GHRNET: GUIDED HIERARCHICAL REFINEMENT NETWORK FOR STEREO MATCHING

Bin Tan, Kai Chen, and Jian Yao†, Jie Li

School of Remote Sensing and Information Engineering, Wuhan University, P.R. China
†Email: jian.yao@whu.edu.cn Web: http://cvrs.whu.edu.cn/

ABSTRACT

Recently, deep convolutional neural networks (CNNs) have been well developed in the task of stereo matching. However, most existing CNNs-based methods cannot accurately estimate details especially for some tiny objects and shape boundaries. To solve this problem, in this paper, we propose a guided hierarchical refinement network (GHRNet) with a specially designed guidance block for details refinement. Specifically, an initial low-resolution disparity map is first estimated. It is then fed into a hierarchical architecture to recover disparity details from coarse to fine. The foremost contribution of our proposed method is an embedded guidance block, which fully makes use of clues from input images to refine predicated disparity map. We evaluate our network on several popular datasets. The experimental results demonstrate that the proposed network can significantly improve disparity details, and the overall result also achieves the state-of-the-art performance.

Index Terms— Stereo matching, hierarchical neural networks, disparity refinement, guidance block.

1. INTRODUCTION

Stereo matching is a classical problem whose main goal is to estimate disparities between a pair of rectified images. The output disparity map is essential for many applications, including autonomous vehicles [1], indoor localization [2] and 3D reconstruction [3]. Most traditional methods can be divided into four steps, including matching cost computation, cost aggregation, disparity estimation and disparity refinement [4]. However, matching cost computed with hand-crafted features is not robust for challenge scenes and thus limits the performance of traditional methods.

Recent researches show that CNNs can be applied to stereo matching. Early methods [5, 6] leverage the strong feature representation ability of CNNs to compute matching cost on image patches instead of using hand-crafted features. However, such local patch-based methods do not exploit global context information in images, and thus the disparity accuracy is limited seriously. Therefore, some end-to-end methods [7, 8, 9, 10] are proposed which exploit rich context information in input images, and greatly improve the disparity estimation accuracy. However, these end-to-end methods still fail to capture tiny objects. The small structures and boundaries are often undesirably lost or blurred in the final output disparity map.

Most recently, several guided networks [11, 12, 13, 14] have shown their superiority for details preserving. Khamis *et al.* [15] directly refer to the input RGB images as a color guide, and leverage an edge-aware upsampling function to refine disparity map. In contrast, many methods resort to some extra networks to generate the required guidance for stereo matching. SegStereo [12] employs a segmentation network to predict semantic segmentation maps from input images, which then are served as the guidance to enhance disparity details. Similarly, EdgeStereo [13] combines a disparity estimation network with an edge detection network. The extracted edges then are used to regulate disparity values in the boundary region. However, StereoNet [15] and EdgeStereo [13] just loosely concatenate the guidance information with the initial disparity map. Such a loosely coupled guidance scheme is easily affected by various noises existing in the provided guidance information, such as errors in edge detection procedure, and blurs contained in input RGB images. Besides, employing an extra network (e.g., networks for segmentation or edge detection) to provide required stereo matching guidance increases the complexity of the whole framework, and also makes the training process more difficult.

In this paper, we focus on the detail refinement problem in stereo matching, and propose a guided hierarchical refinement network (GHRNet). After estimating an initial disparity map in low resolution, a novel guidance block is embedded to improve disparity details. Instead of simply concatenating the RGB images with the initial disparity map, our guidance block first filters the RGB images as well as the initial disparity map with a series of convolution layers, and then concatenates their filtered features to refine the disparity map. A hierarchical scheme is also adopted to progressively improve the disparity map from coarse to fine. Compared with existing methods, our proposed method has two advantages. Firstly, since our guidance block makes full use of image clues and actually achieves a tight coupling between the guidance information and the initial disparity map, it performs stably even when input images have severe noises (e.g., image blur). Besides, compared with the Seg-

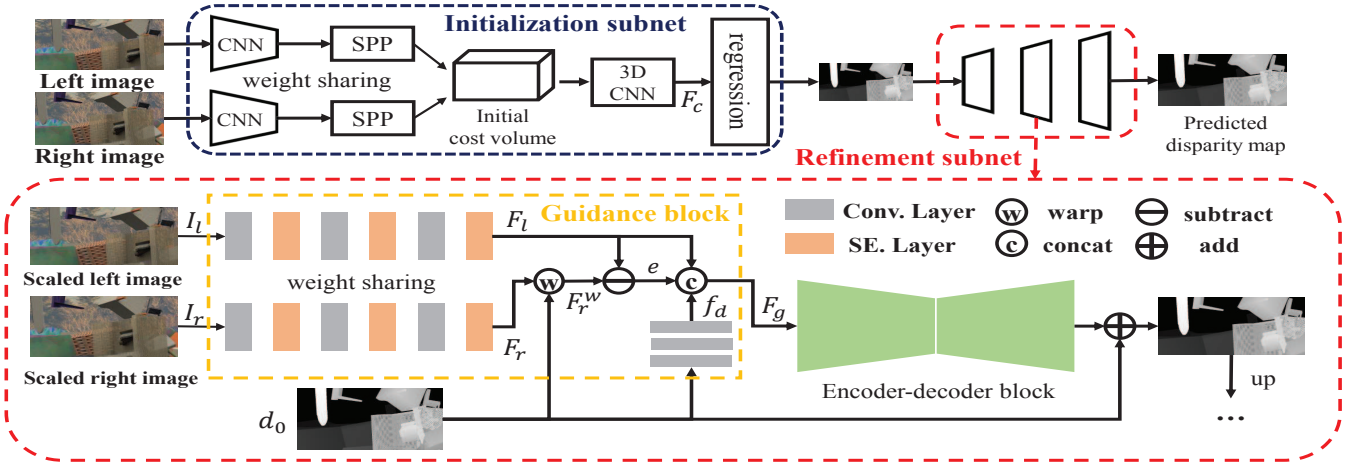


Fig. 1. The architecture overview of our proposed network. It consists of two subnets: an initialization subnet and a refinement subnet. The left and right images are fed to the first subnet to output an initial disparity map. Then, the initial disparity map is fed to the second subnet and refined from coarse to fine.

Stereo [12] and the EdgeStereo [13], our method does not bring any additional tasks, hence it is much easier to train. Experimental results on several datasets demonstrate that our method significantly improves disparity details and achieves the state-of-the-art performance overall.

2. OUR METHOD

The proposed GHRNet consists of two subnets: an initialization subnet and a refinement subnet. The initialization subnet which is most similar to PSMNet [8] takes the stereo images as input to generate an initial disparity map with low resolution and is briefly explained in Section 2.1. Then a refinement subnet is used to hierarchically refine the initial disparities. The process of disparity refinement with guidance block in one scale is described in Section 2.2. The hierarchical architecture of the refinement subnet is described in Section 2.3. The whole architecture of our GHRNet is shown in Fig. 2 and more detailed parameters of the network are shown in supplementary materials.

2.1. Initialization subnet

The initialization subnet can be divided into three steps. The first step is to extract deep features from the input image pair. Both left and right images are fed into a siamese network which contains three 3×3 convolution layers and several residual blocks [16]. The size of the output deep features is $W/8 \times H/8$, where W and H denote the width and the height of original input images. Then, to exploit multi-scale context information from deep features, a SPP module [17] is applied to output the incorporated deep features as shown in Fig. 2.

Secondly, a correlation operation proposed in [7] is used to build a low-resolution 4D cost volume from the incorporated deep features. Several 3D convolution layers are applied to aggregate the cost volume and output a 3D aggregated cost volume F_c with the size of $W/8 \times H/8 \times D/8$, where D denotes the maximum searching range of disparities.

Finally, we use a soft argmin regression function as used in [7, 8, 15, 18] to estimate the initial disparity map with the size of $W/8 \times H/8$ from F_c.

2.2. Disparity refinement with guidance block

The refinement subnet has a hierarchical architecture as described in Section 2.3. In this section, we only explain the process of disparity refinement with guidance block in scale k with the resolution of $W/2^k \times H/2^k$ as shown in Fig. 2.

Given an initial disparity map d_0 in scale k, we first downsample the input left and right images to the corresponding size, defined as I_l and I_r. Then, d_0, I_l and I_r are fed into an embedded guidance block to product a guided volume F_g. At last, an encoder-decoder block is used to fuse guidance information in F_g and learn residual disparities, especially in small structures and boundaries.

Considering that directly serving RGB images as a color guide do not fully exploit the guidance information in detail structures and is easily affected by image noises, we first pass I_l and I_r through three convolution layers respectively and generate robust clues, defined as F_l and F_r. Then, an error map, $e = |F_l - F_r^w|$, as used in [18, 13] is introduced to serve as a soft guidance. Here, F_r^w is computed by warping F_r according to the initial disparity map d_0. A large value in e indicates that there may be occlusion areas or incorrect matching. Instead of loosely concatenating guidance with disparities as [13, 15], the initial disparity map d_0 is preprocessed by three convolution layers that output a 1-dimensional feature map, defined as f_d. Finally, f_d, F_l and e are concatenated together to build a tightly coupled guided volume F_g.

However, features output by convolution layers are usually redundant. Therefore, to enhance useful clues in guidance features F_l and F_r, we employ a squeeze and excitation layer (SELayer) [19] to reweight feature maps after each

convolution layer in our guidance block. Here, we represent the feature volume as $F = [f_1, f_2, f_3, ..., f_C]$, where $f_i \in \mathbb{R}^{W_F \times H_F}$ and C is the channel numbers of F. A global average pooling operation is used to squeeze F across feature channels and output a vector $v \in \mathbb{R}^C$. Then, some convolution layers are applied to v, as

$$u = \mathbf{W} * v, \tag{1}$$

where $\mathbf{W}$ is the convolution filter, $*$ denotes a convolution operation and $u \in \mathbb{R}^C$ is a weight vector. Finally, the original feature volume can be reweighted by the weight vector, as

$$\hat{F} = [f_1 \cdot u_1, f_2 \cdot u_2, f_3 \cdot u_3, ..., f_C \cdot u_C], \tag{2}$$

where $f_i \in F$ and $u_i \in u$. By leveraging SELayers, the guidance block is of the ability to learn which features are useful to disparity refinement and such features that are meaningless can be weakened.

After building a guided volume F_g, an encoder-decoder block is applied to fuse different guidance information. To focus more attention of the network on recovering detail structures in disparities, we learn a residual disparity map from the encoder-decoder block followed by a 3×3 convolution layer without using batch normalization and any activation. The refinement disparity map is finally calculated by adding residual disparities to the initial disparity map.

2.3. Hierarchical refinement architecture

The refinement subnet has three refinement scales in this paper. Starting from the scale k, we first refine the initial disparity map with an embedded guidance block as described in Section 2.2 and then upsample the refined result by a factor of 2. The upsampled result is served as the initial disparity map for the next scale $k-1$. By repeating this process, we finally output a refined disparity map in full resolution. By using such hierarchical scheme, the residual disparities are easy to learn in each scale and the disparity map can be refined with more details from coarse to fine.

Besides, a robust smooth L1 loss is used to train our network. The loss function of our network can be defined as,

$$L = \sum_k smooth_{L1}(d_k - \hat{d}), \tag{3}$$

where d_k is the estimated disparities of scale k and $\hat{d}$ is the ground truth disparities.

3. EXPERIMENTS

3.1. Implementation details

We evaluated our GHRNet on the Scene Flow [9] and KITTI 2015 [20] dataset. The Scene Flow dataset is a synthetic dataset containing 35454 training and 4370 testing image pairs. The KITTI 2015 dataset is a real word dataset containing 200 training and 200 testing image pairs. Our model was implemented using Pytorch and used the Adam optimizer [21] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. All image pairs were cropped to the size of $H = 256$ and $W = 512$ during training. The maximum disparity range was set to 192. We trained our models in three phases on the Scene Flow dataset. In the first phase, we only trained the initialization subnet with a learning rate of 0.001 for 15 epochs. In the second phase, we fixed the initialization subnet, and the refinement subnet was trained with a learning rate of 0.001 for the first 10 epochs and 5×10^{-4} for the remaining 5 epochs. Finally, in the last phase, we trained the whole model with a learning rate of 1×10^{-4} for 5 epochs. For the KITTI 2015 dataset, we finetuned the model pre-trained on the Scene Flow dataset with a decayed learning rate for 1000 epochs. The learning rate was set to 1×10^{-4} at the first 100 epochs and was reduced by a half at the next 200 epochs. At the remaining 700 epochs, the learning rate was fixed to be 1×10^{-5}. The batch size was set to 12 for training on three GTX1080-Ti GPUs. It has taken about 20 hours to train the model on the Scene Flow dataset and 9 hours on the KITTI 2015 dataset.

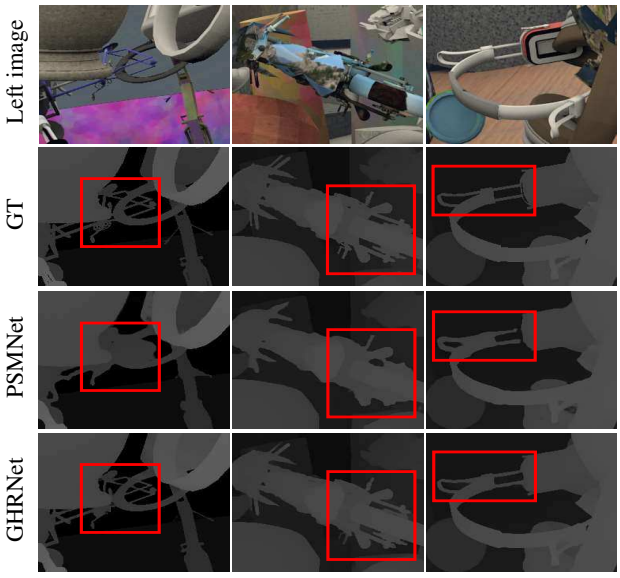


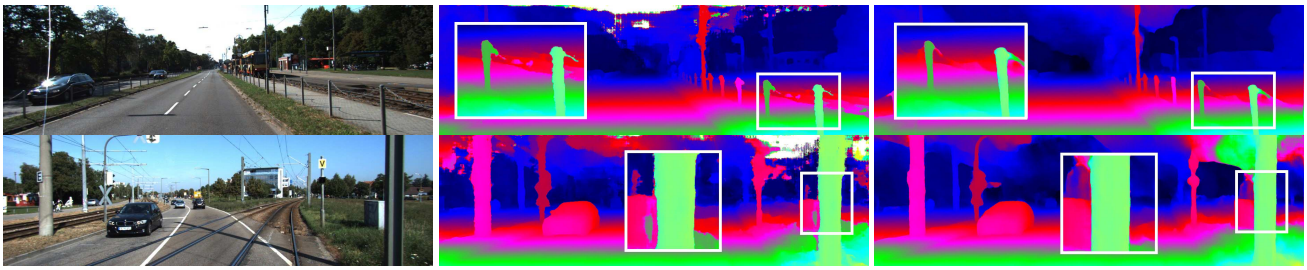
Fig. 2. Results of our GHRNet on the Scene Flow dataset compared to PSMNet [8].

3.2. Benchmark results

We first compared our method on the Scene Flow dataset with some state-of-the-art methods, including CRL [10], DispNetC [9], GC-Net [7], PSM-Net [8], StereoNet [15], iResNet [18] and EdgeStereo [13]. The comparisons are presented in Table 1. The end-point-error (EPE) of our method is 0.68 which is smaller than other methods. Besides, we also use the percentage of disparities with their EPE larger than t pixels ($> t$px), denoted as t-px error. As shown in Table 1, the 3-px error of our method is just slightly larger than iResNet [18] while it outperforms the remaining methods. This is because iResNet [18] repeatedly uses their refinement subnet to improve the accuracy of disparities while the refinement subnet in our network is only use once. Some qualitative results of our network compared with PSMNet [8]

	PSMNet	StereoNet	EdgeStereo	CRL	iResNet	DispNet	GC-Net	GHRNet (Ours)
EPE	1.09	1.10	1.11	1.32	1.40	1.68	2.51	**0.68**
$>$ 3px	-	-	4.97	6.20	**4.57**	8.61	7.20	4.60

Table 1. Comparisons of stereo matching methods on the Scene Flow dataset.



Left image GC-Net [7] GHRNet (Ours)

Fig. 3. Results of disparity estimation on the KITTI 2015 dataset. The first column shows the input left image. The disparity estimation results of GC-Net [7] and our method are shown in the second and third columns. Some details are magnified and showed in white boxes.

	All (%)			Noc (%)		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
PSMNet [8]	**1.86**	4.62	**2.32**	**1.71**	4.31	**2.14**
iResNet-i2 [18]	2.25	**3.4**	2.44	2.07	**2.76**	2.19
EdgeStereo [13]	2.27	4.18	2.59	2.12	3.85	2.40
GC-Net [7]	2.21	6.16	2.87	2.02	5.58	2.61
MC-CNN [5]	2.89	8.88	3.88	2.48	7.64	3.33
GHRNet (Ours)	2.48	4.29	2.78	2.30	3.78	2.55

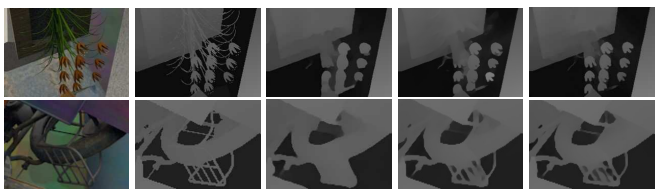
Table 2. Results on the KITTI 2015 dataset.

are shown in Fig. 3. Disparity maps estimated by our network have more details than PSMNet.

For the KITTI 2015 dataset, we trained our model on all 200 training image pairs. We use the percentage of erroneous pixels in background (D1-bg), foreground (D1-fg) and all pixels (D1-all) of all regions (All) and non-occluded regions (Noc). As shown in Table 2, our method achieves a competitive performance compared with some state-of-the-art methods. Some qualitative results of our method compared with GC-Net [7] on the KITTI 2015 dataset are shown in Fig. 4. The results show that our method estimates disparities with more shape boundaries than GC-Net.

3.3. Ablation experiments

We conducted several ablation experiments on the Scene Flow dataset to validate the embedded guidance block. As



Left image GT (a) (b) (c)

Fig. 4. Ablation results on noisy data. Here, (a), (b) and (c) represent the setting of "w/o guidance", "guidance w/o SE" and "guidance w/ SE", respectively .

metric / model	clean			noise		
	EPE	$>$ 1px	$>$ 3px	EPE	$>$ 1px	$>$ 3px
w/o guidance	0.78	9.9	5.0	1.10	12.3	6.0
guidance w/o SE	0.70	9.7	4.8	0.93	11.5	5.8
guidance w/ SE	**0.68**	**9.3**	**4.6**	**0.90**	**10.8**	**5.4**

Table 3. Evaluation of our method with different settings on the Scene Flow dataset. Here, "w/" represents "with" and "w/o" represents "without".

shown in Table 3, "clean" represents the original data as used in Section 3.2 and "noise" represents the data with image blur. Table 3 illustrates that settings with the guidance block outperform the setting without guidance block. Besides, our network achieves the best performance by adding SELayers to the guidance block. Some qualitative results on noise data are shown in Fig. 5 and further prove the robustness of our guidance block in disparity details estimation on noise data.

4. CONCLUSION

In this work, we propose a hierarchical stereo matching network with an embedded guidance block for details refinement. An initial disparity map is estimated first and then refined by a hierarchical architecture which recovers disparity details from coarse to fine. With the help of our guidance block which fully uses the clues from input images, our network performs stably even there are noises in images. The experimental results demonstrate that the proposed network can significantly improve disparity details, and achieves the state-of-the-art performance overall.

Acknowledgement

This work was partially supported by the National Key Research and Development Program of China (No. 2017YFB1302400) and the National Natural Science Foundation of China (No. 41571436).

5. REFERENCES

[1] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *The IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008, pp. 3946–3952. 1

[2] R. Mur-Artal and J. D. Tards, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017. 1

[3] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D reconstruction in real-time," in *The IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 963–968. 1

[4] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002. 1

[5] J. bontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, no. 1-32, pp. 2, 2016. 1, 4

[6] W. Luo, A. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5695–5703. 1

[7] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-end learning of geometry and context for deep stereo regression," pp. 66–75, 2017. 1, 2, 3, 4

[8] J. Chang and Y. Chen, "Pyramid stereo matching network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418. 1, 2, 3, 4

[9] N. Mayer, E. Ilg, P. Husser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048. 1, 3, 4

[10] J. Pang, W. Sun, J. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching.," in *ICCV Workshops*, 2017, vol. 7. 1, 4

[11] Y. Li, J. Huang, N. Ahuja, and M. Yang, "Deep joint image filtering," in *The European Conference on Computer Vision (ECCV)*, 2016, pp. 154–169. 1

[12] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," *The European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[13] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," *arXiv preprint arXiv:1803.05196*, 2018. 1, 2, 3, 4

[14] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, "Learning dynamic guidance for depth image enhancement," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 712–721. 1

[15] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *The European Conference on Computer Vision (ECCV)*, 2018, pp. 8–14. 1, 3, 4

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *The European Conference on Computer Vision (ECCV)*, 2014, pp. 346–361. 2

[18] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2811–2820. 3, 4

[19] J. Hu, L. Shen, and G. Su, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[20] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070. 3

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 4