BidNet: Binocular Image Dehazing without Explicit Disparity Estimation

Yanwei Pang[1], Jing Nie[1], Jin Xie[1], Jungong Han[2]*, Xuelong Li[3]
[1]Tianjin Key Laboratory of Brain-inspired Intelligence Technology,
School of Electrical and Information Engineering, Tianjin University, China
[2]University of Warwick, UK [3]Northwestern Polytechnical University, China
[1]{pyw,jingnie,jinxie}@tju.edu.cn, [2]jungong.han@warwick.ac.uk, [3]li@nwpu.edu.cn

Abstract

Heavy haze results in severe image degradation and thus hampers the performance of visual perception, object detection, etc. On the assumption that dehazed binocular images are superior to the hazy ones for stereo vision tasks such as 3D object detection and according to the fact that image haze is a function of depth, this paper proposes a Binocular image dehazing Network (BidNet) aiming at dehazing both the left and right images of binocular images within the deep learning framework. Existing binocular dehazing methods rely on simultaneously dehazing and estimating disparity, whereas BidNet does not need to explicitly perform time-consuming and well-known challenging disparity estimation. Note that a small error in disparity gives rise to a large variation in depth and in estimation of haze-free image. The relationship and correlation between binocular images are explored and encoded by the proposed Stereo Transformation Module (STM). Jointly dehazing binocular image pairs is mutually beneficial, which is better than only dehazing left images. We extend the Foggy Cityscapes dataset to a Stereo Foggy Cityscapes dataset with binocular foggy image pairs. Experimental results demonstrate that BidNet significantly outperforms state-of-the-art dehazing methods in both subjective and objective assessments.

1. Introduction

Haze is an important factor for degrading image quality and decreasing the performance of computer vision tasks such as object detection [23, 25, 2, 14] and semantic image segmentation [24, 19, 43]. Therefore, image dehazing plays an important role in developing effective computer vision systems. In the dehazing literature [20, 22], the hazing process is usually modeled as an atmosphere scattering model,

$$I(x) = J(x)t(x) + A(1 - t(x)), \quad (1)$$

*Corresponding author

Figure 1. Sample image dehazing results using the proposed BidNet. Top-left: Input left foggy image. Bottom-left: Input right foggy image. Top-right: Dehazed left image. Bottom-right: Dehazed right image.

where $I(x)$ denotes the intensity of pixel x in the hazy image, $J(x)$ is the clear image, $t(x)$ represents the transmission map, and A denotes the global atmospheric light intensity; moreover, there is $t(x) = e^{-\beta d(x)}$ with β and $d(x)$ being the atmosphere scattering parameter and the distance between the camera and the scene, respectively.

According to Eq. 1, image haze is a function of depth. The correlation of the binocular images could help predict the depth [41], which demonstrates binocular images are beneficial for the dehazing task. To overcome the binocular image degradation caused by haze, directly and separately applying single image dehazing methods [29] on left foggy image and right foggy image could not obtain satisfying results, especially for heavy haze, because this kind of methods make no use of the correlation of the binocular images. It is expected that binocular image dehazing will facilitate image-based 3D applications, such as 3D object detection [13, 27].

Existing binocular image dehazing methods [34, 21] rely on simultaneously performing dehazing and disparity estimation. These methods are insightful for developing new binocular image dehazing methods. Nevertheless, this kind of methods has three drawbacks: (1) It is well known that for a given small error in disparity, the error in depth increases with disparity [40]. Because it is required for image

dehazing to estimate transmission maps and the transmission map is an exponential function of depth, the error in disparity also leads to large error in estimating transmission maps and hamper haze-free images. (2) State-of-the-art deep learning based disparity estimation methods are time-consuming because they have to construct a 4D cost volume and then apply 3D convolutions. (3) It only outputs left dehazed images instead of binocular dehazed image pairs. In this paper, we propose a novel deep learning based Binocular image dehazing Network (BidNet), which is capable of utilizing the collaborative information contained in the left and right images without explicitly performing the time-consuming and challenging disparity estimation.

There is no specific dataset containing binocular foggy images for deep learning based binocular image dehazing. Marius *et al*. leverage their fog simulation pipeline to create a Foggy Cityscapes dataset [32] by adding fog to urban scenes from the Cityscapes dataset [4]. We extend the Foggy Cityscapes dataset to a Stereo Foggy Cityscapes dataset, which consists of binocular foggy image pairs. The key point is to utilize the disparity and the given camera parameters to compute the distance between the camera and the left scene, and the distance between the camera and the right scene. In this process, we apply the complete pipeline [32] which adds synthetic fog to real, clear-weather images using incomplete depth information.

The novelties and contributions of the paper are summarized as follows:

(1) A novel framework, termed BidNet, of binocular image dehazing is proposed which is capable of utilizing correlation between left and right images to dehaze binocular image pairs without estimating disparity. It can avoid the large error caused by imprecise disparity estimation.

(2) Inspired by non-local networks [38], a simple yet effective mechanism is proposed and embedded in the BidNet to introduce useful information in the feature maps of right images into the feature maps of left images. It is implemented by computing a stereo horizontal non-local correlation matrix and multiplying the non-local correlation matrix with the feature maps of the right image. The process of embedding is efficient because the size of the correlation matrix is one-order less than that of traditional non-local networks. Analogously, the useful information of feature maps of the left image can be embedded to those of the right one.

(3) Given the input of the left and right images, one can only dehaze either left image or right image using the above dehazing framework. But we find that simultaneously dehazes left and right hazy images can produce better dehazing results by taking into account both left and right images for formulating the dehazing loss function.

(4) A Stereo Foggy Cityscapes dataset is developed by extending from the Foggy Cityscapes dataset. Experimental results show that the proposed BidNet significantly outperforms the state-of-the-art dehazing methods in terms of both subjective and objective assessment. In addition, our BidNet generalizes and performs well for the real stereo foggy scenes. It is expected that more accurate 3D information can be obtained from the dehazed binocular images.

2. Related work

In this section, we briefly review several major works for single image dehazing and stereo image tasks.

2.1. Single Image Dehazing

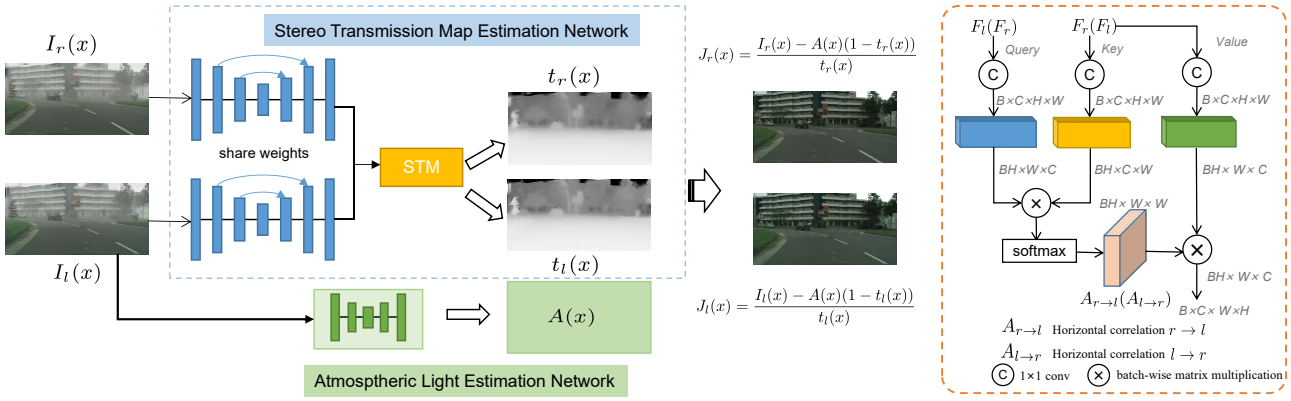
Existing dehazing methods mainly are classified to two categories: hand-crafted prior based dehazing methods and deep learning based dehazing methods.

Hand-crafted prior based dehazing Dehazing methods involves the estimation of the atmospheric light, the transmission map and the haze-free image. Early dehazing methods [35, 6, 5, 44] employed hand-crafted priors based on the statistics of clean images to estimate the transmission map, then used the atmospheric scattering model to recover the haze-free results. Tan *et al*. [35] enhanced the visibility of hazy images through local max contrast. He *et al*. [6] proposed the dark channel prior (DCP) to compute the transmission map. The color-line prior [5] is introduced due to the discovery that pixels of image patches typically exhibit a one-dimensional distribution. The color attenuation prior is adopted in [44] for the development of a supervised learning method for image dehazing.

Deep learning based dehazing With the development of CNNs, deep learning based dehazing methods have been made remarkable progress. Deep learning based dehazing methods could be roughly divided into two categories: model-based dehazing methods and model-free dehazing methods. The model-based dehazing methods [29, 1, 42, 26, 11] are based on the atmospheric scattering model. These methods utilize CNNs to estimate a transmission map, followed by estimation of atmospheric light through traditional methods or CNNs. Finally, the haze-free image is obtained as :

$$J(x) = \frac{I(x) - A(x)(1 - t(x))}{t(x)}. \tag{2}$$

MSCNN [29] first uses a coarse-scale network to predict a holistic transmission map based on the entire image and then employ a fine-scale network to refine it locally. Zhang *et al*. [42] developed a densely connected pyramid dehazing network to jointly learn the transmission map, the atmospheric light and haze-free images for capturing their relations. HRGAN [26] introduces a generative adversarial network for visual haze removal. AOD-Net [11] introduces a reformulation of Eq. 1 to bypass the estimation of the transmission map and the atmospheric light intensity.



(a) The overall architecture of BidNet

(b) Stereo Transformation Module (STM)

Figure 2. (a) Overall architecture of our Binocular image dehazing Network (BidNet). BidNet inputs the binocular foggy image pair and outputs the haze-free binocular image pair. (b) The structure of the Stereo Transformation Module (STM). STM is proposed to explore and encode the relationship between the binocular image pair.

Recently, end-to-end CNNs have been designed to directly learn the clean image from a hazy input for dehazing without replying on the atmospheric scattering model [30, 28, 17]. Gated Fusion Network [30] builds on the principle of image fusion, and is learned to produce the sharp image directly. GridDehazeNet [17] is an end-to-end trainable CNNs consisting of three modules: pre-processing, backbone, and post-processing for single image dehazing.

2.2. Stereo Image Tasks

Stereo matching Stereo matching is reconstructing the scene in 3D. Stereo matching is decomposed into three important steps: feature extraction, matching cost aggregation and disparity prediction [41]. Cost Volume is widely applied in stereo matching [3, 9, 16] to capture long-range dependency in stereo images. Cost Volume is obtained by concatenate left feature maps with their corresponding right feature maps across all disparities to obtain a 4D cost volume. To achieve higher efficiency, other two methods [18, 7] use the inner product between the two representations to compute the matching score.

Stereo image super-resolution Super-resolution aims to reconstruct high-resolution images from their low-resolution counterparts. Wang *et al*. [37] proposed a parallax-attention stereo super-resolution network to incorporate the information from a stereo image pair. Motivated by this, we propose a stereo transformation module to integrate the information from the binocular foggy image pairs.

Stereo vision aided dehazing Recently, using binocular images in dehazing methods has been proposed [21, 15, 34]. These methods attempt to combine the tasks of stereo matching and image dehazing. The method [15] jointly estimates scene depth and recover the clear latent image from a foggy video sequence. Song *et al*. [34] proposed a multi-task network simultaneously estimating a clear image and disparity from a stereo hazy image pair, which demonstrates that stereo matching and dehazing can be synergistically formulated by incorporating depth information from transmission maps into the stereo matching process, and vice versa. These dehazing methods input the stereo image hazy pairs but only estimates the left haze-free images.

3. Method

In this section, we describe the proposed Binocular image dehazing Network (BidNet), which inputs binocular foggy image pair and estimates the transmission maps, the atmospheric light, and simultaneously dehazes the binocular image pairs. The architecture of the BidNet is illustrated in Fig. 2(a). A Stereo Transformation Module (STM) is introduced to explore and encode the correlation between binocular images. BidNet does not need to explicitly perform time-consuming and well-known challenging disparity estimation.

Next, we would introduce a Stereo Transmission Map Estimation Network (STMENet) (Sec. 3.1), a Atmospheric Light Estimation Network (ALENet) (Sec. 3.2), dehazing via the physical scattering model (Sec. 3.3) as well as the loss function (Sec. 3.4).

3.1. Stereo Transmission Map Estimation Network

The Stereo Transmission Map Estimation Network (STMENet) could be divided into three parts: weight-shared feature extraction module, Stereo Transformation module (STM), and refinement module.

Weight-Shared Feature Extraction Module As shown in Fig. 2(a), the shared feature extraction module is a encoder-decoder structure. Tab. 1 shows the detailed structure. The left image and the right image respectively input the weight-shared feature extraction module. The images firstly go

through a pre-processing layer to learn better input features. The learned left (& right) features are passed through four 3×3 convolutional layers with stride 2. The channels of four convolutional layers are increasing as 32, 48, 64, and 96. We then apply four bilinear interpolation followed with 3×3 convolutional layers to the down-sampled features. ReLU and BN are followed by the convolutional layer. Concatenations are then employed with features across scales (s=2, 4, 8) corresponding to the same dimension. Through the bottom-up and top-down structure, the obtained left features (F_l) and right features (F_r) are discriminative.

Stereo Transformation Module (STM) The left features and right features from the weight-shared module only integrate the information of their own. The relationship and correlation between the binocular image pair are not utilized. We design a Stereo Transformation Module (STM) to transform the depth information through learning the horizontal correlation between the left and right features. Fig. 2(b) shows the structure of STM. Because the binocular image pair are aligned in the vertical dimension, the STM only need to learn the horizontal correlation between them. Inspired by the non-local network [38], we compute the response at a position as a weighted sum of the features at all positions along the horizontal dimension, which could capture the long range dependencies that contain disparity (depth) information. The STM has two inputs: $F_l \in \mathbb{R}^{B\times C\times H\times W}$ and $F_r \in \mathbb{R}^{B\times C\times H\times W}$. The convolutional operations with the kernel size 1×1 (W_θ^l, W_ψ^r and W_γ^r) are used to transform F_l and F_r to obtain the embeddings θ_l, ψ_r and γ_r:

$$\theta_l = W_\theta^l(F_l), \psi_r = W_\psi^r(F_r), \gamma_r = W_\gamma^r(F_r), \quad (3)$$

The stereo horizontal correlation matrix $A_{r\to l}$ is computed by the batch-wise multiplication between the reshaped $\theta_l \in \mathbb{R}^{(BH)\times W\times C}$ and the reshaped $\psi_r \in \mathbb{R}^{(BH)\times C\times W}$ followed with the activation of softmax:

$$A_{r\to l} = softmax(\theta_l \times \psi_r), \quad (4)$$

The output ($S_l \in \mathbb{R}^{B\times C\times H\times W}$) of STM for the left transmission map estimation is computed as:

$$S_l = W_o(cat(A_{r\to l} \times \gamma_r, F_l)), \quad (5)$$

where cat means concatenation operation, W_o denotes convolutional layers with the filter size of 1×1 to fuse the information and reduce the channels.

The computation of the stereo horizontal correlation matrix $A_{l\to r}$ and the out ($S_r \in \mathbb{R}^{B\times C\times H\times W}$) are the analogous process, just exchange the place of the two inputs: F_l and F_r. As shown in Tab. 1, S_l and S_r separately pass through a 3×3 convolutional layer to estimate the left transmission map and the right transmission map.

Refinement Module The estimated transmission maps from STM still lack of global structural information. Spatial pyramid pooling is parameter-free and very efficient. We employ spatial pyramid pooling to introduce multi-scale contextual information to refine the transmission maps, which could enhance the robustness. The detailed structure is demonstrated in Tab. 1. We use three average pooling layers with kernel sizes as 3, 7, and 15 and strides as 1. The pooling layers transform the initial estimated transformation maps into a global representation enhanced set. Then, these transformed maps with the initial estimated transformation maps are aggregated through a concatenation and go to a 1×1 convolutional layer to fuse the features. The outputs are the refined transmission maps.

3.2. Atmospheric Light Estimation Network

Atmospheric light Estimation Network (ALENet) aims to estimate atmospheric light A in Eq. 2. As shown in Fig. 2(a), ALENet is also an encoder-decoder structure without skip connection across the feature scales. It consists of a 3×3 convolutional layer as pre-processing, three Conv-BN-Relu-Pool blocks as encoder, three Up-Conv-BN-Relu blocks as decoder, and finally a 3×3 convolutional layer estimating the atmospheric light A shown in Tab. 1. A stereo image pair has the same atmospheric light A. Therefore, the ALENet only inputs the left images for prediction.

3.3. Dehazing via The Physical Scattering Model

As shown in Fig. 2(a), haze-free left images and haze-free right images are computed by Eq. 2. Eq. 2 makes sure the whole network is jointly optimized. The direct computed haze-free binocular images have some noise. We add a image refinement module, which is a light-weight dense block. The light-weight dense block has four 3×3 convolutional layers, whose input is the concatenation of the feature maps produced before in the block. The numbers of input channels are 3, 8, 16, and 24, but the numbers of the output channels are all 8. Finally, a 1×1 convolutional layer is employed for estimating refined haze-free binocular images.

3.4. Losses

The loss function of the BidNet measures the error of the estimated binocular images, transmission maps, and atmospheric light. The errors for both left and right images are taken into account in the loss function so that it is mutually beneficial to simultaneously dehaze both images. Specifically, the loss L_J for haze-free images is defined as

$$L_J = \left\|\hat{J}_l - J_l\right\|_2^2 + \left\|\hat{J}_r - J_r\right\|_2^2 + \left\|\hat{J_{rl}} - J_l\right\|_2^2 + \left\|\hat{J_{rr}} - J_r\right\|_2^2, \quad (6)$$

where $\hat{J}_l$ ($\hat{J}_r$) is the estimated left (right) image. $\hat{J_{rl}}$ ($\hat{J_{rr}}$) is the estimated refined left (right) image. J_l (J_r) is the ground truth left (right) image.

Name	Setting	Input	Output
Stereo Transmission Map Estimation Network			
Weight-Shared Feature Extraction Module			
pre-processing	$\begin{bmatrix}3\times3, 16\\3\times3, 16\end{bmatrix}$	$256\times256\times3$	$256\times256\times16$
ublock1_a	3×3, 32, s=2	$256\times256\times16$	$128\times128\times32$
ublock1_b	3×3, 48, s=2	$128\times128\times32$	$64\times64\times48$
ublock1_c	3×3, 64, s=2	$64\times64\times48$	$32\times32\times64$
ublock1_d	3×3, 96, s=2	$32\times32\times64$	$16\times16\times96$
ublock 2_d	$\begin{bmatrix}upsample, s=2\\3\times3, 64\\\bigoplus\text{ublock1\_c}\end{bmatrix}$	$16\times16\times96$	$32\times32\times64$
ublock 2_c	$\begin{bmatrix}upsample, s=2\\3\times3, 48\\\bigoplus\text{ublock1\_b}\end{bmatrix}$	$32\times32\times64$	$64\times64\times48$
ublock 2_b	$\begin{bmatrix}upsample, s=2\\3\times3, 32\\\bigoplus\text{ublock1\_a}\end{bmatrix}$	$64\times64\times48$	$128\times128\times32$
ublock 2_a	$\begin{bmatrix}upsample, s=2\\3\times3, 16\end{bmatrix}$	$128\times128\times32$	$256\times256\times16$
Stereo Transformation Module			
The STM is detailed in Fig. 2(b).			
pre_layer	3×3, Tanh	$256\times256\times16$	$256\times256\times1$
Refinement Module			
t_refine	$\begin{bmatrix}\text{Avg Pool kernel=3, s=1} & \text{Avg Pool kernel=7, s=1} & \text{Avg Pool kernel=13, s=1}\end{bmatrix}$ 3×3, 1, Sigmoid	$256\times256\times1$	$256\times256\times1$
Atmospheric Map Estimation Network			
pre-processing	3×3	$256\times256\times3$	$256\times256\times16$
ublock down	$\begin{bmatrix}3\times3, 16\\pool, s=2\end{bmatrix}\times3$	$256\times256\times16$	$32\times32\times16$
ublock up	$\begin{bmatrix}upsample, s=2\\3\times3, 16\end{bmatrix}\times3$	$32\times32\times16$	$256\times256\times16$
pre_layer	3×3	$256\times256\times16$	$256\times256\times1$

Table 1. The detailed architecture of our BidNet. If not specifically noted, BN and ReLU are followed by the convolutional layers. Except the weight-shared feature extraction module, the rest weights in left branch and the right branch are not shared. $\bigoplus$ denotes concatenation and a 3×3 convolutional layer to reduce the channels. Upsample denotes bilinear interpolation.

The loss L_t for transmission maps is defined as

$$L_t = \left\|\hat{t_l} - t_l\right\|_2^2 + \left\|\hat{t_r} - t_r\right\|_2^2 + \left\|\hat{t_{rl}} - t_l\right\|_2^2 + \left\|\hat{t_{rr}} - t_r\right\|_2^2, \tag{7}$$

where $\hat{t_l}$ ($\hat{t_r}$), $\hat{t_{rl}}$ ($\hat{t_{rr}}$), and t_l (t_r) are the estimated left (right) transmission map, the estimated refined left (right) transmission map, and the ground truth left (right) transmission map respectively.

The loss L_a for the atmospheric light is defined as

$$L_a = \left\|\hat{A} - A\right\|_2^2, \tag{8}$$

where $\hat{A}$ is the estimated atmospheric light, A is the ground truth atmospheric light.

Perceptual loss based on high-level features extracted from pretrained network is wildly used in image super-resolution [8]. In addition, perceptual losses measure image visual similarities more effectively than pixel-wise loss. Inspired by this, we introduce a perceptual loss to increase perceptual similarities between restored haze-free images and realistic images. The perceptual loss leverages multi-scale features extracted from a pre-trained deep neural network to quantify the visual difference between the estimated image and the ground truth. In our methods, we use the VGG16 [33] pre-trained on ImageNet [31] as the loss network and extract the features from Conv3_3 in the VGG16. The perceptual loss is defined as:

$$L_P = \frac{1}{C_f H_f W_f} \sum_{c=1}^{C_f} \sum_{h=1}^{H_f} \sum_{w=1}^{W_f} ||\phi_{c,w,h}(\hat{J}) - \phi_{c,w,h}(J)||_2^2, \tag{9}$$

where C_f, H_f and W_f specify the dimension of the respective feature maps within the VGG-16 network. $\hat{J}$ denotes the predicted left (& right) images and J represents the clear left (& right) images. The effect of ϕ is to obtain the feature maps from VGG16.

The total loss is defined by combining the following four loss functions:

$$L = L_t + L_a + L_J + 0.04\ \times L_p, \tag{10}$$

where L_t is used to train the STMENet. L_a is used in ALENet for learning to predict the atmospheric light. L_J and L_p are MSE loss and perceptual loss respectively. L is employed to make the whole network be jointly optimized.

4. Stereo Foggy Cityscapes Dataset

The Cityscapes dataset [4] is composed of large stereo video sequences recorded in streets from 50 different cities. The dataset has 5,000 images and each image has 1024×2048 pixels. There are 2,975 images in training set, 500 images in validation set and 1,525 images in test set. We apply synthetic fog to these real, clear-weather stereo image pairs using incomplete depth information as in [32]. According to [32], we could obtain the distance map for left images as:

$$\hat{d}(i,j) = B \times f_x \times ds(i,j)^{-1}, \tag{11}$$

$$d_l(i,j) = \hat{d}(i,j) \times (f_x^2 + (i-c_x)^2 + (j-c_y)^2)^{\frac{1}{2}} \times (f_x)^{-1}, \tag{12}$$

where f_x, (c_x, c_y) denote focal length, camera center, which are camera parameters for Cityscapes dataset and both expressed in pixel coordinates. B is the camera baseline distance. $\hat{d}(i,j)$ denotes the depth map, $ds(i,j)$ is the disparity map and $d_l(i,j)$ represents the left distance map. This depth estimation in Eq. 11 usually contains a large amount of severe artifacts and large holes. Following [32], We use stereoscopic inpainting [36] methods to handle the discrete depth problem, which performs distance completion at the level of superpixels, and introduces a novel, theoretically grounded objective for the superpixel-matching optimization that is involved. Then we generate left foggy images for Cityscapes dataset according Eq. 12 and Eq. 1.

In order to generate right foggy images, we need to obtain the right distance map. If the size of input image is

(a) (b) (c) (d) (e) (f)

Figure 3. Example images of our generated Stereo Foggy Cityscapes dataset. Top row: left images, Bottom row: corresponding right images. (a) and (d): clear binocular image pairs; (b) and (e): transmission maps; (c) and (f) foggy binocular image pairs.

$H \times W$, the size of the right distance map is also $H \times W$. As we know, the matching points in a stereo pair have the same depth between the camera and the imaging plane. Assuming (i, j) is a point in the right distance map, the right distance map could obtain as,

$$d_r(i, \lfloor j - ds(i,j) \rfloor) = \hat{d}(i,j) \times (f_x)^{-1} \times (f_x^2 + (i - c_x)^2 + (j - ds(i,j) - c_y)^2)^{\frac{1}{2}}. \quad (13)$$

The obtained right distance map (d_r) computed by Eq. 13 is highly noisy and incomplete. Following [32], We also use stereoscopic inpainting [36] methods to handle it. Then according to Eq. 1, we get the foggy right images.

We generate the random atmospheric light A = [a], where $a \in (0.7, 1.0)$ and use $\beta \in [0.005, 0.01, 0.02]$for each image. In this way, there are 8,925 binocular foggy image pairs in training set, 1500 binocular foggy image pairs in validation set, and 4,575 binocular foggy image pairs in test set for the Stereo Foggy Cityscapes dataset. Fig. 3 are two synthetic examples of binocular foggy image pairs.

5. Experiments

In this section, we implement the proposed method on the proposed Stereo Foggy Cityscapes dataset to demonstrate the effectiveness of the BibNet. We compare our BidNet with four single image dehazing methods: DehazeNet [1], MSCNN [29] , AOD-Net [11], and GridDehazeNet [17]. We also compare our BidNet with the binocular dehazing method [34], which is a joint learning framework for simultaneous stereo matching and dehazing. In addition, we do an ablation study to demonstrate the effectiveness of our embedding stereo transformation module.

5.1. Implementation

The proposed BibNet is end-to-end trainable without the need of pre-training for sub-modules. We train the network with RGB image patches of size 256×256. The Adam optimizer [10] is used with a batch size of 16, where β_1 and β_2 take the default values of 0.9 and 0.999, respectively. The initial learning rate is set to 0.01. The experiments are carried out on the Stereo Foggy Cityscapes dataset. The training is performed on the training set with 8925 binocular foggy image pairs and the evaluation is done on val set with 1500 binocular foggy image pairs. We train the network for 30 epochs in total and reduce the learning rate every 10 epochs. The training is carried out on two NVIDIA GeForce GTX 1070, and one GPU is used for testing.

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

We perform the evaluation on the proposed Stereo Foggy Cityscapes dataset. The ground truth images and the ground truth transmission maps are available, enabling us to evaluate the performance qualitatively and quantitatively.

Qualitative Results Fig. 4 shows qualitative comparison on the Stereo Foggy Cityscapes val set. BidNet is compared against the recent state-of-the-art single image dehazing methods [29, 17] and the binocular dehazing method [34], which is a Simultaneous Stereo Matching and Dehazing Network (SSMDN). Specially, in terms of GriddehazeNet, we finetune the outdoor model pre-trained on the Outdoor Training Set of RESIDE [12] on the Stereo Foggy Cityscapes dataset for fair comparison. In addition, we re-implement and train the SSMDN on the Stereo Foggy Cityscapes training dataset. Fig. 4 only shows results of five examples which consists of the left foggy images, the left haze-free images dehazed by existing image dehazing methods and our proposed BibNet, and the ground truth images. The first and second foggy examples have thin fog with $\beta = 0.005$ and $\beta = 0.01$ respectively. The rest foggy examples have thick fog with $\beta = 0.02$.

As revealed in Fig. 4, for the degradation due to thin fog ($\beta = 0.005$ and $\beta = 0.01$), MSCNN [29] (observed on the first and second row) tend to darken some regions (notice the cloud in the sky) and blurs the boundaries and texture (notice the trees). GridDehazeNet [17], SSMDN [34] and our method have the most competitive visual results. However, by looking closer, we observe that there is some remaining haze in the images dehazed by GridDehazeNet and SSMDN. In contrast, our method is able to generate realistic colors while better removing haze.

For the degradation due to thick fog ($\beta = 0.02$), it is very challenging (observed on the last three rows). MSCNN is darker than it should be and remains some haze, which is not desirable (observed on the second column). As shown

(a)Foggy Images (b) MSCNN [29] (c)GriddehazeNet [17] (d)SSMDN [34] (e)BidNet(ours) (f)GT

Figure 4. Qualitative comparisons on Stereo Foggy Cityscapes val set.

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
DehazeNet [1]	14.9705	0.4872	15.0384	0.5044
MSCNN [29]	18.9947	0.8595	19.0298	0.8628
AODNet [11]	15.4468	0.6316	15.5508	0.6463
GridDehazeNet* [17]	23.72	0.9226	23.74	0.9247
SSMDN* [34]	22.3753	0.9120	-	-
Ours BidNet	**25.5748**	**0.9438**	**25.6728**	**0.9451**

Table 2. Quantitative comparisons on Stereo Foggy Cityscapes val set. We compare the average values of PSNR and SSIM for each method. The symbol "*" means that we finetune the model or re-implement the methods on the Foggy Stereo Cityscapes train set.

in the third, fourth, and fifth rows, the dehazed results of MSCNN have some remaining haze. The colors of the car region of the result (observed on the third row) and the road of the result (observed on the fourth row) of MSCNN produce color shifts. GridDehazeNet generates relatively clear results, while the results in the third and fourth rows still have some remaining haze as shown in Fig. 4. In addition, there are some texture blur in the fourth line for the results of GriddehazeNet. The degradation for the region of sky even worse in the images dehazed by SSMDN. In contrast, the dehazed results by our BidNet are clear and the details of the scenes are enhanced moderately. Overall, our method has clear quantitative improvements over the state-of-the-art image dehazing methods. Importantly, our method performs better in the thick foggy scene.

Quantitative Results Tab. 2 compares our BidNet with DehazeNet [1], MSCNN [29], AODNet [11], GridDehazeNet [17] and SSMDN [34] in terms of PSNR and SSIM values on the Stereo Foggy Cityscapes val set. For better comparison, we use the single image dehazing methods to dehaze left images and right images separately. Our results are simultaneously estimated. From Tab. 2, our BidNet outperforms the state-of-the-art methods by a large margin. For the metric of SSIM, BidNet is 0.021 dB better than the second-best GridDehazeNet for both left images and right images. In addition, BibNet obtains a significant improvement of 1.8 dB and 1.9 dB in terms of PSNR value, over GridDehazeNet for left images and right images respectively. For the metric of PSNR, BidNet outperforms the binocular dehazing method, SSMDN [34], by 3 dB, which demonstrates the superiority of our BidNet.

Methods	Left		Right	
	PSNR	SSIM	PSNR	SSIM
Concatenation	23.6023	0.9217	23.6920	0.9234
STM	**25.5748**	**0.9438**	**25.6728**	**0.9451**

Table 3. Comparisons of the way how to utilize the correlation between the binocular images on Stereo Foggy Cityscapes val set.

5.3. Ablation Study

The ablation study is performed on the Stereo Foggy Cityscapes val set. The PSNR results and the SSIM results are averaged on left dehazed images or right dehazed images. In order to demonstrate the effectiveness of the STM, we perform an experiment replacing the STM by just making a concatenation of left features and right features. From Tab. 3, when using the concatenation instead of the STM, the dehazing results decrease 1.97 dB and 1.98 dB for left dehazed images and right dehazed images in terms of PSNR. The values of SSIM also reduce more than 0.2 compared with employing the STM, which demonstrates that our STM makes full use of the correlations between the binocular image pair.

We perform an ablation study involving the following four experiments: (1) when jointly estimating haze-free left images and haze-free right images, we discard the refine-

(a) (b) (c) (d) (e) (f)

Figure 5. Qualitative results on Stereo Foggy Cityscapes val Dataset. (a) and (d) stereo foggy images. (b) and (e)stereo haze-free images dehazed by BibNet. (c) and (f) ground truth (stereo clear images).

Outputs	L_{rt}	L_p	Left		Right	
			PSNR	SSIM	PSNR	SSIM
Stereo foggy img		✓	22.5501	0.9141	22.2598	0.9098
Stereo foggy img	✓		23.8823	0.9315	23.5926	0.9308
Left foggy img	✓	✓	24.2875	0.9397	-	-
Stereo foggy img	✓	✓	**25.5748**	**0.9438**	**25.6728**	**0.9451**

Table 4. The inputs are binocular foggy image pairs and ablation experiments are conducted to explore the effects of the refinement module in STENet, the perceptual loss, and jointly estimating the right haze-free images with the left haze-free images. $L_{rt} = \left\|\hat{t_{rl}} - t_l\right\|_2^2 + \left\|\hat{t_{rr}} - t_r\right\|_2^2$, L_{rt} denotes the loss for predict refined transmission map. $\hat{t_{rl}}$ ($\hat{t_{rr}}$) is the estimated refined transmission maps in the STENet. L_p is the perceptual loss.

ment module in the STMENet and the loss for predicting refined transmission map estimation; (2) when the perceptual loss is not used, jointly estimate haze-free left images and haze-free right images; (3) it is trained to only estimate the left haze-free images and all loss about the right images are removed; (4) BidNet. Tab. 4 shows that the refinement module in the STMENet is important for the performance of dehazing. Without the perceptual loss, the dehazed results decreased 1.69 dB and 2.08 dB in terms of PSNR for left and right images respectively. Comparing the results in the third line and forth line in Tab. 4, we could find that the performance of jointly estimating haze-free left images and haze-free right images is better than only training a model to estimate haze-free left images.

5.4. Evaluation on Real Dataset

To demonstrate the generalization ability of the BidNet in real scenes, we evaluate the proposed method on several real-world binocular hazy images from Drivingstereo dataset [39]. Drivingstereo dataset is a large-scale dataset for stereo matching in real autonomous driving scenarios. It selects 2000 frames with 4 different weathers (sunny, cloudy, foggy, rainy) for specific requests. There are 500 frames with foggy weather from sequences are selected. For the 500 foggy images, the corresponding clear images are not available. We leverage the fog simulation pipeline described in Sec.4 to add fog to the sunny and cloudy sequences in Drivingstereo dataset, and then finetune our BidNet on these synthetic stereo foggy images. We test our model on the 500 real binocular foggy images from the

Figure 6. Examples evaluated on Drivingstereo Dataset [39].

Drivingstereo dataset. Fig. 6 shows three examples dehazed by our BibNet, which demonstrates the proposed method generalizes well in the real stereo foggy scenes.

Speed: For 400×881 images, BidNet takes 0.23s dehaze the binocular pair on a NVIDIA GeForce GTX 1070.

6. Conclusion

We have proposed a novel dehazing framework: Binocular image dehazing Network (BidNet). It inputs binocular foggy image pairs and aims at recovering the haze-free binocular image pairs. BidNet could explore the correlations between the binocular image pairs to improve the performance of image dehazing. BibNet employs a Stereo Transformation Module to learn the horizontal correlation between the binocular image pairs and embeds the information from the other image in a binocular image pair, which does not need estimate disparity explicitly. In addition, we have extended the Foggy Cityscapes dataset to a Stereo Foggy Cityscapes dataset for binocular image dehazing task. Experimental results on synthetic and real datasets demonstrate the effectiveness of the proposed BidNet.

Acknowledgments: The work is supported by the National Key R&D Program of China (Grant # 2018AAA0102800 and 2018AAA0102802) and National Natural Science Foundation of China (Grant # 61632018).

References

[1] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 25(11):5187–5198, 2016.

[2] Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. Hierarchical shot detector. In *ICCV*, 2019.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[5] Raanan Fattal. Dehazing using color-lines. *ACM Transactions on Graphics (TOG)*, 34(1):13, 2014.

[6] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2011.

[7] Zequn Jie, Pengfei Wang, Yonggen Ling, Bo Zhao, Yunchao Wei, Jiashi Feng, and Wei Liu. Left-right comparative recurrent model for stereo matching. In *CVPR*, 2018.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[9] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017.

[12] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *TIP*, 28(1):492–505, 2019.

[13] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, 2019.

[14] Yazhao Li, Yanwei Pang, Jianbing Shen, Jiale Cao, and Ling Shao. Netnet: Neighbor erasing and transferring network for bettersingle shot object detection. In *CVPR*, 2020.

[15] Z. Li, P. Tan, R. T. Tan, D. Zou, Steven Zhiying Zhou, and L. Cheong. Simultaneous video defogging and stereo reconstruction. In *CVPR*, 2015.

[16] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *CVPR*, 2018.

[17] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, 2019.

[18] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[19] Shuai Ma, Yanwei Pang, Jing Pan, and Ling Shao. Preserving details in semantics-aware context for scene parsing. *SCIENCE CHINA Information Sciences*, 63(2):120106, 2020.

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

[21] Jeong-Yun Na and Kuk-Jin 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*, pages 15–19, 10 2018.

[22] Srinivasa G Narasimhan and Shree K Nayar. Chromatic framework for vision in bad weather. In *CVPR*, 2000.

[23] Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Enriched feature guided refinement network for object detection. In *ICCV*, 2019.

[24] Yanwei Pang, Yazhao Li, Jianbing Shen, and Ling Shao. Towards bridging semantic gap to improve semantic segmentation. In *ICCV*, 2019.

[25] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *ICCV*, 2019.

[26] Yanwei Pang, Jin Xie, and Xuelong Li. Visual haze removal by a unified generative adversarial network. *TCSVT*, 2018.

[27] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: From monocular to stereo 3d object detection. In *CVPR*, 2019.

[28] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *CVPR*, 2019.

[29] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016.

[30] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, 2018.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, Sep 2018.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Taeyong Song, Youngjung Kim, Changjae Oh, and Kwanghoon Sohn. Deep network for simultaneous stereo matching and dehazing. In *BMVC*, 2018.

[35] Robby T Tan. Visibility in bad weather from a single image. In *CVPR*, 2008.

[36] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *CVPR*, 2008.

[37] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *CVPR*, 2019.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[39] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019.

[40] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.

[41] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, 2019.

[42] He Zhang and Vishal M. Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018.

[43] Zhijie Zhang and Yanwei Pang. Cgnet: cross-guidance network for semantic segmentation. *SCIENCE CHINA Information Sciences*, 63(2):120104, 2020.

[44] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *TIP*, 24(11):3522–3533, 2015.