24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Dedge-AGMNet:an effective stereo matching network optimized by depth edge auxiliary task

Weida Yang[1] and **Xindong Ai**[1] and **Zuliu Yang**[1] and **Yong Xu**[2] and **Yong Zhao**[1*]

Abstract. To improve the performance in ill-posed regions, this paper proposes an atrous granular multi-scale network based on depth edge subnetwork(Dedge-AGMNet). According to a general fact, the depth edge is the binary semantic edge of instance-sensitive. This paper innovatively generates the depth edge ground-truth by mining the semantic and instance dataset simultaneously. To incorporate the depth edge cues efficiently, our network employs the hard parameter sharing mechanism for the stereo matching branch and depth edge branch. The network modifies SPP to Dedge-SPP, which fuses the depth edge features to the disparity estimation network. The granular convolution is extracted and extends to 3D architecture. Then we design the AGM module to build a more suitable structure. This module could capture the multi-scale receptive field with fewer parameters. Integrating the ranks of different stereo datasets, our network outperforms other stereo matching networks and advances state-of-the-art performances on the Sceneflow, KITTI 2012 and KITTI 2015 benchmark datasets.

1 INTRODUCTION

Visual perception is a fundamental problem that focuses on the capability to obtain accurate results in a 3D scene. Depth estimation is an important part of perception, which has various essential applications, such as autonomous driving, dense reconstruction, and robot navigation. As a type of passive depth sensing techniques, stereo matching estimates the disparity from rectified image pairs.

The classical pipeline for disparity estimation involves finding corresponding points based on matching cost and post-processing. With the development of deep learning, learning-based methods acquire cues from classical ones, they are embedded in different modules that attempt to obtain a better result. However, because of the discontinuous inference process and the shallow features, the early CNN-based methods capture a terrible performance in ill-posed regions. Nowadays, the end-to-end disparity estimation network is proposed to improve the performance.

Currently, there are two main methods to optimize the networks in ill-posed regions. The first approach captures the additional features and constraints using auxiliary networks, such as semantic segmentation and edge detection subnetworks[5, 16, 19]. However, the semantic segmentation tasks are incapable to distinct the overlap instance with the same label. And the classical edge information contains a large number of noise edges. Those issues induce disparity estimation misjudgments. Secondly, some network utilize a set of stacked 3D convolution modules[2, 5] or parallel structures [1] to capture multi-scale context information. These methods are useful but greatly increase computational consumption and memory resources.

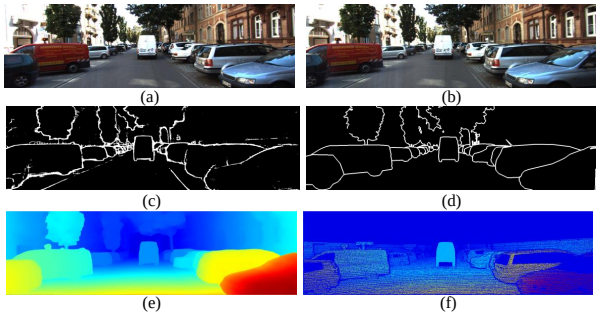

Figure 1: (a)&(b) the left and right images from KITTI 2015; (c)&(e) the predicted depth edge and disparity map from training set; (d)&(f) the ground-truth of the depth edge and the disparity estimation.

In view of the above problems, this paper proposes a multi-task learning network called Dedge-AGMNet that effectively alleviates the drawbacks of both previous methods. We generate depth edge ground-truth and propose the depth edge auxiliary network. Sharing the feature extraction module with the stereo matching main network, the auxiliary branch provides the depth edge constraints. For effective multi-task interactions, we design the Dedge-SPP that embeds the depth edge features to the main branch. Compared with traditional edge detection, the proposed network substantially reducing the noise edges.

The paper proposes a novel module, called the AGM module. Referring to Res2Net[6], we extract the granular convolution from its block and extend to the 3D representation. Retaining the advantages of multi and large scale receptive field, we employ the parallel structure to trade-off the running latency and the scale of the receptive field. The main contributions of this work are summarized as follows:

- We propose the multi-task learning network Dedge-AGMNet that optimizes the feature extraction module with hard sharing parameter, and utilizes the Dedge-SPP to incorporate depth edge cues into disparity estimation pipeline.
- The AGM module is designed to capture the multi-scale information while requiring fewer parameters at a reduced computational cost.
- Our method achieves state-of-the-art accuracy on the Sceneflow

[1] The Key Laboratory of Integrated Microsystems, Shenzhen Graduate School, Peking University, China. email:weida.yang@pku.edu.cn, yongzhao@pkusz.edu.cn, *corresponding author
[2] Harbin Insititute of Technology

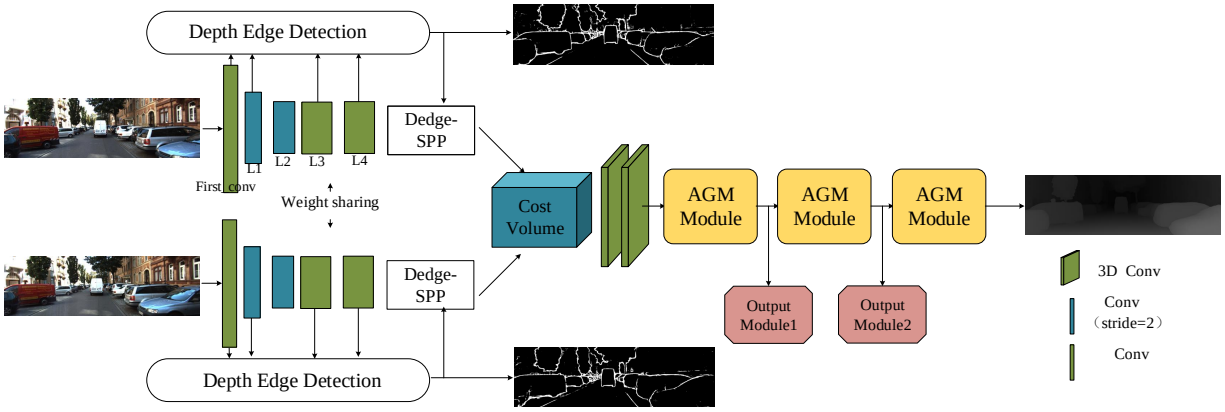

Figure 2: The pipeline of the proposed atrous granular multi-scale network based on depth edge subnetwork(Dedge-AGMNet).

dataset, KITTI 2012 and KITTI 2015 stereo benchmark.

2 RELATED WORK

2.1 Stereo Matching

Depth from stereo has been widely studied for a long time in the literature. The traditional stereo matching methods[14] have been proposed for four steps: matching cost computation[22], cost aggregation[12], optimization[15], and disparity refinement. Recently, convolutional neural networks have become popular in solving this problem. Zbontar and LeCun [22] were the first to use CNN for matching cost computation. Luo et al.[12] designed a novel Siamese network to treat the computation of matching cost as a multi-label classification, which computes the inner product between the left and the right feature maps. Seki et al.[15] raised the SGM-Net that predicts SGM penalties for regularization.

Inspired by other pixel-wise labeling tasks, the end-to-end neural networks have been proposed using the fully-convolution network[11] for disparity estimation. Mayer et al.[13] designed the first end-to-end disparity estimation network, DispNet, which utilizes the encoder-decoder structure with short-cut connections for second stage processing. Kendall et al.[10] raised GCNet, a cost volume formed by concatenating the feature maps to incorporate contextual information. This network applies the 3D encoder-decoder architecture to regularize the cost volume. To find correspondences in ill-posed regions, Chang and Chen[2] proposed the PSMNet to regularize cost volume using stacked multiple hourglass networks in conjunction with intermediate supervision.

Currently, Chabra[1] proposed a depth refinement architecture that helps the fusion system to produce geometrically consistent reconstructions, and utilized 3D dilated convolutions to construct hourglass architecture. Meanwhile, XianZhi Du et al.[5] designed a similar construction with three atrous multi-scale modules, while it is useful to aggregate rich multi-scale contextual information from cost volume. Base on the PSMNet [2], both of them achieved state of the art on the different stereo datasets.

2.2 Multi-scale Features

The multi-scale feature is an important factor in the pixel predicted tasks, such as semantic segmentation and disparity estimation. Because ambiguous pixels require a diverse range of contextual information, ASPP[3] was designed to concatenate various feature maps with multi-scale receptive fields. To further quest the importance of the receptive field, Yang et al.[20] proposed Dense ASPP to concatenate a set of different atrous convolutional layers densely. The approach encourages feature reusing by constructing a similar structure with the DenseNet[9]. Instead of representing the multi-scale features in a layer-wise manner, Gao et al.[6] designed novel architecture, called Res2Net. The network uses hierarchical residual-like connections in a single block to represent it at a granular level. Controlling the same computational resources as ResNet block, Res2Net achieved more accurate results.

2.3 Multi-task Learning network

Focus on improving the accuracy of the ill-posed regions where the single stereo matching networks are difficult to predict. Yang et al.[19] proposed SegStereo to embed the semantic features, and regularized semantic cues as the loss term. Xianzhi Du et al.[5] utilized foreground-background segmentation map to improve disparity estimation. This paper believed that better awareness of foreground objects would lead to a more accurate estimation. Song et al.[17, 16] proposed EdgeStereo which composes a disparity estimation subnetwork and an edge detection subnetwork. By combining the advantages of the semantic segmentation and edge detection, we propose the depth edge detection auxiliary network.

3 Dedge-AGMNet

The proposed Dedge-AGMNet is composed of a depth edge detection branch and a disparity estimation branch. The depth edge subnetwork provides geometric knowledge and constraints without adding irrelevant edges. We also utilize the granular convolution to design a more efficient 3D aggregate filtering module. It is worth noting that our network only estimates the disparity map but not predicts depth edge in the inference process, which decreases the parameters significantly.

3.1 Network Architecture

The structure of the proposed Dedge-AGMNet is shown in Fig.2. The network consists of five parts, feature extraction, depth edge prediction and embedding, cost volume construction, 3D aggregation, and disparity prediction.

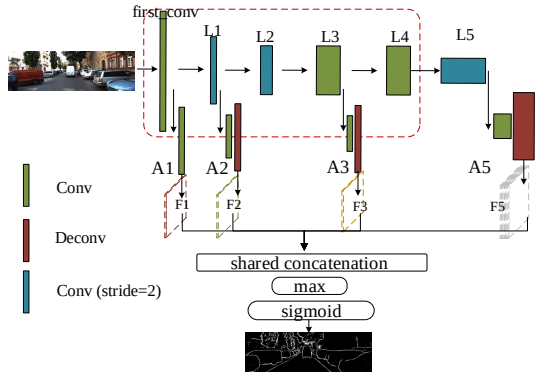

Figure 3: The architecture of depth edge detection branch. The red dashed box denotes the sharing feature extraction module.

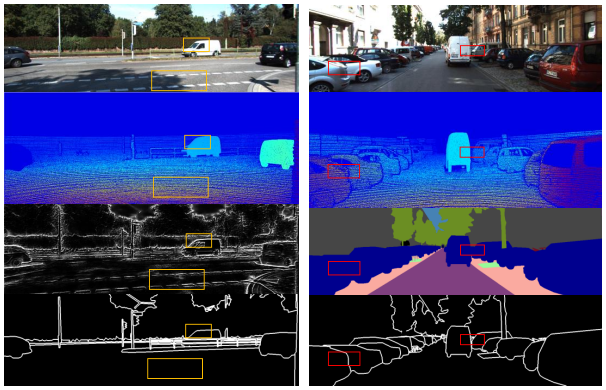

Figure 4: The first row denotes left images, the second row presents the disparity ground-truth, the third row shows the classical edge and the semantic segmentation map, and the last row displays the depth edge maps. The yellow box presents the noise edge performance in the smooth region. The red box compares the performance in the disparity change region.

For the feature extraction from both subnetworks, we retain the ResNet-like structure used in PSMNet[2] except that the first down-sampling operation occurs at L1 but not first_conv.

We present the depth edge subnetwork with corresponding loss function to provide geometrical constraints for the shared features. In addition, instead of SPP[2], the Dedge-SPP module is constructed to fuses the geometric knowledge from depth edge subnetwork. The details are described in Section 3.2.

The cost volume consists of two parts: a concatenation volume and a distance volume, which is explained in Section 3.4. We process the cost volume with a pre-hourglass module and three stacked AGM modules. And details are described in Section 3.3.

In the disparity estimation subnetwork, the three stacked AGM modules are connected to output modules to predict disparity maps. The details of the output modules and the loss functions are described in Section 3.4.

3.2 Depth edge auxiliary task

3.2.1 Validity analysis & Generation of dataset

Without additional knowledge or constraints, it is difficult to find correct correspondences in ill-posed regions. The classical edge subnetwork[17, 16] is beneficial, but it captures considerable edge noises, such as object pattern and inner edges. This non-semantic information heavily interferes with the disparity estimation. Semantic segmentation subnetworks[19, 5] are commonly used. However, semantic boundaries always lack the edge for overlap instances that have the same label, it induces disparity estimation misjudgment. As shown in Fig.4, depth edge combines the advantages of classical edge and semantic map, it segments different individuals accurately without edge noise.

In the autonomous driving scene, a single foreground object always could be considered the same depth, this paper utilizes the binary instance bounds to represent the depth edges for the foreground object. We employ binary semantic boundaries to compensate for the lacking background edges. In summary, we generate the depth edge map by mining the instance&semantic ground-truth in stereo datasets.

3.2.2 Structure of subnetwork

As shown in Fig.3, except to share parameters in the feature extraction module, the auxiliary network adds L5(a similar structure with L4[2]) to capture more features. Different from the classical edge detection network[18], our network does not predict the depth edge based on bottom side features. But those features are useful to provide detailed edge information, we utilize the shared concatenation[21] to fuse multi-frequency features, and only predict the edge at the last stage.

The bottom features $F = \{F_1, F_2, F_3\}$ are output from the feature re-extraction module$(A1, A2, A3)$, and the top features with K channels are represented by F_5. The shared concatenation is as follows:

$$\{F_5(1), F, F_5(2), F, ..., F_5(K), F\}$$

the depth edge branch adopts a similar architecture as CASENet but contains several key modifications.

- The different task. Fist, compared with the semantic edge, the depth edge ground-truth consists of the boundaries from the instance and semantic map. It contains more useful edge information. Besides, we simplify the task from multi-label to binary label, which decreases the task complexity substantially. And drive the auxiliary branch to pay more attention to the edge details but not the classication.
- Fewer parameters. Since the limitation of the parameters and computational cost, in contract to CASENet[21], we adopt about 1/8 channels in the feature extraction module. However, we believe that the simplied task could utilize fewer channels to capture the required features.
- Similar to CASENet, Our network handles more channels(K) for top features. But instead of building the relationship between each channel and corresponding label, we believe that more channels mean greater importance. Compared to the different probability from corresponding channels, the subnetwork selects the highest probability as the predicted probability for depth edge.

Besides, since the classical edge lacks semantic information, EdgeStereo[16] only shares parameters in shallow layers to capture low-frequency features. In constrast, The depth edge contains the semantic and instance information, our network still shares parameters in the high layers. The network could employ semantic information to suppress the interference of non-depth noise edges. We will illustrate it in Section 4.3.

3.2.3 *The incorporation of the networks*

This paper designs the depth edge loss function to optimize the sharing feature extraction. Because the depth edge label is a binary representation, we use the binary cross-entropy loss instead of multi-label[21]. It denoted as $L_{edge}P(X_i;W)$ and Y_i denote the predicted probability and ground-truth for the image pixel X_i.

$$L_{edge}\left(X_i,W\right)=\begin{cases}\alpha\times\log(1-P(X_i;W)), & if\ Y_i=0\\ \beta\times\log P(X_i;W), & if\ Y_i=1\end{cases} \tag{1}$$

in which

$$\begin{aligned}\alpha&=\frac{|Y^+|}{|Y^+|+|Y^-|}\\ \beta&=\frac{|Y^-|}{|Y^+|+|Y^-|}\end{aligned} \tag{2}$$

where $|Y^+|$ and $|Y^-|$ represent the number of positive samples and negative samples, respectively.

Since the disparity discontinuity point is always on the depth boundaries. the depth edge gradient is more consistent with the change of the disparity map, and $L_{dedge-disp}$ is presented as follows:

$$L_{dedge-disp}=\frac{1}{N}\sum_{i,j}|\partial_x d_{i,j}|\,e^{-\gamma|\partial_x\xi_{i,j}|}+|\partial_y d_{i,j}|\,e^{-\gamma|\partial_y\xi_{i,j}|} \tag{3}$$

where N denotes the number of pixels, γ is the loss intensity, ∂d and $\partial\xi$ present the disparity and the depth edge map gradient, respectively.

Whats more, we concatenate the depth edge features with the output of L4 to modify the SPP[2]. Dedge-SPP is designed to share the geometric knowledge with the disparity estimation branch.

3.3 Atrous granular multi-scale module

3.3.1 *Structure of AGM module*

We propose AGM-module. As shown in Fig.5, the AGM module combines the advantages of the hourglass and the parallel structure. The hourglass structure could reduce the feature size reasonably, we utilize the short-cut connection to transmit shallow features. Besides, the parallel structure with the dilated granular convolution boosts the performance significantly. Compared to the standard convolution, granular convolution captures multi-scale context information that requires fewer parameters. Meanwhile, the parallel structure balances the running latency and the scale of receptive field.

3.3.2 *Granular convolution*

The blue dashed box of Fig.5 illustrates the details of granular convolution. It shows that the number of receptive fields in granular convolution is approximate G times than the standard convolution. The granular convolution divides the input features into several groups, the output features of the previous group are input to the next group of filters along with another group of input feature maps. The features map $v=(v_1,v_2,...,v_G),v_i\in R^{W\times H\times c/G(groupnumber)}$. $\sum$ denotes the concatenation operator and $<,>$ denotes standard convolution. We formulate granular convolution as follows:

$$\begin{aligned}v'&=w_{pw}\sum_{g=1}^{G}\hat{v}'_g\\ &=w_{pw}\sum_{g=1}^{G}\sum_{i=1}^{g}<w_1...<w_i v_{(g-i)}>>\end{aligned} \tag{4}$$

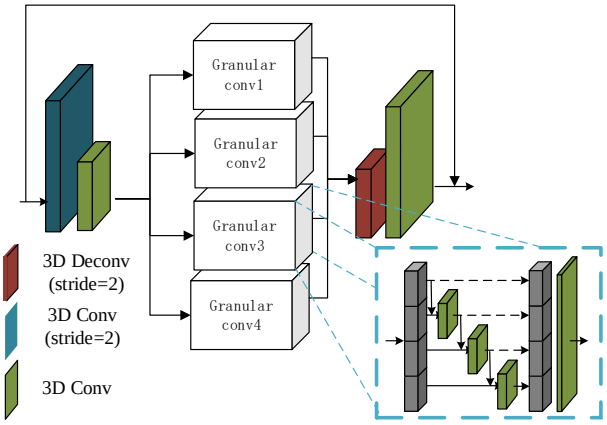

Figure 5: The architecture of AGM module. The blue dashed box illustrates the granular convolution.

where the weight $w=(w_1,w_2,...,w_G)$, $w_i\in R^{\frac{c}{G}\times\frac{c}{G}\times s\times s}$. And the w_{pw} denotes the weight of point-wise convolution weight.

Keeping the channel and size of the feature map, the parameters of standard and granular convolution is shown below:
Standard convolution:

$$N_{standard}=C_{out}\times C_{in}\times s\times s=C^2\times s^2$$

Granular convolution:

$$\begin{aligned}N_{granular}&=C_{in}\times C_{out}\times\frac{s}{C}\times\frac{s}{C}\times(G-1)+C_{out}\times C_{out}\\ &=C_{in}\times C_{out}\times s\times s\times(\frac{G-1}{G\times G}+\frac{1}{s\times s})\\ &\approx\frac{1}{G}\times N_{standard}\end{aligned} \tag{5}$$

3.3.3 *Running latency*

Granular convolution utilizes the internal cascade structure to capture the multi and large receptive field. However, As one layer is splitted into two or more sequential layers, the latency increases progressively. Therefore, this structure increases running latency inevitably. K and G denote the number of sequential layers and groups, respectively. The running latency of the cascade and parallel structures are shown as follow:

$$RL_{parallel}=G-1=1/K\times R_{cascade}$$

In summary, contrasted with standard convolution, the computational cost of granular convolution is about $1/G$ times. The hyper-parameter $G=K=4$, AGM module utilizes the parallel structure to trade-off the running latency and the scale of the receptive field.

3.4 Cost volume

We designed the cost volume by stacking the concatenation module and the distance module. The former provides the overall information of the features, which is formed by concatenating left feature maps with their corresponding right feature maps[10].and the latter calculates the difference between the two at disparity level to provide feature similarity information[1].

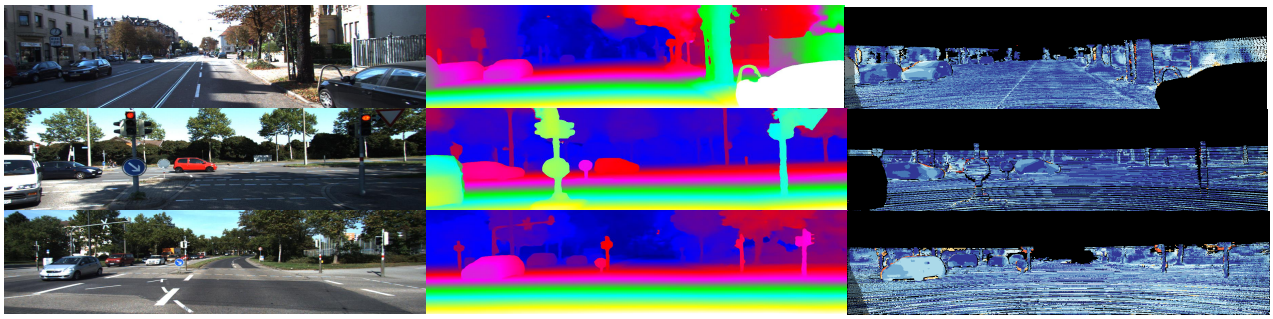

Figure 6: Results on the KITTI 2015 test sets. From left: left stereo image, disparity map, error map.

3.5 Output module and loss function

The output module contains two stacked 3D convolution layers and the upsampling operator. The volume c_d from the output module is converted into a probability volume with a softmax function $\sigma(.)$ along the disparity dimension. The predicted disparity $\hat{d}$ calculated as follows:

$$\hat{d} = \sum_{d=0}^{D_{max}} d \times \sigma(-c_d) \tag{6}$$

The predicted disparity maps from the three output modules are denoted as $\hat{d_1}$, $\hat{d_2}$, $\hat{d_3}$, and L_{disp} is as follows:

$$L_{disp} = \sum_{i=1}^{3} \lambda_i \times Smooth_{L_1}\left(\hat{d_i} - d^*\right) \tag{7}$$

where λ denotes the coefficients and d^* represents the ground-truth disparity map.Therefore, by combining L_{disp}, L_{edge} and the related loss $L_{edgedisp}$, we have

$$L_{total} = L_{disp} + a \times L_{edge} + (1-a) \times L_{edge_disp} \tag{8}$$

4 EXPERIMENT

In this section, we train the proposed model on the Sceneflow, Cityscapes and KITTI datasets, but evaluate it only on the Sceneflow and KITTI datasets. The disparities of the Cityscapes dataset are obtained by SGM algorithm[8] but not the ground truth. The paper presents datasets and network implementation in Section 4.1 and Section 4.2. And illustrates the effectiveness of each module in Section 4.3 and Section 4.4. Then, the evaluation results on the different datasets are presented.

4.1 Datasets and evaluation metric

4.1.1 Stereo dataset

Sceneflow is a large scale synthetic dataset containing three subsets(Flyingthings3D, Driving and Monkaa), which provides approximately 35000 training and 4000 testing stereo image pairs of size 960×540. It consists of left and right images, complete ground-truth disparity maps and segmentation images. This paper adopts end-point-error (EPE) as the evaluation metric.

Cityscapes is an urban scene-understanding dataset. This dataset provides 3475 rectified stereo pairs, fined annotated segmentation maps and corresponding disparity maps precomputed by SGM.

KITTI 2012 and KITTI 2015 are both driving scene datasets. KITTI 2012 provides 194 training and 195 testing image pairs, while KITTI 2015 contains 200 training and 200 testing image pairs. With a size of 1240×376, both datasets provide sparse disparity maps. Twenty image pairs have remained as the validation set. The main evaluation metric for KITTI 2015 is D1-all error, which computes the percentage of pixels for which the estimation error is $\geq 3px$ or $\geq 5\%$ from the ground-truth disparity. The main evaluation criterion for KITTI 2012 is Out-Noc, which computes the percentage of pixels for which the estimation error is $\geq 3px$ for all non-occluded pixels.

4.1.2 Depth edge dataset

Sceneflow & Cityscapes & KITTI 2015 According to the method proposed in Section 3.2.3, this paper generates the ground-truth of depth edges for their corresponding dataset, respectively.

4.2 Network implementation

The Dedge-AGMNet architecture is implemented with PyTorch. All the models are trained using the Adam optimizer($\beta_1 = 0.09$ $\beta_2 = 0.999$). We use 4 Nvidia TITAN XP GPUs when training the models, and the batchsize is fixed to 8. The images are randomly Cropped to 512×256. The coefficients of disparity outputs are set as follows:$\lambda_1 = 0.5, \lambda_2 = 0.7, \lambda_3 = 1.0$. In $L_{edgedisp}$, $\gamma = 0.5$.

The training process of our network contains two steps. For the first step, we pre-train Dedge-AGMNet only on the Sceneflow dataset. The initial learning rate is set to 0.001, then down-scaled by 2 every 2 epochs from epoch 10 to 16. The maximum disparity (D_{max}) is set to 192. Besides, we fine-tune the pre-trained model with stepped learning rates of 0.001 for 300 epochs on KITTI 2012/2015. Furthermore, we extend the training to 70 epochs on Sceneflow to get the final results.

For the second step, we combine Sceneflow and Cityscapes as the pre-trained dataset. And employ the same training strategy to obtain the compared result. Finally, our network is fine-tuned with learning rates of 0.001 for 600 epochs and 0.0001 for another 400 epochs to capture the final results.

4.3 Effectiveness of depth edge network

As shown in the graph of Fig.7, with more and deeper shared layers in the feature extraction, the EPE decreases significantly on Sceneflow. To prove the effectiveness and generalization of depth edge auxiliary task, this paper embeds the depth edge subnetwork into

MODEL	MODULE				RESULT		
	Hourglass	Cost volume	Hard parameter sharing	SPP	Parameters	Sceneflow(EPE)	KITTI 2015(D1-all)
PSM	[2]	[2]	-	[2]	5.27M	0.884	1.67
AGMNet	✓	[2]	-	[2]	3.85M	0.801	1.62
AGMNet	✓	✓	-	[2]	3.88M	0.754	1.56
Dedge-AGMNet	✓	✓	✓	[2]	3.88M	0.648	1.57
+Cityscapes	✓	✓	✓	[2]	3.88M	-	1.45
Dedge-AGMNet	✓	✓	✓	✓	3.98M	0.645	1.54
+Cityscapes	✓	✓	✓	✓	3.98M	-	1.38

Table 1: Ablation study on the Sceneflow test set and the KITTI 2015 validation set. The symbol '✓' denotes the module we proposed. '+Cityscapes' denotes that the network pre-trains on the hybrid dataset, which contains Sceneflow and Cityscapes dataset.

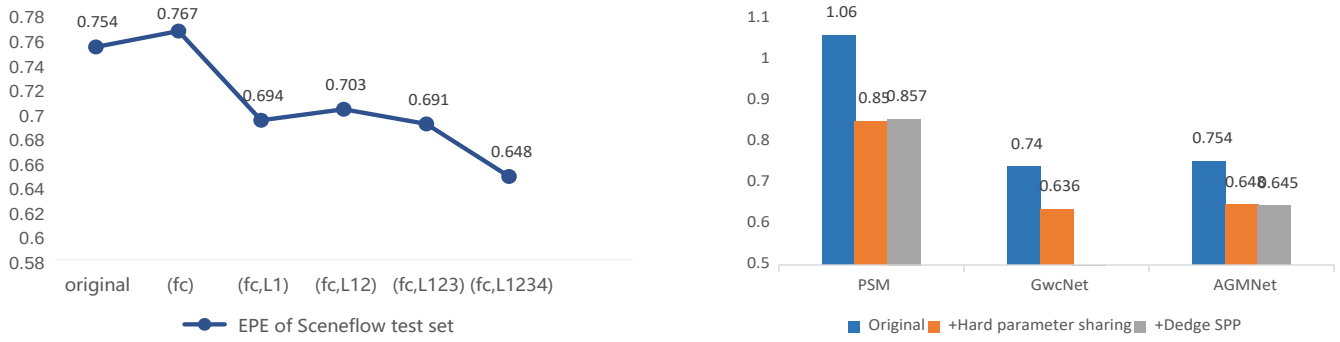

Figure 7: **The graph**: the relationship between the performance and the shared layers. 'fc' denotes first_conv. **The histogram**: Embedding the Depth edge auxiliary subnetwork into PSMNet[2], GwcNet[7] and our network.The blue columns are the original results, the orange columns show the results that sharing parameters in feature extraction module. And gray columns mean that adding the hard parameter sharing mechanism and Dedge-SPP together.

a	0	0.2	0.5	0.8	1
D1-all	**1.38**	1.42	1.48	1.45	1.47
EPE	**0.62**	0.64	0.65	0.65	0.65

Table 2: Control experiment for the weight a of loss function. We computed the D1-all and the EPE on the KITTI 2015 validation set.

PSMNet[2],GwcNet [7] and our AGMNet. As shown in the histogram, utilizing the hard parameter sharing mechanism, the depth edge subnetwork could optimize the feature extraction module. The EPE of Sceneflow is reduced about $15\% \sim 20\%$. It is worth emphasizing that this module does not add any parameters and computational cost any more. On Sceneflow dataset, the Dedge-SPP module does not improve the accuracy of the disparity map remarkably.

As shown in Table 1, Dedge-SPP plays a significant role to improve the accuracy on KITTI 2015. Without estimating the depth edge in the inference process, Dedge-SPP only increases few parameters(0.1M) to obtain the depth edge features. Besides, while adding Cityscapes dataset, the 3-px error is reduced clearly on KITTI validation set. This dataset guides the auxiliary subnetwork to learn more accurate features of depth edge on city scenes.

In the pre-training process, we fixed the hyper-parameter $a = 0.5$ and select the optimal setting for the weight a on KITTI 2015. As shown in Table 2, when $a = 0$, the validation set could obtain the best performance. It demonstrates that, for the fine-tuned network, smoothen the disparity map play a more important role than learning the depth edge feature on KITTI 2015.

Model	Hourglass	Dilation rate				Sceneflow(EPE)
PSMNet	[2]	-				0.889
AGMNet	✓	1	4	8	-	0.836
AGMNet	✓	1	2	3	4	0.823
AGMNet	✓	1	2	4	8	0.821
AGMNet	✓	1	4	8	16	**0.801**
AGMNet	✓	1	4	16	32	0.842

Table 3: The AGMNet network is dened as the version that only replaces the hourglass structure. All the models are trained with the same learning strategy.

4.4 Best setting of AGM module

The experimental results in Table 3 show that when the dilation rate is set to an appropriate range, the parallel structure with four granular convolutions can achieve better results than three. We conclude that the AGM module with dilation rates of 1, 4, 8, and 16 provides optimal performance. All the AGM-base networks outperform PSMNet. Under the best AGM module settings, the EPE is reduced by 9.3% on Sceneflow dataset.

4.5 Result

As shown in Table 1, the result shows that the module we proposed has a certain effect to promote the network. Compared to PSMNet[2], our proposed network has a 27.0% reduction on Sceneflow and 17.4% on the KITTI 2015 validation dataset.

mthod	$> 2px(\%)$		$> 3px(\%)$		$> 4px(\%)$		$> 5px(\%)$		Mean Error	
	Noc	All	Noc	All	Noc	All	Noc	All	Noc	All
GC-Net[10]	2.71	3.46	1.77	2.30	1.36	1.77	1.12	1.46	0.6	0.7
SegStereo[19]	2.66	3.19	1.68	2.03	1.25	1.52	1.00	1.21	0.5	0.6
EdgeStereo[17]	-	-	1.73	2.18	1.30	1.64	1.04	1.32	-	-
PSMNet[2]	2.44	3.01	1.49	1.89	1.12	1.42	0.90	1.15	0.5	0.6
EdgeStereo-v2[16]	2.32	2.88	1.46	1.83	1.07	1.34	0.83	1.04	**0.4**	**0.5**
Dedge-AGMNet	**2.02**	**2.56**	**1.26**	**1.64**	**0.95**	**1.24**	**0.77**	**1.01**	**0.4**	**0.5**

Table 4: Comparison with the top publiced methods on the KITTI stereo 2012 test set.

Method	All(%)			Non-Occluded(%)			Runtime
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	(s)
GC-Net[10]	2.21	6.16	2.87	2.02	5.58	2.61	0.9s
PSMNet[2]	1.86	4.62	2.32	1.71	4.31	2.14	0.41s
SegStereo[19]	1.88	4.07	2.25	1.76	3.70	2.08	0.6s
EdgeStereo[17]	1.87	3.61	2.16	1.72	3.41	3.00	0.7s
EdgeStereo-v2[16]	1.84	3.30	2.08	1.69	2.94	1.89	**0.32s**
AGMNet	1.66	4.30	2.10	1.53	3.89	1.92	0.84s
Dedge-AGMNet	**1.54**	**3.37**	**1.85**	**1.41**	**2.98**	**1.67**	0.9s

Table 5: Comparison with the top publiced methods on the KITTI stereo 2015 test set.

Mod.	EPE	Mod.	EPE	Mod.	EPE
GC-Net	2.51	SegStereo	1.45	PSMNet	1.09
CSPN	0.78	AMNet32	0.74	ours	**0.520**

Table 6: Comparison with the top publiced methods on the Sceneflow test set.

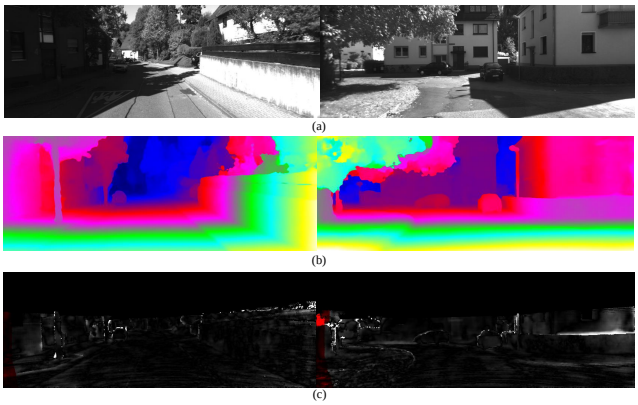

Figure 8: Results on the KITTI 2012 test sets. (a) denotes left stereo image, (b) denotes disparity map and (c) presents the error map.

Sceneflow: We compared the performance of PSMNet with other state-of-the-art methods, such CSPN[4], AMNet32[5]. As shown in Table 6, Dedge-AGMNet ranks first compared to other published papers, which shows the effectiveness of the depth edge auxiliary task thoroughly.

KITTI 2012 and 2015: Our approach achieves state-of-the-art performances on KITTI 2012 and KITTI 2015 benchmark datasets.Utilizing the best hyper-parameter setting selected in the experiment, we train our model for 1000 epochs on KITTI 2015. Then estimate the disparity maps for the 200 testing images. According to the online leaderboard, as shown in Table 5, the D1-all for the Dedge-AGMNet is 1.85%, which ranks in the **fourth** place. Similarly, we calculate the disparity for the KITTI 2012 test set. As shown in Table 4, the result ranks **fourth**, too.

Fig.6 and Fig.8 give qualitative results on the KITTI 2012 and 2015 test sets, which demonstrates that our network produces high-quality results in ill-regions.

5 CONCLUSION

In this paper, we propose the Dedge-AGMNet, a stereo matching network optimized by depth edge. This paper expounds on the superiority of the auxiliary depth edge task and generates the depth edge ground-truth innovatively. Dedge-AGMNet contains two main modules: Depth edge subnetwork and AGM module. We utilize the hard parameter sharing mechanism to joint optimize the feature extraction module. And design Dedge-SPP to fuse the depth edge features. The proposed AGM module provides multi-scale context information while consuming fewer computational resources. The ablation study demonstrates the effectiveness of the above modules. In our experiment, Dedge-AGMNet achieves state-of-the-art performances and outperforms other multi-task learning models. The proposed network ranks in the first place on Sceneflow and fourth place on both KITTI 2012 and 2015. In the future, we plan to apply the depth edge auxiliary task on a real-time stereo matching network.

ACKNOWLEDGEMENTS

This work is supported by Science and Technology Planning Project of Shenzhen(JCYJ20180503182133411).

REFERENCES

[1] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs, 'Stereodrnet: Dilated residual stereonet', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11786–11795, (2019).

[2] Jia-Ren Chang and Yong-Sheng Chen, 'Pyramid stereo matching network', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, (2018).

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, 'Rethinking atrous convolution for semantic image segmentation', *arXiv preprint arXiv:1706.05587*, (2017).

[4] Xinjing Cheng, Peng Wang, and Ruigang Yang, 'Learning depth with convolutional spatial propagation network', *arXiv preprint arXiv:1810.02695*, (2018).

[5] Xianzhi Du, Mostafa El-Khamy, and Jungwon Lee, 'Amnet: Deep atrous multiscale stereo disparity estimation networks', *arXiv preprint arXiv:1904.09099*, (2019).

[6] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr, 'Res2net: A new multi-scale backbone architecture', *arXiv preprint arXiv:1904.01169*, (2019).

[7] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, 'Group-wise correlation stereo network', 3273–3282, (2019).

[8] Heiko Hirschmuller, 'Accurate and efficient stereo processing by semi-global matching and mutual information', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 807–814. IEEE, (2005).

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, 'Densely connected convolutional networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, (2017).

[10] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry, 'End-to-end learning of geometry and context for deep stereo regression', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75, (2017).

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, 'Fully convolutional networks for semantic segmentation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, (2015).

[12] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun, 'Efficient deep learning for stereo matching', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5695–5703, (2016).

[13] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, 'A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4040–4048, (2016).

[14] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger, 'A multi-view stereo benchmark with high-resolution images and multi-camera videos', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3260–3269, (2017).

[15] Akihito Seki and Marc Pollefeys, 'Sgm-nets: Semi-global matching with neural networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231–240, (2017).

[16] Xiao Song, Xu Zhao, Liangji Fang, and Hanwen Hu, 'Edgestereo: An effective multi-task learning network for stereo matching and edge detection', *arXiv preprint arXiv:1903.01700*, (2019).

[17] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang, 'Edgestereo: A context integrated residual pyramid network for stereo matching', in *Asian Conference on Computer Vision*, pp. 20–35. Springer, (2018).

[18] Saining Xie and Zhuowen Tu, 'Holistically-nested edge detection', in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, (2015).

[19] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia, 'Segstereo: Exploiting semantic information for disparity estimation', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 636–651, (2018).

[20] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang, 'Denseaspp for semantic segmentation in street scenes', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, (2018).

[21] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam, 'Casenet: Deep category-aware semantic edge detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5964–5973, (2017).

[22] Jure Zbontar, Yann LeCun, et al., 'Stereo matching by training a convolutional neural network to compare image patches.', *Journal of Machine Learning Research*, **17**(1-32), 2, (2016).