for learning matching errors for stereo computation," in *Proc. Brit. Mach. Vis. Conf.*, 2004, p. 2.

[9] 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.

[10] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5248–5257.

[11] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.

[12] M. Gerrits and P. Bekaert, "Local stereo matching with segmentation-based outlier rejection," in *Proc. 3rd Can. Conf. Comput. Robot Vis. (CRV)*, Jun. 2006, p. 66.

[13] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 807–814.

[14] Y. Li, Q. Xie, H. Huang, and Q. Chen, "Research on a tool wear monitoring algorithm based on residual dense network," *Symmetry*, vol. 11, no. 6, p. 809, Jun. 2019.

[15] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a surface defect detection algorithm based on MobileNet-SSD," *Appl. Sci.*, vol. 8, no. 9, p. 1678, Sep. 2018.

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[17] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[20] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Netw.*, vol. 121, pp. 461–473, Jan. 2020.

[21] C. Tian, Y. Xu, L. Fei, and K. Yan, "Deep learning for image denoising: A survey," in *Proc. Int. Conf. Genetic Evol. Comput.* Singapore: Springer, 2018, pp. 563–572.

[22] N. Dhungel, G. Carneiro, and A. P. Bradley, "Fully automated classification of mammograms using deep residual neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 310–314.

[23] W. Lotter, G. Sorensen, and D. Cox, "A multi-scale cnn and curriculum learning strategy for mammogram classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Singapore: Springer, 2017, pp. 169–177.

[24] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, nos. 1–32, p. 2, 2016.

[25] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 972–980.

[26] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[27] 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 (ICCVW)*, Oct. 2017, pp. 887–895.

[28] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 573–590.

[29] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *Proc. Asian Conf. Comput. Vis.* Singapore: Springer, 2018, pp. 20–35.

[30] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3273–3282.

[31] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for End-To-End stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.

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

[33] P. Brandao, E. Mazomenos, and D. Stoyanov, "Widening siamese architectures for stereo matching," *Pattern Recognit. Lett.*, vol. 120, pp. 75–81, Apr. 2019.

[34] H. Lu, H. Xu, L. Zhang, Y. Ma, and Y. Zhao, "Cascaded multi-scale and multi-dimension convolutional neural network for stereo matching," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.

[35] Z. Zhu, M. He, Y. Dai, Z. Rao, and B. Li, "Multi-scale cross-form pyramid network for stereo matching," 2019, *arXiv:1904.11309*. [Online]. Available: http://arxiv.org/abs/1904.11309

[36] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.
[37] C.-W. Xie, H.-Y. Zhou, and J. Wu, "Vortex pooling: Improving context representation in semantic segmentation," 2018, *arXiv:1804.06242*. [Online]. Available: http://arxiv.org/abs/1804.06242
[38] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "StereoDRNet: Dilated residual StereoNet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11786–11795.
[39] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "MSDC-Net: Multi-scale dense and contextual networks for automated disparity map for stereo matching," 2019, *arXiv:1904.12658*. [Online]. Available: http://arxiv.org/abs/1904.12658
[40] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.
[41] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180
[42] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1857–1866.
[43] 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. 2881–2890.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3061–3070.
[46] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2811–2820.
[47] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 636–651.
[48] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011.
[49] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (PDS): Toward applications-friendly deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5871–5881.
[50] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4641–4650.
[51] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 4.
[52] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 231–240.

XIAOWEI YANG received the bachelor's degree in measurement and control technology and instrument specialty from the Nanyang Institute of Technology, Nanyang, China, in 2013, and the M.E. degree from the School of Mechanical Engineering, Guizhou University, Guiyang, China, in 2016, where he is currently pursuing the Ph.D. degree. His research interests include pattern recognition, computer vision, deep learning, and 3-D reconstruction.

LIN HE received the bachelor's degree in precision machinery design and manufacture specialty from the Chengdu University of Science and Technology, Chengdu, China, in 1987, the degree in mechanical manufacturing and automation subject from the Guizhou Institute of Technology, Guiyang, China, in 1990, and the Ph.D. degree in mechanical manufacturing and automation subject from Shandong University, Jinan, China, in 2003. From 1990 to 1998, he worked as a Research Assistant at the Department of Mechanical Engineering, Guizhou Institute of Technology (now Guizhou University). From 1998 to 2003, he worked as the Vice Dean/Associate Professor at the Department of Mechanical Engineering, Guizhou University. From 2004 to 2008, he worked as the Deputy Director for science and technology at Guizhou University. From 2008 to 2014, he worked as the Dean of Guizhou Education University, where he also worked as the Dean, from 2014 to 2017. Since 2017, he has been working as the President of the Liupanshui Normal College, Liupanshui, China. His research interests and areas include the advanced processing technology and equipment, automatic control, and image processing.

YONG ZHAO (Member, IEEE) received the B.S. degree in mathematics from Guizhou University, Guiyang, China, in 1985, the M.E. degree from Northwestern Polytechnic University, Xi'an, China, in 1988, and the Ph.D. degree in engineering from the Research Institute of Automation, Southeast University, Nanjing, China, in 1991. From 1991 to 1994, he worked as a Research Assistant at the Department of Biomedical Engineering, Zhejiang University. From 1994 to 1997, he worked as the Vice Dean/Associate Professor at the Department of Electrical Engineering, Hangzhou University (now Zhejiang University). From 1997 to 2000, he worked as a Postdoctoral Fellow at the Center for Signal Processing and Communications, Electrical and Computer Engineering, Concordia University, Montreal, Canada. From 2000 to 2004, he worked as a Senior Software Engineer at Honeywell Corporation, Ottawa, Canada. Since 2004, he has been with the Peking University Shenzhen Graduate School, Shenzhen, China, where he is currently an Associate Professor. His research interests and areas of publication include video codecs, signal processing, and machine learning.

HAIWEI SANG received the bachelor's degree in computer science from Henan Normal University, Xinxiang, China, in 2011, and the M.E. degree in computer science from Guizhou University, Guiyang, China, in 2014, where he is currently pursuing the Ph.D. degree in software engineering. His research interests include pattern recognition, computer vision, and image processing.

ZU LIU YANG received the bachelor's degree in microelectronics science and engineering from the Hefei University of Technology, Hefei, China, in 2017. He is currently pursuing the M.E. degree in software engineering, with a specialization in multimedia technology, with Peking University. His research interests include image processing, machine learning, and deep learning.

XIAN JING CHENG received the bachelor's degree in computer science from the Shenyang Institute of Engineering, Shenyang, China, in 2011, and the master's degree in computer science from Nanjing Normal University, Nanjing, China, in 2015. He is currently pursuing the Ph.D. degree in soft engineering with Guizhou University. His research interests include pattern recognition and computer vision.