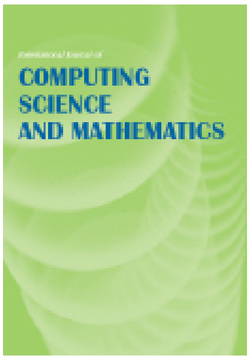

International Journal of
COMPUTING
SCIENCE
AND MATHEMATICS

International Journal of Computing Science and Mathematics

ISSN online: 1752-5063 - ISSN print: 1752-5055
https://www.inderscience.com/ijcsm

IRPSM-net: Information retention pyramid stereo matching network

Yun Zhao, Jiahui Tang, Xing Xu, Xiang Zhou

DOI: 10.1504/IJCSM.2023.10055624

Article History:

Received:	02 November 2020
Last revised:	28 March 2021
Accepted:	23 July 2021
Published online:	20 April 2023

IRPSM-net: Information retention pyramid stereo matching network

Yun Zhao and Jiahui Tang

School of Information and Electronic Engineering,
Zhejiang University of Science and Technology,
Hangzhou, 310023, China
Email: zy_super0201@163.com
Email: 18751072631@163.com

Xing Xu*

School of Mechanical and Energy Engineering,
Zhejiang University of Science and Technology,
Hangzhou, 310023, China
Email: xuxing3220@163.com
*Corresponding author

Xiang Zhou

School of Information and Electronic Engineering,
Zhejiang University of Science and Technology,
Hangzhou, 310023, China
Email: zhouxiang9591@gmail.com

Abstract: In order to prevent the lack of information in the stereo matching process and improve the disparity map accuracy. The information retention pyramid stereo matching network (IRPSM-Net) was proposed a novel architecture that can relieve the limitation of accuracy and retention the original information of the image. The proposed network consisted an information retention pyramid module (IRPM) without batch normalisation to retain the image information. And the training process was optimised by group normalisation, which further improves the effect of stereo matching. The ablation experiments show that our method can effectively improve the accuracy of 0.17% in the threshold 3 pixels of KITTI2012 stereo dataset and 0.09% in the whole region of KITTI2015 stereo dataset. It showed that the improvement of IRPSM-Net can effectively improve the quality of the generated disparity map.

Keywords: stereo matching; multi-scale; information retention pyramid; group normalisation.

Reference to this paper should be made as follows: Zhao, Y., Tang, J., Xu, X. and Zhou, X. (2023) 'IRPSM-net: Information retention pyramid stereo matching network', *Int. J. Computing Science and Mathematics*, Vol. 17, No. 1, pp.50–66.

Biographical notes: Yun Zhao is a Professor of School of Information and Electronic Engineering, Zhejiang University of Science and Technology. She received her PhD from Zhejiang University.

Jiahui Tang received his ME in Information and Electronic Engineering from Zhejiang University of Science and Technology, China, in 2021. His research interests include internet of vehicles, intelligent transportation, intelligent driving.

Xing Xu is a Professor of School of Mechanical and Energy Engineering, Zhejiang University of Science and Technology. He received his PhD from Zhejiang Sci-Tech University.

Xiang Zhou received his ME in Information and Electronic Engineering from Zhejiang University of Science and Technology, China, in 2021. His research interests include internet of vehicles, intelligent transportation, intelligent driving.

1 Introduction

Binocular stereo matching has been continuously developed over the past few decades. It has been extensively used in different 3D imaging fields, such as Ling et al. (2019) and Zeng et al. (2019) in robotics, Chen et al. (2020) in autonomous vehicles and Lai et al. (2019) medical fields. Higher stereo matching accuracy is required in the field of automatic driving. The purpose of stereo matching was to use two cameras to capture the same scene in different perspectives and find the corresponding pixels from another image. The three-dimensional information of scene could be established according to the disparity of the two pictures. A traditional pipeline of stereo matching generally contains four steps: matching cost computation, cost aggregation, disparity estimation and disparity refinement. During the related researches, the traditional stereo matching method got the low accuracy in the areas of textureless, discontinuities, occlusions and illumination differences.

With the development of deep learning convolution networks and artificial intelligence in recent years, the traditional stereo matching methods have transferred into deep learning stereo matching methods. Many researches have bent their efforts for deep learning to improve stereo matching. Deep stereo matching method in early stage only combined part of the steps with deep learning, and most of them were used in the step of feature extraction to get a cost volume. Siamese neural network was introduced by Zbontar and LeCun (2015) to calculate matching costs (MC-CNN). Traditional semi-global block matching (SGM) proposed by Hirschmuller (2007) played a great influence on contemporary stereo matching methods. It has the characteristics of fast speed and a very excellent effect among the traditional algorithms. However, under the background of neural networks, the advantages of SGM algorithm could not be well maintained. Efficient stereo matching network (ESMNet) was proposed by Guo et al. (2019) to fasten stereo matching by adding super-resolution perception of low resolution and discontinuous depth. An unsupervised network was proposed by Zou et al. (2018) to obtain more robust disparity evaluation results, which could perform disparity estimation of depth and flow in a video sequence. A method that could generate stereo images from a single image was proposed by Lo et al. (2020), considering the translational rotation of the object and modifying the appearance flow network, which had achieved good results in the field of monocular imaging. Displet proposed by Guney and Geiger (2015) overcame the ambiguity problem in stereo matching by using semantic segmentation and object recognition. Another network DispNet was proposed by Mayer et al. (2016) together with a large synthetic dataset Scene Flow, which solves the shortcomings to obtain the large dataset from the real-world. In the current research stage, various methods were proposed to solve the problem of perfectly extracting context information to generate the cost volume of cost aggregation. Deepprunner was proposed by Duggal et al. (2019) used a differentiable patchmatch module, this module was proposed by (Barnes et al., 2009) could select valid disparity when generating cost volume without global disparity evaluation to promote the speed of generating disparity. Group-wise correlation stereo network (GWC-Net) was proposed by Guo et al. (2019) extracted the features by splitting channel dimension into multiple groups, then the cost proposals obtained by packing the correlation maps between each group to get the cost volume.

Since Kendall et al. (2017) GC-Net directly incorporated cost aggregation into the training channel through 3D convolution, people had more novel ideas for the improvement of 3D convolution. In the loss strategy of the cost aggregation process, such as smooth L1 loss function proposed by Girshick (2015) was very popular in object detection field, most stereo matching neural networks such as Nguyen and Jeon (2019) and Nie et al. (2019) used a smooth L1 loss function to evaluate the gap between the disparity and the ground truth disparity. Sang et al. (2019) proposed a novel loss strategy in Multi-scale context attention network (MCANet) to perform hard disparity point mining online, which improved the accuracy of disparity generating. Chang and Chen (2018) proposed PSMNet by using a method of spatial pyramid pooling (SPP) which proposed by He et al. (2015). And the whole network using the SPP for feature extraction and stacked hourglass network (Newell et al., 2016) for cost aggregation. In the spatial pyramid pooling network, the use of small-size feature maps can ensure that the contour of the final generated disparity map. However, the normalisation operation may destroy the feature information of the original network in the process of feature extraction and influent the accurate disparity map obtained by stereo matching.

To solve such a problem, we referred to some ideas in the field of super resolution (Wang et al., 2018) and proposed the novel information retention pyramid stereo matching network (IRPSM-Net) in the research. The network used the proposed information retention pyramid module (IRPM) to extract features from the binocular images, so that the acquired feature map can retain more original information of the image and obtained a four-dimensional cost volume. In addition, a three-dimensional stacked hourglass network optimised by group normalisation (Wu and He, 2018) was used to process the cost volume and obtained the probability of disparity by softmax. Finally, the predicted disparity map was obtained through regression processing. The results showed that IRPSM-Net was competitive against among the advanced methods on Scene Flow, KITTI2012 (Geiger et al., 2012) and KITTI2015 (Menze and Geiger, 2015) datasets. The main contributions are listed below:

- *Firstly*. The residual in residual dense block (RRDB) without normalisation was introduced in the proposed IRPSM-Net to retain the information during the feature extraction process.
- *Secondly*. An additional largest average pooling feature map was introduced to make the IRPSM-Net more sensitive about the large object in the scenes and prevent the small feature map affecting the completeness of disparity.
- *Thirdly*. The group normalisation was used to instead of the batch normalisation to further optimise the three-dimension of the stacked hourglass to catch a more accuracy result.

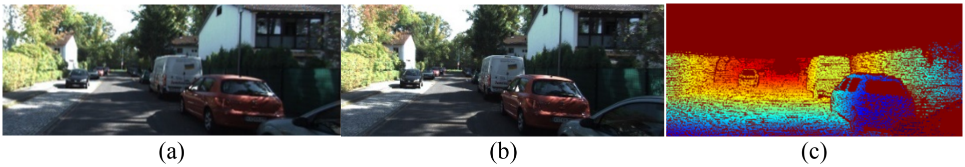
2 Materials

2.1 Datasets

In this research, Scene Flow and KITTI was used to train IRPSM-Net. The large and synthetic dataset Scene Flow always used for pre-training before fine-tuning on real-world dataset. It could effectively solve the problems of overfitting and lack of generalisation caused by training on small real-world dataset. The Scene Flow datasets used the open source 3D creation software Blender to process complex moving 3D objects. Virtual imaging sensors were used for 3D video capture and render the acquisition results into tens of thousands of frames. The size of the imaging sensor is Height = 32.0 mm and Width = 18.0 mm. Most scenes use a virtual focal length of 35.0 mm and parameters of experimental images size is Height = 540 pixels and Width = 960 pixels. The dataset provides clean-pass versions and final-pass versions. The clean-pass version contains lighting and shading effects scenes. Moreover, final-pass version contains the clean-pass scene and the motion blur and defocus blur scene. There are 40,024 pairs of disparity images in the Scene Flow dataset. It was divided into three parts: FlyingTing3D dataset consisting of 26,760 pairs of suspended stereo objects images, Monkaa dataset is consisting of 8864 pairs of static stereo cartoon monkey images and Driving dataset consisting of 4400 pairs of synthetic stereoscopic traffic scene images. Since the coordinates of objects in the virtual scene are determined, all pixels in the disparity map will regression during the training.

The KITTI dataset used a pair of two grayscale imaging sensors PointGray Flea2 (10 Hz, resolution: 1392 × 512 pixels, opening angle: 90°× 35°) installed on the car to capture stereo video of road scenes. Each video obtains the ground truth of optical flow and disparity from the 3D laser sensors Velodyne HDL (10 Hz, 64 laser beams, range: 100 m). KITTI datasets was presented in 2012 and expanded in 2015. Half of the KITTI dataset provided the ground truth disparity obtained by the laser equipment to researchers for training. And another half dataset without the ground truth disparity used for testing the different methods of stereo matching on KITTI online evaluation. KITTI2012 consists 194 pairs of training images and 195 pairs of testing images. After semi-dense ground truth calibration, the resolution is 1240 × 376 pixels. KITTI2015 consists of 200 pairs of training images and 200 pairs of testing images, and the resolution is also 1240 × 376 pixels. Compared with other datasets, the KITTI dataset is real and non-synthetic and contains various scenes in rural, urban, and highways with occlusion, discontinuous, specular area, repeated textures, and textureless. It is extremely challenging for the current various stereo matching methods dataset. The dataset was divided into multiple different disparity maps for comprehensive evaluation of various stereo matching methods. In KITTI2012 'All' means that all pixels were considered in error estimation in the scene, while 'Noc' means that only considered the pixels in the non-occluded regions. The endpoint error (EPE) proposed by Gidaris and Komodakis (2017) of threshold more than two, three and five pixels were measured for evaluation. In KITTI2015, the 'D1-bg", "D1-fg", and "D1-all' were the percentage of outliers averaged over the background regions, foreground regions, and all ground truth pixels respectively. The example of the KITTI datasets were shown in Figure 1.

Figure 1 Examples of KITTI stereo dataset. (a) the left image of the scene. (b) the right image of the scene and (c) the disparity image of the scene (see online version for colours)



(a) (b) (c)

2.2 *Equipments*

IRPSM-Net was trained by using two processor of Intel Xeon Silver 4210 10 Core @ 2.20G Hz, the 32GB memory, four NVIDIA GTX 2080Ti graphic processing unit with total 44GB video memory, Ubuntu18.04LST operating system, and the software such as anaconda, python3.6, cuda10.0, pytorch1.3.0, and gcc5.3. The network was testing on the colab platform provided by google, the platform processor is Intel (R) Xeon (R) CPU @ 2.30G Hz, 12. 72GB memory. Tesla P100 graphic processing unit with 16GB video memory. Ubuntu 18.04 LTS operating system. The software system uses python3.6, cuda10.0, pytorch1.4.0, and gcc5.3. When training, all models were optimised using Adam Kingma and Ba (2014) ($\beta 1 = 0.9$, $\beta 2 = 0.999$), model trained for 10 epochs in Scene Flow and used 300 epochs for pre-training, and 50 epochs for fine-tuning in KITTI. Considering video memory of the device, the batchsize was set to 8 which could ensure the loss could be smoothly converged with sufficient memory. The images in the KITTI dataset were very scarce, so it was necessary to pre-training in Scene Flow dataset to obtain richer disparity features, and then performed the KITTI dataset training.

3 Methods

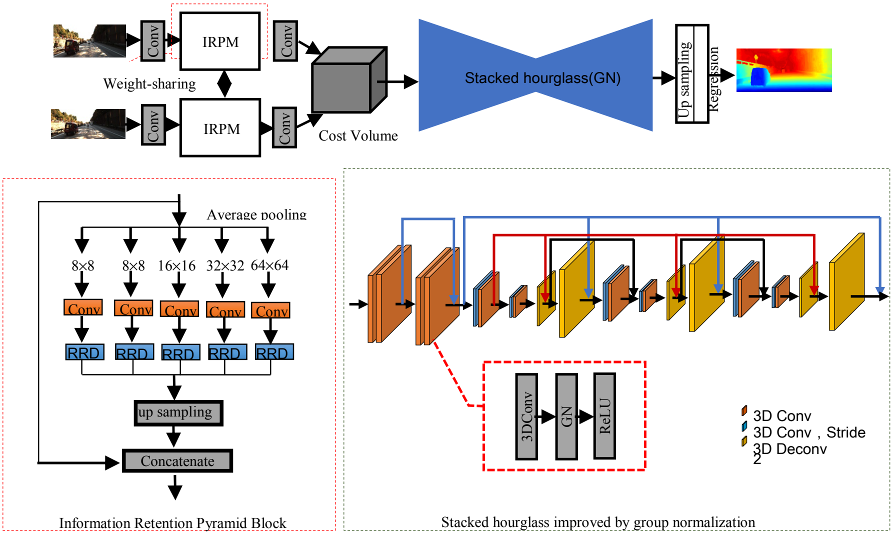
3.1 *Information retention pyramid stereo matching network*

3.1.1 *Network structure*

In this research, a novel IRPSM-Net structure was proposed which based on the two-stage improvement of the pyramid stereo matching. In the step of feature extraction. It used the novel stereo matching architecture with the IRPM to obtain better four-dimensional cost volume. And in cost aggregation step, group normalisation was used to optimise the stacked hourglass network to solve the problem that the small batchsize always failed to obtain the optimal model.

The network structure was shown in Figure 2. The input of the network was two strictly rectified images. And through two weight-sharing networks for feature extraction. The weight-sharing network connected multiple convolutional networks and dilated convolutional networks. Several residual modules were also introduced to suppress the instability of gradient exploding and gradient vanishing caused by deepening the network. In addition, the IRPM was added to increase the storage of the original image information. It contained an additional kernel of 8×8 average pooling layer in order to increase the sensitivity of the large-scale features. The residual in residual dense block (Wang et al., 2018) was beneficial to retain the part of epipolar constraint of the original image pair and improve the effect of feature extraction. For the step of cost aggregation network, the 4-dimensional cost volume was used group normalisation to improve the original stacked hourglass network (Chang and Chen, 2018). Batch normalisation was replaced by group normalisation after 3-dimensional convolutional layers. The hourglass network used two down sampling and two deconvolution structures to learn the 3-dimensional features of the cost volume. In the process of three stacked hourglass, the residual skip-connection structure was applied between the same size four-dimensional feature to increase network robustness and reduce the risk of network degradation. Batch normalisation reduced the feature distribution between each image by normalising multiple samples. It can allow the training process use larger learning rate to speed up the convergence of the network without causing gradient explosions. In tasks which take low requirement of calculation, batch normalisation allowed large samples in each regression. It can get high-accuracy results in the field of detection or identification. However, the stereo matching task of the three-dimensional stacked hourglass network structure with a large computational burden often fails to obtain the expected results. Therefore, group normalisation was used to improve the defects of the batch normalisation in stacked hourglass network. It adjusted the number of samples for normalisation by grouping the cost volume in the channel, and the accuracy of the network can get rid of the limitation of batchsize.

Figure 2 Architecture overview of proposed information retention pyramid stereo matching network (IRPSM-net) (see online version for colours)



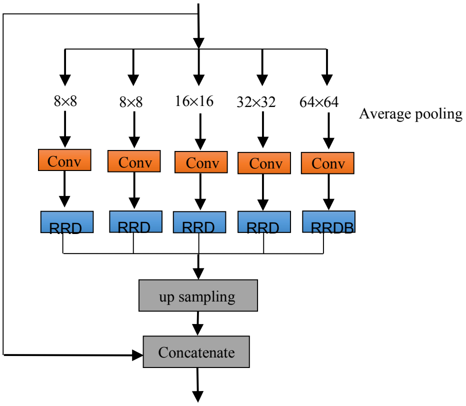
Information Retention Pyramid Block

Stacked hourglass improved by group normalization

3.1.2 Information retention pyramid block for feature extraction

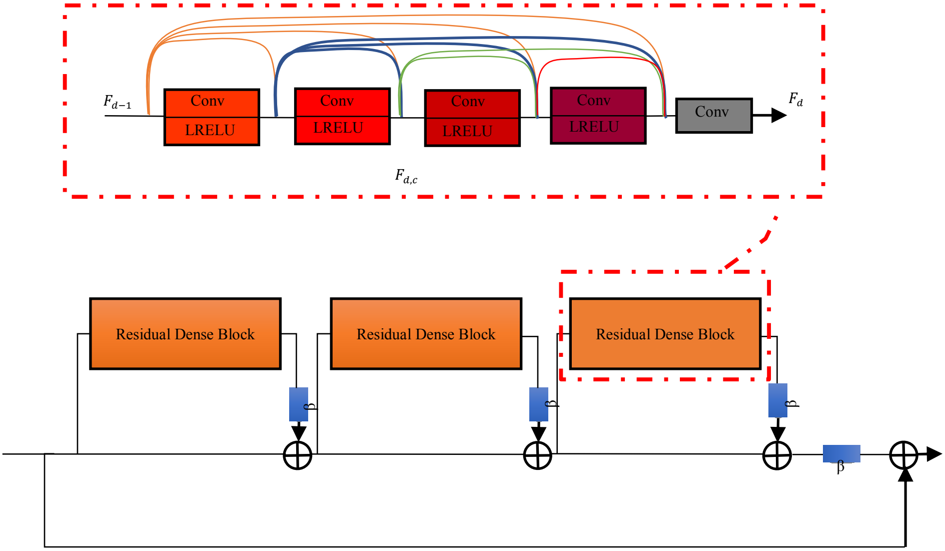
In many fields of deep learning, different scales of feature maps were considered to establish connections between objects and their subcategories. It can make tasks such as detection or recognition more accurate. In the field of stereo matching, multi-scale methods can solve the problem of mismatched points due to occlusion or insufficient texture by considering context information. Since the difference of the foreground and background regions in traffic scene, considering single-size features will lead to the shortcomings of detail loss or overall incompleteness in obtained disparity map. The different kernel sizes were used in spatial pyramid pooling layers to obtain features. These features will up sample to the same size and fuse to ensure the final feature more accuracy. An IRPM based on the classic spatial pyramid pooling structure was proposed. The residual in residual dense module without the batch normalisation was introduced to improve the network. The removal of batch normalisation can preserve the differences between image regions. It retained characteristics of epipolar constraints of the binocular view and improved the effect of cost aggregation. However, removing batch normalisation may cause the risk of gradient disappearance or gradient explosion. The residual in residual dense structure (Wang et al., 2018) made the network more stable to solve the problem. During the experiment, it was found that the small-scale feature map after up sampling contains low information of original images, which was also not conducive to the preservation of epipolar constraints. An additional large-scale average pooling feature map was used to strengthen the preservation of information. The proposed IRPM was shown in Figure 3. A size of 128 × 64 × 128 cost volume was input after multi-layer convolution and residual block. And then the kernel of 64 × 64, 32 × 32, 16 × 16, 8 × 8 average pooling layers were performed. For five feature maps get from the forward steps, the convolution of 3 × 3 kernels change the 128 channels to 32 channels. The residual in residual dense block was applied before up sampling to make the network more stable. Finally, the up sampled to same size five feature maps were concatenated with 128 and 64 channels feature before the average pooling. And the convolution of 3 × 3 kernels was used to get the final 32 × 64 × 128 feature map. The feature map retains the information of the four different scales and original image size. Two feature maps obtained from the left and right views was compared to form cost volume and performed cost aggregation to generate a disparity map.

Figure 3 Information retention pyramid module (see online version for colours)



For four scales feature maps, batch normalisation was removed to preserve the differences between regions. However, the network without batch normalisation is prone to gradient explosion and gradient disappearance during the training process. Therefore, a residual in residual dense block was adopted to ensure the stability of the network. RRDB module connected three dense residual blocks with a skip-layer structure. It consisted of four dense connected convolution-LReLU modules to consider the information from shallow to deep. The details of this module were shown in Figure 4, where β is the parameter used to control residual scaling, F_{d-1} represents the input of the *d*th residual block, $F_{d,c}$ is the output feature map at convolution c and F_d represents the output of the *d*th residual block.

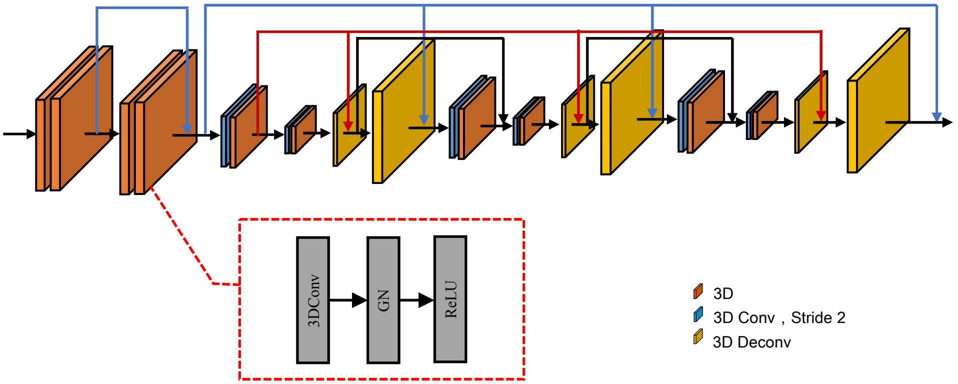
Figure 4 Residual in residual dense module (see online version for colours)



3.1.3 *The group normalisation for stacked hourglass network*

The two-dimensional stacked hourglass network contained multiple encoding-decoding structures, and the skip-layer between same feature map strengthens the connection with the shallow and deep to improve the performance of the network. But for the three-dimensional convolutional stacked hourglass network, the skip-layer connection will cause a huge computational burden. Therefore, stereo matching cannot deal with multiple samples in the regression at this stage. Compared with other tasks in areas where the batch size can be set to 64 or 128, most stereo matching networks just set the low batch size to 8 or 16. It can reduce the risk of insufficient video memory and improve the stability of the network during training. It was very unfavourable for the convergence of the stereo matching network using batch normalisation and obtaining the best model. In order to solve value of batchsize influenced stereo matching network convergence to obtain the accurate disparity map. Different from batch normalisation normalised in batch dimension, group normalisation divided the input data into multiple group channels and normalised the data in each group dimension. The stacked hourglass with group normalisation solved the limitation of small number of batches. It ensured high accuracy disparity when the network was complex or the research platform was insufficiently memory. The stacked hourglass network structure was improved by group normalisation as shown in Figure 5.

Figure 5 Stacked hourglass network structure based on group normalisation (see online version for colours)



3.2 *Loss function*

In order to evaluate the gap between prediction disparity map and the ground truth value during the training process and perform regression calculations. A smooth L1 loss function was used to train IRPSM-net. Smooth L1 was widely used in the field of object detection. It combines the advantages of L1 loss function and L2 loss function. It has a fast convergence and insensitive to the speed outliers. The loss function was defined as equations (1) and (2) shown.

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

$$Smooth_{L_1(x)}=\begin{cases}0.5x^2, & |x|1\\ |x|-0.5, & |x|1\end{cases} \tag{2}$$

where N is the number of pixel values that need to be evaluated, $\boldsymbol{d_i}$ is the ground truth disparity, and $\hat{\boldsymbol{d}}_i$ is the disparity estimated by the network.

3.3 *Performance estimation*

The algorithm was applied on the KITTI dataset for endpoint-error (EPE) of disparity to get the average error pixels. As shown in equation (3), the error was obtained by calculating the average Euclidean distance between the estimated disparity and the ground truth. It was evaluated between multiple pixel ranges and got the disparity error rate of EPE greater than *t* pixels (> tpx) as a percentage. If the target pixel disparity EPE was less than *t* pixels, the pixel was considered to be correct.

$$EPE = \frac{1}{N}\sum_{i=i}^{N} \| d_i - \hat{d}_i \| \quad (3)$$

The endpoint error was calculated by comparing the distance between the estimated disparity map $\boldsymbol{d}_i$ and the ground truth disparity map $\hat{\boldsymbol{d}}_i$. For the KITTI2012 dataset, the task was calculated as the percentage of the non-occluded (Noc) and all (All) pixels with EPE greater than 2px, 3px and 5px. For the KITTI2015 dataset, the percentages of background (bg), foreground (fg), and all region(all) disparity outliers D1 were reported. The outliers were defined as pixels with a disparity error greater than 3 pixels or 5% ground truth.

4 Results

4.1 Various ablation experiments

For each part of improvement in the IRPSM-net, we will evaluate the results to verify the feasibility of improvement. At the beginning of pyramid stereo matching network was used as a reference. Two groups of comparative experiments were designed in terms of the average pooling size. One group was added the 8 × 8 kernel average pooling to get the largest feature map, and the other was added the 4 × 4 pooling to get a larger scale feature map. It can verify the scale of feature map that we selected. After verifying what kind of pooling layer can be added to achieve the highest accuracy, the improved information retention module was added to the stereo matching feature extraction process. It was used to prove the hypothesis that our proposed module can increase the stereo matching performance for image information retention. Finally, the batch normalisation of all two-dimensional and three-dimensional convolutional layers was improved by group normalisation. In order to solve the shortcoming of batch normalisation with small batchsize cannot obtain the optimal model. In the evaluation of KITTI2012 and KITTI2015, our improvements have further improved the accuracy of the disparity map. Comparing three experiments with different scale pooling layers: the original four scale SPP, 4+1 scale SPP added the 8 × 8 kernel pooling layer, 5 scale SPP added the 4 × 4 kernel pooling layer. It also compared the experiments of adding residual in residual dense modules and group normalisation layers. As shown in Tables 1 and 2, for retaining more information from the adding largest scale pooling features map. The three-pixel accuracy of KITTI2012 evaluation had increased by 0.08%, and the overall accuracy of KITTI2015 evaluation also increased by 0.04%. However, However, in adding the larger size feature map ablation experiment, the accuracy of three-pixel KITTI2012 and overall 2015 evaluation all reduced by 0.02% compared with an additional layer of the largest size feature map. Because the larger feature map was already very close to the size of the original image by using the 4 × 4 kernel of the average pooling layer. As a result, the training was more sensitive to large-scale features, causing the model less effective in processing details, and always appearing objects blurred in the disparity map. After the additional average pooling layer was confirmed to improve the accuracy, we added the information retention module to reach more accuracy results. The information retention module used a residual in residual dense block without normalisation operations. It retained the original image features as much as possible before the upsampling step to reduce the spatial information loss and improve the characterisation ability of the image. Further improve the accuracy of KITTI2015 evaluation by 0.03% and the accuracy of the three-pixel KITTI2012 evaluation by 0.07%. Finally, group normalisation was used to further optimise the network, so that the network would not be affected by the value of batchsize when training. After that, the overall accuracy of KITTI2015 evaluation result was increased by 0.02% again, and the accuracy of three-pixel KITTI2012 evaluation was increased by 0.02%.

Table 1 Different improvement strategies are evaluated in KITTI2015

	Improved					*Evaluation results/%*	
Network	*4 scale SPP*	*4+1 scale SPP*	*5 scale SPP*	*RRDB*	*GN*	*All region*	*Noc region*
PSMNet	✓					2.33	2.14
IRPSM-Net		✓				2.29	2.09
			✓			2.31	2.12
		✓		✓		2.26	2.06
		✓		✓	✓	2.24	2.04

Table 2 Different improvement strategies are evaluated in KITTI2012

	Improved					*Threshold division/%*		
Network	*4 scale SPP*	*4+1 scale SPP*	*5 scale SPP*	*RRDB*	*GN*	*2px*	*3px*	*5px*
PSMNet	✓					3.14	1.96	1.19
IRPSM-Net		✓				3.01	1.88	1.12
			✓			3.04	1.90	1.15
		✓		✓		2.88	1.81	1.11
		✓		✓	✓	2.86	1.79	1.07

4.2 Data comparison of multiple methods

In addition to IRPSM-Net ablation experiment comparison. The results of IRPSM-Net were also compared with the results of advanced networks, such as MC-CNN-art, ESMNet, GCNet, DispNetC, PSMNet and Deeppruner. The comparison shows that the accuracy of the information-retaining stereo matching network has great advantages over other networks. From Table 3, we can see some important indicators in KITTI2012. Such as evaluation of two, three and five pixel thresholds, IRPSM-Net had an error rate of 2.86%, 1.79% and 1.07% in the evaluation of all pixels, which was the smallest among the comparison network. Compared with the MC-CNN-art and DispNetC networks in the early exploration stage of the end-to-end stereo matching network, IRPSM-Net got higher accuracy by 2.59%, 1.84%, 1.32% and 5.25%, 2.86%, 1.32% in three thresholds evaluation. For the first end-to-end stereo matching network GCNet, the accuracy of the three thresholds evaluation of IRPSM-Net was also higher by 0.60%, 0.51%, and 0.39%. And for some advanced stereo matching networks proposed in recent years, such as ESMNet, PSMNet, and Deeppruner, the accuracy of IRPSM-Net in all regions of the three pixels was higher by 0.74%, 0.17% and 0.07%. For mean error of each pixel among the entire image, the IRPSM-Net, PSMNet and Deeppruner had an evaluation result of the least 0.5 pixels in the non-occluded and all regions.

Table 3 Multiple methods evaluation results on KITTI2012

Network	*>2px (%)*		*>3px (%)*		*>5px (%)*		*Mean error(px)*	
	Noc	*All*	*Noc*	*All*	*All*	*Noc*	*All*	*Noc*
MC-CNN-art	3.90	5.45	2.43	3.63	1.64	2.39	0.7	0.9
ESMNet	3.65	4.30	2.08	2.53	1.11	1.41	0.6	0.7
GC-Net	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7
DispNetC	7.38	8.11	4.11	4.65	2.05	2.39	0.9	0.7
PSMNet	2.58	3.14	1.52	1.96	0.93	1.20	0.5	0.5
Deeppruner	2.41	2.94	1.49	1.86	1.11	1.38	0.5	0.5
IRPSM-Net	2.29	2.86	1.38	1.79	0.82	1.07	0.5	0.5

Table 4 Multiple methods evaluation results on KITTI2015

Network	*All (%)*			*Noc (%)*		
	D1-bg	*D1-fg*	*D1-all*	*D1-bg*	*D1-fg*	*D1-all*
MC-CNN-art	2.89	8.88	3.89	2.48	7.64	3.33
ESMNet	2.57	4.86	2.95	2.41	4.30	2.72
GC-Net	2.21	6.16	2.87	2.02	5.58	2.61
DispNetC	4.32	4.41	4.34	4.11	3.72	4.05
PSMNet	1.98	4.40	2.34	1.77	4.10	2.15
Deeppruner	2.10	3.68	2.36	1.95	3.33	2.18
IRPSM-Net	1.72	4.86	2.24	1.57	4.37	2.04

It can be seen from Table 4 of KITTI2015 evaluation. Compared with MC-CNN-art and DispNetC, the accuracy of the overall evaluation in the non-occluded and all regions was higher by 1.65%, 1.29% and 2.10%, 2.01% in IRPSM-Net, which has great advantages. Compared with the first stereo matching network GCNet, the accuracy of the IRPSM-Net was 0.63% and 0.57% higher in the overall evaluation of non-occluded and all regions. The experiment found that although the IRPSM-Net in all region evaluation accuracy was 0.71%, 0.10%, and 0.12% higher than ESMNet, PSMNet and Deeppruner which were proposed in recent years, and 0.68 %, 0.11% and 0.14% higher in non-occluded regions.

However, IRPSM-Net only surpasses MC-CNN-art and GCNet in the foreground. Because the addition of a large-scale spatial pooling layer, which makes the objects more able to consider the context in the background. It means the IRPSM-Net ignored the details of the large target. Therefore, the error rate of IRPSM-Net has the highest accuracy in the background evaluation of KITTI2015, even 0.26% higher than the best PSMNet. And for effective comparison with other methods, the results of PSMNet and Deepprunner in Tables 3 and 4 were implemented on the same platform as IRPSM-NET. The remaining results were the best experimental data uploaded by the original author on KITTI official website.

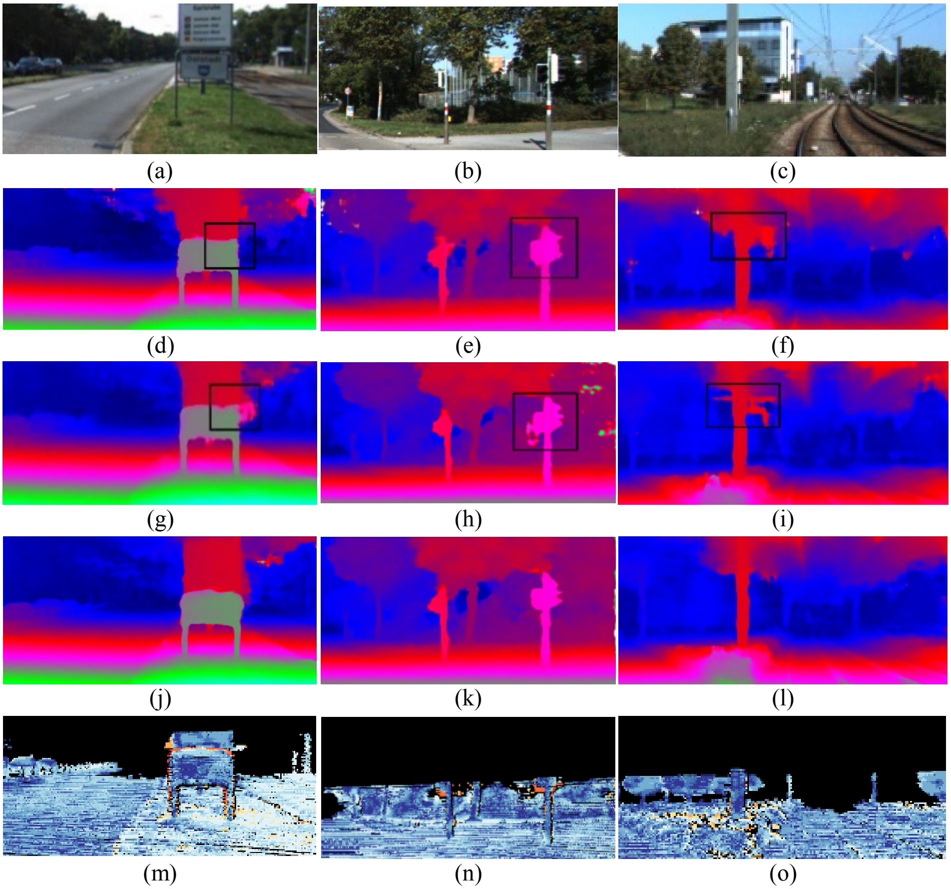
4.3 Visual comparison of multiple methods

The results of IRPSM-Net were uploaded to the KITTI official website for prediction, and visually compared with GCNet and PSMNet in Figure 6. In the error map, the blue parts indicate the correct points in the evaluation, the yellow parts indicate the error points in the evaluation, and the black part indicate the points that were not evaluated. Black boxes were used to identify the lack of details of objects in the other networks. Comparing GCNet and IRPSM-Net disparity map, the disparity map of GCNet generated many white error regions. Although the part of the white regions was not be considered in the accuracy calculation in the KITTI evaluation, it can be seen from the details of the image that the overall integrity of the GCNet image was lower than IRPSM-Net. And compared with the black box marked in the Figure 6(d-i), GCNet was hard to distinguish the details of complex objects, it only can obtain the rough outline. At the same time, comparing PSMNet and IRPSM-Net disparity map, the disparity maps generated by PSMNet and IRPSM-Net were equally complete, but the shape of foreground object in PSMNet will be affected by the background. Among the three scenes, IRPSM-Net can generate the contours of smaller objects in the disparity map accurately. Because the additional layer of large-scale pooling feature maps and information retention modules were used in IRPSM-Net, so that objects in the background can make great use of context information to ensure the integrity of objects in the disparity map.

5 Discussion

In this research, IRPSM-Net aimed to use an improved pyramid pooling feature extraction module to retain the original information of the image to improve the accuracy of the results. The improvement of the information retention strategy was divided into two aspects. The addition of largest average pooling features was considered to reduce the small-scale feature maps impact on results after sampling. The removing normalisation residual in residual dense module was adopted to highlight the information in the original image during the pooling process. During the experiment, modifying the number and size of feature maps in the pyramid network can improve the quality of the disparity map, but it may also increase the burden on the network. The removal of the normalisation operation was not advisable in the common network such as detection and identification, it will affect the network convergence speed and affect the regression of the task. A residual in residual dense module to prevent network instability. At the same time, stereo matching was a task with strict epipolar constraints. Removing the normalisation can effectively retain some features of the original image and achieve good results. In this experiment, we grouped the three-dimensional convolution into 8 groups in normalisation. The result confirmed that can effectively optimise the convolution effect and solve the error caused by small batches.

Figure 6 Visual comparison of the four methods on three pictures of the KITTI dataset. (a–c) are the examples of three scenes. (d–f) are the disparity map of GC-net evaluation in three scenes. (g–i) are the disparity map of PSMNet evaluation in three scenes. (j–l) are the disparity map of IRPSM-net evaluation in three scenes. (m–o) are the error map result of IRPSM-net evaluation in three scenes (see online version for colours)


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

6 Conclusion

In the course of the research, in order to solve the problem that the end to end stereo matching network always lost information and effect the accuracy of the result in the process of multiple up and down samplings. An IRPSM-net was proposed. The IRPSM-Net aimed to use the residual in residual dense block without the batch normalisation and the addition largest pyramid pooling feature map to retain more original image information. Furthermore, the batch normalisation in the three-dimensional stacked hourglass network was improved by the group normalisation, which solves the problem that the too small batchsize cannot obtain the optimal model after training. Then IRPSM-Net was trained and tested on the KITTI2012 and KITTI2015 datasets. The overall evaluation results showed that IRPSM-Net proposed in the research obtained an error rate of 1.79% in the threshold 3 pixels of KITTI2012. And the overall evaluation achieved an error rate of 2.24% in the whole region of KITTI2015. It improved the accuracy of 0.17% in KITTI2012 and 0.09% in KITTI2015 compared with the PSMNet.

In future work, the research will be focus on fastening the network speed, while maintaining the accuracy of the network. The stereo matching method could be applied to the binocular vision system in the field of vehicles and get the distance between the object and the vehicle when detecting. Finally, the research could be further improving the matching accuracy to adapt the complex road situation.

Acknowledgements

This work was supported by the National key research and development program of China (No. 2019YFE0126100), Key R & D projects in Zhejiang Province (No. 2019C54005).

References

Barnes, C., Shechtman, E., Finkelstein, A. and Goldman, D.B. (2009) ‘PatchMatch: A randomized correspondence algorithm for structural image editing’, *ACM Trans. Graph.*, Vol. 28, No. 3, pp.1–11.

Chang, J-R. and Chen, Y-S. (2018) ‘Pyramid stereo matching network’, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp.5410–5418.

Chen, F., Yu, H. and Ha, Y. (2020) ‘Quality estimation and optimization of adaptive stereo matching algorithms for smart vehicles’, *ACM Transactions on Embedded Computing Systems* (*TECS*), Vol. 19, No. 2, pp.1–24.

Duggal, S., Wang, S., Ma, W-C., Hu, R. and Urtasun, R. (2019) ‘Deeppruner: learning efficient stereo matching via differentiable patchmatch’, *Proceedings of the IEEE International Conference on Computer Vision*, pp.4384–4393.

Geiger, A., Lenz, P. and Urtasun, R. (2012) ‘Are we ready for autonomous driving? the Kitti vision benchmark suite’, *2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE*, pp.3354–3361.

Gidaris, S. and Komodakis, N. (2017) ‘Detect, replace, refine: deep structured prediction for pixel wise labeling’, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5248–5257.

Girshick, R. (2015) 'Fast R-CNN', *Proceedings of the IEEE International Conference on Computer Vision*, pp.1440–1448.

Guney, F. and Geiger, A. (2015) 'Displets: resolving stereo ambiguities using object knowledge', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4165–4175.

Guo, C., Chen, D. and Huang, Z. (2019) 'Learning efficient stereo matching network with depth discontinuity aware super-resolution', *IEEE Access*, Vol. 7, pp.159712–159723.

He, K., Zhang, X., Ren, S. and Sun, J. (2015) 'Spatial pyramid pooling in deep convolutional networks for visual recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 9, pp.1904–1916.

Hirschmuller, H. (2007) 'Stereo processing by semiglobal matching and mutual information', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 2, pp.328–341.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A. and Bry, A. (2017) 'End-to-end learning of geometry and context for deep stereo regression', *Proceedings of the IEEE International Conference on Computer Vision*, pp.4165–4175.

Kingma, D.P. and Ba, J. (2014) *Adam: A Method for Stochastic Optimization*. arXiv preprint arXiv: 1412.6980.

Lai, X., Xu, X., Zhang, J., Fang, Y. and Huang, Z. (2019) 'An efficient implementation of a census-based stereo matching and its applications in medical imaging', *Journal of Medical Imaging and Health Informatics*, Vol. 9, No. 6, pp.1152–1159.

Ling, X., Zhao, Y., Gong, L., Liu, C. and Wang, T. (2019) 'Dual-arm cooperation and implementing for robotic harvesting tomato using binocular vision', *Robotics and Autonomous Systems*, Vol. 2019, No. 114, pp.134–143.

Menze, M. and Geiger, A. (2015) 'Object scene flow for autonomous vehicles', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3061–3070.

Newell, A., Yang, K. and Deng, J. (2016) 'Stacked hourglass networks for human pose estimation', *European Conference on Computer Vision*. Springer, pp.483–499. Nguyen, T.P. and Jeon, J.W. (2019) 'Wide context learning network for stereo matching', *Signal Processing: Image Communication*, Vol. 78, pp.263–273.

Nie, G-Y., Cheng, M-M., Liu, Y., Liang, Z., Fan, D-P., Liu, Y. and Wang, Y. (2019) 'Multi-level context ultra-aggregation for stereo matching', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3283–3291.

Sang, H., Wang, Q. and Zhao, Y. (2019) 'Multi-scale context attention network for stereo matching', *IEEE Access*, Vol. 7, pp.15152–15161.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Change Loy, C. (2018) 'ESRGAN: Enhanced super-resolution generative adversarial networks', *European Conference on Computer Vision*, pp.63–79.

Wu, Y. and He, K. (2018) 'Group normalization', *Proceedings of the European Conference on Computer Vision*, Munich, Germany, pp.3–19.

Zbontar, J. and LeCun, Y. (2015) 'Computing the stereo matching cost with a convolutional neural network', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1592–1599.

Zeng, J-s., Xue, W-k., Xu, B-f. and Lang, M-m. (2019) 'Research on robot positioning and grasping technology based on binocular vision', *Modular Machine Tool and Automatic Manufacturing Technique*, Vol. 1, pp.35–46.

Zou, Y., Luo, Z. and Huang, J-B. (2018) 'Df-net: unsupervised joint learning of depth and flow using cross-task consistency', *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.36–53.