Stereo Refinement Dehazing Network

Nie, Jing; Pang, Yanwei; Xie, Jin; Pan, Jing; Han, Jungong

Published in:
IEEE Transactions on Circuits and Systems for Video Technology

DOI:
10.1109/TCSVT.2021.3105685

Publication date:
2022

Citation for published version (APA):
Nie, J., Pang, Y., Xie, J., Pan, J., & Han, J. (2022). Stereo Refinement Dehazing Network. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(6), 3334-3345. https://doi.org/10.1109/TCSVT.2021.3105685

Stereo Refinement Dehazing Network

Journal:	*IEEE Transactions on Circuits and Systems for Video Technology*
Manuscript ID	TCSVT-06805-2021
Manuscript Type:	Transactions Papers - Regular Issue
Date Submitted by the Author:	02-Jul-2021
Complete List of Authors:	Nie, Jing Pang, Yanwei; Tianjin University, School of Electrical and Information Engineering Xie, Jin Han, Jungong; Aberystwyth University, computer science Pan, Jing; Tianjin University of Technology and Education,
EDICS:	Multi-Image Fusion and Restoration < 1.7.2 ☐ Restoration and Enhancement < 1.7 ☐ Image/Video Quality Enhancement < 1 ☐ IMAGE/VIDEO PROCESSING

TCSVT-06095-2021

Stereo Refinement Dehazing Network

Response to the Reviewers

REJECT AND INVITE TO RESUBMIT

Firstly, we would like to thank the Reviewers for their valuable comments and insightful advises on our paper, and for the important opportunity to resubmit the paper. We have greatly benefited from the ideas and recommendations. Based on the reviewer team's inputs, we have been able to prepare a much better manuscript. We provide below a detailed account on the changes that we have made in response to comments of the reviewers. The corresponding changes in the resubmitted manuscript are marked in *blue* color. Apart from addressing these concerns from the editor and reviewers, we have also further polished the presentation, and the changed parts are labelled in *red* in the resubmitted version.

To Reviewer # 1

1. In Section II, a recent work is missing, can the author explain the advantage of the proposed method over the method below:

   [R1] Song, T., Kim, Y., Oh, C. et al. Simultaneous Deep Stereo Matching and Dehazing with Feature Attention. IJCV, 2020.

   Response: Thanks for your comment. The paper [R1] is a journal extension of the method SSDMN [8], in the early submission we have compared SSDMN with our method in Section II and compared their dehazing performance in the experimental part (Section IV). In addition, we add the method [R1] and explain the differences and advantages of our proposed SRDNet in Section II. The method [R1], called SSDFA by us, is a multi-task method that estimates a clear image and disparity simultaneously from a hazy stereo image pair. It introduces an attentional feature fusion in order to integrate depth-related features effectively from the matching cost and haze transmission. Disparity estimation is time-consuming and the estimation from haze images is a more challenging problem. Furthermore, achieving the two tasks optimal jointly is hard. In contrast, we concentrate on the dehazing task without predicting disparity and obtain better dehazing performance. In addition, SSDFA is still a dehazing network based on the physical model. Our SRDNet directly restores haze-free stereo images, which is not limited in the computational relation of the physical model and generalizes real hazy scenes well.

2. Except to the perceptual quality for high-level vision tasks on the foggy kitti dataset, the authors should also present the dehazing performance such as the metrics of PSNR.

   Response: Thanks for your advice. Following your advice, we evaluate the dehazing performance on the foggy KITTI dataset in terms of PSNR and SSIM in Section IV.E. On the Stereo Foggy KITTI validation set, our SRDNet improves the PSNR values from 11.89 dB and 10.96 dB to 22.83 dB and 22.14 dB in terms of the left view and the right view respectively as presented in Tab. 1. For the metric of the SSIM, our SRDNet boosts 0.2301 and 0.2303 for the left view and the right view, respectively.

   Table 1. Comparing the dehazing performance of the foggy inputs and the outputs restored by our SRDNet on the KITTI validation set.

	Left		Right	
	PSNR	SSIM	PSNR	SSIM
Foggy	11.89	0.5766	10.96	0.5697
SRDNet	22.83	0.8067	22.14	0.8066

3. For high-level vision tasks, only SRDNet results are presented. Please also include other networks for a fair

comparison.

Response: Thanks for your advice. We compare our SRDNet and BidNet which both belong to the stereo dehazing methods, in 3D detection task. In terms of each metric of AP_{3D} in the different hazy scenes, Tab. 2 (*i.e.,* Tab. VII in the manuscript) shows that our method obtains higher accuracy and has better perceptual quality.

Table 2. 3D Detection performance comparisons on the KITTI validation set for car with all three settings: Light Haze, Medium Haze and Heavy Haze. Specially, APE, APM and APH are used to evaluate the performance of easy, moderate and hard sets respectively.

Haze	AP_{3D}	Hazy	BidNet	SRDNet
Light	AP_E	40.75	46.34	**47.06**
	AP_M	25.79	31.17	**32.37**
	AP_H	23.75	25.59	**26.90**
Medium	AP_E	34.26	44.0	**44.97**
	AP_M	21.49	27.2	**31.13**
	AP_H	19.83	25.3	**25.97**
Heavy	AP_E	24.14	40.0	**40.37**
	AP_M	14.89	26.35	**27.16**
	AP_H	13.84	21.50	**21.82**

4. Some Typos:

(1) Page3Line50: The weight-sharing coarse dehazing network (WSCDN) enjoys a (an) encoder-decoder structure and is shared by the left view and the right view.

(2) Page6Line49: The values of SSIM also reduce more than 0.0018 dB (delete dB) compared with employing the GCSR module.

(3) Page7Line51: Our SRDNet achieves an obvious gain with 3.36 dB in the PSNR and 0.054 dB (delete dB) in the SSIM compared with the MSBDN.

Response: Thank you very much for pointing out these problems. According to your suggestion, in the resubmitted version, we have completely solved the acronym problems you mentioned. Furthermore, we have polished our manuscript and marked them red.

To Reviewer #2

1. (1) The author only summarizes the two shortcomings of the previous methods, but does not explain the improvement of their method in terms of these shortcomings. For example, how to solve the generalization problem?

Response: Thanks for your comment. In paragraph 2, line 58, right column, page 1, of the resubmitted version, we have described how our method can solve the shortcomings including the generalization problem. For the sake of clarity, we re-state the two shortcomings followed by describing how our method is capable of solving the corresponding shortcomings. Specifically, there are two shortcomings in the previous stereo dehazing methods:

① They simultaneously restore clear images and predict disparity. Disparity estimation is time-consuming and the estimation from haze images is a more challenging problem. A small error in disparity gives rise to a large variation in depth and in estimation of haze-free image. Furthermore, achieving the two tasks optimal jointly is hard, it is preferable to not directly utilizing disparity for haze removal. Although BidNet constructs the matrix in horizontal dimension to mining the information from the cross view, when the width of the input image gets larger, the needed memories for the matrix construction are very large. It can be observed from Tab. 4.

② The previous stereo dehazing methods are based on the atmosphere scattering model. The performance of the restored images is over dependent on the joint accurate estimations of the transmission map and the atmospheric light. The atmosphere scattering model is crude and cannot fully express the complicated real scenes [R2] [R3].

[R2] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in Proc. IEEE Conf. Computer Vision and Pattern Recognition,2019, pp. 8160–8168.

[R3] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in Proc. IEEE Conf. Computer Vision, 2019, pp. 7314–7323.

To address the above two shortcomings, ①our SRDNet concentrates on the dehazing task and does not predict disparity, which makes the dehazing task optimal. Without applying the matrix, the SRDNet concatenates the left features with the right features and mines the depth information through a stereo feature extractor. When the width of input gets large, the improvement for the need of the computational memories is far lower than the matrix construction. We also design a guided channel and spatial refinement (GCSR) module separating the features to choose the useful information for different views, which impresses the negative effect of the inaccurate information and the irrelevant information. ② Our SRDNet is an end-to-end deep learning model that directly restore haze-free stereo images and is not dependent on the joint accurate estimations of the transmission map and the atmospheric light. In addition, our method is not limited in the computational relation of the physical model and can model more complicated and redundant computational relation to fit and generalize real foggy scenes.

In addition, experimental results demonstrate the superiority of the proposed method beyond two types of methods: SSMDN and BidNet.

(2) The contribution is just a two-stage dehazing net, which is somewhat limited.

Response: Thanks for your comment. In the last three paragraphs of Section I of the resubmitted version, we have re-summarized our core contributions. Importantly, we have also clarified and emphasized the challenges behind the contributions. The challenges are mainly divided into two aspects:

On the one hand, it is not effective that directly apply a two-stage dehazing net of the domain of single image dehazing into the scenario of stereo image dehazing because of lack of mechanism to adopting stereo information helpful for dehazing. In the domain of stereo image dehazing, it is challenging to design an effective two-stage dehazing net in a coarse-to-fine way.

On the other hand, it is challenging to make use of stereo information to positively refine the coarse dehazed images because that the stereo information such as disparity/depth/distance is not accurate and employing the inaccurate information can even damage the dehazed results. Therefore, it is considerately difficult to utilize stereo information for designing a stereo dehazing framework immune to the negative effect of the inaccurate information.

To summarize, our contribution lies in how to deal with the challenges in designing a two-stage dehazing net in the relatively new domain of stereo image dehazing.

It is the proposed SRDNet (Stereo Refinement Dehazing Network) that effectively deal with the above-mentioned challenges. Our SRDNet is not a simple two-stage dehazing net. It incorporates a weight-sharing coarse dehazing network (WSCDN) and a guided separated refinement network (GSRN). The GSRN learns the residues for the corresponding views to refine the coarse dehazed image pair through a stereo feature extractor and a guided channel and spatial refinement (GCSR) module. The stereo feature extractor makes full use of the information of cross views. The GCSR module separates the features for different views and predicts the corresponding residues, which impresses the negative effect of the inaccurate information and the irrelevant information. We also construct a two-stage dehazing net by replacing the

GCSR module by a 3×3 convolutional layer. This simple two-stage dehazing net is compared with our method in Tab. 3 (*i.e.,* Tab. IV in the manuscript).

Table 3. The ablation study on the Stereo Foggy Cityscapes validation set.

	Methods	Left		Right	
		PSNR	SSIM	PSNR	SSIM
Case1	Single-stage: WSCDN w/o GSRN	28.55	0.9648	28.34	0.9648
Case2	Two-stage: 3×3 Convolution	28.70	0.9681	28.33	0.9681
Case3	Two-stage: GCSR module	**30.27**	**0.9704**	**30.11**	**0.9699**

From Tab. 3, a simple two-stage dehazing net work (*i.e.* case2) only outperforms the single-stage dehazing net (*i.e.* case1) by a little margin. Instead, our two-stage network (*i.e.* case3) outperforms the other two cases by a large margin. Specially, the dehazing results decrease 1.57 dB and 1.78 dB for left dehazed images and right dehazed images from the perspective of PSNR once removing our designed GCSR module. In addition, compared with case2 (params: 753.23k), our network (*i.e.* case3) only adds negligible overheads (params: 759.56k).

The contribution of our method can be mainly divided into three aspects:

We propose a stereo refinement dehazing network (SRDNet) to directly recover the clean stereo images in a coarse-to-fine fashion, which is the first attempt to address the stereo image dehazing progressively. The SRDNet makes full use of information collaboratively encoded in the cross views meanwhile without employing disparity or correlation matrix. Our SRDNet is not limited to the simple physical model, and learns a more complicated model that better matches the real foggy scenes. It is of great importance to eliminate the performance degradation of stereo based 3D detectors caused by the foggy inputs.

The SRDNet incorporates a weight-sharing coarse dehazing network (WSCDN) and a guided separated refinement network (GSRN). The WSCDN removes a part of haze and obtains a coarse dehazed image pair. The GSRN learns the residues for the corresponding views to refine the coarse dehazed image pair through a stereo feature extractor and a guided channel and spatial refinement (GCSR) module. The stereo feature extractor makes full use of the information of cross views. The GCSR module separates the features for different views and predicts the corresponding residues, which impresses the negative effects of the inaccurate information and the irrelevant information.

Experimental results demonstrate that our proposed SRDNet surpasses previous state-of-the-art image dehazing methods by a large margin both quantitatively and qualitatively. Specially, our method outperforms the sota stereo dehazing method by 4.70 dB and 4.44 dB for the binocular image pair on the Stereo Foggy Cityscapes dataset in terms of the PSNR. Moreover, our SRDNet could be a preprocessing step of the stereo image-based 3D object detection and boost the 3D detection accuracy in hazy scenes. By appending the SRDNet, the average precision improves by 16.23% in the heavy haze condition on the KITTI Val dataset for easy sets.

2. (1)The contribution parts. First, why the method without employing disparity or correlation matrix is better? How to solve the shortcomings by the proposed two-stage networks?

Response: Thanks for your comment. The disadvantages of the method employing disparity or correlation matrix are: (a) Firstly, disparity prediction is a challenging task and achieving the two tasks optimal jointly is hard. A small error in disparity gives rise to a large variation in depth and in estimation of haze-free image. In hazy scenes, it is hard to estimate the correct disparity map or the correct correlation matrix. (b) Although the computation of the matrix is only in horizontal dimension, when the width of the input image is large, the needed memories for the matrix multiplication are large too. In

order to have a clear understanding about the needed memories of the horizontal matrix, we replace the horizontal matrix by a convolution with the kernel size of 3 in a simple net. As presented in Tab. 4, when the input size gets large, the need of the memory of the horizontal matrix [R4] improves a lot.

[R4] Y. Pang, J. Nie, J. Xie, J. Han, and X. Li, "Bidnet: Binocular image dehazing without explicit disparity estimation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2020.

Table 4. In order to have a clear understanding about the needed memories of the horizontal matrix, we replace the horizontal matrix by a convolution with the kernel size of 3 in some simple nets and test their needed memories on GeForce GTX 1070 GPU with batchsize of 1. . The simple net consists of a convolution with the size of (3,64,3,3)(*i.e.* input channel, output channel, kernel width, kernel height), a convolution with the size (64,64,3,3) (or a horizontal matrix as in BidNet), and a convolution with the size (64,3,3,3). When the input size gets large, the need of the memory of the horizontal matrix improves a lot.

Input_size	Input_channel	Module (one branch)	Memory	
256×256	64	Conv3×3	661M	×1
	64	Horizontal matrix	1013M	×1.53
512×256	64	Conv3×3	773M	×1
	64	Horizontal matrix	2129M	×2.75
512×512	64	Conv3×3	965M	×1
	64	Horizontal matrix	3669M	×3.80

In contrast, our SRDNet concentrates on the dehazing task and does not predict disparity, which makes the dehazing task optimal. Without applying the matrix, the SRDNet concatenates the left features with the right features and mines the depth information through a stereo feature extractor. When the width of input gets large, the improvement for the need computational memories is far lower than the matrix construction. In order to choose the useful information for each view and not introduce confusing information, we also design a guided channel and spatial refinement (GCSR) module to separates the features of each view. From Tab. I in the manuscript, the dehazing performance of our SRDNet outperforms other methods by a large margin, which also demonstrates the effectiveness of our method.

(2) Maybe the proposed method can solve these problems, but the introduction part is not clear, and should be rewritten.

Response: Thanks for your advice. We rewrite and update the introduction in a clear and organized way. For better understanding, the changed parts are labelled in *blue* in the resubmitted version.

3. The results on Drivingstereo dataset only compare with MSBDN on the quantitative results. More comparisons are needed to make the results on the real data set convincing.

Response: Thanks for your comment. We compare quantitative results of more methods in Tab. 5 (*i.e.* Tab. III of the manuscript) on Drivingstereo dataset.

Table 5. PSNR and SSIM comparisons on foggy synthetic Drivingstereo dataset.

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
BidNet	22.96	0.8765	23.06	0.8785
MSBDN	24.75	0.8949	-	-
GCANet	27.41	0.8962	-	-
Ours	**28.01**	**0.9017**	**28.13**	**0.9065**

Additionally, we provide more results in Fig. 1 (*i.e.,* Fig.6 in the resubmitted manuscript) to make the results on the real data set convincing. Fig. 1 gives qualitative comparisons of our dehazed results with the state-of-the-arts: MSBDN, BidNet and GCANet on the real hazy images from the Drivingstereo dataset. It can be observed that there still exists quite a lot of haze in the results of MSBDN. Color distortion is introduced by BidNet. Compared with GCANet, our method performs better in the regions of the sky of row-4 and the road of row-5. Our method has visually appealing results.

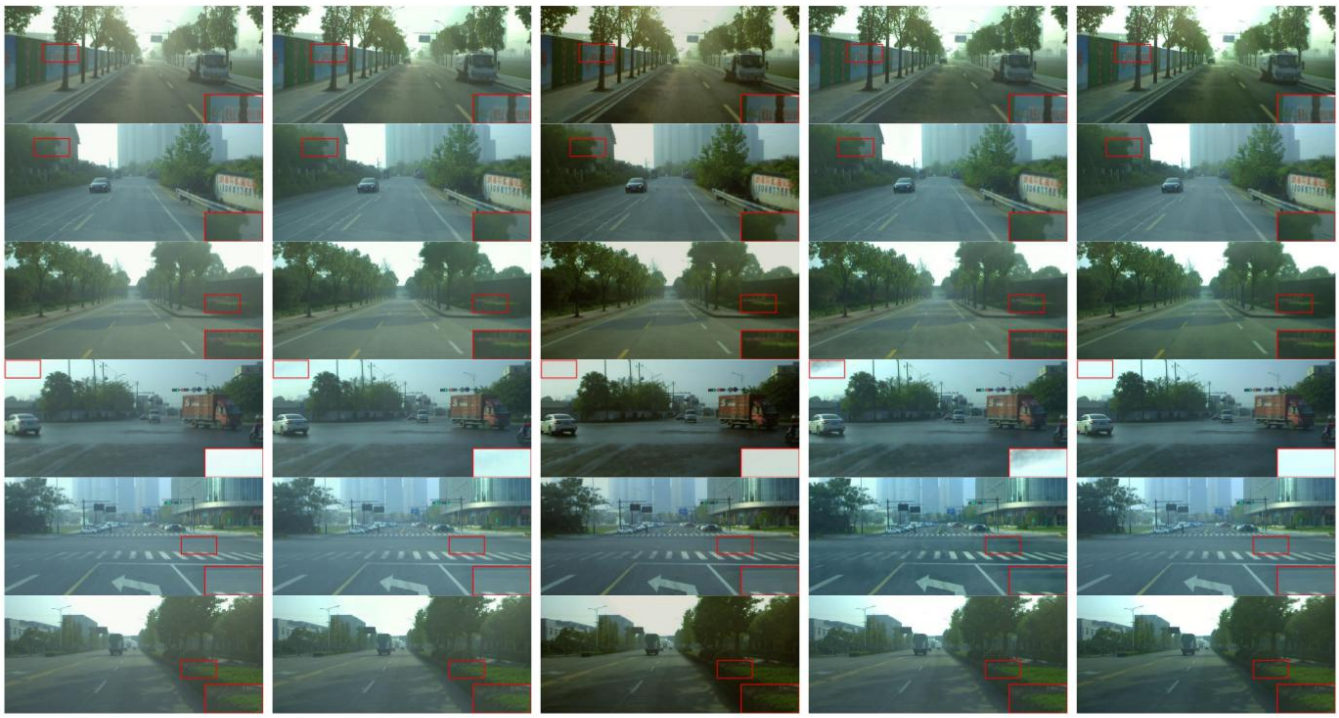


Figure 1. Evaluation on real foggy stereo images from the Drivingstereo Dataset. We only present the left dehazed images.

To Reviewer #3

1. For GCSR, only the feature from each view is used as the input. Why the residue information of each view is not fed to GCSR like the SFE? Please clarify it.

Response: Thanks for your comment. This is a typo in the Fig.2(b) of the manuscript, we correct the figure and the corresponding description in the manuscript. We use the residue information to conduct the channel refinement and the feature from each view is used in the spatial refinement as shown in Fig.2 (*i.e.,* Fig.2(b) in the resubmitted manuscript).

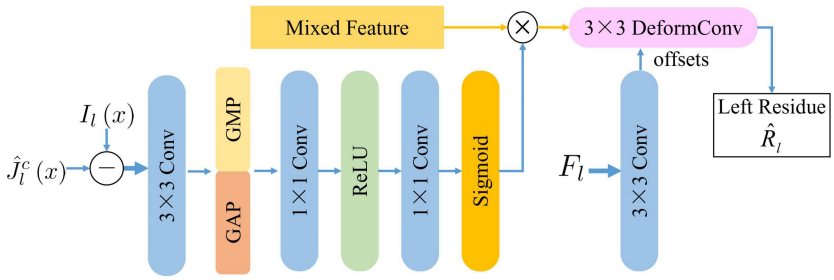


Figure 2. Guided channel and spatial refinement (GCSR) module

2. Why the max and average pooling are both used in GCSR and the basic block, but only the max-pooling is used in WSCDN? Please clarify the reasons and add more ablation experiments.

Response: Thanks for your comments. There may exists an ambiguity. In the GCSR module and the basic block, we utilize the global average pooling (GAP) and the global max pooling (GMP) to gather global statistic information and discriminative information, respectively. Our WSCDN consists of the Basic Blocks and some max-pooling operator with stride 2 to extract features. The max-pooling with stride 2 in WSCDN is used to expand the receptive field. We modify the description and the figures to eliminate the ambiguity in the resubmitted manuscript.

In addition, we add some ablation experiments to explore the effects of the global max-pooling and the global average-pooling in the Basic block in Tab.6 (*i.e.,* Tab. IV in the resubmitted manuscript). It shows that combing the GAP and the GMP could extract more abundant information and obtain the best dehazing performance.

Table 6. The ablation study on the Stereo Foggy Cityscapes validation set, which exploring the effects of the global average pooling (GAP) and the global max pooling (GMP) in the Basic block.

Basic Block	Left view		Right view	
	PSNR	SSIM	PSNR	SSIM
GAP	29.76	0.9696	29.54	0.9695
GMP	29.65	0.9693	29.60	0.9697
GAP+ GMP	**30.27**	**0.9704**	**30.11**	**0.9699**

3. In Section IV-C, the authors should also add some subjective comparisons between the dehazing results of WSCDN and the proposed two-stage method.

Response: Thanks for your comment. We add some subjective comparisons between the dehazing results of WSCDN and the results of the GSRN in Fig. 3 (*i.e.,* Fig.5 of the resubmitted manuscript), which demonstrates that the GSRN indeed refines the dehazing results.

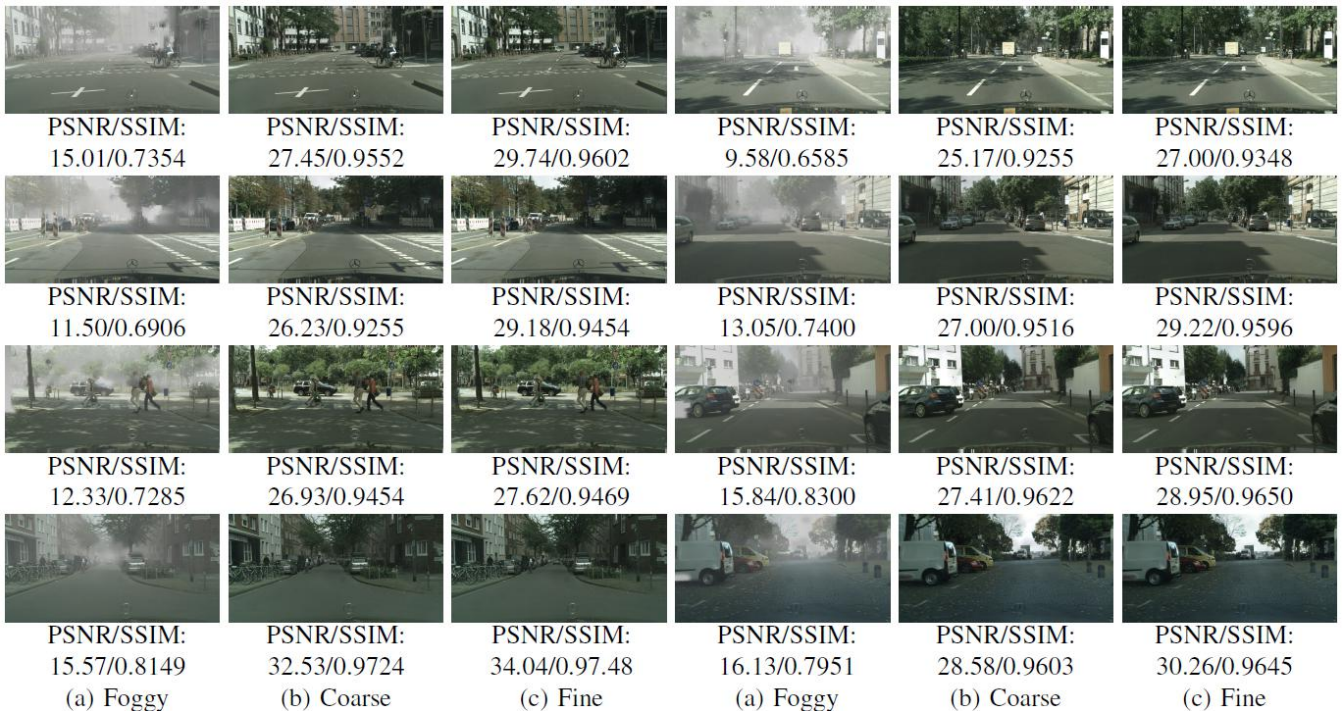


Figure 3. Subjective comparisons between the coarse dehazed results of the WSCDN and the refined results of the GSRN. We present the dehazed results

of left images. The right images dehazed by our method have analogous results.

4. In the caption of Fig. 6, the citation for Stereo Foggy KITTI Dataset is mistaken.

Response: Thanks for your comment. We have modified the citation in our resubmitted manuscript.

5. It would be better to add the comparison of parameters and FLOPS for different methods.

Response: Thanks for your advice. We add the comparison of parameters and FLOPS for different methods in Tab. 7 (*i.e.,* Tab. II in the manuscript). From Tab. 7, our SRDNet achieves a better trade-off between the performance and the computational cost when comparing with the methods: SSMDN, GCANet, MSBDN, and BidNet.

Table 7. Parameters and FLOPS comparison of our method with other methods when the input resolution is 512 ×1024. For clear comparison, the FLOPS of monocular dehazing methods are doubled (×2) because they are applied in the left view and the right view separately. We also present the value of the PSNR on the Stereo Foggy Cityscapes validation set in terms of the left view.

Method	Params	FLOPS	PSNR(L)
SSMDN	75.16M	82.18G	22.37
GCANet	702.82k	121.76G×2	27.66
MSBDN	31.35M	196.19G×2	24.73
BidNet	323.06k	26.86G	25.57
SRDNet(ours)	759.56k	63.47G	30.27

Stereo Refinement Dehazing Network

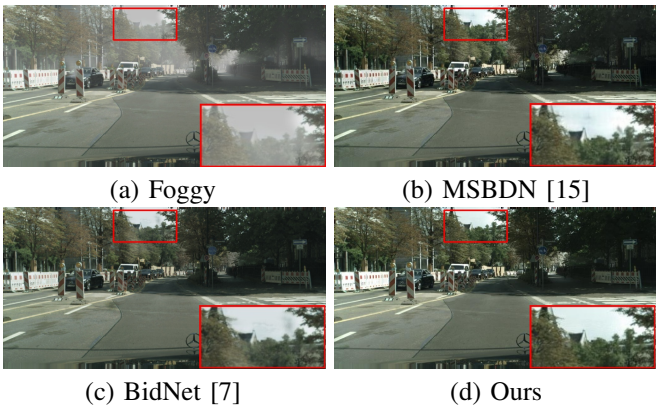
Jing Nie, Yanwei Pang, *Senior Member, IEEE*, Jin Xie, Jing Pan, and Jungong Han, *Senior Member, IEEE*

***Abstract*—The performance of stereo vision tasks degrades when haze exists in the input stereo image pair. Independently applying single image dehazing algorithm on left and right images is not optimal. To overcome the problem, we propose an effective framework, called SRDNet, for simultaneously dehazing stereo images. The main idea of SRDNet is to make full use of the stereo information from cross views improving dehazing performance. It does not explicitly employ the disparity estimation and the correlation matrix. SRDNet comprises two parts: a weight-sharing coarse dehazing network (WSCDN) and a guided separated refinement network (GSRN). The WSCDN is utilized to predict a coarse dehazed image pair. Then the GSRN is introduced to predict the residues for different views by extracting the fused information of cross views and separating the features of different views with a guided channel and spatial refinement module. The residues are added to the coarse dehazed pair so as to make refinement and remove the remained haze. The experimental results demonstrate that our proposed SRDNet surpasses previous image dehazing methods by a significant margin both quantitatively and qualitatively. Moreover, our SRDNet could be a preprocessing step of the stereo image-based 3D object detection and boost the 3D detection accuracy in hazy scenes.**

I. Introduction

Stereo vision has numerous advantages over monocular vision. For example, stereo vision is able to provide more precise depth and three-dimensional information of the objects and scenes [1], [2], [3], [4], [5], [6]. Therefore, stereo vision is widely applied in practical applications such as advanced driving assistance system, self-driving vehicles, unmanned surface vessel, and human-machine intelligence. However, the visibility of the stereo images and the scene understanding ability of stereo vision are deteriorated when haze occurs. The low-level vision tasks such as stereo dehazing [7], [8], [9], [10] and stereo deraining [11], [12], [13] are very necessary and have attracted increasing research attention in the computer vision community, which restores the stereo images from the corrupted inputs. Moreover, stereo images provide more information from cross views, which could boost the performance of the dehazing methods since they are depth related.

There are two strategies for dehazing stereo images. The first and straightforward strategy is independently dehazing the left image and right image captured by a binocular vision system. It can be accomplished by applying existing excellent single image dehazing methods such as GCANet [14], and MSBDN [15]. However, the single image dehazing methods do not utilize the relationship between the binocular images. Therefore, directly applying single image dehazing methods is not optimal for dehazing binocular images. The second strategy [7], [8], [10] is stereo image dehazing methods, utilizing the depth information contained in the stereo image pairs to help predict the dehazed images, which demonstrates the superiority of the stereo images. Na *et al*. [10] and Song *et al*. [8] explicitly estimated disparity and merged the intermediate features for disparity estimation into a dehazing network. Because disparity estimation from haze images is a more challenging problem and achieving the two tasks optimal jointly is hard, it is preferable to not directly utilizing disparity for haze removal. BidNet [7] dehazes the binocular images by mining the correlation between left and right images through constructing the matrix without explicitly estimating disparity, which achieves the state-of-the-art. Although the computation of the matrix is only in horizontal dimension, when the width of the input image gets large, the needed computational resources and the need of memories for the matrix multiplication are very large. The above stereo image dehazing methods are based on the atmosphere scattering model [16] and utilize the depth information contained in the stereo image pairs to help predict the transmission maps. Though of success, the performance of the model-based methods is over dependent on the joint accurate estimations of the transmission map and the atmospheric light. The model is too crude to fit the real foggy scenes and the dehazed results of the model-based methods for the real-world hazy images are unsatisfactory.



(a) Foggy (b) MSBDN [15] (c) BidNet [7] (d) Ours

Fig. 1. Visual comparison between our method and the state-of-the-art MSBDN [15] and BidNet [7] methods for a hazy image from the Stereo Foggy Cityscapes dataset [7].

To address the above issues, we design a stereo refinement dehazing network (SRDNet) in this paper to directly transform hazy stereo images to haze-free stereo images in a coarse-to-fine manner. Our SRDNet is not limited in the computational relation of the physical model and can model

J. Nie, Y. Pang and J. Xie are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. (E-mail: {jingnie,pyw, jinxie}@tju.edu.cn)

J. Pan is with the School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China (E-mail: jingpan23@gmail.com)

J. Han is with the School of Computer Science, Aberystwyth University, UK, (E-mail: jungonghan77@gmail.com)

more complicated and redundant computational relation to fit and generalize real foggy scenes. Our SRDNet employs neither disparity nor correlation matrix. It concentrates on the dehazing task and does not predict disparity, which makes the dehazing task optimal. Firstly a part of haze are removed by our designed weight-sharing coarse dehazing network called WSCDN for both the left view and the right view. Then we could generate the residues between the coarse dehazed stereo image pair and the input foggy stereo image pairs, which are combined with the features from the WSCDN together as the input to a guided separated refinement network (GSRN). The GSRN is composed of a stereo feature extractor and a guided channel and spatial refinement (GCSR) module. Without applying the matrix, the SRDNet concatenates the left features with the right features and mines the depth information through a stereo feature extractor. When the width of the input gets larger, the improvement for the need of the computational resources and memories is far lower than the matrix construction. The stereo feature extractor makes full use of the information of cross views. The GCSR module separates the features for different views and predicts the corresponding residues for the coarse haze-free image pairs. The residues help refine the coarse dehazed stereo image pairs to remove the remained haze. To summarize, our contributions are three-fold as below:

(1) We propose a stereo refinement dehazing network (SRDNet) to directly recover the clean stereo images in a coarse-to-fine fashion, which is the first attempt to address the stereo image dehazing progressively. The SRDNet makes full use of information collaboratively encoded in the cross views meanwhile without employing disparity or correlation matrix. Our SRDNet is not limited to the simple physical model, and learns a more complicated model that better matches the real foggy scenes. It is of great importance to eliminate the performance degradation of stereo based 3D detectors caused by the foggy inputs.

(2) The SRDNet incorporates a weight-sharing coarse dehazing network (WSCDN) and a guided separated refinement network (GSRN). The WSCDN removes a part of haze and obtains a coarse dehazed image pair. The GSRN learns the residues for the corresponding views to refine the coarse dehazed image pair through a stereo feature extractor and a guided channel and spatial refinement (GCSR) module. The stereo feature extractor makes full use of the information of cross views. The GCSR module separates the features for different views and predicts the corresponding residues, which impresses the negative effects of the inaccurate information and the irrelevant information.

(3) Experiments demonstrate that our proposed SRDNet surpasses previous state-of-the-art image dehazing methods by a large margin both quantitatively and qualitatively. Specially, our method outperforms the sota stereo dehazing method by 4.70 dB and 4.44 dB for the binocular image pair on the Stereo Foggy Cityscapes dataset in terms of the PSNR. Moreover, our SRDNet could be a preprocessing step of the stereo image-based 3D object detection and boost the 3D detection accuracy in hazy scenes. By appending the SRDNet, the average precision improves by 16.23% in the heavy haze condition on the KITTI Val dataset for easy sets.

II. Related works

Haze deteriorates the visibility and quality of images and introduces low contrast, blurring and so on [17], [18], [19], which leads the poor performance of image based tasks including object detection [20], [21], [22], classification [23], [24], tracking [25], [26] and person re-identification [27], [28], *etc*.. Image dehazing is a highly ill-posed problem and very challenging. In this section, we first describe single image dehazing method and then review stereo image dehazing methods.

A. Single Image Dehazing Methods

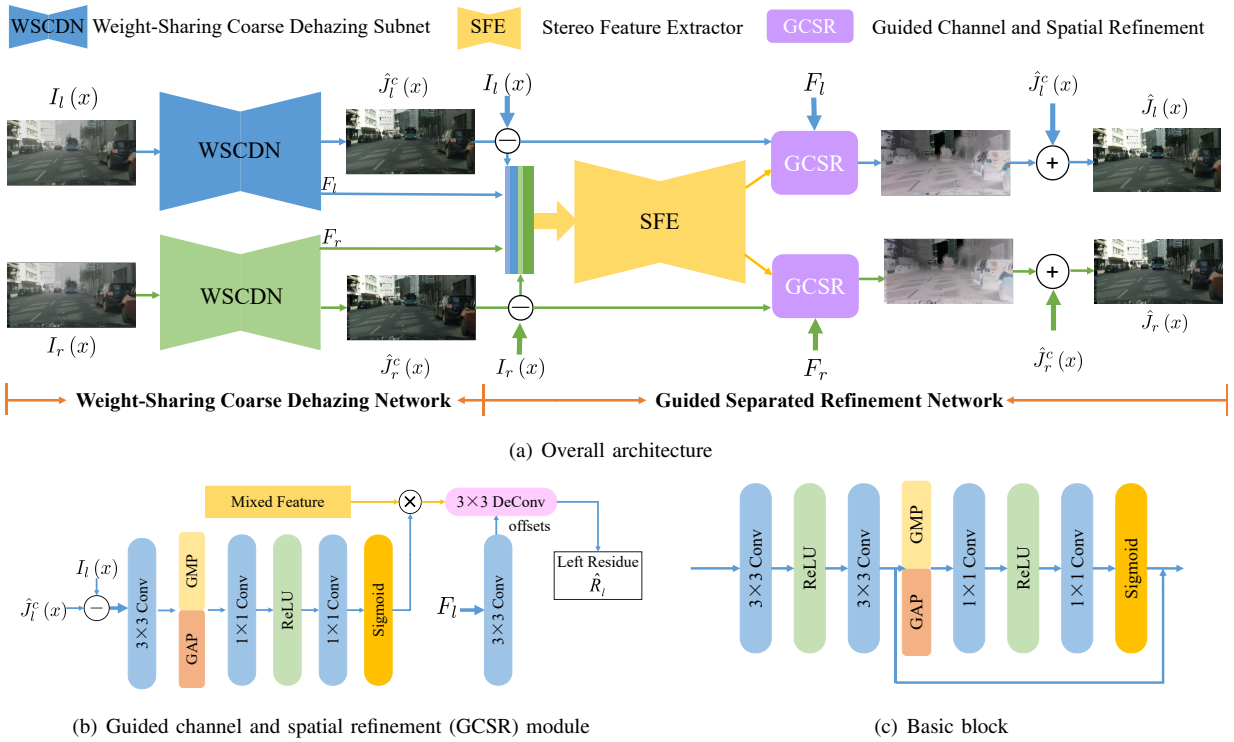
Single image dehazing methods can be divided into two categories: prior-based approaches and learning-based approaches. Most dehazing methods rely on the atmosphere scattering model formulated as:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

where $I(x)$ and $J(x)$ denote the hazy image and the clear image respectively. A is the global atmospheric light intensity, and $t(x)$ represents the transmission map. $t(x)$ is a function of depth: $t(x) = e^{-\beta d(x)}$, in which β and $d(x)$ are the atmosphere scattering parameter and the distance, respectively.

Prior-based approaches [29], [30], [31], [32], [33] employ strong priors as extra constraints to estimate the transmission maps and the global atmospheric lights, and then compute the haze-free results according to the atmosphere scattering model mentioned above. In order to boost the visibility of hazy images, Tan *et al*. [33] proposed to maximize the local contrast. The dark channel prior (DCP) [29] is put forward to estimate the transmission maps and restore the clean outdoor images. Berman *et al*. [32] developed an effective non-local path prior for single image dehazing.

Recently, most deep monocular dehazing methods achieve great success. The previous learning-based approaches [34], [35], [36], [37] also rely on the atmosphere scattering model, which first utilize the CNN to estimate transmission maps and atmospheric lights, and then restore clear images. Zhang *et al*. [18] regard the image dehazing problem as a iterative progress: first divide a hazy image into different regions and then optimize the atmospheric light and transmission simultaneously and iteratively based on local physical features. Several recent works [38], [14], [39], [40], [15], [41] reduce the image dehazing problem to an image-to-image translation problem. The Enhanced Pix2pix Dehazing Network (EPDN) [39] utilizes a generative adversarial network augmented with a well-designed enhancer to restore clear images directly. The Gated Context Aggregation Network (GCANet) adopts the smoothed dilated technique and fuses multi-level features for haze removal. The GCANet learns the residue between the clear image and the input foggy one. In contrast, the learning target of our guided separated refinement network is the residues between the coarse dehazed stereo image pairs and the final haze-free ones. GriddehazeNet [40] is an enhanced GridNet [42] with residual dense blocks [43]



(a) Overall architecture

(b) Guided channel and spatial refinement (GCSR) module

(c) Basic block

Fig. 2. The architecture of our SRDNet, the guided channel and spatial refinement module and the basic block.

for dehazing. A Multi-Scale Boosted Dehazing Network (MS-BDN) [15] applies the boosting strategy and the error back-projection technique to improve the performance of dehazing. To solve the widely diffusing caused by the haze, an end-to-end Pyramid Global Context Network (PGCNet) [17] is proposed to learn the global context. Zhang *et al*. [19] designed two modules adaptively fusing multi-level features to keep fine details and extract semantics.

B. Stereo Image Dehazing Methods

Stereo images processing has attracted increasing attention due to the advantages such as providing comparable depth accuracy. The methods based on stereo images have made great progress, such as 3D object detection [5], [4], [6] and stereo matching [44], [45], [46]. There are 18 stereo based 3D object detectors in the leaderboard of 3D detection evaluation on the KITTI website in the past two years. More information are provided from cross views by stereo images that have thus been utilized to improve the quality of various low-level tasks, including super-resolution [47], stereo image deraining [11], [12], [13] and stereo image dehazing [7], [8], [9], [10]. Li *et al*.presented a method recovering the clear images from foggy videos, which jointly predicts scene depths. They regarded that the stereo matching and the dehazing can reinforce each other. Song *et al*. [8] and Yun *et al*. [10] both proposed the deep-learning based multi-task methods that estimate a clear latent image and disparity simultaneously from a hazy stereo image pair. The intermediate features for disparity estimation are fused into a dehazing network to enhance each other. Song *et al*. [48] extends the work [8] by introducing an attentional feature fusion in order to integrate depth-related features effectively from the matching cost and haze transmission. Their attentional feature fusion consists of a channel/spatial attention fusion and a gated fusion. The channel/spatial attention fusion is separately conducted on the stereo features or on the transmission features, which is in a self-learning way. The gated fusion is to adaptively fuse the stereo features and the transmission features through a learning weight map. Differently, we propose a guided channel and spatial refinement (GCSR) module to extract features for the respective view from the mixed stereo features. The GCSR module is composed of a guided channel refinement and a guided spatial refinement, which are not in a self-learning way and instead is learned by the guidance of the residue information from the WSCDN and the feature from each view. Recently, BidNet [7] is proposed and does not explicitly estimate disparity. It explicitly computes correlation matrix of left and right features which is closely related to disparity. By contrast, our SRDNet employs neither disparity nor correlation matrix. The above methods based on the atmosphere scattering model are too simple to fit the real foggy scenes.

To address the aforementioned issues, this paper designs a stereo refinement dehazing network to directly recover the clean stereo pair from the foggy input pair, which utilizes the information from cross views and is more effective than single image dehazing methods in stereo tasks without estimating disparity.

III. METHODS

The visibility of the stereo images and the scene understanding ability of stereo vision are degraded when haze exists.

Stereo dehazing could be a preprocessing step of the high-level vision tasks such as the stereo image-based 3D object detection. Different from that existing stereo dehazing methods are based on the physical model, in this paper, we propose an end-to-end stereo refinement dehazing network (SRDNet) to directly recover the clean stereo images from the foggy input pair. It restores the haze-free stereo images in a coarse-to-fine manner: firstly learn a coarse haze-free image pairs and then learn the residues to refine the coarse images through excavating the depth information from the stereo image pairs.

A. Overall Architecture

Fig. 2(a) shows the overall architecture of the SRDNet which contains two parts: a weight-sharing coarse dehazing network (WSCDN) and a guided separated refinement network (GSRN). The WSCDN is utilized to predict a coarse dehazed image pair directly, whose weights are shared between the left images and the right images. The coarse dehazed stereo pair is removed most of haze. The residues between the coarse dehazed stereo image pair and the input foggy stereo image pairs are combined with the left and right features from the WSCDN, which are input to the GSRN. The GSRN utilizes a stereo feature extractor to fuse the information of cross views instead of predicting disparity. In addition, the GSRN designs a channel and spatial refinement module to separate the features for different views to predict the residues. The predicted residues refines the coarse dehazed images, which removes the remained haze and obtains clearer image pairs.

B. Weight-Sharing Coarse Dehazing Network

The weight-sharing coarse dehazing network (WSCDN) enjoys an encoder-decoder structure and is shared by the left view and the right view. The encoder of the WSCDN inputs the left foggy image I_l or the right foggy image I_r and stacks basic blocks and max-pooling with stride 2 to extract the features. The encoder first utilizes a basic block to learn better input features and then downsamples the input features through a max-pooling followed with a basic block, which is repeated 4 times and iteratively enlarges the receptive field. The decoder accordingly applies the bilinear interpolation with one basic blocks by 4 times to restore the detailed structure. The same level feature maps between the encoder and the decoder are concatenated to preserve spatial information at each resolution. The final feature maps F_l and F_r output from the weight-shared decoders of the left view and the right view will be further fed into two separated 3×3 convolutional layer to restore the coarse dehazed image pair: $\hat{J}_l^c$ and $\hat{J}_r^c$, which are removed most of haze.

Basic Block: As shown in Fig. 2(c), a basic block is composed of a 3×3 convolutional layer with ReLU and another 3×3 convolutional layer with self gated mechanism. The self gated mechanism is to learn the channel-wise weight $\mathcal{G}_s$ for the input feature F_{in} to recalibrate the features adaptively. We choose the global max pooling (GMP) and the global average pooling (GAP) along spatial dimension to obtain the global spatial information, which is formulated as Eq. 2. Then we use two 1×1 convolutional layers followed with ReLU and sigmoid respectively to further fuse the useful information and generate channel-wise weights which are used to multiply with the input feature to recalibrate the input feature along the channel dimension.

$$P_c = Concat(\mathcal{P}_{max}(F_{in}), \mathcal{P}_{avg}(F_{in})), \tag{2}$$

$$\mathcal{G}_s = \sigma(\mathcal{C}_{1\times1}(\delta((\mathcal{C}_{1\times1}(P_c))))), \tag{3}$$

where $Concat$ means concatenating the outputs of the global max pooling $\mathcal{P}_{max}$ and the global average pooling $\mathcal{P}_{avg}$. σ and δ are the sigmoid function and the ReLU function respectively.

Finally, the output feature are obtained below:

$$F_o = \mathcal{G}_s \odot F_{in} + F_{in}. \tag{4}$$

where $\odot$ refers to channel-wise product.

C. Guided Separated Refinement Network

In order to utilize the information from cross views and refine the predicted coarse dehazed stereo images $\hat{J}_l^c$ and $\hat{J}_r^c$, we design a guided separated refinement network, called GSRN. The GSRN first uses a stereo feature extractor to extract stereo mixed features and fuse the information from cross views. Then a guided channel and spatial refinement (GCSR) module is designed to guide to separate the features for different views. The separated features predict the separated residues $\hat{R}_l$ and $\hat{R}_r$ for refining the left coarse dehazed image and the right coarse dehazed image respectively.

Stereo Feature Extractor: For the stereo feature extractor, we combine the left feature F_l, the right feature F_r out from the WSCDN, the original stereo foggy images I_l and I_r, and the predicted coarse dehazed stereo images $\hat{J}_l^c$ and $\hat{J}_r^c$ as the input S_{in}, which is formulated as:

$$S_{in} = Concat(F_l, (\hat{J}_l^c - I_l), F_r, (\hat{J}_r^c - I_r)), \tag{5}$$

Specially, we input the residues between the coarse dehazed stereo image pair and the input foggy stereo image pair to add the cues of the already detected haze. The stereo feature extractor has the similar structure as the WSCDN. It contains three Basic Block-MaxPooling and three bilinear interpolation-Basic Block, in which skip connection are applied with features across scales (s = 2, 4, 8) corresponding to the same dimension. The extractor only downsamples the input with stride 8 to keep more detail information compared with the WSCDN. Through the stereo feature extractor, the mixed feature F_m is obtained and includes the information from cross views.

Guided Channel and Spatial Refinement Module: If the mixed feature F_m output from the stereo feature extractor is directly utilized to predict the residues for the coarse dehazed stereo pair, some confusing information would be introduced. Therefore, we design a guided channel and spatial refinement (GCSR) module to guide the network to learn the respective residue for the corresponding view. As shown in Fig. 2(b), our GCSR module consists of two steps: guided channel refinement and guided spatial refinement.

The guided channel refinement is similar with the self gated mechanism in the basic block. The difference is that the guided

channel refinement learns the channel weights G_{cr}^{left} for the mixed feature by learning from the left feature F_{cl} instead of learning from itself. The left feature F_{res_l} is learned from the coarse residues $(\hat{J}_l^c - I_l)$ through a 3×3 convolution. The detailed process for predicting the residue of left view is given below:

$$P_c^{left} = Concat(\mathcal{P}_{max}(F_{res\_l}), \mathcal{P}_{avg}(F_{res\_l})), \quad (6)$$

$$\mathcal{G}_{rc}^{left} = \sigma(\mathcal{C}_{1\times1}(\delta((\mathcal{C}_{1\times1}(P_c))))), \quad (7)$$

$$F_{cr}^{left} = \mathcal{G}_{cr}^{left} \odot F_m. \quad (8)$$

where F_{cr}^{left} denotes the left feature after guided channel refinement by the left feature. Analogously, we replace the left feature F_{res_l} by the right feature F_{res_r} learned from the coarse residues $(\hat{J}_r^c - I_r)$ to guide channel refinement and obtain the refined right feature F_{cr}^{right}.

As for the guided spatial refinement, we use a 3×3 convolutional layer to learn the spatial offsets Δp_k^{left} from the left feature F_l in terms of the left view. As the kernel offsets in the deformable convolution operator, Δp_k^{left} augments the regular sampling grid G at position p_0 obtaining a refined feature F_{sr}^{left}, as follows:

$$F_{sr}^{left}(p_0) = \sum\nolimits_{p_k \epsilon G} w_r \cdot F_{cr}^{left}(p_0 + p_k + \Delta p_k^{left}). \quad (9)$$

where F_{cr}^{left} is the input feature to be sampled, and G is a regular grid (*i.e.*If the kernel is 3×3 with dilation 1, $G = (-1,-1),(-1,0),...,(0,1),(1,1)$) sampling the input features. p_k is a position of G, whose corresponding convolutional weight is w_r. Finally, we apply another 3×3 convolutional layer on the spatial refined feature F_{cr}^{left} to predict the left residue $\hat{R}_l$. The process of learning the residue $\hat{R}_r$ for the right view is the analogous process. The final dehazed stereo image pair $\hat{J}_l$ and $\hat{J}_r$ are generated as follows:

$$\hat{J}_l = \hat{J}_l^c + \hat{R}_l, \hat{J}_r = \hat{J}_r^c + \hat{R}_r, \quad (10)$$

D. Loss Function

Our SRDNet is trained by adopting the smooth L1 loss and the perceptual loss [49]. The total loss L is defined as:

$$L = L_{coarse}^{left} + L_{coarse}^{right} + L_{residue}^{left} + L_{residue}^{right}, \quad (11)$$

$$L_{coarse} = L_S(\hat{J}_c, J) + L_P(\hat{J}_c, J), \quad (12)$$

$$L_{residue} = L_S(\hat{R}, J - \hat{J}_c) + L_P(\hat{R}, J - \hat{J}_c). \quad (13)$$

where L_S and L_P are the smooth L1 loss and the perceptual loss respectively. $\hat{J}_c$ and J are the predicted coarse dehazed image and the ground truth respectively. $\hat{R}$ is the predicted residue for the corresponding coarse dehazed image.

IV. Experiments

It is a great challenge to collect a large-scale foggy stereo dataset including real-world foggy stereo images and their clear counterparts for learning-based stereo dehazing methods. To address this problem, Pang *et al*. [7] extended the Foggy Cityscapes dataset to a Stereo Foggy Cityscapes dataset with 8,925 stereo foggy image pairs in the training set and 1500 stereo foggy image pairs in the validation set. The dataset is produced by setting the global atmosphere light ranging from 0.7 to 1.0 and the scatter parameter $\beta \epsilon [0.005, 0.01, 0.02]$. In this work, we utilize the synthetic Stereo Foggy Cityscapes training set to train the model and then utilize the validation set to test our method.

TABLE I
STATE-OF-THE-ART DEHAZING METHODS COMPARISON ON THE STEREO FOGGY CITYSCAPES VALIDATION SET. THE SYMBOL "*" MEANS THAT WE TRAIN THE METHODS ON THE STEREO FOGGY CITYSCAPES TRAINING SET. THE REST RESULTS ARE REPORTED IN [7].

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
SSMDN [8]	22.37	0.9120	-	-
GCANet* [14]	27.66	0.9534	-	-
MSBDN* [15]	24.73	0.9395	-	-
BidNet [7]	25.57	0.9438	25.67	0.9451
SRDNet(ours)	**30.27**	**0.9704**	**30.11**	**0.9699**

A. Training Details

We train the SRDNet on Pytorch with the size 256×256 and augment the training with randomly vertical flip. We set the training batch size as 8 and the total number of epochs as 30. We use Adam optimizer [50], where β_1 and β_2 are set as the default values: 0.9 and 0.999 respectively. We employ the cosine annealing strategy [51] to adjust the learning rate from the initial value 1×10^{-3} to 0. The cosine function is formulated as:

$$l_t = \frac{1}{2}(1 + cos(\frac{t\pi}{T}))l_0 \quad (14)$$

where the total number of batches is T. l_0 and l_t are the initial learning rate and the learning rate at the batch t respectively. The training is carried on 2 TitanX GPUs and only one GPU is used for testing.

B. Comparison with State-of-the-art Methods

The proposed network is tested on the synthetic Stereo Foggy Cityscapes validation set for qualitative and quantitative comparisons with the state-of-the-arts that include SSMDN [8], GCANet [14], MSBDN [15] and BidNet [7]. We exploit the metrics of PSNR and SSIM [52] to evaluate the performance of restored images. Besides, we compare parameters and FLOPS for different methods. For fair comparisons, we re-train GCANet and MSBDN according to their provided training details in their papers on the same Stereo Foggy Cityscapes training set and evaluate them on the same Stereo Foggy Cityscapes validation set as ours. It is worthy noting that we test all methods with the image size of 1024×512.

Quantitative Results: Tab. I shows the quantitative comparison on the Stereo Foggy Cityscapes validation set between our SRDNet with SSMDN [8], GCANet [14], MSBDN [15] and BidNet [7] in terms of the PSNR and the SSIM. The single image dehazing methods only restore the left images. The stereo image dehazing methods: BidNet and our SRDNet obtain dehazed the left images and dehazed the right images simultaneously. It can be found that our proposed SRDNet surpasses all four different state-of-the-art methods by a wide

Fig. 3. Qualitative comparisons on the Stereo Foggy Cityscapes validation set. We present the dehazing results of left images.

margin. Our method is 2.61 dB and 0.017 better than the second-best GCANet [14] in terms of PSNR and SSIM values respectively for the left images. Our SRDNet outperforms BidNet with significant gains of 4.70 dB and 4.44 dB for the metric of PSNR for the left view and the right view respectively. We also add the comparison of parameters and FLOPS for different methods in Tab. II. For the fair and clear comparison, the flops of monocular dehazing methods are doubled ($\times 2$) because they are applied in the left view and the right view separately in stereo foggy scenes. From Tab. II, our SRDNet achieves a better trade-off between the performance and the computational cost when comparing with the methods: SSMDN, GCANet, MSBDN, and BidNet.

Qualitative Results: Fig. 3 shows qualitative state-of-the-arts [14], [15], [7] comparison with the presented SRDNet on the Stereo Foggy Cityscapes validation set. Fig. 3 only shows eight examples which consists of the left foggy images, the left haze-free images dehazed by existing image dehazing methods and our proposed SRDNet, and the ground truth images. Four upper rows are examples with thin fog and the rest examples

Fig. 4. Stereo images haze removal examples on the Stereo Foggy Cityscapes val Dataset. (a) and (d) are Stereo foggy image pairs. (b) and (e) are Stereo haze-free results of our SRDNet. (c) and (f) are the Ground truth.

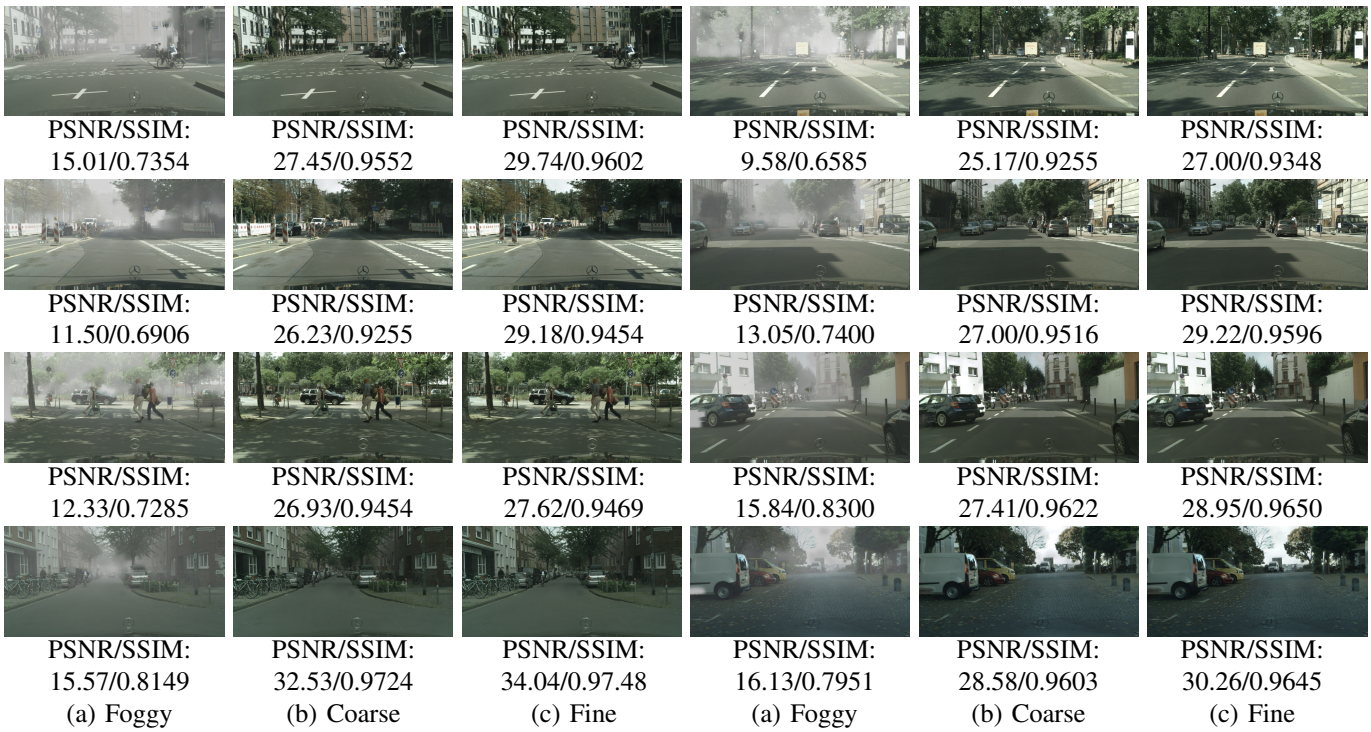


Fig. 5. Subjective comparisons between the coarse dehazed results of the WSCDN and the refined results of the GSRN. We present the dehazed results of left images.

have thick fog. We can observe that the MSBDN can not remove the haze entirely, especially the row-5. The processing power of the BidNet at the sky in the first three rows is unsatisfactory. The GCANet recovers images with excessive brightness relative to ground truth. In addition, the sky in the first row and the building in the fifth row for the GCANet still remains a great amount of haze. In contrast, our method achieves better and visually appealing results. Analogously, the corresponding right images dehazed by our method are also appealing. In addition, we present some haze-free stereo image pairs of our method in Fig. 4.

C. Ablation Study

We conduct the ablation study on the Stereo Foggy Cityscapes validation set. Tab. III shows the impacts of the WSCDN and our GCSR module. Without the GSRN, we use the WSCDN directly to restore the clear stereo pair, the values are reduced by 1.72 dB and 1.77 dB in terms of the PSNR from Tab. III. It demonstrates that only dehazing once is not optimal and using our GSRN could indeed refine the dehazing results. To demonstrate the effectiveness of the GCSR module, we perform an experiment replacing the GCSR module by the 3×3 convolutional layer. As shown in Tab. III, the dehazing results decrease 1.57 dB and 1.78 dB for left dehazed images and right dehazed images from the perspective of PSNR.

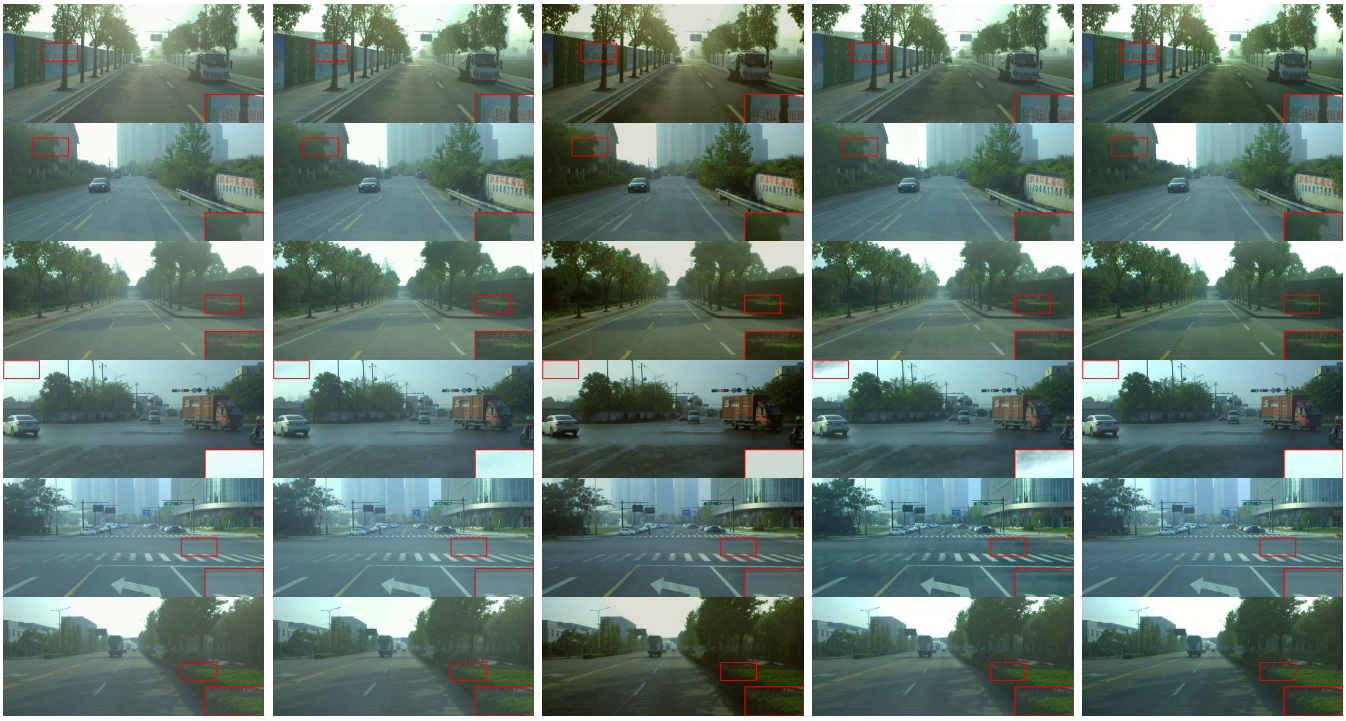


Fig. 6. Evaluation on real foggy stereo images from the Drivingstereo Dataset [53]. We only present the left dehazed images.

TABLE II
PARAMETERS AND FLOPS COMPARISON OF OUR METHOD WITH OTHER METHODS WHEN THE INPUT RESOLUTION IS 512 × 1024. FOR CLEAR COMPARISON, THE FLOPS OF MONOCULAR DEHAZING METHODS ARE DOUBLED (×2) BECAUSE THEY ARE APPLIED IN THE LEFT VIEW AND THE RIGHT VIEW SEPARATELY. WE ALSO PRESENT THE VALUE OF THE PSNR ON THE STEREO FOGGY CITYSCAPES VALIDATION SET IN TERMS OF THE LEFT VIEW.

Method	Params	FlOPS	PSNR
SSMDN [8]	75.16M	82.18G	22.37
GCANet [14]	702.82k	121.76G×2	27.66
MSBDN [15]	31.35M	196.19G×2	24.73
BidNet [7]	323.06k	26.86G	25.57
SRDNet(ours)	759.56k	63.47G	**30.27**

TABLE III
THE ABLATION STUDY ON THE STEREO FOGGY CITYSCAPES VALIDATION SET.

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
WSCDN w/o GSRN	28.55	0.9648	28.34	0.9648
Convolution	28.70	0.9681	28.33	0.9681
GCSR module	**30.27**	**0.9704**	**30.11**	**0.9699**

The values of SSIM also reduce more than 0.0018 compared with employing the GCSR module, which shows that the concatenated stereo features contains confusing information and our GCSR module could separate the useful information belong to the left image and the information belong to the right image. We add some ablation experiments to explore the effects of the max-pooling and the average-pooling in the Basic block of WSCDN in Tab. IV. From Tab. IV, it shows that combing the global average pooling (GAP) and the global max pooling (GMP) could extract more abundant information, which gathers global statistic information and discriminative information, respectively.

TABLE IV
THE ABLATION STUDY ON THE STEREO FOGGY CITYSCAPES VALIDATION SET, WHICH EXPLORING THE EFFECTS OF THE GLOBAL AVERAGE POOLING (GAP) AND THE GLOBAL MAX POOLING (GMP) IN THE BASIC BLOCK.

Basic Block	Left		Right	
	PSNR	SSIM	PSNR	SSIM
GAP	29.76	0.9696	29.54	0.9695
GMP	29.65	0.9693	29.60	0.9697
GAP and GMP	**30.27**	**0.9704**	**30.11**	**0.9699**

TABLE V
PSNR AND SSIM COMPARISONS ON FOGGY SYNTHETIC DRIVINGSTEREO DATASET.

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
BidNet [7]	22.96	0.8765	23.06	0.8785
MSBDN [15]	24.75	0.8949	-	-
GCANet [14]	27.41	0.8962	-	-
SRDNet(ours)	**28.01**	**0.9017**	**28.13**	**0.9065**

We add some subjective comparisons between the dehazing results of WSCDN and the results of the GSRN in Fig. 5 of Section IV-C, which demonstrates that the GSRN indeed refines the dehazing results.

D. Real-world Hazy Images

Moreover, for evaluations on real-world images, we use the stereo foggy images from Drivingstereo dataset [53]. The Drivingstereo dataset is a large-scale dataset including stereo

TABLE VI
DETECTION PERFORMANCE (AP) COMPARISONS ($IoU > 0.7$) ON THE STEREO FOGGY KITTI VALIDATION SET AS FOR CAR WITH ALL FIVE SETTINGS: HEAVY + S AND HEAVY + SR ARE SHORT FOR HEAVY + STEREO R-CNN AND HEAVY + SRDNET FOLLOWED BY STEREO R-CNN, RESPECTIVELY; SIMILARLY FOR THE OTHER GROUPS. "GROUND-TRUTH" IS REPRESENTING THE DETECTION RESULTS OF STEREO R-CNN ON THE CLEAR VALIDATION SET.

metric	Setting	Heavy + S	Heavy + SR	Medium + S	Medium + SR	Light + S	Light + SR	Goundtruth
AP_{3d}	AP_E	24.14	**40.37**	34.26	**44.97**	40.75	**47.06**	54.91
	AP_M	14.89	**27.16**	21.49	**31.13**	25.79	**32.37**	37.72
	AP_H	13.84	**21.82**	19.83	**25.97**	23.75	**26.90**	32.17
AP_{bv}	AP_E	28.89	**50.30**	43.61	**56.58**	51.58	**59.06**	68.16
	AP_M	21.74	**34.97**	28.99	**36.32**	34.45	**42.24**	48.98
	AP_H	15.85	**29.13**	22.81	**34.53**	28.07	**35.73**	48.16

TABLE VII
3D DETECTION PERFORMANCE COMPARISONS ON THE KITTI VALIDATION SET FOR CAR WITH ALL THREE SETTINGS: LIGHT HAZE, MEDIUM HAZE AND HEAVY HAZE. SPECIALLY, AP_E, AP_M AND AP_L ARE USED TO EVALUATE THE PERFORMANCE OF EASY, MODERATE AND HARD SETS RESPECTIVELY.

Haze	AP_{3D}	Hazy	BidNet [7]	SRDNet
Light	AP_E	40.75	46.34	**47.06**
	AP_M	25.79	31.17	**32.37**
	AP_H	23.75	25.59	**26.90**
Medium	AP_E	34.26	44.0	**44.97**
	AP_M	21.49	27.2	**31.13**
	AP_H	19.83	25.3	**25.97**
Heavy	AP_E	24.14	40.0	**40.37**
	AP_M	14.89	26.35	**27.16**
	AP_H	13.84	21.5	**21.82**

image pairs in real autonomous driving scenarios, in which 2000 frames with 4 different weathers (sunny, cloudy, foggy, rainy) are selected for specific needs. In terms of the 500 foggy stereo image pairs, their corresponding clear pairs are not available.

We leverage the fog simulation pipeline described in [7] to add fog to the sunny and cloudy sequences in the Drivingstereo dataset, and randomly divide the dataset into the training set and the validation set. We generate the random atmospheric light from 0.7 to 1.0 and set $\beta \in [0.005, 0.01, 0.02]$ for each stereo image pair. We finetune our model, BidNet [7], MSBDN [15] and GCANet [14] pre-trained by the Stereo Foggy Cityscapes training set on the generated foggy Drivingstereo training set containing 2400 stereo pairs and evaluate them on the generated validation set containing 800 foggy stereo pairs. Tab. V compares the dehazing performance of our SRDNet and other methods on the synthetic foggy Drivingstereo validation set. Our SRDNet achieves an obvious gain with 3.36 dB in the PSNR and 0.054 in the SSIM compared with the MSBDN. In terms of the SSIM value, our method outperforms the second best method GCANet by 0.0055.

Fig. 6 gives qualitative comparison of our dehazed results with the state-of-the-arts: MSBDN [15], BidNet [7] and GCANet [14] on the real hazy images from the Drivingstereo dataset. It can be observed that there still exists quite a lot of haze in the results of MSBDN. Color distortion is introduced by BidNet. Compared with GCANet, our method performs better in the regions of the sky of row-4 and the road of row-5. Our method has visually appealing results. The right images dehazed by our method have analogous results.

E. Perceptual Quality for High-level Vision Tasks

As the stereo dehazing algorithms are usually used as the pre-processing step for high-level computer vision tasks such as 3D object detection, the accuracy of 3D object detection can be treated as an indirect indicator of the stereo dehazing quality. We adopt the accuracy of stereo image-based 3D object detection on the KITTI dataset to evaluate the perceptual quality of our dehazing method. KITTI dataset [54] is a challenge benchmark for evaluating the performance of 3D object detection, which is divided into training set and validation set with 3,712 images and 3,769 images respectively. In order to generate foggy stereo images for the KITTI dataset, we first estimate the depth map for each image by a stereo matching method PSMNet [55], and then use the depth map to synthesize foggy stereo images using the fog simulation pipeline described in [7]. This synthetic dataset is referred to as the Stereo Foggy KITTI dataset in this work. We produce the atmospheric light randomly from 0.7 to 1.0 and use $\beta \in [0.02, 0.04, 0.06]$ for each stereo image pair. Hence, there are 11,136 stereo foggy image pairs for training, and 11,307 stereo foggy image pairs for validation. We first train the SRDNet on the Foggy Stereo KITTI training set following Sec. IV-A. On the Stereo Foggy KITTI validation set, our SRDNet improves the PSNR values from 11.89 dB and 10.96 dB to 22.83 dB and 22.14 dB in terms of the left view and the right view respectively. For the metric of the SSIM, our SRDNet boosts 0.2301 and 0.2303 for the left view and the right view respectively. For the 3D detection accuracy, we choose Stereo R-CNN [4] pretrained on the KITTI clear training set to evaluate the dehazed results of our methods in light (β=0.02), medium (β=0.04), and heavy (β=0.06) foggy scenes. Specially, the Stereo R-CNN model uses ResNet101 [56] and FPN [57] as the backbone.

Generally, the metrics of 3D detection and 3D localization performance are Average Precision for 3D box (AP_{3d}) and birds eye view (AP_{bv}). AP_E, AP_M and AP_L are the average precision of easy, moderate and hard sets divided according to the KITTI setting, respectively. Tab. VI compares the 3D detection accuracy (AP_{3d}) only Stereo R-CNN and SRDNet concatenated with Stereo R-CNN in foggy scenes using IoU = 0.7 on the Stereo Foggy KITTI validation set, which proves that our SRDNet as the pre-process for the detector can stably boost the accuracy in the conditions of light, medium, and heavy haze. Specifically, the heavy haze degrades AP_{3d} by 30.77%, 22.83% and 18.33% across the easy, moderate and

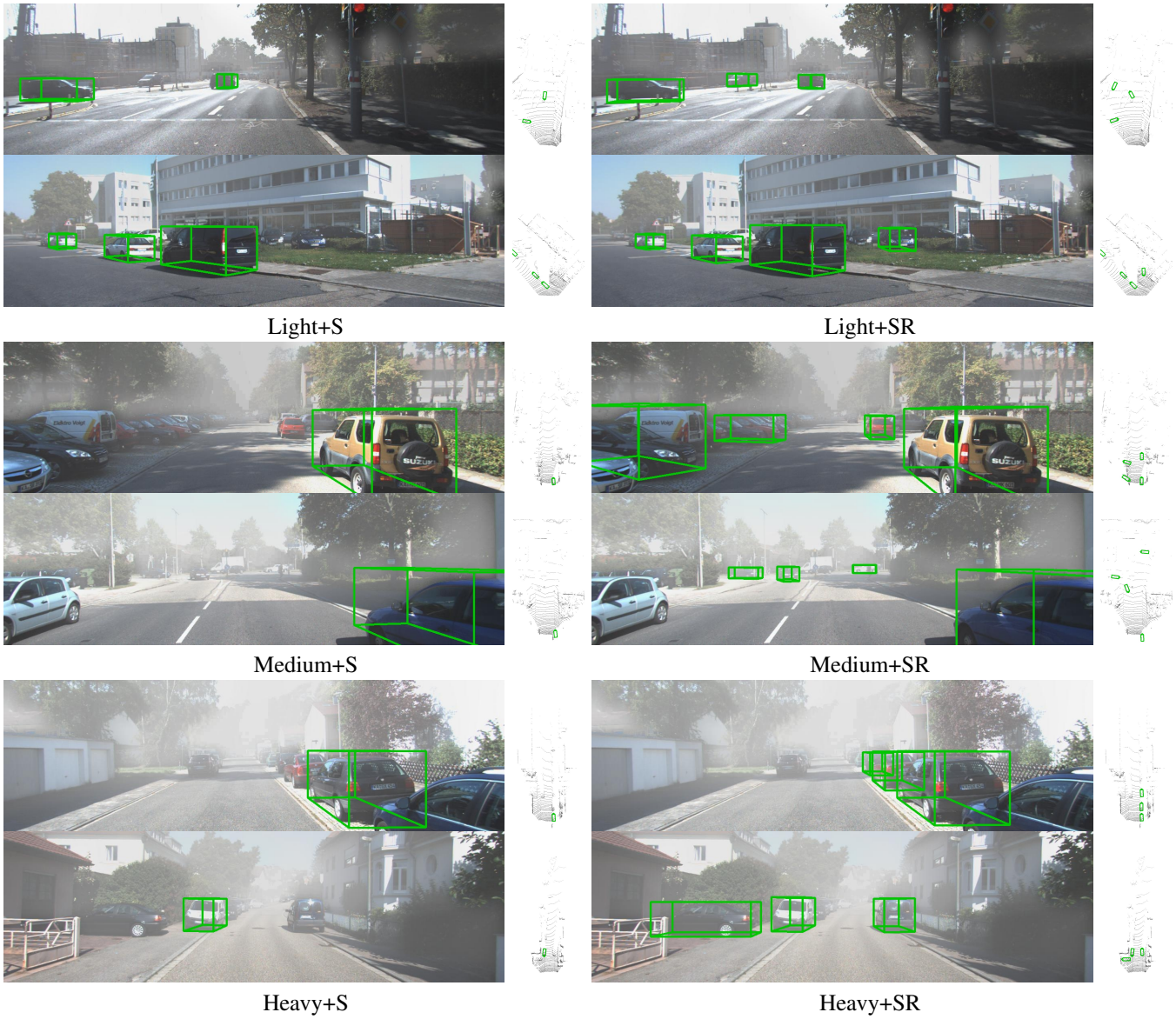


Fig. 7. Qualitative 3D detection results of Stereo R-CNN [4] on the Stereo Foggy KITTI Dataset [54]. Heavy + S and Heavy + SR are short for Heavy + Stereo R-CNN and Heavy + SRDNet followed by Stereo R-CNN, respectively; similarly for the other groups.

hard sets. By appending the SRDNet, the AP_E, AP_M and AP_L improve by 16.23%, 12.27% and 7.98% respectively in the heavy hazy circumstance. Further, as shown in Tab. VI, the haze degrades the 3D localization accuracy (AP_{bv}) of the Stereo R-CNN. After concatenating our SRDNet, the AP_{bv} for birds eye view obtains a notable absolute gain, demonstrating the high perceptual quality of our stereo dehazing method. We compare our SRDNet and BidNet in 3D detection task, which both belong to the binocular dehazing methods. In terms of each metric of AP_{3D} in the light hazy scenes, the medium hazy scenes, and the heavy hazy scenes, Tab. VII shows that our method obtains higher accuracy and has better perceptual quality.

Fig. 7 gives some stereo detection visual results of Stereo R-CNN in the conditions of light, medium, and heavy haze. The birds eye view images projected from the 3D box are also presented. When the haze gets heavier, there are more objects that are missed by the Stereo R-CNN. After appending the SRDNet, the missed objects are correctly detected and located. Our SRDNet is flexible and can pre-process the foggy stereo inputs for up-to-date stereo based 3D object detectors, which eliminates the degradation of the foggy inputs.

V. Conclusion

In this paper, we have proposed a stereo refinement dehazing network directly restoring the haze-free stereo image pair without disparity estimation in a coarse-to-fine manner. It is composed of two parts : a weight-sharing coarse dehazing network restoring a coarse dehazed image pair; and a guided separated refinement network designed to predict the residues for different views. The guided separated refinement network fuses information of cross views and separates the features of different views through a guided channel and spatial refinement module. The residues are added to the coarse dehazed

pair in order to make refinement and remove the remained haze. Extensive evaluations demonstrate the superiority of the proposed network against state-of-the-art methods on the synthetic dataset. In addition, our SRDNet generalizes well for the real stereo foggy scenes. Furthermore, our SRDNet as the pre-process of the stereo image-based 3D object detection can boost its accuracy in the hazy scenes.

REFERENCES

[1] Z. Lu, J. Wang, Z. Li, S. Chen, and F. Wu, "A resource-efficient pipelined architecture for real-time semi-global stereo matching," *IEEE Trans. Circuits and Systems for Video Technology*, 2021.

[2] C. Xu, C. Wu, D. Qu, F. Xu, H. Sun, and J. Song, "Accurate and efficient stereo matching by log-angle and pyramid-tree," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[3] L. Li, S. Zhang, X. Yu, and L. Zhang, "Pmsc: Patchmatch-based superpixel cut for accurate stereo matching," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 679–692, 2018.

[4] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

[5] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2020.

[6] Z. Qin, J. Wang, and Y. Lu, "Triangulation learning network: From monocular to stereo 3d object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2019.

[7] Y. Pang, J. Nie, J. Xie, J. Han, and X. Li, "Bidnet: Binocular image dehazing without explicit disparity estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2020.

[8] T. Song, Y. Kim, C. Oh, and K. Sohn, "Deep network for simultaneous stereo matching and dehazing," in *British Machine Vision Conference*, 2018.

[9] Z. Li, P. Tan, R. T. Tan, D. Zou, Steven Zhiying Zhou, and L. Cheong, "Simultaneous video defogging and stereo reconstruction," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[10] J.-Y. Na and K.-J. Yoon, "Stereo vision aided image dehazing using deep neural network," in *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*, 10 2018, pp. 15–19.

[11] A. Yamashita, Y. Tanaka, and T. Kaneko, "Removal of adherent waterdrops from images acquired with stereo camera," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 400–405.

[12] J. Kim, J. Sim, and C. Kim, "Stereo video deraining and desnowing based on spatiotemporal frame warping," in *Proc. IEEE International Conf. Image Processing*, 2014, pp. 5432–5436.

[13] Z. Kaihao, L. Wenhan, R. Wenqi, W. Jingwen, Z. Fang, M. Lin, and L. Hongdong, "Beyond monocular deraining: Stereo image deraining via semantic understanding," in *Proc. European Conference on Computer Vision*, 2020.

[14] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1375–1383.

[15] D. Hang, P. Jinshan, H. Zhe, L. Xiang, W. Fei, and Y. Ming-Hsuan, "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2020.

[16] E. J. McCartney, "Optics of the atmosphere: scattering by molecules and particles," *New York, John Wiley and Sons, Inc., 1976. 421 p.*, 1976.

[17] D. Zhao, L. Xu, L. Ma, J. Li, and Y. Yan, "Pyramid global context network for image dehazing," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[18] Y. Zhang, P. Wang, Q. Fan, F. Bao, X. Yao, and C. Zhang, "Single image numerical iterative dehazing method based on local physical features," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3544–3557, 2020.

[19] X. Zhang, T. Wang, W. Luo, and P. Huang, "Multi-level fusion and attention-guided cnn for image dehazing," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[20] J. Cao, Y. Pang, S. Zhao, and X. Li, "High-level semantic networks for multi-scale object detection," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3372–3386, 2020.

[21] J. Nie, Y. Pang, S. Zhao, J. Han, and X. Li, "Efficient selective context network for accurate object detection," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[22] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Enriched feature guided refinement network for object detection," in *Proc. IEEE Conf. Computer Vision*, 2019.

[23] N. Han, J. Wu, X. Fang, W. K. Wong, Y. Xu, J. Yang, and X. Li, "Double relaxed regression for image classification," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 307–319, 2020.

[24] M. Jubran, A. Abbas, A. Chadha, and Y. Andreopoulos, "Rate-accuracy trade-off in video classification with deep convolutional neural networks," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 145–154, 2020.

[25] Z. Sun, J. Chen, C. Liang, W. Ruan, and M. Mukherjee, "A survey of multiple pedestrian tracking based on tracking-by-detection framework," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[26] B. Ramesh, S. Zhang, H. Yang, A. Ussa, M. Ong, G. Orchard, and C. Xiang, "e-tld: Event-based framework for dynamic object tracking," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[27] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[28] S. Lian, W. Jiang, and H. Hu, "Attention-aligned network for person re-identification," *IEEE Trans. Circuits and Systems for Video Technology*, 2020.

[29] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," vol. 33, no. 12, pp. 2341–2353, 2011.

[30] R. Fattal, "Single image dehazing," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 72–92, 2008.

[31] ——, "Dehazing using color-lines," *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 13–31, 2014.

[32] D. Berman, S. Avidan *et al.*, "Non-local image dehazing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 1674–1682.

[33] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[34] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Trans. Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[35] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. European Conference on Computer Vision*, 2016, pp. 154–169.

[36] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 3194–3203.

[37] Y. Liu, J. Pan, J. Ren, and Z. Su, "Learning deep priors for image dehazing," in *Proc. IEEE Conf. Computer Vision*, 2019, pp. 2492–2500.

[38] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 3253–3261.

[39] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019, pp. 8160–8168.

[40] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE Conf. Computer Vision*, 2019, pp. 7314–7323.

[41] Y. B. X. X. Xu Qin, Zhilin Wang and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *AAAI*, 2020.

[42] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," in *British Machine Vision Conference*, 2017.

[43] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2018.

[44] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2020.

[45] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[46] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

[47] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.

[48] T. Song, Y. Kim, C. Oh, H. Jang, N. Ha, and K. Sohn, "Simultaneous deep stereo matching and dehazing with feature attention," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 799–817, 2020.
[49] J. Justin, A. Alexandre, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. European Conference on Computer Vision*, 2016.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[51] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[53] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019.
[54] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.
[55] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[57] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.