PMYNet: A Light-weight Network Base on Image Channel Separation for Stereo Matching

Yang Chen[1,a] and Zongqing Lu[1,b]

[1]*Department of Electronic Engineering Graduate School at Shenzhen, Tsinghua University ShenZhen, China 518055*

a. cy-17@mails.tsinghua.edu.cn, b. luzq@sz.tsinghua.edu.cn

Keywords: Stereo matching, color, CNN, light-weight.

Abstract: With the continuous development of neural network research, the network has higher and higher requirements for memory and computational power, because the network has been designed to be more and more complex with tons of parameters. We compose a light-weight network with a pyramid module of YUV channel and attention block, which made a reasonable trade-off between the complexity of the network and the accuracy of the results.

1. Introduction

Given a pair of images captured by two cameras in different viewpoints, the process to obtain the depth information by matching corresponding points is known as stereo matching. The basis of stereo matching is to estimate the shift between pixels called disparity of the rectified image pair. Stereo matching is a classical problem in the domain of computer vision, which arose widespread concern in both industry and academia. And it is not widely applied just in autonomous driving but also in 3D reconstruction, object recognition and road detection [1] etc.

With the development of photographic technology, the images turn colored from gray ones. Meanwhile, in the field of using convolution neural networks to study stereo matching, there is a mainstream trend to continuously deepen the network to improve the learning ability of the network [2, 3], which has increased the amount of network parameters and computational complexity. In practical applications, a network with too many parameters and too complex calculations sometimes is inconvenient to apply, so it is necessary to propose a lightweight network.

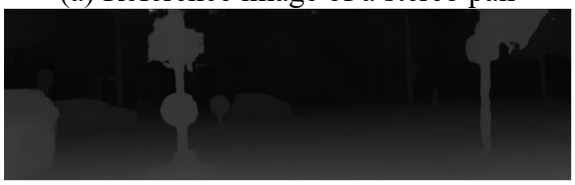
However, observing numerous common networks, we find that the three-channel combination of the colored image makes the convolution operation more complicated than the single-channel convolution operation.

Based on these facts above, we raise some reasonable questions: is the effect of the three-channel combined image operation proportional to the increase in complexity? Is it possible to sacrifice a small amount of arithmetic precision for a lighter network and simpler computation? In our work, we manage to give a plausible solution of the above problems.

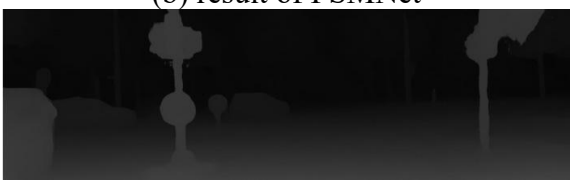
In our work, we proposed a novel and light-weight architecture with less parameters, consisting of two main modules: the pyramid module of YUV channels and attention block. As for the first module, we convert the images into YUV color space and compute the YUV channels separately through three different branches of network and aggregates the outputs of the branches to obtain the

final estimation. For this part, we try to aggregate the 3 channels information innovatively. As for the attention block, we add it to our network to gain better performance of accuracy.

(a) Reference image of a stereo pair



(b) result of PSMNet



(c) result of our work

Figure 1: A pair of output images obtained by PSMNet and our work on KITTI 2015 validation set.

Our contribution can be summarized in 3 points:

1. We propose a novel module to reduce the parameters of network, which computed the image channels separately.

2. We reasonably assume that the information of Y is much richer than the UV channel, and the information of the RGB channel is more evenly distributed among the three channels. Based on this guess, we perform different arithmetic operations on different channels: upscale the information-rich channels so that they can transmit more information to the network. And our guesses were verified by experiments.

3. We train our network with Squeeze-and-Excitation (SE) attention block which can help our network make a better balance between accuracy and computational complexity. And the experiment results prove that our work sacrifices only a small amount of accuracy for a considerable reduction in the number of parameters.

2. Related Work

Since the stereo match being of great value in the computer vision domain, numerous peers of the realm spare no effort to do research on it.

A Geiger and P Lenz [13] collected the KITTI datasets for the autonomous driving field usage, not only for stereo matching but also optical flow and scene flow, which offer a bunch of valid and generic data captured from real driving scenario.

Jia-Ren Chang and Yong-Sheng Chen [5] proposed a novel network architecture including spatial pyramid pooling module [6,7] and dilated convolution [8,9] to aggregate context information from different scales, so that the relationship between an object and its detailed information in sub-regions can be learned.

Bleyer M et al. [10] discussed the role of color in global stereo matching approaches, and concluded that color some-times even has consistently led to performance degradation. In [11], Mikel Galar et al. affirmed the positive effect of the methods based on RGB space but only in the case of using appropriate aggregation functions. We propose the pyramid module of YUV channels, and apply them on PSMNet's architecture. Inspired by the discussion of the effectiveness in [10] and [11], we design the pyramid module of YUV channels to apply various computation through different branches of network to leverage the color information.

3. Method

We present a novel architecture, which consists of the pyramid module of YUV channels for more reasonable aggregation of color information, which can be used to reduce the amount of network parameters.

3.1. Network Architecture

Our network is based on PSMNet's backbone. The architecture of our work is illustrated in Figure 2.

The pyramid module of YUV channels, as shown in Figure 2, is applied before the Feature Extraction module aiming to leverage the color information separated in YUV channels. In contrast to the PSMNet [5] and other studies, we separate the two input images in 3 channels of YUV space and then compute the left and right image information of one single channel respectively through the module. Then CNN blocks and dilated spatial pyramid module are applied on the output of three channels. We fed the 3 pairs of feature maps into cost volume and following blocks to obtain the disparity regression.

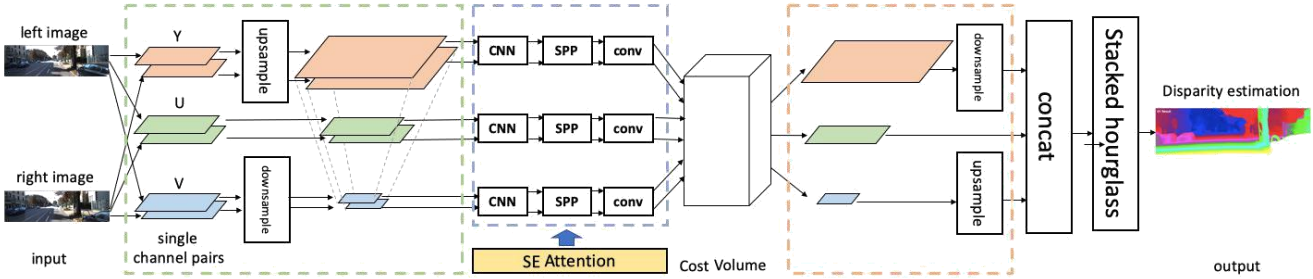


Figure 2: Network architecture of our work.

Figure 3: Schematic diagrams of the image after the separation of the RGB & YUV channel. The left column is the set of RGB channel separation images, and the right column is the set of YUV channel separation images.

3.2.Pyramid Module of YUV Channels

With the development of photographic technology, the images turn colored from gray ones. According to mentioned in [8], the use of RGB space in stereo matching outperforms the methods based on gray scale images [9] because the former one leverage the color information from the three channels. Hence, the methods based on RGB space effectively avoids the false matches. However, the improvement relies heavily on the appropriate aggregation function, since the improper handling of the extra information might have an adverse effect [7]. Meanwhile, compared with the methods on gray scale, the use of color image just improves the results slightly [12] considering its much more complex computation in some particular condition. In our work, we manage to give a plausible solution of the above problems. The stereo matching is a rather complicated task, and needs large amount of computation. Based on this, we are motivated to find solutions to reduce the complexity of computation.

Compared with a gray image, a color image can provide useful information for stereo matching [12] in some cases. For example, pixel A has two disparity candidates B and C, which have the same intensity value. B has the same color with A while C does not. It's obvious that color images would gain a better performance. However, the color images also make the computation more complex as the result of the 3-channel information. We doubt if the improvements brought by 3-channel information deserve the increase of computational complexity.

In traditional methods of stereo matching, we compute the images encoded in RGB color space, which is combined by 3 channels. And 3D convolution is used to complete the computation. So we manage to compute the 3 channels separately so that we can avoid using 3D convolution and reduce the complexity of computation. Moreover, images can also be encoded in YUV color space. In YUV space, Y represents luminance and (U,V) represent chrominance [13]. Human eyes are more sensitive to luminance and relatively insensitive to position and color. As shown in Figure 3, we can intuitively assume that the information of Y is much richer than that of the UV channel, and the information in the RGB space is more evenly distributed among the three channels. Y channel.is separated from the chromatic channels to reduce the lighting variation effect, and contains most of the image information. So out of reasonable speculation, luminance information may play a greater role in the disparity estimation.

To reduce the computational complexity and remain the rich information, we designed a novel architecture composed by pyramid module of RGB channels. As is shown in Figure 2, we compute the 3 channels of YUV separately so that we can reduce the complexity of computation.

First, we convert the image from RGB to YUV. And we compress the different channels of image into different scales to aggregate the global context information with the feature maps of more detailed information. We upscale the Y channel to fed more luminance information to the network. Meanwhile, we remain the scale of U and downscale the V channel information to suppress the chroma information. With these processes, we reduce the parameters and increase the proportion of the channel which contains more "useful" information in operations.

3.3. Attention Block

When something come into human's eyes, the visual system won't process the whole scene at once but process the important part of the scene first. This is an important mechanism to make human to catch more information with less efforts.

As mentioned in the related work section, Squeeze-and-Excitation (SE) [4] block is proved to be effective to improve the experiment results. Spired by this, we add the SE block to network

respectively to increase the accuracy of our network, which can make up for the accuracy lost due to the reduction in the number of parameters.

4. Experiment & Result

Our network has been experimented on two data sets: Scene Flow and KITTI 2015 [13]. And the experiment's results show that our method greatly reduces the amount of network parameters. At the same time, after improving the network effect by attention block, only a small amount of accuracy is sacrificed.

4.1. Datasets

4.1.1. Sceneflow

The sceneflow dataset is a dataset com-posed of synthetic images, including flyingthings-3d, Driving, Monkaa, which contains more than 39000 stereo frames in 960×540 pixel resolution. We mainly experiment on flyingthings-3d. Flyingthings-3d is divided into two parts, train and test, trainset is 5 times larger than testset, while the two parts are completely consistent in other aspects.

4.1.2. KITTI 2015

The KITTI 2015 is a dataset of the real-world traffic situation scenario images captured by a driving vehicle. The image size is 376×1240. And 200 pairs of training stereo images with sparse ground-truth disparities and 200 pairs of images without ground-truth for testing are concluded in the dataset.

4.2. Experiments for Ablation Study on Scene Flow

In order to verify the specific role of the module, we conducted a comparative experiment: keeping various parameters unchanged, we experimented on the PSM backbone network and the network with the YUV module added. The experimental results are shown in Table 1: the network with Pyramid Module of YUV Channels gains the end-point error at 1.473. Compared with the PSM backbone network, the 3-channel pyramid module also gains a good performance, which sacrificed a little accuracy at an acceptable level for the considerable parameter reduction.

Table 1: Experimental Results on SceneFlow Datase.

PSM backbone	Pyramid Module of YUV Channels			Flyingthings 3d Dataset		
	Upsampled channel	Unchanged channel	Downsampled channel	Senet attention	Number of parameters	end-point error
P					5224768	1.147
P	V	U	Y		2081752	84.206
P	Y	U	V		2081752	1.473
P	Y	U	V	P	2082650	1.158

And when we changed the channels into VUY order and make the V channel feed the biggest amount of information, the accuracy of the experimental results fell sharply to an unacceptable level at 84.206. And this result proves that the Y channel contains more “useful” information.

Furthermore, compared with PSM backbone network, the experimental results show that the accuracy of the network with the attention block increased by 0.315.

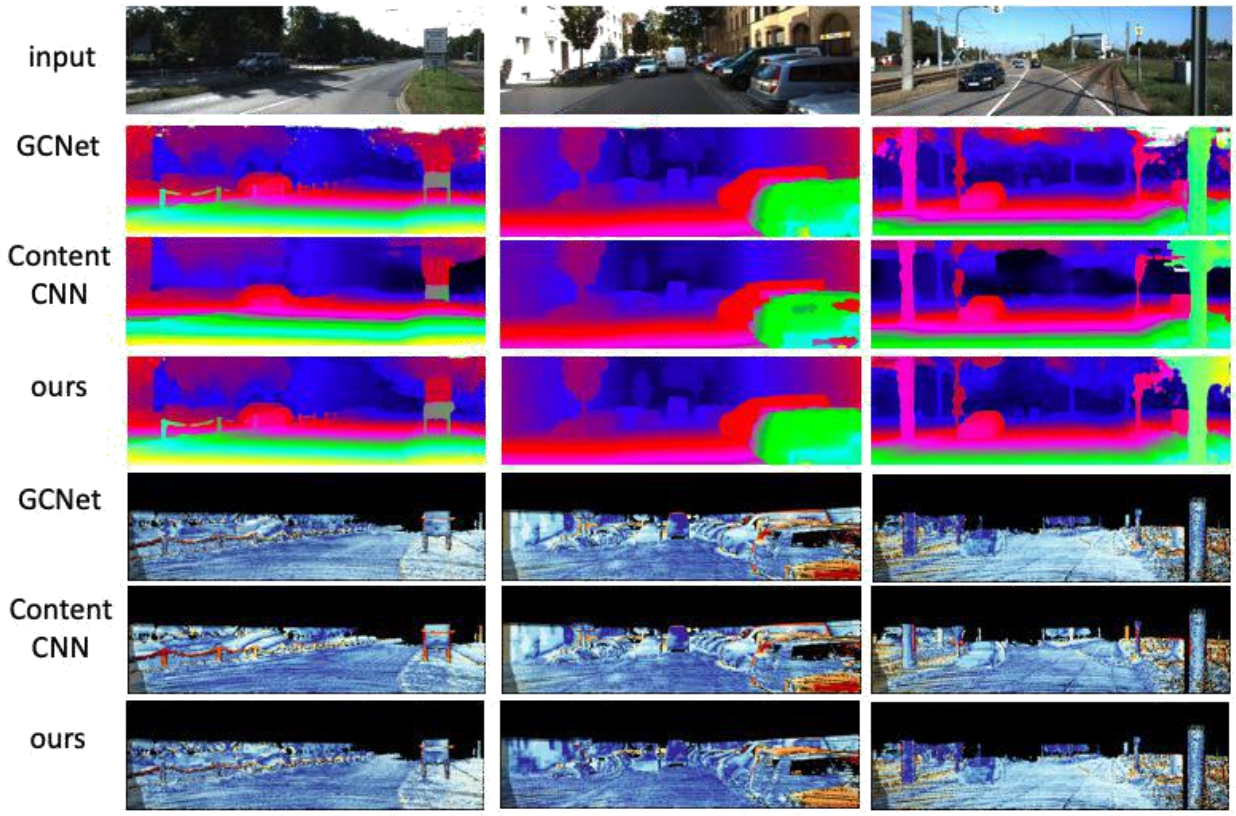


Figure 4. The results of disparity estimation on KITTI 2015 test images. The first row is the set of left input images. And the second to forth rows are the output disparity estimations of GC-Net [15], Content-CNN [14], our work and in a top-down order. Accordingly, the fifth to seventh rows are the error maps in the same order.

4.3. Experiments on KITTI 2015

We constructed our work named Pyramid Module of YUV channels Network (PMYNet) on the KITTI 2015 dataset [13] and submitted the model on KITTI evaluation server, and we obtained the result of 2.63% 3-pixel error. And it turns out that our work also surpassed some prior studies as showed in Table 2 according to the official evaluation table. As an illustration, the “All” columns present the error estimation over all pixels. And the “D1” means the percentage of stereo disparity outliers in first frame, when the “bg”, “fg”, “all” present the percentage of error estimation averaged only over background, foreground and all ground truth regions.

Table 2: Results on KITTI Evaluation Server (Submitted on Jan.29 2020).

Methods	All(%)		
	D1-bg	D1-fg	D1-all
PSM[5]	1.86	4.62	2.32
ContentCNN[14]	3.73	8.58	4.54
SGM[16]	2.66	8.64	3.66
DispNetC[17]	4.32	4.41	4.34
PMYNet(ours)	2.63	5.72	3.15

Figure 1 shows a pair of output images obtained by PSMNet and our work. Figure 4 shows some output image pairs including the result images and corresponding error maps of ourwork, Content-CNN [14] and GC-Net [15] given by the KITTI evaluation server. It can be figured out that our method works better over some ill-posed areas in Figure 1 and Figure 4. As is shown in Table 1, the parameter of our work is 2082650, which is much more less than the 5224768 of PSMNet, which means the reduction in computational complexity.

5. Conclusions

The use of color images has performed well on traditional stereo matching task, but it has brought complex computation due to the redundant information of the all three channels. According to results above, our work is proved to utilize the color information and reduce the computation. However, the output accuracy of our work still has room for improvement. In the future, we plan to exploit more about the network to improve its accuracy.

Acknowledgments

This work was financially supported by the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (JCYJ20170817161056260).

References

[1] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual corre-spondence embedding model for stereo matching costs," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 972–980.

[2] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2014.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[5] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "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, 2015.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.

[9] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convo-lutions," arXiv preprint arXiv:1511.07122, 2015.

[10] M. Bleyer and S. Chambon, "Does color really help in dense stereo matching," in Proceedings of the international symposium 3D data processing, visualization and transmission. Citeseer, 2010.

[11] M. Galar, A. Jurio, C. Lopez-Molina, D. Paternain, J. Sanz, and Bustince, "Aggregation functions to combine rgb color channels in stereo matching," Optics express, vol. 21, no. 1, pp. 1247–1257, 2013.

[12] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1721–1730.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[14] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5695–5703.

[15] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75.

[16] A. Seki and M. Pollefeys, "Sgm-nets: Semi-global matching with neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 231–240.

[17] Z. Zhu, M. He, Y. Dai, Z. Rao, and B. Li, "Multi-scale cross-form pyramid network for stereo matching," in 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2019, pp. 1789–1794.