Computing and Informatics, Vol. 43, 2024, 505–528, doi: 10.31577/cai_2024_2_505

CGCANET: CONTEXT-GUIDED COST AGGREGATION NETWORK FOR ROBUST STEREO MATCHING

Wenmei Sun, Yuan Zheng

Inner Mongolia University
College of Computer Science
Huhhot, China
e-mail: zhengyuan@imu.edu.cn

Abstract. Stereo matching methods based on Convolutional Neural Network (CNN) have achieved a significant progress in recent years. However, they still cannot work well on generalization performance across a variety of datasets due to their poor robustness. In view of this, we aim to enhance the robustness in three main steps of stereo matching, namely cost computation, cost aggregation, and disparity refinement. For cost computation, we propose an atrous pyramid grouping convolution (APGC) module, which combines local context information with multi-scale features generated from CNN backbone, aiming to obtain a more discriminative feature representation. For cost aggregation, we provide a multi-scale cost aggregation (MSCA) module, which sufficiently and effectively fuses multiple cost volumes at three different scales into the 3D hourglass networks to improve initial disparity estimation. In addition, we present a disparity refinement (DR) module that employs the color guidance of left input image and several convolutional residual blocks to obtain a more accurate disparity estimation. With such three modules, we propose an end-to-end context-guided cost aggregation network (CGCANet) for robust stereo matching. To evaluate the performance of the proposed modules and CGCANet, we conduct comprehensive experiments on the challenging SceneFlow, KITTI 2015 and KITTI 2012 datasets, with a consistent and competitive improvement over the existing stereo matching methods.

Keywords: Stereo matching, cost computation, cost aggregation, disparity refinement

Mathematics Subject Classification 2010: 68-T45

1 INTRODUCTION

Stereo matching is indispensable task for many computer vision applications. It takes two rectified images as the input and attempts to compute the disparity of every pixel by matching the corresponding pixels along conjugate epipolar lines. For higher-level computer vision tasks such as 3D reconstruction, robot navigation and autonomous driving, it is crucial to obtain dense and accurate disparity maps.

The traditional stereo matching methods usually were formulated as a multi-stage optimization problem, containing the following four steps: cost computation, cost aggregation, disparity estimation and disparity refinement [1, 2]. Firstly, cost computation aims to calculate the pixel-wise costs by evaluating similarity measure between left input image and right input image via absolute differences or relative differences. Secondly, cost aggregation is to filter out incorrect cost values and improve the quality of cost calculation by incorporating priors and contextual matching costs. Next, disparity estimation is supposed to obtain an initial disparity map by searching for the minimum matching cost. Finally, disparity refinement aims to refine the initial disparity map due to the existence of occlusions, reflections and noise.

The existing CNN-based stereo matching methods mostly follow such a pipeline of the traditional methods, which consist of three main modules, namely cost computation module, cost aggregation module and disparity refinement module. The cost computation module generally employs several convolutional layers to extract features from input images and then construct the cost volumes. Cost aggregation module aims to aggregate these cost volumes for initial disparity estimation, while disparity refinement module aims to refine the initial disparity to obtain an accurate disparity estimation.

Existing CNN-based stereo matching methods generally have poor generalization performance from synthetic data to real-world scenes due to their poor robustness. As shown in Figure 1, the existing CNN-based methods including PSMNet [3] are pretrained on SceneFlow dataset [4] that is a synthetic data and then tested on real KITTI datasets [5, 6], and moreover we observe that PSMNet produces poor disparity predictions, referring to the red and black boxes in this Figure. We believe that improving robustness in cost computation module, cost aggregation module and disparity refinement module will ultimately boost the performance of stereo matching.

Within these three modules, cost aggregation module plays an important role for improving the performance since it integrates multi-scale feature information to generate an initial disparity. The existing state-of-the-art methods employ multi-scale cost volumes for cost aggregation shown in Figures 2 a) and 2 b). From Figure 2, we find that the fusion of multi-scale cost volumes is not sufficient and effective. In view of this, we propose a multi-scale cost aggregation (MSCA) module shown in Figure 2 c), where multi-scale cost volumes information is fully utilized and aggregated into all 3D hourglass networks to produce an accurate initial disparity.

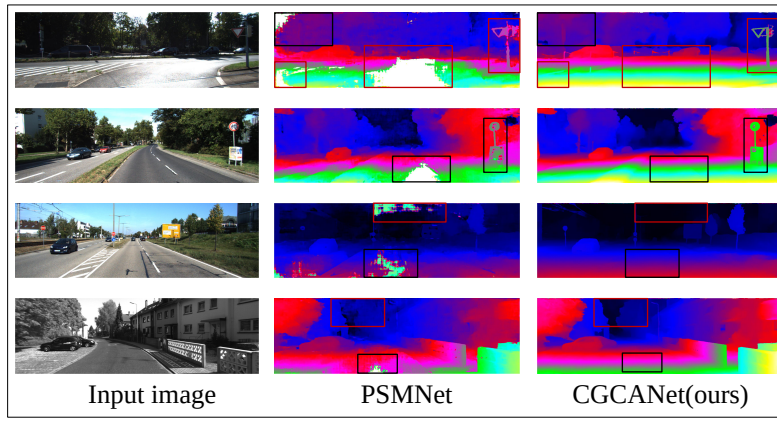

Figure 1. Visualization of the disparity results. Models are pre-trained on a synthetic SceneFlow dataset [4] and tested on real-world KITTI datasets [5, 6]. Obviously, our CGCANet can produce more accurate disparity estimations than PSMNet [3] on the same test dataset, referring to the red and black boxes in this figure.

Besides, a more discriminative feature representation, especially when the textureless, occluded and reflective regions appear, is certainly required for constructing the cost volumes. In view of this, we propose an atrous pyramid grouping convolution (APGC) module. APGC module not only uses several grouping convolutions with different dilation rates to capture local context information, but also combines the features generated from CNN backbone at different scales. To obtain more accurate depth estimation, an initial disparity is supposed to be further refined, and hence we propose a disparity refinement (DR) module. By means of the color guidance of left input image, DR module adopts several convolutional residual blocks to produce the final disparity estimation. With the proposed APGC, MSCA and DR modules, we present an end-to-end context-guided cost aggregation network (CGCANet) for robust stereo matching.

We summarize our contributions as follows:

- We observe that local context information is helpful for stereo matching and hence we propose a APGC module. It fuses both local context and multi-scale features generated from CNN backbone, aiming to achieve a more discriminative feature representation for cost computation.
- We believe that a sufficient and effective fusion of multi-scale cost volumes is helpful for cost aggregation. We propose a MSCA module that makes full use of cost volume information at three different scales by fusing them into all hourglass networks to improve the initial disparity estimation.
- We present an end-to-end trainable convolutional neural network for robust stereo matching, named CGCANet. It consists of the proposed APGC, MSCA and DR modules for feature extraction, cost aggregation and disparity refine-

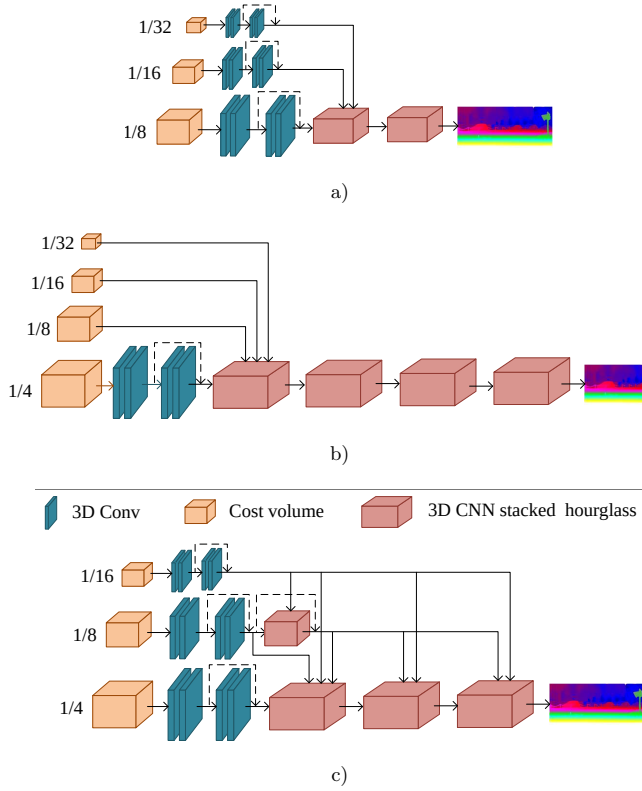

Figure 2. Comparison between our CGCANet and existing methods in terms of cost aggregation. a) CFNet [7]. b) MSMD-Net [8]. c) Our CGCANet. Different from CFNet and MSMD-Net, our CGCANet aims to make full use of multi-scale cost volumes and aggregate them into all 3D hourglass networks to produce an accurate initial disparity.

ment respectively, aiming to obtain more accurate and robust disparity estimation.

- We evaluate the proposed modules and CGCANet on the challenging SceneFlow, KITTI 2012 and KITTI 2015 datasets. The comprehensive experiments demonstrate their superiority and effectiveness.

2 RELATED WORK

2.1 Feature Extractor/Feature Representation for Cost Computation

As a pioneering work, PSMNet [3] adopts a spatial pyramid pooling (SPP) for feature extraction. SPP module is able to learn the relationship between objects and their sub-regions to effectively capture context information using several dilated

convolutions. To further improve the performance of PSMNet, Cheng et al. [9] propose a Convolutional Spatial Propagation Network, which is capable of learning an affinity matrix directly from images for both feature aggregation and spatial propagation of features. PyConv [10] presents a pyramid of grouping convolutions, where the filters of various types with different sizes and depths are utilized at different levels to capture the details at multiple levels in the scene.

For cost computation, DispNetC [4] is the first end-to-end training network for disparity estimation, which employs a 1D correlation layer to generate the cost volume by measuring the similarity of left and right features. Another representative method is called GC-Net [11] that is the first to use 3D convolutional networks to produce 4D cost volume by directly concatenating left and right features. As an extension to the concatenation-based methods, GWCNet [12] proposes to construct the cost volume via group-wise correlations which provides an efficient representations for measuring feature similarities and does not lose too much information unlike full correlation [4].

Inspired by PyConv [10], we propose an atrous pyramid grouping convolution (APGC) module. Different from PyConv [10], APGC module not only employs the same small convolution kernel as well as different dilation rates for extracting context information, but it also combines the generated features from the backbone network at different levels, with the aim of achieving a more robust and discriminative feature representation. After feature extraction, we follow the same cost computation method as GWCNet [12].

2.2 Cost Aggregation for Initial Disparity Estimation

PSMNet [3] proposes a 3D hourglass architecture, which can integrate top-down and bottom-up feature information through shortcut connections to leverage the context for cost volume regularization. The latest works provide some solutions to integrate multi-scale cost volumes to generate a high-quality disparity map. AANet [13] correlates left and right image features at corresponding scales to construct a multi-scale 3D cost volumes, and then feeds them to several stacked adaptive aggregation modules. CFNet [7] fuses multiple low-resolution dense cost volumes to enlarge the receptive field and extracts a robust structural representation for initial disparity estimation. MSMD-Net [8] introduces an encoder-decoder structure to directly fuse the cost volumes at different scale to generate initial disparity maps.

Different from CFNet [7] and MSMD-Net [8], which only fuse the cost volumes at multiple scales into the first hourglass network, our MSCA module sufficiently and effectively combines the cost volumes at multiple scales with all hourglass networks to generate more accurate and robust initial disparity maps.

2.3 Disparity Refinement Based Deep Stereo Matching

The disparity refinement has been integrated into many end-to-end deep stereo matching methods in recent years. CRL [14] is a two-stage architecture where the

first stage extends DispNet [4] to obtain an initial disparity map and the second stage is responsible for computing residual corrections to refine the initial disparity map. EdgeStereo [15] is an effective multi-task learning network that achieves a well-performed disparity estimation due to the combination of edge cues. MSMD-Net [8] introduces a warped correlation cost volume to generate a fine-grained disparity searching range to guide the disparity refinement. LWSN [16] designs a color guidance refinement step with depthwise separable convolutions to refine the predicted initial disparity, increasing the accuracy of disparity estimation.

Similar to LWSN [16], we employ the color guidance of left input image. Different from LWSN [16], we first concatenate the left input image and initial disparity map, and then employ several convolutional residual blocks to refine disparity estimation.

3 METHODOLOGY

In this section, we first elaborate APGC module for obtaining a more discriminative feature representation, MSCA module for sufficiently and effectively aggregating multi-scale cost volumes and DR module for refining disparity estimation in turn. Then we introduce the network architecture of the proposed CGCANet.

3.1 APGC Module

Considering that local context information is critical for cost computation, PyConv [10] proposes a pyramid of different types of kernels to capture context information, where grouping convolutions are used to obtain the kernels of different depths. PyConv is able to process the inputs using the filters with different kernels and then capture local features at different levels. The work in [17] shows that the dilated convolutions can enlarge the receptive fields which is helpful to gather local context information. Inspired by these works, we propose an atrous pyramid grouping convolution (APGC) module, where each level of the pyramid employs the same 3×3 grouping convolutions with different dilation rates, as shown in Figure 3.

In addition, we argue that the effective feature fusion can further improve the robustness of disparity estimation. Accordingly, we carefully design a strategy for feature fusion. Given an input image $X \in R^{H \times W \times 3}$, we feed X into CNN backbone for feature extraction, and the outputs are $conv1$, $conv2$, $conv3$ and $conv4$ feature maps. As shown in Figure 3, $conv4$ is first processed by a pyramid of grouping dilated convolutions mentioned above to obtain four feature maps, namely $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ and $\mathcal{F}_4$. Then, such four feature maps are concatenated and added to $conv4$ feature map. Finally, a concatenation operation is performed on the obtained feature maps and both $conv2$ and $conv3$ feature maps. The output $Y \in R^{H/4 \times W/4 \times 320}$ of APGC module is formulated by

$$\mathcal{F}_i = Conv^{3\times3}_{Group\_Dilated}\left(conv4\right), \quad i \in [1,2,3,4], \tag{1}$$

$$Y = Cat\left[Cat\left[\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4\right] + conv4, conv3, conv2\right], \tag{2}$$

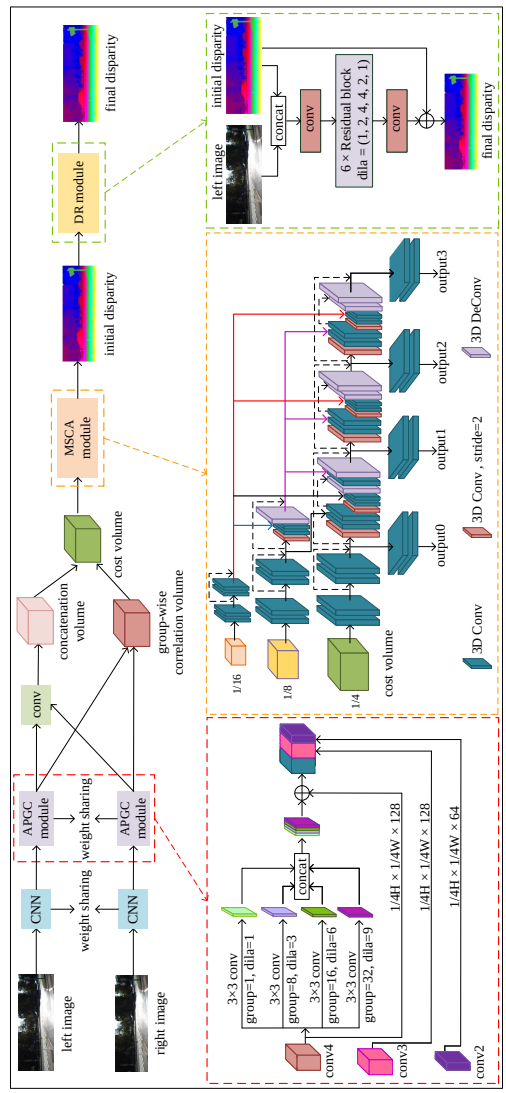

Figure 3. Overview of the proposed CGCANet for stereo matching, which consists of APGC module, MSCA module and DR module for feature extraction, cost aggregation and disparity refinement, respectively. Different from PyConv [10], our APGC module aims to combine local context with multi-scale feature maps, obtaining a more discriminative feature representation. Different from CFNet [7] and MSMD-Net [8], our MSCA module can aggregate multi-scale cost volumes into all hourglass networks by adding several concatenation operations denoted by the red, blue and purple color lines, with the purpose of generating more accurate initial disparity. Different from LWSN [16], our DR module adopts the color guidance of left input image and several convolutional residual blocks with different dilation rates to improve the disparity estimation.

where $Conv^{3\times3}_{Group\_Dilated}(\cdot)$ means performing a 3×3 grouping convolution with a dilation rate and $Cat[,]$ denotes a concatenation operation.

Different from PyConv [10], APGC module not only employs the convolutions with different dilation rates but also combines the multi-scale features from backbone network. Obviously, APGC module helps to obtain a more discriminative feature representation for cost computation by combining local context information with multi-scale features from CNN backbone. We conduct extensive ablation studies on the APGC module in the part of Experiments.

3.2 MSCA Module

Cost Volume Construction. We employ the discriminative feature representation produced by APGC module to construct the combination of cost volumes. Similar to GWCNet [12] and CFNet [7], the combination of cost volumes contains the concatenation and group-wise correlation of cost volumes, which can be formulated as

$$C_{concat}(d,x,y,f) = f_l(x,y) \,||\, f_r(x-d,y), \tag{3}$$

$$C_{gwc}(d,x,y,g) = \frac{1}{N_c/N_g} \langle f_l^g(x,y), f_r^g(x-d,y) \rangle, \tag{4}$$

$$C_{combination} = C_{concat} \,||\, C_{gwc}, \tag{5}$$

where $C(d,x,y,\cdot)$ denotes the cost volume at location (x,y) for disparity candidate d, and f and f^g denote the feature and the g^{th} feature group after APGC module, respectively. $||$ denotes the vector concatenation operation at the feature dimension, $\langle,\rangle$ represents the inner product, and N_c and N_g represent the channel number of the feature and the number of the grouping, respectively.

MSCA Module. Some existing methods, such as CFNet [7] and MSMD-Net [8], aim to fuse the multi-scale cost volumes. However, these methods still suffer from insufficient and ineffective information fusion. In view of this, we propose a multi-scale cost aggregation (MSCA) module.

As shown in Figure 3, MSCA module first uses 3D convolutions with $stride = 2$ to continuously downsample the 1/4 cost volume of an input image to generate both 1/8 and 1/16 cost volumes. Then, it employs four 3D convolution layers with skip connections to regularize these three cost volumes. For 1/4 cost volume, we arrange three 3D hourglass networks that combine both 1/8 and 1/16 cost volume information, where the shortcut connections are preserved to avoid information loss. Besides, we add a simple hourglass network for 1/8 cost volume that fuses 1/16 cost volume information. Finally, two 3D convolution layers are used to obtain a 1-channel 4D cost volume and then such a 4D cost volume is upsampled to the original input image resolution, which correspond to $output0$, $output1$, $output2$ and $output3$ in MSCA module. To generate initial disparity maps, a soft-argmin operation in [11] is adopted.

Different from CFNet [7] and MSMD-Net [8], our CGCANet is to make full use of three cost volumes at different scales and aggregate them into all 3D hourglass networks by adding several concatenation operations denoted by the red, blue and purple color lines in Figure 3. With such these operations, a 3D hourglass network at 1/8 cost volume fuses 1/16 cost volume information, and three 3D hourglass networks at 1/4 cost volume sufficiently combine both 1/8 and 1/16 cost volume information. Obviously, our MSCA module makes the cost aggregation of multi-scale cost volumes more sufficient and effective, helping obtain a more accurate and robust initial disparity. We conduct extensive ablation studies on the MSCA module in the part of Experiments.

3.3 DR Module

The initial disparity generated by MSCA module needs to be refined. LWSN [16] adopt a color guidance refinement to improve disparity estimation. Inspired by LWSN [16], we propose a disparity refinement (DR) module. It refines the initial disparity via the color guidance of left input image and employs six convolutional residual blocks to perform feature refinement for generating more accurate disparity estimation, as shown in Figure 3. The dilation rates in such six convolutional residual blocks are set to 1, 2, 4, 4, 2 and 1. Different from LWSN [16], our CGCANet employs several convolutional residual blocks to refine disparity estimation after concatenating the left input image and initial disparity map. Obviously, our DR module adopts the color guidance of left input image and several convolutional blocks with different dilation rates to improve the final disparity estimation. We conduct extensive ablation studies on the DR module in the part of Experiments.

Name	**Layer Setting**	**Output**
conv0_1	3×3, 32	$H \times W \times 32$
conv0_2	3×3, 32	$H \times W \times 32$
conv0_3	3×3, 32	$H \times W \times 32$
conv1	$\begin{bmatrix} 3 \times 3, & 32 \\ 3 \times 3, & 32 \end{bmatrix} \times 3$	$H/2 \times W/2 \times 32$
conv2	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 18$	$H/4 \times W/4 \times 64$
conv3	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 3$, dila $= 1$	$H/4 \times W/4 \times 128$
conv4	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 3$, dila $= 2$	$H/4 \times W/4 \times 128$

Table 1. Settings of CNN backbone network for feature extraction

3.4 Network Architecture

Figure 3 illustrates the architecture of the proposed CGCANet. The left and right input images are fed to a weight-sharing CNN backbone network for feature extraction, where the settings of the backbone network are listed in Table 1. The feature maps after the CNN backbone network are then processed by the proposed APGC module, MSCA module and DR module in turn to generate final disparity maps.

Disparity Regression. Following the work in [11], we use the disparity regression to estimate the continuous disparity maps. As reported in [11], such a disparity regression is more robust than classification-based stereo matching methods. The predicted disparity $\tilde{d}$ is calculated by a soft-argmin function

$$\tilde{d} = \sum_{d=0}^{d_{max}} d_l \cdot \sigma\left(-C_d\right), \tag{6}$$

where d_l and $\sigma\left(-C_d\right)$ denote the possible disparity level and the corresponding probability, respectively. The probability of each disparity d is calculated by the predicted cost C_d via a softmax operation denoted by $\sigma\left(\cdot\right)$.

Loss. We adopt smooth L_1 loss to train the proposed CGCANet. Smooth L_1 loss is widely used for bounding box regression due to its robustness and low sensitivity to outliers. The disparity maps generated from our CGCANet are denoted by $\hat{d}_0$, $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_3$ and $\hat{d}_4$, where ($\hat{d}_0$, $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_3$) correspond to (*output*0, *output1*, *output*2, *output*3) of MSCA module. The training loss of our CGCANet is defined as

$$L = \sum_{i=0}^{i=4} \lambda_i \cdot Smooth\_L_1\left(d^* - \hat{d}_i\right), \tag{7}$$

$$Smooth\_L_1\left(x\right) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise}, \end{cases} \tag{8}$$

where λ_i represents the weight for the i^{th} prediction of disparity map and d^* denotes the ground-truth disparity map.

Obviously, the training loss is the weighted sum of five prediction losses, where four prediction losses are from the MSCA module and the remaining one loss is from the DR module. We conduct such ablation studies to explore which of these five losses has the greatest impact and how they are supposed to be combined for improving performance of stereo matching. The results show that using the last prediction loss produced by DR module yields a best performance when only one prediction loss is used. As for the fusion strategy of these five prediction losses, the results show that a best disparity estimation is achieved when the combination of $(\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is set to $(0, 1, 1, 1, 1)$. In other words, the best performance can be obtained when using the last four prediction losses.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

SceneFlow [4] is a large scale synthetic dataset for optical flow and stereo matching, which contains 35, 454 training and 4, 370 testing image pairs with the image size of 960×540. It is composed of Flyingthings3D, Driving and Monkaa with dense and accurate ground-truth for training. Here we use it to pretrain our CGCANet.

KITTI 2012 [5] is the first real-world dataset with street views captured from a driving car. It contains 194 stereo image pairs for training with sparse ground-truth disparities obtained by LiDAR and another 195 stereo image pairs for testing without ground-truth disparities. Each image size is $1\,240 \times 376$. Here we use 180 training image pairs as a training set and the rest as a validation set.

KITTI 2015 [6] is an extended real-world dataset with street views captured from a driving car. It contains 200 stereo image pairs for training with sparse ground-truth disparities obtained by LiDAR and another 200 stereo image pairs for testing without ground-truth disparities. Each image size is $1\,240 \times 376$. Here we use 180 training image pairs as a training set and the rest as a validation set.

Evaluation Metrics. We compute end-point error (EPE) on SceneFlow test set and KITTI validation sets, which is the mean of average disparity error in pixels. For KITTI 2012 validation set, the percentages of three-pixel error rate, namely *3-px*, is reported for all pixels. For KITTI 2015 validation set, the percentage of disparity outliers error rate, namely *D1-all*, is evaluated for all pixels.

4.2 Implementation Details

We use PyTorch framework to implement our CGCANet network and employ Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The batch size is fixed to 8 for training on one NVIDIA RTX A6000. We set the maximum disparity value as $D_{max} = 192$ during training and testing. Some pixels are excluded in the loss computation if the disparity is larger than the limits set in our experiments.

For SceneFlow dataset, we train our model from scratch for 36 epochs in total. The initial learning rate is set to 0.001 and down-scaled by 2 after epoch 16, 24, and 30. To evaluate the methods on SceneFlow test set, the trained model is directly used for testing.

For KITTI datasets, we merge KITTI 2012 and KITTI 2015 training sets as a KITTI mix training set. We use the model pretrained with SceneFlow dataset to finetune on KITTI mix training set for 600 epochs, where the learning rate of the finetuning process begins at 0.001 and decreases by 10 after epoch 300. Then, we use the model trained with KITTI mix dataset to finetune on KITTI 2012 and KITTI 2015 datasets for 100 epochs, where the learning rate is 0.001. Moreover,

we prolong the training process to 1000 epochs on KITTI mix dataset to obtain the final model for KITTI submission.

4.3 Ablation Studies

Ablation Study on APGC Module. We first discuss the impact of the feature fusion strategy on APGC Module, and the results are listed in Table 2. The results show that the best performance is achieved on KITTI mix, KITTI 2012 and KITTI 2015 datasets when combining both *conv*4, *conv*2 and *conv*3, which demonstrates that a sufficient fusion of multi-scale features from the backbone network certainly improves the performance of stereo matching.

Add conv4	Cat conv2	Cat conv3	KITTI mix EPE↓	KITTI mix D1-all↓	KITTI 2015 EPE↓	KITTI 2015 D1-all↓	KITTI 2012 EPE↓	KITTI 2012 3-px↓
	✓	✓	0.367	0.694	0.452	0.851	0.456	1.137
✓		✓	0.365	0.711	0.463	0.879	0.452	1.122
✓	✓		0.355	0.687	0.458	0.881	0.442	1.102
✓			0.364	0.705	0.462	0.917	0.460	1.233
✓	✓	✓	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**

Table 2. Ablation study of feature fusion strategy of APGC module on KITTI validation sets

Dilation Rates				KITTI mix EPE↓	KITTI mix D1-all↓	KITTI 2015 EPE↓	KITTI 2015 D1-all↓	KITTI 2012 EPE↓	KITTI 2012 3-px↓
1	1	1	1	0.363	0.697	0.444	0.846	0.430	1.075
1	2	4	6	0.362	0.752	0.455	0.854	0.436	1.058
1	**3**	**6**	**9**	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**
1	4	8	12	0.366	0.689	0.444	0.815	0.454	1.124
1	6	12	18	0.354	0.696	0.445	0.826	0.430	1.052

Table 3. Ablation study of dilation rates of APGC module on KITTI validation sets

Next, we explore the influence of the dilation rates used in APGC module and the results are listed in Table 3. As can be seen from Table 3, the best performance is obtained on KITTI mix, KITTI 2012 and KITTI 2015 datasets when the dilation rates are set to 1, 3, 6 and 9. It is worth noting that too small or too large dilation rates will reduce the stereo matching performance. Therefore, it is crucial to select an appropriate dilation rates to obtain the reasonable receptive field.

Finally, we compare the proposed APGC module with the popular context extraction modules, including PyConv [10], SPP module [18] and ASPP module [19]. As shown in Table 4, our APGC module achieves better performance improvement than other methods on KITTI mix, KITTI 2012 and KITTI 2015 datasets, which verifies the effectiveness of our APGC module.

Method				KITTI mix		KITTI 2015		KITTI 2012	
APGC	PyConv	SPP	ASPP	EPE↓	D1-all↓	EPE↓	D1-all↓	EPE↓	3-px↓
	✓			0.360	0.699	0.457	0.860	0.440	1.001
		✓		0.361	0.703	0.446	0.801	0.427	1.112
			✓	**0.346**	0.672	0.446	0.830	0.438	1.034
✓				0.348	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**

Table 4. Comparison of PyConv [10], SPP [18], ASPP [19] and our APGC on KITTI validation sets

Ablation Study on MSCA Module. Table 5 lists the results of ablation studies on MSCA module. First, we explore the influence of four 3D convolution layers used in both 1/8 and 1/16 cost volumes. It is observed that using 3D convolution layer certainly improves the performance on KITTI mix, KITTI 2012 and KITTI 2015 datasets, which indicates that the feature refinements at both 1/8 and 1/16 cost volumes via 3D convolution layers definitely contribute to the accuracy of stereo matching.

Next, we discuss the impacts of different fusion strategies in 3D hourglass networks (see blue, red and purple lines in Figure 3). As shown in the last line of Table 5, a sufficient fusion of multi-scale cost volumes, namely using all fusion strategies denoted by blue, red and purple lines in Figure 3, significantly improves the accuracy of disparity estimation on KITTI mix, KITTI 2012 and KITTI 2015 datasets.

MSCA Module				KITTI mix		KITTI 2015		KITTI 2012	
3D conv	blue	red	purple	EPE↓	D1-all↓	EPE↓	D1-all↓	EPE↓	3-px↓
	✓	✓	✓	0.381	0.784	0.459	0.873	0.446	1.116
	✓		✓	0.387	0.768	0.457	0.878	0.464	1.208
✓	✓			0.373	0.715	0.461	0.856	0.487	1.215
✓			✓	0.375	0.761	0.461	0.869	0.457	1.250
✓		✓	✓	0.352	0.704	0.451	0.837	0.445	1.131
✓	✓		✓	0.353	0.685	0.440	0.812	0.422	0.982
✓	✓	✓		0.370	0.752	0.460	0.880	0.459	1.209
✓	✓	✓	✓	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**

Table 5. Ablation study of the detailed structure of MSCA module on KITTI validation sets. 3D *conv* denotes using four 3D convolutions in 1/8 and 1/16 cost volumes. *blue*, *red* and *purple* denote the different fusion strategies of multi-scale cost volumes shown in Figure 3.

Ablation Study on DR Module. First, we discuss the effect of convolutional residual blocks on DR module. As shown in the first two columns of Table 6, using six convolutional residual blocks obtains better performance than using three convolutional residual blocks on KITTI mix, KITTI 2012 and KITTI 2015

datasets, which verifies that the feature refinement via more convolutional residual blocks certainly helps improve disparity estimation.

Next, we explore the influence of the dilation rates used in convolutional residual blocks. As can be seen from the third column of Table 6, more accurate disparity estimations are achieved on KITTI mix, KITTI 2012 and KITTI 2015 datasets when using the dilation rates of $(1, 2, 4, 4, 2, 1)$, which indicates that the symmetric dilation rate setting is more suitable for the proposed DR module.

$3 \times$ conv + $3 \times$ residual blocks	$6 \times$ residual blocks	Dilation Rates (1, 1, 2, 2, 4, 4)	Dilation Rates (1, 2, 4, 4, 2, 1)	KITTI mix EPE ↓	KITTI mix D1-all ↓	KITTI 2015 EPE ↓	KITTI 2015 D1-all ↓	KITTI 2012 EPE ↓	KITTI 2012 3-px ↓
✓			✓	0.353	0.685	0.440	0.812	0.422	0.982
	✓	✓		0.359	0.704	0.454	0.829	0.432	1.085
	✓		✓	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**

Table 6. Ablation study of the detailed structure of DR module on KITTI validation sets

Ablation Study on Network Structure of CGCANet. Here we discuss the importance of each component in our CGCANet, including APGC module, MSCA module and DR module. Table 7 gives the results of ablation studies, where the baseline model that does not use both APGC module and DR module and contains only 1/4 cost volume is shown in the first row.

APGC Module	MSCA Module 1/4	MSCA Module 1/8	MSCA Module 1/16	DR Module	KITTI mix EPE ↓	KITTI mix D1-all ↓	KITTI 2015 EPE ↓	KITTI 2015 D1-all ↓	KITTI 2012 EPE ↓	KITTI 2012 3-px ↓
	✓				0.372	0.759	0.480	0.862	0.452	1.120
✓	✓				0.367	0.744	0.456	0.847	0.452	1.076
	✓			✓	0.370	0.740	0.454	0.837	0.445	1.079
✓	✓			✓	0.365	0.734	0.449	0.823	0.442	1.057
✓	✓	✓			0.357	0.730	0.452	0.833	0.440	1.066
	✓	✓	✓		0.355	0.681	0.448	0.857	0.427	1.093
✓	✓	✓	✓		0.354	0.664	0.446	0.817	0.420	0.995
	✓	✓	✓	✓	0.351	0.673	0.443	0.817	0.420	1.007
✓	✓	✓	✓	✓	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**

Table 7. Ablation study of network structure of the proposed CGCANet on KITTI validation datasets

For APGC module, the results demonstrate that using APGC module yields more accurate disparity estimations on KITTI mix, KITTI 2012 and KITTI 2015 datasets. This benefits from that APGC module contributes to a more discriminative feature representation for cost computation by fusing both local context information and multi-scale features generated from CNN backbone.

For MSCA module, the results indicate that introducing 1/8 and 1/16 cost volume information into 1/4 cost volume obviously improves disparity estimations

on KITTI mix, KITTI 2012 and KITTI 2015 datasets, compared to the baseline model. The reason lies in that MSCA module makes a sufficient and effective cost aggregation of multiple cost volumes at different scales to produce more accurate and robust initial disparity.

For DR module, the results show that using DR module obtains more accurate disparity estimations on KITTI mix, KITTI 2012 and KITTI 2015 datasets, which indicates that a sufficient and reasonable refinement for the initial disparity is an indispensable step to improve the performance of stereo matching.

As shown in the last row of Table 7, the best performance is achieved on KITTI mix, KITTI 2012 and KITTI 2015 datasets when using the proposed APGC module, MSCA module and DR module simultaneously, which verifies the effectiveness of the proposed CGCANet for stereo matching.

Ablation Study on Loss Weights. We discuss the influences of the weights λ_i used in the training loss (see Equation (7)) and the results are listed in Table 8. Although the fifth loss in Equation (7) plays a strong role for stereo matching, the best improvement is obtained on KITTI mix, KITTI 2012 and KITTI 2015 datasets when the loss weights $(\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are set to $(0, 1, 1, 1, 1)$.

λ_0	λ_1	λ_2	λ_3	λ_4	**KITTI mix**		**KITTI 2015**		**KITTI 2012**	
					EPE↓	**D1-all**↓	**EPE**↓	**D1-all**↓	**EPE**↓	**3-px**↓
1	0	0	0	0	0.936	3.099	0.977	3.009	1.075	3.351
0	1	0	0	0	1.095	2.397	1.233	2.953	1.119	2.651
0	0	1	0	0	0.638	1.799	0.698	2.067	0.704	2.158
0	0	0	1	0	0.425	0.990	0.512	1.151	0.491	1.309
0	0	0	0	1	0.389	0.766	0.463	0.864	0.452	1.149
0	0	0	1	1	0.378	0.703	0.467	0.907	0.464	1.125
0	0	1	1	1	0.373	0.739	0.459	0.883	0.445	1.044
0	**1**	**1**	**1**	**1**	**0.348**	**0.659**	**0.435**	**0.776**	**0.417**	**0.923**
1	1	1	1	1	0.371	0.751	0.455	0.896	0.436	1.003

Table 8. Ablation study of weights λ_i of the training loss on KITTI validation sets

4.4 Comparison with State-of-the-Art Methods

We compare our CGCANet with the existing state-of-the-art methods on KITTI test datasets, where we use the best trained model of our CGCANet to calculate the disparity estimations submitted to the KITTI serving for evaluation.

Results on KITTI 2015. Table 9 reports the comparison results on KITTI 2015 test dataset. As we can see, our CGCANet outperforms most existing methods in all metrics, which demonstrates the effectiveness not only of the proposed CGCANet but also of the carefully designed APGC module, MSCA module and DR module. Specifically, our CGCANet reaches a 1.70 % D1-all error rate,

which surpasses PSMNet [3], GWCNet [12] and CFNet [7] by 0.62 %, 0.41 % and 0.18 %, respectively. Figure 4 gives some qualitative results of the disparity maps estimated by PSMNet [3], GWCNet [12], CFNet [7] and our CGCANet, where both disparity maps and error maps are downloaded from the KITTI evaluation server. It can be seen that our CGCANet can produce more accurate disparity maps than other methods, even in some challenging image regions including textureless regions as well as object boundaries. Specifically, as shown in the first and second examples of Figure 4, there are high dynamic range and low-light scenarios respectively, which indicates that our CGCANet is robust to deal with these scenarios.

Method	**All pixels (%)**			**Non-occluded pixels (%)**		
	D1-bg↓	**D1-fg↓**	**D1-all↓**	**D1-bg↓**	**D1-fg↓**	**D1-all↓**
MC-CNN [20]	2.89	8.88	3.89	2.48	7.64	3.33
GC-Net [11]	2.21	6.16	2.87	2.02	5.58	2.61
CRL [14]	2.48	3.59	2.67	2.32	3.12	2.45
PSMNet [3]	1.86	4.62	2.32	1.71	4.31	2.14
SegStereo [21]	1.88	4.07	2.25	1.76	3.70	2.08
GWCNet-gc [12]	1.74	3.93	2.11	1.61	3.49	1.92
HD^3 [22]	1.70	3.63	2.02	1.56	3.43	1.87
AANet+ [13]	1.65	3.96	2.03	1.49	3.66	1.85
SCV-Stereo [23]	1.67	3.78	2.02	1.52	3.47	1.84
SGNet [24]	1.63	3.76	1.99	1.46	3.40	1.78
GANet [25]	1.55	3.82	1.93	1.40	3.37	1.73
CasStereo [26]	1.59	4.03	2.00	1.43	3.55	1.78
CAL-Net [27]	1.59	3.76	1.95	1.45	3.42	1.77
Abc-Net [28]	1.47	4.20	1.92	1.33	3.81	1.74
GMStereo [29]	1.49	3.14	1.77	1.34	2.97	1.61
CamLiRAFT [30]	1.48	3.46	1.81	1.34	3.11	1.63
TemporalStereo [31]	1.61	2.78	1.81	1.52	2.58	1.70
CFNet [7]	1.54	3.56	1.88	1.43	3.25	1.73
AMNet [32]	1.53	3.43	1.84	1.39	3.20	1.69
CGCANet (ours)	**1.40**	**3.21**	**1.70**	**1.28**	**3.03**	**1.57**

Table 9. Comparison with existing state-of-the-art methods on KITTI 2015 test set. We report the percentages of the 3-pixel errors in background (D1-bg), foreground (D1-fg) and all pixels (D1-all) in non-occluded and all regions for evaluation.

Results on KITTI 2012. Table 10 reports the comparison results on KITTI 2012 test dataset. As can be seen, our CGCANet outperforms the existing methods in most metrics, which verifies the effectiveness not only of the proposed CGCANet but also of the carefully designed APGC module, MSCA module and DR module. Specifically, our CGCANet achieves a 1.55 % overall 3-pixel error rate, which surpasses PSMNet [3], GWCNet [12] and CFNet [7] by 0.34 %, 0.24 % and 0.03 %, respectively. Figure 5 shows some qualitative results of the disparity maps estimated by PSMNet [3], GWCNet [12], CFNet [7] and our

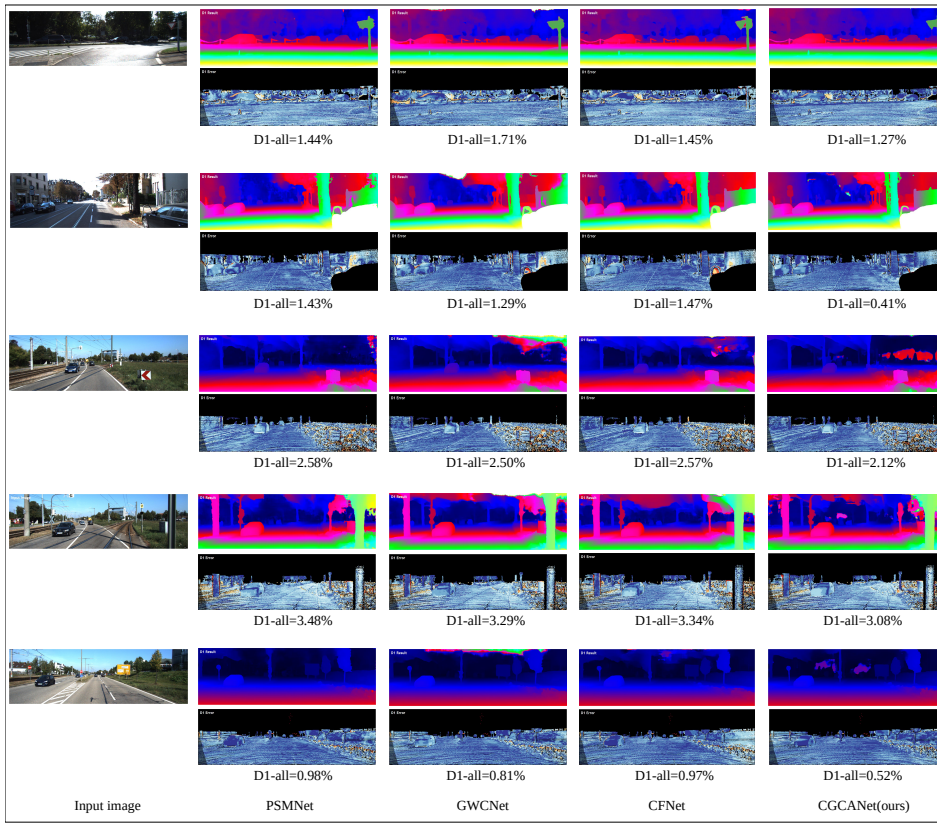

Figure 4. Comparison of visualization results on KITTI 2015 test set. The left panel shows the left input image of the stereo image pair. The right panel gives the predicted disparity maps and error maps obtained by the PSMNet [3], GWCNet [12], CFNet [7] and our CGCANet.

CGCANet, where both disparity maps and error maps are downloaded from the KITTI evaluation server. Obviously, our CGCANet is able to generate more accurate disparity maps than other methods. Notice that our CGCANet is more robust in some challenging scenarios, such as low-light and high dynamic range scenarios.

These results on both KITTI 2015 and KITTI 2012 test sets show an obvious improvement in obtaining more accurate and robust disparity maps, which indicates the superiority of our CGCANet when comparing to the existing stereo matching methods.

Generalization performance. To evaluate the cross-domain generalization of the stereo matching methods, we conduct the corresponding experiments and compare our CGCANet with the existing methods. Table 11 lists the results where

Methods	> 2px (%)		> 3px (%)		> 4px (%)		> 5px (%)		Mean Error (px)	
	Noc↓	All↓	Noc↓	All↓	Noc↓	All↓	Noc↓	All↓	Noc↓	All↓
MC-CNN [20]	3.90	5.45	2.43	3.63	1.90	2.85	1.64	2.39	0.7	0.9
PVStereo [33]	4.55	5.25	1.98	2.47	1.32	1.70	1.01	1.33	0.7	0.9
GC-Net [11]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	0.6	0.7
CoT-Stereo [34]	3.72	4.40	1.82	2.32	1.28	1.68	1.03	1.37	0.7	0.8
DispNetC [4]	7.38	8.11	4.11	4.65	2.77	3.20	2.05	2.39	0.9	1.0
PSMNet [3]	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15	0.5	0.6
SegStereo [21]	2.66	3.19	1.68	2.03	1.25	1.52	1.00	1.21	0.5	0.6
BGNet [35]	2.78	3.35	1.62	2.03	1.16	1.48	0.90	1.16	0.5	0.6
CoEx [36]	2.54	3.05	1.55	1.93	1.15	1.42	0.91	1.13	0.5	0.5
PGNet [37]	2.14	2.81	1.32	1.79	1.00	1.32	0.82	1.11	0.5	0.5
GWCNet-gc [12]	2.16	2.71	1.32	1.79	1.00	1.32	0.80	1.03	0.5	0.5
SGM-Net [38]	3.60	5.15	2.29	3.50	1.83	2.80	1.60	2.36	0.7	0.9
EdegStereo [15]	2.32	2.88	1.46	1.83	1.07	1.34	0.83	1.04	0.4	0.5
SSPCVNet [39]	2.47	3.09	1.47	1.90	1.08	1.41	0.87	1.14	0.5	0.6
AANet+ [13]	2.30	2.96	1.55	2.04	1.20	1.58	0.98	1.30	0.4	0.5
HD^3 [22]	2.00	2.56	1.40	1.80	1.12	1.43	0.94	1.19	0.5	0.5
GANet [25]	1.89	2.50	**1.19**	1.60	0.91	1.23	0.76	1.02	0.4	0.5
CFNet [7]	1.90	2.43	1.23	1.58	0.92	1.18	**0.74**	**0.94**	0.4	0.5
CGCANet (ours)	**1.88**	**2.39**	1.20	**1.55**	**0.91**	**1.18**	0.75	0.97	**0.4**	**0.5**

Table 10. Comparison with existing state-of-the-art methods on KITTI 2012 test set. We report the percentages of the 2-pixel, 3-pixel, 4-pixel, 5-pixel errors in non-occluded (Noc) and all regions (All) for evaluation.

all models are pretrained on the synthetic SceneFlow dataset and then tested on both real-world KITTI 2015 and KITTI 2012 datasets. It can be seen that our CGCANet outperforms several existing methods including PSMNet [3] and GWCNet [12], which confirms the superioty of our CGCANet in cross-domain generalization performance.

5 CONCLUSION

In this paper, we propose a novel end-to-end network, named CGCANet, for the stereo matching task. Inspired by the traditional stereo matching methods, the proposed CGCANet contains three main steps, including cost computation, cost aggregation, and disparity refinement. For cost computation, a new APGC module is presented that extracts local context information in a way of a pyramid and then combines with the features at different scales to obtain a more discriminative feature representation. For cost aggregation, we provide a MSCA module to fuse multi-scale cost volume information into the hourglass networks with the aim of improving the initial disparity estimation. For disparity refinement, we present the new DR module that is able to generate more accurate disparity estimations

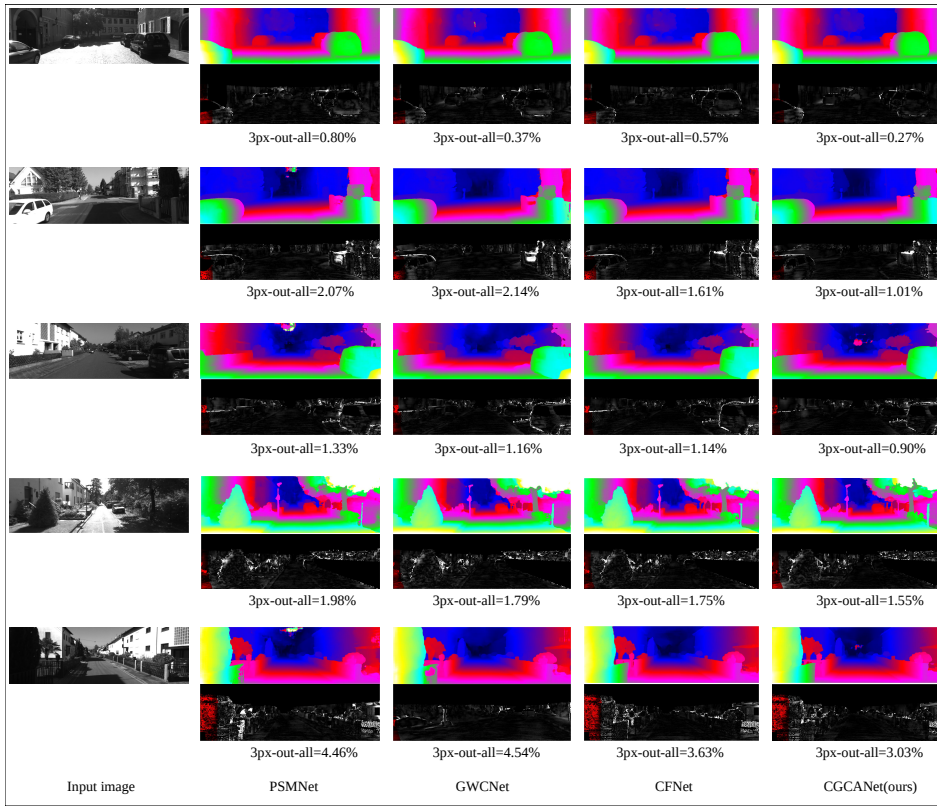

Figure 5. Comparison of visualization results on KITTI 2012 test set. The left panel shows the left input image of the stereo image pair. The right panel gives the predicted disparity maps and error maps obtained by the PSMNet [3], GWCNet [12], CFNet [7] and our CGCANet.

Methods	**KITTI 2015** **D1-all** ↓	**KITTI 2012** **3-px** ↓
HD^3 [22]	26.5	23.6
PSMNet [3]	16.3	15.1
GWCNet [12]	12.2	12.0
CasStereo [26]	11.9	11.8
GANet [25]	11.7	10.1
DSMNet [40]	6.5	6.2
CGCANet (ours)	**6.19**	**5.62**

Table 11. Generalization performance on KITTI 2015 and KITTI 2012 datasets. All methods are first trained on SceneFlow dataset and then directly tested on KITTI training sets.

via both the color guidance of left input image and the feature refinement of several residual blocks. Experimental results on commonly used datasets show the superiority and generalization performance of our CGCANet, which verifies the effectiveness of the proposed APGC module, MSCA module and DR module. In future work, we will propose a lightweight version of our CGCANet where the model parameters will be reduced while the model accuracy will not be significantly reduced, with the goal of applying to more scenarios including real-time and resource-limited scenarios.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under the Grant No. 61962043.

REFERENCES

[1] Scharstein, D.—Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision, Vol. 47, 2002, No. 1, pp. 7–42, doi: 10.1023/A:1014573219977.

[2] Hirschmuller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, 2008, No. 2, pp. 328–341, doi: 10.1109/TPAMI.2007.1166.

[3] Chang, J. R.—Chen, Y. S.: Pyramid Stereo Matching Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418, doi: 10.1109/CVPR.2018.00567.

[4] Mayer, N.—Ilg, E.—Häusser, P.—Fischer, P.—Cremers, D.—Dosovitskiy, A.—Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4040–4048, doi: 10.1109/CVPR.2016.438.

[5] Geiger, A.—Lenz, P.—Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[6] Menze, M.—Geiger, A.: Object Scene Flow for Autonomous Vehicles. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061–3070, doi: 10.1109/CVPR.2015.7298925.

[7] Shen, Z.—Dai, Y.—Rao, Z.: CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13906–13915, doi: 10.1109/CVPR46437.2021.01369.

[8] Shen, Z.—Dai, Y.—Rao, Z.: MSMD-Net: Deep Stereo Matching with Multi-Scale and Multi-Dimension Cost Volume. CoRR, 2020, doi: 10.48550/arXiv.2006.12797.

[9] Cheng, X.—Wang, P.—Yang, R.: Learning Depth with Convolutional Spatial Propagation Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, 2020, No. 10, pp. 2361–2379, doi: 10.1109/TPAMI.2019.2947374.

[10] Duta, I. C.—Liu, L.—Zhu, F.—Shao, L.: Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition. CoRR, 2020, doi: 10.48550/arXiv.2006.11538.

[11] Kendall, A.—Martirosyan, H.—Dasgupta, S.—Henry, P.—Kennedy, R.—Bachrach, A.—Bry, A.: End-to-End Learning of Geometry and Context for Deep Stereo Regression. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 66–75, doi: 10.1109/ICCV.2017.17.

[12] Guo, X.—Yang, K.—Yang, W.—Wang, X.—Li, H.: Group-Wise Correlation Stereo Network. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3268–3277, doi: 10.1109/CVPR.2019.00339.

[13] Xu, H.—Zhang, J.: AANet: Adaptive Aggregation Network for Efficient Stereo Matching. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1956–1965, doi: 10.1109/CVPR42600.2020.00203.

[14] Pang, J.—Sun, W.—Ren, J. S. J.—Yang, C.—Yan, Q.: Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 878–886, doi: 10.1109/ICCVW.2017.108.

[15] Song, X.—Zhao, X.—Fang, L.—Hu, H.—Yu, Y.: EdgeStereo: An Effective Multi-Task Learning Network for Stereo Matching and Edge Detection. International Journal of Computer Vision, Vol. 128, 2020, No. 4, pp. 910–930, doi: 10.1007/s11263-019-01287-w.

[16] Wang, J.—Duan, Z.—Mei, K.—Zhou, H.—Tong, C.: A Light-Weight Stereo Matching Network with Color Guidance Refinement. In: Sun, F., Liu, H., Fang, B. (Eds.): Cognitive Systems and Signal Processing (ICCSIP 2020). Springer, Singapore, Communications in Computer and Information Science, Vol. 1397, 2021, pp. 481–495, doi: 10.1007/978-981-16-2336-3_46.

[17] Yu, F.—Koltun, V.: Multi-Scale Context Aggregation by Dilated Convolutions. CoRR, 2015, doi: 10.48550/arXiv.1511.07122.

[18] He, K.—Zhang, X.—Ren, S.—Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, 2015, No. 9, pp. 1904–1916, doi: 10.1109/TPAMI.2015.2389824.

[19] Chen, L. C.—Papandreou, G.—Schroff, F.—Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR, 2017, doi: 10.48550/arXiv.1706.05587.

[20] Žbontar, J.—LeCun, Y.: Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. Journal of Machine Learning Research, Vol. 17, 2016, Art. No. 65, `http://jmlr.org/papers/v17/15-535.html`.

[21] Yang, G.—Zhao, H.—Shi, J.—Deng, Z.—Jia, J.: SegStereo: Exploiting Semantic Information for Disparity Estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture

Notes in Computer Science, Vol. 11211, 2018, pp. 660–676, doi: 10.1007/978-3-030-01234-2_39.

[22] Yin, Z.—Darrell, T.—Yu, F.: Hierarchical Discrete Distribution Decomposition for Match Density Estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6037–6046, doi: 10.1109/CVPR.2019.00620.

[23] Wang, H.—Fan, R.—Liu, M.: SCV-Stereo: Learning Stereo Matching from a Sparse Cost Volume. 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 3203–3207, doi: 10.1109/ICIP42928.2021.9506214.

[24] Chen, S.—Xiang, Z.—Qiao, C.—Chen, Y.—Bai, T.: SGNet: Semantics Guided Deep Stereo Matching. In: Ishikawa, H., Liu, C. L., Pajdla, T., Shi, J. (Eds.): Computer Vision – ACCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12622, 2021, pp. 106–122, doi: 10.1109/10.1007/978-3-030-69525-5_7.

[25] Zhang, F.—Prisacariu, V.—Yang, R.—Torr, P. H. S.: GA-Net: Guided Aggregation Net for End-to-End Stereo Matching. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 185–194, doi: 10.1109/CVPR.2019.00027.

[26] Gu, X.—Fan, Z.—Zhu, S.—Dai, Z.—Tan, F.—Tan, P.: Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2492–2501, doi: 10.1109/CVPR42600.2020.00257.

[27] Chen, S.—Li, B.—Wang, W.—Zhang, H.—Li, H.—Wang, Z.: Cost Affinity Learning Network for Stereo Matching. ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2120–2124, doi: 10.1109/ICASSP39728.2021.9413556.

[28] Li, X.—Fan, Y.—Lv, G.—Ma, H.: Area-Based Correlation and Non-Local Attention Network for Stereo Matching. The Visual Computer, Vol. 38, 2022, No. 11, pp. 3881–3895, doi: 10.1007/s00371-021-02228-w.

[29] Xu, H.—Zhang, J.—Cai, J.—Rezatofighi, H.—Yu, F.—Tao, D.—Geiger, A.: Unifying Flow, Stereo and Depth Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, 2023, No. 11, pp. 13941–13958, doi: 10.1109/TPAMI.2023.3298645.

[30] Liu, H.—Lu, T.—Xu, Y.—Liu, J.—Wang, L.: Learning Optical Flow and Scene Flow with Bidirectional Camera-LiDAR Fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 46, 2024, No. 4, pp. 2378–2395, doi: 10.1109/TPAMI.2023.3330866.

[31] Zhang, Y.—Poggi, M.—Mattoccia, S.: TemporalStereo: Efficient Spatial-Temporal Stereo Matching Network. 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023, pp. 9528–9535, doi: 10.1109/IROS55552.2023.10341598.

[32] Du, X.—El-Khamy, M.—Lee, J.: AMNet: Deep Atrous Multiscale Stereo Disparity Estimation Networks. CoRR, 2019, doi: 10.48550/arXiv.1904.09099.

[33] Wang, H.—Fan, R.—Cai, P.—Liu, M.: PVStereo: Pyramid Voting Module for End-to-End Self-Supervised Stereo Matching. IEEE Robotics and Automation Letters, Vol. 6, 2021, No. 3, pp. 4353–4360, doi: 10.1109/LRA.2021.3068108.

[34] Wang, H.—Fan, R.—Liu, M.: Co-Teaching: An Ark to Unsupervised Stereo Matching. 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 3328–3332, doi: 10.1109/ICIP42928.2021.9506283.

[35] Xu, B.—Xu, Y.—Yang, X.—Jia, W.—Guo, Y.: Bilateral Grid Learning for Stereo Matching Networks. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12492–12501, doi: 10.1109/CVPR46437.2021.01231.

[36] Bangunharcana, A.—Cho, J. W.—Lee, S.—Kweon, I. S.—Kim, K. S.—Kim, S.: Correlate-and-Excite: Real-Time Stereo Matching via Guided Cost Volume Excitation. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021, pp. 3542–3548, doi: 10.1109/IROS51168.2021.9635909.

[37] Chen, S.—Xiang, Z.—Qiao, C.—Chen, Y.—Bai, T.: PGNet: Panoptic Parsing Guided Deep Stereo Matching. Neurocomputing, Vol. 463, 2021, pp. 609–622, doi: 10.1016/j.neucom.2021.08.041.

[38] Seki, A.—Pollefeys, M.: SGM-Nets: Semi-Global Matching with Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6640–6649, doi: 10.1109/CVPR.2017.703.

[39] Wu, Z.—Wu, X.—Zhang, X.—Wang, S.—Ju, L.: Semantic Stereo Matching with Pyramid Cost Volumes. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7483–7492, doi: 10.1109/ICCV.2019.00758.

[40] Zhang, F.—Qi, X.—Yang, R.—Prisacariu, V.—Wah, B.—Torr, P.: Domain-Invariant Stereo Matching Network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12347, 2020, pp. 420–439, doi: 10.1007/978-3-030-58536-5_25.

Wenmei Sun received her Bachelor's degree from the Shenyang University of Chemical Technology in 2014 and the Master's degree from the Inner Mongolia University in 2023. Her main research interests include deep learning, computer vision and stereo matching.

Yuan Zheng is currently Associate Professor in the College of Computer Science at the Inner Mongolia University, Hohhot, China. She received both the Bachelor's and Master's degrees from the Harbin Institute of Technology (HIT), and her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2006, 2008 and 2014, respectively. Her main research interests include deep learning, computer vision and pattern recognition.