Dehazing Evaluation: Real-World Benchmark Datasets, Criteria, and Baselines

Shiyu Zhao, Lin Zhang, *Senior Member, IEEE*, Shuaiyi Huang, Ying Shen, *Member, IEEE*, and Shengjie Zhao, *Senior Member, IEEE*

***Abstract*— On benchmark images, modern dehazing methods are able to achieve very comparable results whose differences are too subtle for people to qualitatively judge. Thus, it is imperative to adopt quantitative evaluation on a vast number of hazy images. However, existing quantitative evaluation schemes are not convincing due to a lack of appropriate datasets and poor correlations between metrics and human perceptions. In this work, we attempt to address these issues, and we make two contributions. First, we establish two benchmark datasets, i.e., the BEnchmark Dataset for Dehazing Evaluation (BeDDE) and the EXtension of the BeDDE (exBeDDE), which had been lacking for a long period of time. The BeDDE is used to evaluate dehazing methods via full reference image quality assessment (FR-IQA) metrics. It provides hazy images, clear references, haze level labels, and manually labeled masks that indicate the regions of interest (ROIs) in image pairs. The exBeDDE is used to assess the performance of dehazing evaluation metrics. It provides extra dehazed images and subjective scores from people. To the best of our knowledge, the BeDDE is the first dehazing dataset whose image pairs were collected in natural outdoor scenes without any simulation. Second, we provide a new insight that dehazing involves two separate aspects, i.e., visibility restoration and realness restoration, which should be evaluated independently; thus, to characterize them, we establish two criteria, i.e., the visibility index (VI) and the realness index (RI), respectively. The effectiveness of the criteria is verified through extensive experiments. Furthermore, 14 representative dehazing methods are evaluated as baselines using our criteria on BeDDE. Our datasets and relevant code are available at https://github.com/xiaofeng94/BeDDE-for-defogging.**

***Index Terms*— Benchmark dataset, dehazing evaluation metrics, dehazing baselines, FR-IQA.**

I. Introduction

HAZE, fog, and mist lead to low visibility due to their scattering and absorption of light. Although they are different atmospheric phenomena with heterogeneous characteristics and components [1], they can result in similar atmospheric visibility impairment in images and thus degrade the performance of related vision algorithms or systems, e.g., classification, detection, segmentation, and advanced driver assistance systems (ADASs). Consequently, researchers have shown great enthusiasm for dehazing and a great number of relevant approaches [2]–[6] have been presented. However, due to a lack of appropriate metrics and benchmark datasets consisting of natural hazy images and clear references, how to evaluate the performance of these methods remains an open issue.

The widely adopted evaluation schemes can be categorized into three classes. The first class relies on readers' subjective judgments on dehazed images. However, these kinds of schemes restrict themselves to a limited number of evaluation images and tend to result in contradictions among different readers. The second class adopts no-reference image quality assessment (NR-IQA) metrics [11], [12] which are specially designed for evaluating dehazing methods. However, NR-IQA is still an open issue, and its metrics are less reliable than full reference image quality assessment (FR-IQA) metrics. The third class is the most prevalent one. It simulates hazy images from clear images based on Koschmieder's law [13] and then employs FR-IQA metrics, such as the peak-signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [10], to evaluate dehazing algorithms. However, this strategy is also questionable. First, these kinds of schemes usually adopt indoor images with scene depth. However, as mentioned in [13], Koschmieder's law assumes that the surface of the earth is a uniform horizontal plane and that the linear dimensions of an object are small compared to its distance from an observer. Apparently, indoor scenes can hardly satisfy these assumptions. Second, there is a certain gap between real hazy images and simulated images, and a dehazing method that fits such images well might not necessarily fit natural images well. Third, the widely used FR-IQA metrics are designed to evaluate general image distortions such as noise and blur, and thus, they might not be the most suitable ways to evaluate dehazing methods.

As mentioned above, although dehazing evaluation has been explored for some time, appropriate datasets and suitable metrics are still lacking. In this work, we attempt to objectively and reasonably address these issues.

First, we overcome the difficulty of collecting real-world images under different weather conditions and provide a

Manuscript received June 6, 2019; revised November 20, 2019 and March 16, 2020; accepted May 13, 2020. Date of publication May 22, 2020; date of current version July 8, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61672380, Grant 61973235, Grant 61936014, and Grant 61972285, and in part by the Natural Science Foundation of Shanghai under Grant 19ZR1461300. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M. Ascenso. *(Corresponding author: Lin Zhang.)*

Shiyu Zhao, Lin Zhang, Ying Shen, and Shengjie Zhao are with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: 1731558@tongji.edu.cn; cslinzhang@tongji.edu.cn; yingshen@tongji.edu.cn; shengjiezhao@tongji.edu.cn).

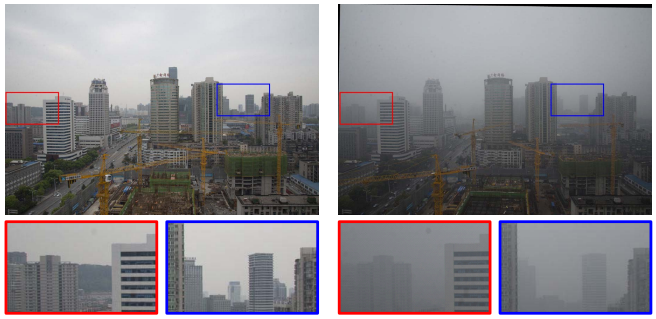
Shuaiyi Huang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: huangshy1@shanghaitech.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.2995264

dataset called the BEnchmark Dataset for Dehazing Evaluation (BeDDE), which had been lacking for a long period of time, for the evaluation of dehazing algorithms. The BeDDE contains 208 pairs of natural images, and each image pair consists of a natural hazy image and a well-aligned clear reference. The raw images of the BeDDE were collected from 23 provincial capital cities in China. For each raw image pair, the hazy image and the corresponding clear image were roughly registered. Due to slight changes in viewpoints and contents during data collection, all raw image pairs are aligned, and then, their common regions of interest (ROIs) are delineated by manually labeled masks. The registered ROIs between the hazy images and their clear references make it possible to explore FR-IQA metrics to assess the quality of dehazing results. Notably, the evaluation with masks is statistically reliable because our masks cover key foreground objects and the BeDDE involves a sufficient number of images. In addition to masks, based on the haze density, we manually classify the hazy images of the BeDDE into three haze levels, "light", "medium" and "heavy". To the best of our knowledge, although some datasets [14]–[16] in this field provide real-world hazy and haze-free image pairs, they collected their hazy images in artificial hazy environments that had certain problematic gaps with natural scenes. The drawbacks of these datasets are further discussed in Sect. II-C. Therefore, as a dehazing evaluation benchmark dataset, the BeDDE is the first dataset whose hazy images and clear references are all collected from natural outdoor scenes.

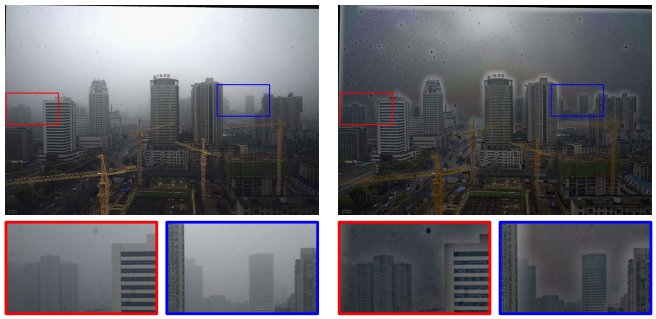
Second, we select 167 hazy images from 12 cities in the BeDDE and generate 1670 dehazed images by feeding the selected images into 10 representative dehazing methods. Among those images, we group the hazy images by cities and group the dehazed images by their corresponding hazy inputs. Additionally, we provide subjective scores (also called the mean opinion scores (MOSs) of different individuals) to indicate the visibility for the hazy images in each hazy group or the realness for the dehazed images in each dehazing group. With all these images and scores, we build the EXtension of the BeDDE (exBeDDE) to assess the performance of dehazing evaluation metrics.

Third, we find that some dehazing results can be good in visibility but have artifacts, while others are akin to natural images but have more haze. If we consider visibility to be more important, the former dehazing results are better and vice versa. Therefore, under these circumstances, it is hard for us to make consistent judgments, and the performance of dehazing methods mainly depends on the given weight between visibility and realness. Typical samples are shown in Fig. 1. The result of the All-in-One Dehazing Network (AOD-Net) [5] is more visually pleasing but has more remaining haze. In contrast, the result of the dark channel prior (DCP) [2] includes less haze but involves halos and color distortions. Several general FR-IQA metrics make contradictory evaluations of these samples, and different people might draw opposite conclusions as well. Therefore, it is difficult to judge dehazing quality via one criterion. To handle this problem, we suggest that dehazing methods should be assessed in two separate aspects, i.e., visibility and realness, and propose two FR-IQA criteria, the visibility index (VI) and the realness index (RI),



A: VI, B: RI, C: VSI
D: GMSD, E: FSIM, F: SSIM
(a) Clear image

A: 0.740, B: **0.973**, C: 0.909
D: 0.177, E: 0.733, F: 0.669
(b) Hazy image



A: 0.842, B: 0.972, C: **0.935**
D: 0.165, E: 0.818, F: **0.679**
(c) AOD-Net [5]

A: **0.878**, B: 0.969, C: 0.917
D: **0.139**, E: **0.861**, F: 0.672
(d) DCP [2]

Fig. 1. Dehazing results should be evaluated in terms of visibility and realness. (a) and (b) are the clear reference and the hazy image, respectively. (c) and (d) are the dehazing results of two methods, the AOD-Net [5] and DCP [2], respectively. Below each image are the scores of 6 metrics. "A"∼"F" represent the 6 metrics, i.e., VI, RI, VSI [7], GMSD [8], FSIM [9], and SSIM [10]. The best value for each metric is highlighted in boldface. Note that VI and RI are the two proposed criteria in this work.

to evaluate dehazing methods in the two aspects, respectively. Fig. 1 offers a glimpse of the effectiveness of the VI and the RI. (b), (c) and (d) increase in visibility and decrease in realness. The VI and RI provide the correct orders. In Sect. V, we conduct extensive experiments on the exBeDDE to test the superiority of the proposed criteria to general FR-IQA metrics and existing no-reference dehazing evaluation metrics. Additionally, evaluations of 14 dehazing methods with our criteria on BeDDE are provided as baselines.

The remainder of this paper is organized as follows. Sect. II reviews the related work. Sect. III introduces our established benchmark datasets BeDDE and exBeDDE in detail. Sect. IV presents the proposed VI and RI for dehazing evaluation. Experimental results are reported in Sect. V. Finally, Sect. VI concludes this paper.

II. Related Work

In this section, we briefly review several well-known FR-IQA metrics and current evaluation schemes for image dehazing. Then, we introduce recent efforts to construct dehazing datasets.

A. Advances in FR-IQA

Pixel-based metrics, e.g., the mean square error (MSE) and PSNR, correlate poorly with human perceptions, and thus,

human visual system (HVS)-based IQA metrics have been explored. The most well-known metric is the SSIM proposed by Wang *et al.* [10]. It considers that the HVS is highly adapted to extract the structural information from the visual scene, and thus, leverages the luminance, contrast and structural information to calculate the similarity. In a later work of Wang *et al.*, multiscale (MS) information was introduced in the SSIM, and MS-SSIM [17] was proposed. In [18], Wang and Li introduced a novel quality score pooling strategy based on information content weighting (IW) and improved the MS-SSIM by proposing the IW-SSIM.

Different from the variants of the SSIM, Zhang *et al.*'s work [9] held the view that the HVS understood an image mainly based on its low-level features and proposed two feature similarity indices, the FSIM and FSIMc, which involved the phase congruency, gradient magnitude features and chrominance features. Later, they replaced the phase congruency features with saliency maps and proposed a new metric named the visual saliency-induced index (VSI) [7]. In Liu *et al.*'s work [19], exploiting the prior that gradients convey important visual information and are crucial to scene understanding, the authors proposed a gradient similarity-based metric (GSM). However, Xue *et al.* [8] found that gradient maps were sensitive to image distortions while different local structures in a distorted image had different degrees of problematic degradations. Accordingly, they proposed the gradient magnitude similarity deviation (GMSD), which calculated the standard deviation of the gradient magnitude similarity (GMS) map as the similarity score. In [20], Zhang *et al.* explored assessing image quality by deep learning and proposed the learned perceptual image patch similarity (LPIPS) metric, which adopted features from pretrained neural networks to compare the distorted image and the reference.

B. Current Dehazing Evaluation Schemes

As mentioned above, there are three classes of evaluation schemes for image dehazing. The first class encourages an article to present dehazed images or other intermediate outputs (e.g., transmission maps) generated by different algorithms, and it resorts to the subjective judgments of readers only. Early dehazing studies [21], [22] preferred this strategy, but they used different hazy images, making the comparison less convincing. In [3], Fattal collected the most frequently used hazy images in previous dehazing studies [23]–[26] and provided a benchmark dataset with 23 hazy images for subjective judgments. However, these images have similar visibility conditions, and a sole evaluation of them might lead to a preference to handle hazy images with certain conditions. Additionally, some dehazed images or outputs are too similar to each other for people to judge.

The second class uses specially designed NR-IQA metrics. In [11], Hautière *et al.* considered the contrast restoration of dehazed images and proposed three indicators, i.e., e, $\overline{r}$, and σ. Specifically, e assesses the ability of a dehazing method to restore the edges. $\overline{r}$ evaluates the quality of contrast restoration by a dehazing method, and σ computes the number of saturated pixels (black or white) in the dehazed image. Later, Choi *et al.* [12] proposed another no-reference assessment method, called the fog aware density evaluator (FADE), which focuses on the characteristics of hazy images including low contrast, faint color, and shifted luminance. Some dehazing studies [27]–[30] adopted these NR-IQA metrics to evaluate their models. However, as mentioned above, those metrics limit themselves to certain kinds of distortions, such as distorted contrast and luminance, whereas different dehazing methods may involve various types of distortions. Moreover, Ma *et al.* [31] found that these no-reference metrics correlated poorly with human perceptions. As a result, for this task, NR-IQA metrics are less reliable.

The third class explores FR-IQA metrics to evaluate dehazing methods. Due to a lack of real-world image pairs, the third class usually simulates hazy images from clear images based on Koschmieder's law [13]. In [3], Fattal used 11 clear images with depth maps provided by [32] to simulate hazy images with ground-truth transmission maps, and the L_1 distance between the estimated transmission map and the ground-truth is calculated as the metric. Some dehazing studies [33]–[36] have adopted Fattal's dataset as their test set. In addition, more studies [5], [6], [37], [38] have employed existing indoor datasets with depth maps, e.g., the NYU2 [39] and Middlebury datasets [40], [41], to handle the lack of essential depth information in the simulation. Then, these studies adopted FR-IQA metrics, such as the MSE, PSNR and SSIM [10], to evaluate dehazing methods on pairs of clear images and restored images. However, there are some remaining issues. First, as mentioned above, indoor scenes do no actually satisfy the premise on which Koschmieder's law is established. Second, the gap between real hazy images and simulated images was ignored. Third, the employed FR-IQA metrics were designed to evaluate general image distortions. However, regarding dehazing evaluation, their effectiveness was not verified.

C. Efforts to Construct Dehazing Datasets

To handle issues in exploiting FR-IQA metrics for the evaluation of dehazing methods, there are a few studies that consider establishing appropriate dehazing datasets. Using SiVIC^{TM} software, Tarel *et al.* constructed two synthetic outdoor datasets, namely, the Foggy Road Image DAtabase (FRIDA) [42] and FRIDA2 [43], to test dehazing methods. The FRIDA and FRIDA2 contained 90 synthetic images of 18 urban road scenes and 330 synthetic images of 66 diverse scenes, respectively, and they provided both homogeneous and heterogeneous fog. However, their images were at low resolutions and did not look realistic. Later, based on Koschmieder's law, Ancuti *et al.* [44] simulated hazy images using clear images and depth maps from both the NYU2 and Middlebury datasets. With these images, they built a dehazing dataset called D-HAZY. Apparently, this dataset followed the same idea of haze simulation as other studies and had the aforementioned problems. To illustrate their drawbacks, typical images of D-HAZY and synthetic images of the FRIDA and FRIDA2 are shown in Fig. 2.

Fig. 2. Typical samples from D-HAZY [44] and synthetic images from the FRIDA [42] and FRIDA2 [43]. (a) and (b) come from the FRIDA and FRIDA2, respectively. (c) and (d) are from D-HAZY. Note that many dehazing studies use images similar to (c) and (d) as training and testing data.

More recently, Li *et al.* [45] established a dataset named REalistic Single-Image DEhazing (RESIDE) which provided indoor and outdoor images with simulated haze for the training and testing of dehazing models. The indoor images also came from the NYU2 and Middlebury datasets. Thus, they had the same problems as D-HAZY did. The outdoor images were collected from the Internet, and their depth maps were estimated from those monocular images using Liu *et al.*'s model [46]. However, depth estimation from a monocular image was highly ill-posed, and thus, the acquired depth maps were unreliable, leading to poor simulation. As compensation for the dataset's drawback, they proposed an indirect evaluation scheme. In their scheme, state-of-the-art object detection algorithms were used to detect the objects of interest on dehazed images that were generated by different dehazing methods from real hazy images, and then, the mean Average Precisions (mAPs) of those detection algorithms were calculated as the scores of the dehazing methods. In [47], Sakaridis *et al.* added synthetic fog to images from Cityscapes [48] and established a dataset named Foggy Cityscapes. However, the depth in Cityscapes was not complete, and thus, the quality of the simulated hazy images could not be guaranteed. Therefore, it is also inappropriate to evaluate on Foggy Cityscapes. To better illustrate their drawbacks, outdoor sample images of RESIDE and Foggy Cityscapes with the corresponding depth maps are exhibited in Fig. 3.

There are some studies [14]–[16], [49] focusing on collecting image pairs from real-world scenes, but most of them rely on artificial hazy scenes that have certain problematic gaps compared with natural hazy scenes. Ancuti *et al.* constructed O-HAZE [14], which provides 45 pairs of hazy outdoor images and corresponding references. They used two haze machines and a fan to produce fog or haze. However, the generated haze can cover only an area much smaller than that covered by natural haze. Moreover, haze machines produce only water vapor, whereas natural haze is a complex compound of vapor and other particles (e.g., aerosol, sulfur dioxide and nitrogen

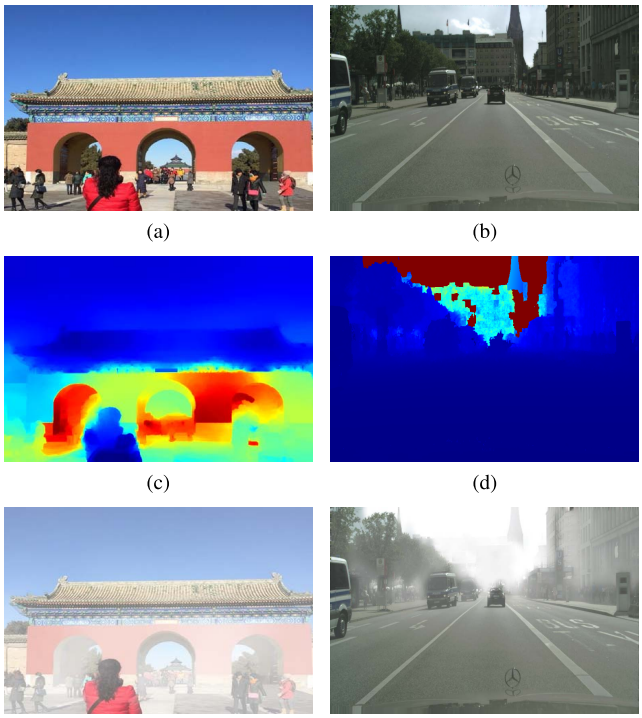


Fig. 3. Outdoor samples from RESIDE [45] and Foggy Cityscapes [47]. The left column is from RESIDE. The right column comes from Foggy Cityscapes. (a) and (b) are the original clear images. (c) and (d) display depth maps. (e) and (f) are simulated images. As shown, the depth map of RESIDE is inaccurate, and that of Foggy Cityscapes is incomplete. Both defects result in poor simulation quality.

dioxide). Unlike vapor, those particles affect the extinction coefficient of the atmosphere by both absorbing and scattering light. Therefore, the hazy images cannot fully represent the true nature of fog or haze. Ancuti *et al.* adopted the same method of constructing O-HAZE to build an indoor dataset, I-HAZE [15], which consists of 35 indoor image pairs. In [16], Bijelic *et al.* employed a delicate fog chamber [50] to simulate haze scenes and established a large dataset with hazy and haze-free image pairs. However, the chamber is too small in size and limited in scene variety compared with real outdoor environments. Additionally, this chamber is only able to generate haze with water droplets and thus has similar problems as O-HAZE and I-HAZE.

In Table I, we summarize the characteristics of the BeDDE and other datasets with paired images to illustrate their differences.

III. The BeDDE & the exBeDDE: Real-World Benchmark Datasets for Dehazing Evaluation

The BeDDE is a real-world dataset containing hazy images and the corresponding clear references for dehazing studies. Its extension, the exBeDDE, includes dehazed images generated by dehazing methods and the MOSs of these images. In this section, we will present overviews of the two datasets and the way in which we establish them.

A. Dataset Overview

The BeDDE contains 208 image pairs collected from 23 provincial capital cities in China. For each city, one clear

(a) Reference image (b) Light (c) Medium (d) Medium (e) Heavy

Fig. 4. The clear reference and hazy images with different visibility conditions taken from Chengdu, a major city in China. (a) is the clear image. (b)∼(e) are hazy images whose visibility conditions become sequentially worse, and their labeled haze levels are presented below.

image and several hazy images from the same place are provided. For each image pair, the hazy image is well aligned with the corresponding clear image via a 2D projective transformation, and a manually labeled mask is provided. This mask is used to delineate regions with the same contents in these two images, which we call the ROI of this pair and which are involved in the scoring of dehazing methods. In addition to clear reference images and masks, another outstanding advantage of the BeDDE is its diversity of visibility conditions. To exploit this advantage, we manually classified 208 hazy images of the BeDDE into three haze levels, "light", "medium" and "heavy", based on their haze densities. To illustrate this classification, Fig. 4 provides several images of Chengdu with their haze levels.

The exBeDDE is the extension of the BeDDE and is designed for the assessment of dehazing evaluation metrics. It contains 167 hazy images from 12 cities in the BeDDE and 1670 dehazed images generated by 10 dehazing methods using these hazy images. These images are divided into groups for different assessment purposes. The hazy images are grouped by the cities where they were taken to measure the ability of visibility evaluation. The dehazed images are grouped by the original hazy images from which they were derived to assess the ability of realness evaluation. For each group, MOSs indicating the quality of the images are provided. The scores of the images in a hazy group are determined by visibility. For dehazed images, their scores are determined mainly by realness and partly by visibility.

B. Pipeline to Establish the BeDDE

There are five steps in the pipeline of the establishment of the BeDDE: data acquisition, image registration, data cleaning, mask labeling, and haze level labeling.

1) Data Acquisition: In this step, an image of a fixed place was collected at a time between 8:00 and 9:00 each day for a period of 40 days. Such collections were conducted simultaneously at 34 provincial capitals in China in one year, and the representative scenes in those cities were chosen as the collection sites. All images were taken by professional photographers. To increase the variety, the photographers were required to take photos with their own cameras and favorable settings, but for each city, the device and setting were fixed and remained the same during the collection. We put these parameters (e.g., the device model, aperture value, exposure time, focal length, ISO speed, metering mode, and white balance mode) in a separate document and release them with our datasets. Owing to the 46 photographers, we acquired 1269 high-resolution images as raw data.

TABLE I

CHARACTERISTICS OF THE BEDDE AND OTHER DEHAZING DATASETS WITH PAIRED IMAGES

Dataset	In/outdoor	Clear scene	Haze
FRIDA [42]	Outdoor	Synthetic	Synthetic
FRIDA2 [43]	Outdoor	Synthetic	Synthetic
D-HAZY [44]	Indoor	Real-world	Synthetic
RESIDE [45]	Both	Real-world	Synthetic
Foggy Cityscapes [47]	Outdoor	Real-world	Synthetic
I-HAZE [15]	Indoor	Real-world	Artificial
O-HAZE [14]	Outdoor	Real-world	Artificial
Bijelic *et al.*'s [16]	Indoor	Real-world	Artificial
BeDDE (Ours)	Outdoor	Real-world	Natural

2) Image Registration: Although the images of a city were taken in the same place, slight changes in viewpoints were inevitable. Therefore, for each city, we chose one image as the reference, which was in overcast weather and provided good visibility. If there was no appropriate reference for a city, we simply dropped all images of this city. Afterwards, we aligned all the other images to this reference by a standard image registration procedure [51], which is composed of keypoint detection, feature extraction and matching, transformation matrix estimation, and transformation application. Specifically, we used speeded up robust features (SURF) [52]. Since there were only slight changes in the viewpoints, we adopted a 2D projective matrix as the transformation model which can be formulated as,

$$[x, y, 1] = [u, v, 1] \cdot T \tag{1}$$

where T is a 3×3 transformation matrix and $[u, v, 1]$ and $[x, y, 1]$ are the homogeneous coordinates of a pixel in images before and after registration, respectively.

3) Data Cleaning: In this step, we first filtered out images whose environmental conditions, such as sun position and color appearance, are obviously different from the selected reference. Thus, except for haze density, the differences in environmental conditions between the hazy images and references were indistinguishable to human eyes. Then, we filtered out undesired images that were poorly aligned with the references. To better visualize the registration quality and facilitate our selection, we created an overlaid image by assigning the grayscale version of the reference and that of the hazy image to different channels of a blank image (zero values in RGB channels). With such an overlay, poorly aligned edges became salient. In Fig. 6, an overlaid image for a poorly aligned pair

(a) 0.1054 (b) 0.0075 (c) 0.5418 (d) 0.1668

Fig. 5. Samples of different hazy scenes with their subjective scores. It is difficult to judge haze levels between two scenes such as (a) and (c) due to their scale varieties. However, we can tell that, in visibility, (a) is better than (b) and that (c) is better than (d). The mean opinion score for each image is provided. A higher score indicates a better quality. Note that our scores are meaningful only in the same group.

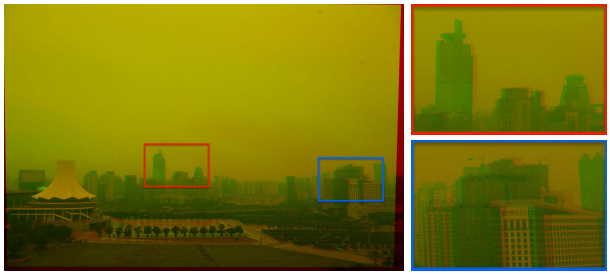


Fig. 6. An overlaid image for a poorly aligned pair. Some poorly aligned edges are highlighted by a red box and a blue box.

is displayed. As we can see, the registration quality is quite easy to judge in the overlaid image. With this technique, we manually filtered out poorly aligned image pairs.

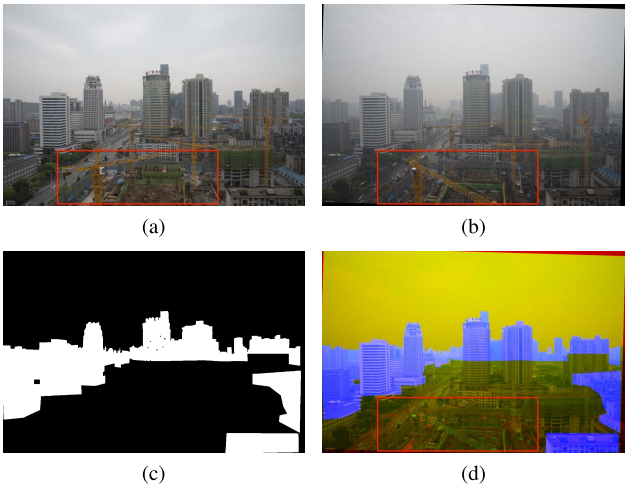
As mentioned above, there were two stages in our selection. In each stage, three persons individually judged the same image pair. Our selection was quite strict and adopted only pairs that were regarded as good samples by all three persons. Eventually, we selected 208 image pairs out of 1269 candidates to establish the BeDDE, with the number of hazy images for each city ranging from 1 to 26.

4) Mask Labeling: Although the images of a city were well aligned, there were still contents that could be different between them, such as vehicles, pedestrians, trees and water. Examples of such differences can be seen in Fig. 7. To handle this problem, we manually labeled a mask to delineate regions with the same contents between two images from a pair, i.e., the ROI of this pair. To acquire high-quality masks, two individuals were involved in the labeling process for each image pair. One labeled the whole mask and the other checked the mask and refined it. In the evaluation phase, we calculated only the score for ROIs to rank dehazing methods.

5) Haze Level Labeling: In this step, based on the haze density of an image, a value (1, 2 or 3) was manually assigned to this image. Each image was scored by ten people and the average score was calculated. Then, the average score was mapped to one of three levels as the haze level label of this image, based on the following rule: $[1, 1.5)$ is mapped to "light", $[1.5, 2.5)$ to "medium" and $[2.5, 3)$ to "heavy".

C. Establishment of the exBeDDE

To assess dehazing evaluation metrics, we need images with different haze levels, dehazed images, and their scores given by people. Fortunately, the BeDDE provides hazy images



(a) (b) (c) (d)

Fig. 7. Examples with different contents in spite of a good alignment. (a) and (b) are the clear reference and the hazy image in a pair of BeDDE, respectively. (c) is the mask of this pair. (d) is an overlaid image for this pair, and the blue region is the ROI of this pair. The red box highlights the major differences between (a) and (b).

with different visibility conditions, and dehazed images can be generated by different dehazing methods. However, human perceptions of haze are vulnerable to the contents of the scene, especially when there are obvious scale varieties in scenes. As shown in Fig. 5, it may be difficult to judge whether (a) is better than (c) in visibility, but it is much easier to order (a) and (b) (or (c) and (d)) by visibility conditions. Following this idea, we built the exBeDDE with different groups of images. Specifically, we selected hazy images from 12 cities with the highest number of images and grouped them by cities as hazy groups. For the dehazed images, we grouped them by the original hazy images as dehazing groups. Such a grouping strategy for dehazed images was reasonable and eased our scoring. The reason is that we just need to compare different dehazing methods using the same hazy image each time, and a superb evaluation metric ought to perform well on our dehazing groups.

Finally, we obtained 12 hazy groups with 167 images and 167 dehazing groups with 1670 images. In each dehazing group, 10 dehazed images were generated by 10 representative dehazing methods, i.e., fast visibility restoration (FVR) [27], DCP [2], Bayesian defogging (BayD) [25], color attenuation prior (CAP) [53], Non-Local image dehazing (NLD) [54], the multiscale convolutional neural network MSCNN [55],

DehazeNet [4], AOD-Net [5], the densely connected pyramid dehazing network (DCPDN) [6], and the gated fusion network (GFN) [56]. The first 5 methods are prior-based approaches without training, while the last 5 ones are learning-based.

After groups were established, we invited 10 volunteers to rank images in each group. Two images of the same group were shown at a time to each subject (a volunteer). If the two images came from a hazy group, the subject was asked to order them by the degree of haze based on his or her own perception, which ensured that the images in hazy groups were ranked by visibility only. If the two images were from a dehazing group, the subject was initially required to order them by the number of distortions and artifacts. If the subject thought both images were very close in realness and was unable to tell the difference between them, he or she was required to judge by the haze degree. In this way, the images of dehazing groups were ranked by realness first and by visibility second. With all the orders from all the subjects, we obtained the overall ranks of the images in each group. We converted the rank of an image in a group into its MOS by the following mapping function:

$$score = \frac{1}{M}\sum_{i=1}^{M}(1-\frac{n_i}{N}) \tag{2}$$

Here, n_i is the ranked order of this image in this group given by the ith subject. M is the number of volunteers. N is the number of images in this group. In one group, the higher the score is, the better the image is. To demonstrate, scores of 4 samples are provided in Fig. 5. Note that our scores are defined in groups and should never be used between groups.

IV. The Visibility Index and the Realness Index

We propose two criteria, the VI and the RI, to assess dehazing results in visibility and realness, respectively. This section concretely discusses them.

A. Our Visibility Index

To better understand our VI, we begin with some essentials about the haze effect in the image. In atmospheric science, under certain constraints, Koschmieder's law [13] is widely used to describe the relationship among the apparent luminance, intrinsic luminance, extinction coefficient, and observing distance. The apparent luminance refers to the intensity of the observed object accepted by our eyes or camera. The intrinsic luminance means the light just reflected by the object. The extinction coefficient is a factor that describes the degree of fog or haze. The observing distance refers to the distance between the observer and the object.

In our case, Koschmieder's law can be defined as,

$$I(\mathbf{x}) = J(\mathbf{x})\,t(\mathbf{x}) + A\,(1-t(\mathbf{x}))\,. \tag{3}$$

Here, $\mathbf{x}$ is a pixel of the image. $I(\mathbf{x})$ and $J(\mathbf{x})$ refer to the apparent luminance and the intrinsic luminance of $\mathbf{x}$, respectively. A is the global skylight which represents ambient light in the atmosphere. $t(\mathbf{x})$ is the transmission of the intrinsic luminance in the atmosphere and can be further modeled as,

$$t(\mathbf{x}) = e^{-\beta d(\mathbf{x})} \tag{4}$$

where β is the extinction coefficient, and $d(\mathbf{x})$ is the observing distance of $\mathbf{x}$.

Our visibility index evaluates the quality of a hazy or dehazed image using the similarity of visibility between the image and its clear reference. Such a similarity is calculated by the transmission and gradients. First, based on Koschmieder's law, the transmission is highly related to the haze degree, and thus, the similarity between transmission maps of the hazy image and the reference can be used to assess the haze level. Given $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$, the transmission values of the clear image and the hazy image at pixel $\mathbf{x}$, the similarity of transmission $S_T(\mathbf{x})$ is defined as,

$$S_T(\mathbf{x}) = \frac{2T_1(\mathbf{x})\cdot T_2(\mathbf{x}) + C_1}{T_1^2(\mathbf{x}) + T_2^2(\mathbf{x}) + C_1} \tag{5}$$

where C_1 is a given positive constant to increase stability.

We calculate the transmission map using DCP [2] which states that in most of the nonsky patches, at least one color channel has some pixels whose intensity is very low and close to zero. Based on DCP, we obtain the dark channel at both sides of Eq. 3 and obtain the result as,

$$I^{dark}(\mathbf{x}) = J^{dark}(\mathbf{x})\,t(\mathbf{x}) + A\,(1-t(\mathbf{x})) \tag{6}$$

where $I^{dark}(\mathbf{x})$ and $J^{dark}(\mathbf{x})$ are the dark channels of images I and J at pixel $\mathbf{x}$, respectively. Since $J^{dark}(\mathbf{x}) \to 0$,

$$t(\mathbf{x}) = 1 - \frac{I^{dark}(\mathbf{x})}{A} \tag{7}$$

Here, $t(\mathbf{x})$ can represent the transmission map of any image. For the acquisition of A, we followed the DCP approach [2] to predict it since when the input image has cloudy sky regions, this method can generate high-quality predictions and fits the BeDDE. With the estimated A and Eq. 7, we obtain $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$ for the clear reference and the hazy image, respectively. Note that we do not adopt a widely used guided filter [57] to smooth transmission maps because we find that it makes no contribution to the performance of the VI and requires extra computation.

In addition to transmission, we find that the gradients of an image decrease as the extinction coefficient β in Eq. 4 increases. Supposing that $\mathbf{x}_1$ and $\mathbf{x}_2$ are two adjacent pixels in an image that are not located on the depth border, by applying Eq. 3 and Eq. 4 to $\mathbf{x}_1$ and $\mathbf{x}_2$, we obtain the following,

$$\begin{cases} I(\mathbf{x}_1) = J(\mathbf{x}_1)\,e^{-\beta d(\mathbf{x}_1)} + A\left(1-e^{-\beta d(\mathbf{x}_1)}\right) \\ I(\mathbf{x}_2) = J(\mathbf{x}_2)\,e^{-\beta d(\mathbf{x}_2)} + A\left(1-e^{-\beta d(\mathbf{x}_2)}\right) \end{cases} \tag{8}$$

Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are adjacent, their observing distances $d(\mathbf{x}_1)$ and $d(\mathbf{x}_2)$ are very close and can be replaced by an approximation value d_{appr}. Then, the difference between $I(\mathbf{x}_1)$ and $I(\mathbf{x}_2)$ can be approximated as,

$$I(\mathbf{x}_1) - I(\mathbf{x}_2) \approx (J(\mathbf{x}_1) - J(\mathbf{x}_2))e^{-\beta d_{appr}}. \tag{9}$$

As shown in Eq. 9, the difference between $I(\mathbf{x}_1)$ and $I(\mathbf{x}_2)$ varies as β varies. Therefore, the value of the gradient can be a good indicator of the extinction coefficient or haze level.

Previous FR-IQA studies have revealed that the gradient modulus (GM) is an effective feature for exploring gradient information. Thus, we adopt it in this work as well. The GM of

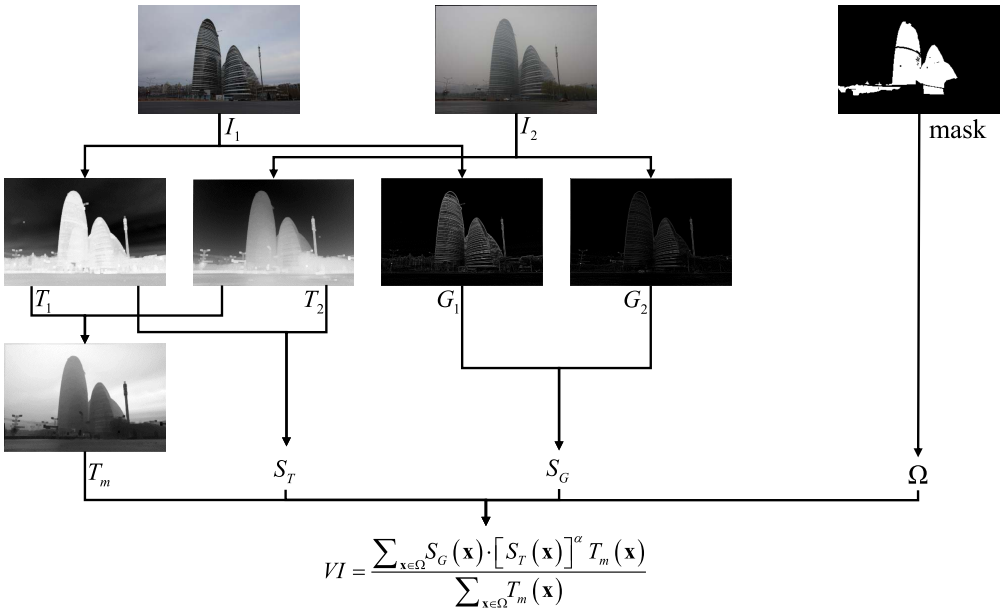


Fig. 8. Flowchart of the computation of our visibility index. I_1 is the clear reference image. I_2 is the hazy image. The mask is used to delineate regions with the same content between I_1 and I_2.

image I is computed as $G(\mathbf{x}) = \sqrt{G_x^2(\mathbf{x}) + G_y^2(\mathbf{x})}$, where $\mathbf{x}$ is a pixel of I, and $G_x^2(\mathbf{x})$ and $G_y^2(\mathbf{x})$ are partial derivatives of I at $\mathbf{x}$. Given $G_1(\mathbf{x})$ and $G_2(\mathbf{x})$, the GM values of the haze-free and hazy images, the similarity at pixel $\mathbf{x}$, $S_G(\mathbf{x})$, is defined as,

$$S_G(\mathbf{x}) = \frac{2G_1(\mathbf{x}) \cdot G_2(\mathbf{x}) + C_2}{G_1^2(\mathbf{x}) + G_2^2(\mathbf{x}) + C_2} \tag{10}$$

where C_2 is another given positive constant.

Finally, we take both transmission and the GM into consideration and combine $S_T(\mathbf{x})$ and $S_G(\mathbf{x})$ to obtain the overall visibility similarity of the hazy image and its clear reference, $S_V(\mathbf{x})$, which is defined as follows,

$$S_V(\mathbf{x}) = S_G(\mathbf{x}) \cdot [S_T(\mathbf{x})]^{\alpha} \tag{11}$$

where α is the parameter designed to adjust the relative importance between transmission and the GM. According to Eq. 3, haze will be more noticeable in regions with smaller transmission values. Therefore, we use $T_m(\mathbf{x}) = max(1 - T_1(\mathbf{x}), 1 - T_2(\mathbf{x}))$ to weight the importance of $S_V(\mathbf{x})$ to obtain the final visibility score of the hazy image. Our visibility index is defined as,

$$VI = \frac{\sum_{\mathbf{x}\in\Omega} S_V(\mathbf{x}) \cdot T_m(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} T_m(\mathbf{x})} \tag{12}$$

where Ω means the ROI delineated by a mask. Fig. 8 illustrates the flowchart of the computation of our VI for visibility evaluation.

B. Our Realness Index

As mentioned above, many dehazing methods introduce artifacts or distortions that degrade image quality. Therefore, although the original hazy images are natural images that do not have these degradations, we should also consider the realness of the dehazing results to thoroughly evaluate a dehazing method. Fortunately, FR-IQA for artifacts or distortions has already been extensively studied. Thus, we follow relevant studies by exploiting the similarity between the dehazed image and the clear reference in feature spaces to evaluate the realness of the dehazed image. Specifically, we adopt two potent features proposed in FR-IQA studies, i.e., phase congruency (PC) [58] and the chrominance information of LMN color space [59], which is optimized for the HVS [60].

PC was first defined in [58] and first introduced into the field of IQA by Zhang *et al.*'s FSIM [9]. In this work, we directly employ the $PC(\mathbf{x})$ (the PC of pixel $\mathbf{x}$) used in the FSIM for dehazing evaluation. However, $PC(\mathbf{x})$ is computed from the Y channel (luminance channel) of the YIQ color space without the chrominance information of RGB images. To involve this information, we follow the VSI [7] and exploit the M and N channels from LMN color space [59], [60]. Therefore, given the clear reference $I_1(\mathbf{x})$ and the hazy image $I_2(\mathbf{x})$, we first convert them into YMN color space as follows,

$$\begin{bmatrix} Y \\ M \\ N \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{13}$$

Then, we follow the FSIM and VSI to compute the similarities of PCs and chrominance features. For PCs, given $PC_1(\mathbf{x})$ of $I_1(\mathbf{x})$ and $PC_2(\mathbf{x})$ of $I_2(\mathbf{x})$, the similarity, $S_{PC}(\mathbf{x})$, is calculated as,

$$S_{PC}(\mathbf{x}) = \frac{2PC_1(\mathbf{x}) \cdot PC_2(\mathbf{x}) + C_3}{PC_1^2(\mathbf{x}) + PC_2^2(\mathbf{x}) + C_3} \tag{14}$$

where C_3 is a positive constant. For chrominance features, supposing that $M_1(\mathbf{x})$ and $N_1(\mathbf{x})$ are computed from $I_1(\mathbf{x})$ and $M_2(\mathbf{x})$ and $N_2(\mathbf{x})$ are derived from $I_2(\mathbf{x})$, the similarity $S_C(\mathbf{x})$

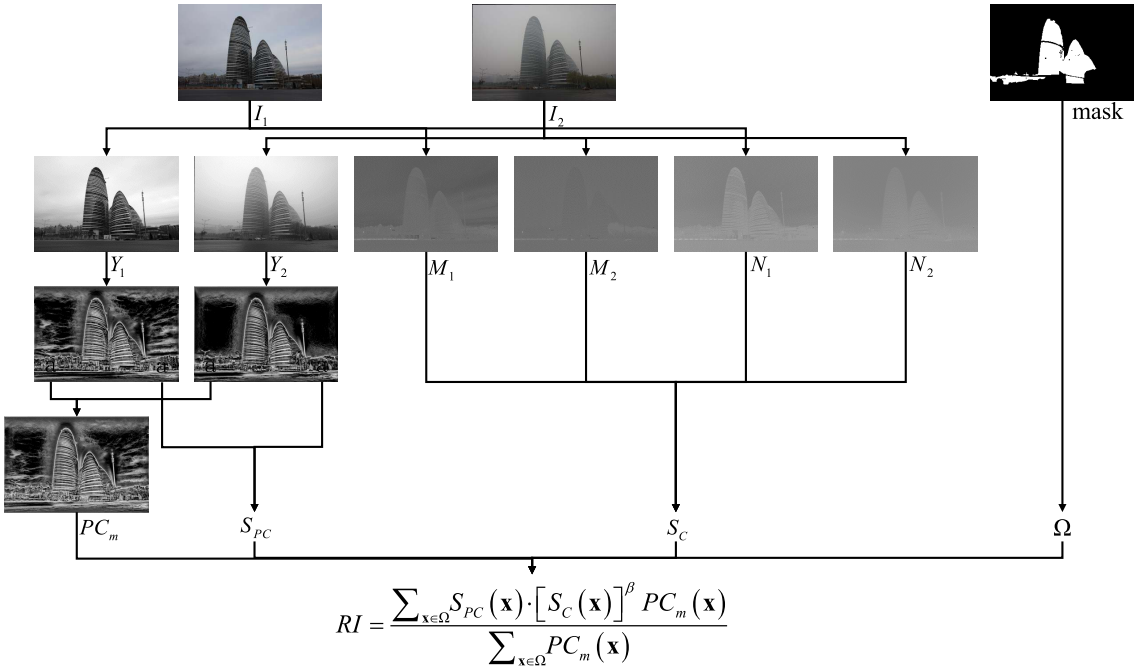


Fig. 9. Flowchart of the computation of our realness index. I_1 is the clear reference image. I_2 is the hazy image. The mask is employed to point out regions with the same content between I_1 and I_2.

is calculated as,

$$S_C(\mathbf{x}) = \frac{2M_1(\mathbf{x}) \cdot M_2(\mathbf{x}) + C_4}{M_1^2(\mathbf{x}) + M_2^2(\mathbf{x}) + C_4} \cdot \frac{2N_1(\mathbf{x}) \cdot N_2(\mathbf{x}) + C_4}{N_1^2(\mathbf{x}) + N_2^2(\mathbf{x}) + C_4} \quad (15)$$

where C_4 is a positive constant. Similar to our VI, $S_{PC}(\mathbf{x})$ and $S_C(\mathbf{x})$ are combined to obtain the overall similarity at point $\mathbf{x}$, $S_R(\mathbf{x})$, as follows,

$$S_R(\mathbf{x}) = S_{PC}(\mathbf{x}) \cdot [S_C(\mathbf{x})]^{\beta} \quad (16)$$

where β is a parameter used to adjust the relative importance between PC and chrominance.

Finally, we adopt $PC_m(\mathbf{x}) = max(PC_1(\mathbf{x}), PC_2(\mathbf{x}))$ introduced in the FSIM as weights for the importance of different positions. Our RI is ultimately defined as,

$$RI = \frac{\sum_{\mathbf{x}\in\Omega} S_R(\mathbf{x}) \cdot PC_m(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} PC_m(\mathbf{x})} \quad (17)$$

where Ω means the ROI delineated by a mask. Fig. 9 demonstrates the flowchart of the computation of the RI for realness evaluation.

V. Experiments and Discussions

A. Experimental Protocol

1) Efficacy of Features in the VI and RI: This experiment aimed to justify the feature selections of our VI and RI. In this experiment, we tested the abilities of different features to evaluate visibility and realness, including the transmission feature and 5 widely adopted features in recent FR-IQA metrics. In each test, one feature was assessed on both the hazy and dehazing groups of the exBeDDE. For each feature, given the feature map of the hazy image $F_1(\mathbf{x})$ and that of the reference $F_2(\mathbf{x})$, we computed the similarity map as,

$$S_F(\mathbf{x}) = \frac{2F_1(\mathbf{x}) \cdot F_2(\mathbf{x}) + C_F}{F_1^2(\mathbf{x}) + F_2^2(\mathbf{x}) + C_F}. \quad (18)$$

where C_F is a positive constant. The final score was calculated by averaging the similarity map for the ROI of this pair.

2) Efficacy of the VI and RI: This experiment was designed to illustrate the superiority of our criteria. In this experiment, we compared them on the exBeDDE with 7 state-of-the-art general FR-IQA metrics and 4 dehazing evaluation NR-IQA metrics. Implementations released by the original authors or official versions of MATLAB were used. If any, the parameters of these metrics were set to default.

3) Dehazing Baselines: In this experiment, we evaluated 14 representative dehazing methods on the BeDDE with the measurement of three metrics, e.g., the VI, RI and LPIPS [20]. Official implementations of those dehazing methods were used. All parameters were set to default, and trained models for CNN-based methods were provided by the original authors. We report the evaluation results of these methods as dehazing baselines so that they can be exploited by future dehazing research.

4) Assessment of the Metrics' Performance: In the first two experiments, the performances of IQA features or indices were assessed by 4 commonly used performance metrics, i.e., the Spearman rank-order correlation coefficient (SRCC), the Kendall rank-order correlation coefficient (KRCC), the Pearson linear correlation coefficient (PLCC), and the root mean squared error (RMSE). The SRCC and KRCC can measure the prediction monotonicity of an IQA index. However, they consider the rank of data points only and ignore the relative distances between data points. Thus, the PLCC and RMSE

TABLE II

COMPARISONS OF DIFFERENT FEATURES IN HAZY GROUPS OF THE EXBEDDE. THE SRCC AND KRCC ARE CONSIDERED AS A WHOLE. THE TOP TWO PERFORMANCE VALUES IN TERMS OF THE SUM OF THE TWO METRICS ARE HIGHLIGHTED IN RED AND BLUE

City		Trans	GM	PC	VS	ChromIQ	ChromMN
Beijing	SRCC	**0.73928**	**0.93214**	0.46071	0.71428	0.33928	-0.04285
	KRCC	**0.52381**	**0.79047**	0.35238	0.52381	-0.27619	-0.04761
Changsha	SRCC	0.73809	**0.92857**	0.73809	**0.78571**	-0.02381	0.00000
	KRCC	0.57142	**0.85714**	0.50000	**0.64285**	-0.07142	0.00000
Chengdu	SRCC	**0.98359**	**0.91589**	0.78598	0.91247	0.36205	0.73401
	KRCC	**0.91384**	**0.75384**	0.61230	0.74153	0.26769	0.58153
Hangzhou	SRCC	0.81818	**0.94545**	0.43636	**0.88181**	0.06363	0.50000
	KRCC	0.70909	**0.85454**	0.30909	**0.78181**	-0.01818	0.41818
Hefei	SRCC	0.55000	**0.91666**	**0.90000**	0.45000	0.81666	0.81666
	KRCC	0.38888	**0.77777**	**0.77777**	0.33333	0.66666	0.66666
Hong Kong	SRCC	**0.95054**	0.87912	0.61538	**0.93956**	0.16483	0.39560
	KRCC	**0.84615**	0.69230	0.48717	**0.82051**	0.12820	0.33333
Lanzhou	SRCC	0.20357	**0.68928**	-0.34642	0.22500	**-0.45357**	-0.38214
	KRCC	0.12381	**0.50476**	-0.25714	0.14285	**-0.27619**	-0.27619
Nanchang	SRCC	0.85770	**0.95355**	0.81324	**0.86067**	0.53359	0.67786
	KRCC	0.67588	**0.84189**	0.61264	**0.71541**	0.36758	0.50197
Shanghai	SRCC	**0.85031**	0.37126	0.52695	**0.67067**	-0.17964	**-0.67067**
	KRCC	**0.76376**	0.32732	0.47280	**0.47280**	-0.10910	**-0.47280**
Shenyang	SRCC	**0.81052**	**0.96691**	0.30075	0.52030	0.53684	0.60150
	KRCC	**0.64210**	**0.86315**	0.21052	0.38947	0.40000	0.46315
Tianjin	SRCC	0.33333	**0.80952**	**0.50000**	-0.30952	-0.09523	0.28571
	KRCC	0.21428	**0.71428**	**0.35714**	-0.28571	-0.07142	0.28571
Wuhan	SRCC	0.75454	**0.85454**	0.50000	0.77272	**0.80909**	0.79090
	KRCC	0.60000	**0.67272**	0.30909	0.56363	**0.67272**	0.63636
Average	SRCC	**0.71580**	**0.84691**	0.51925	0.61864	0.18293	0.30888
	KRCC	**0.58108**	**0.72085**	0.39531	0.48686	0.14002	0.25752

were used to evaluate the distances between subjective scores and the objective scores after a nonlinear regression. Following previous studies [8], [9], the nonlinear regression mapping function suggested in [61] was used in our experiments. It is defined as,

$$f(x) = \beta_1\left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2(x-\beta_3)}}\right) + \beta_4 x + \beta_5 \tag{19}$$

where $\beta_i, i = 1, 2, 3, 4, 5$ are parameters to fit. Since the 4 metrics have become the standard for assessing the performance of IQA indices, we do not discuss them in depth here. More details about them can be found in [8], [18].

5) Implementation Details of the VI and RI: In all experiments, we fixed all hyperparameters of our criterion. That is, for the VI, $C_1 = 0.45$, $C_2 = 160$, and $\alpha = 0.4$. For the RI, $C_3 = 0.85$, $C_4 = 130$, and $\beta = 0.02$. These values are empirically determined based on previous FR-IQA studies [7]–[9].

B. Feature Evaluation

In this experiment, we consider the transmission feature (Trans) and 5 widely used IQA features, i.e., the GM [8], PC [9], visual saliency (VS) [7], the I and Q channels of YIQ color space (ChromIQ) [9], and the M and N channels of LMN color space (ChromMN) [7]. Their performances are assessed on the hazy groups and the dehazing groups of the exBeDDE. The hazy groups were used to assess their abilities to evaluate visibility, while the dehazing groups were used to measure their abilities to evaluate realness. Table II and Table III provide the test results on the hazy groups and the dehazing groups, respectively. For each city and each feature, the KRCC and SRCC are provided. Moreover, the average performance of each feature is supplied. We take the sum of the KRCC and SRCC as the overall performance and sequentially highlight the top two performance values in red and blue.

As shown in Table II, Trans and the GM are the two features with the best fit for the visibility evaluation. This result is not surprising since both features can be well explained by the imaging model, Koschmieder's law, introduced in Sect. IV-A. Additionally, in Table III, PC and the ChromMN obtain the top two performance values in realness evaluation, which is strong evidence supporting our feature selections for the RI. Additionally, both positive and negative values are valid for the KRCC and SRCC, but the judgment of a good feature should be consistent, that is, always positive or negative. Therefore, although the ChromIQ may achieve good results in some cities in Table II, it cannot be regarded as a reasonable feature for evaluating visibility.

C. Metric Comparison

In this experiment, we illustrate the advantages of the proposed two criteria for dehazing evaluation. Since there was no specially designed FR-IQA metric for the evaluation of dehazing methods, we compared our criteria with 7 state-of-the-art FR-IQA metrics designed for the general IQA task and 4 no-reference dehazing evaluation metrics. The FR-IQA metrics are LPIPS [20], the VSI [7], the GMSD [8], the

VI: 1.000, RI: 1.000
LPIPS: 0.000, FADE: 0.000

VI: 0.779, RI: **0.978**
LPIPS: 0.241, FADE: 2.458

VI: 0.863, RI: 0.977
LPIPS: 0.271, FADE: 1.025

VI: 0.881, RI: 0.975
LPIPS: **0.229**, FADE: 1.264

VI: **0.907**, RI: 0.969
LPIPS: 0.391, FADE: **0.655**

VI: 1.000, RI: 1.000
LPIPS: 0.000, FADE: 0.000
(a) Reference image

VI: 0.861, RI: **0.954**
LPIPS: 0.285, FADE: 3.039
(b) Haze

VI: **0.893**, RI: 0.949
LPIPS: 0.351, FADE: 1.070
(c) MSCNN [55]

VI: 0.885, RI: 0.952
LPIPS: **0.282**, FADE: 2.235
(d) DCPDN [6]

VI: 0.885, RI: 0.944
LPIPS: 0.487, FADE: **0.685**
(e) DCP [2]

Fig. 10. Samples and scores of different IQA metrics. The images of (a) are clear references, and the images of (b) are hazy images. (c)∼(e) are dehazing results generated by MSCNN [55], DCPDN [6] and DCP [2]. The VI, RI, LPIPS [20] and FADE [12] scores are below each image. Additionally, for images in the first row of (b)∼(e), GMSD scores are 0.177, 0.138, 0.127 and **0.106**. VSI scores are 0.928, **0.960**, 0.949 and 0.951. FSIMc scores are 0.797, 0.859, 0.877 and **0.893**. FSIM scores are 0.798, 0.861, 0.878 and **0.893**. SSIM scores are 0.726, 0.699, 0.612 and **0.792**. PSNR scores are **23.28**, 17.75, 16.93 and 19.14. For images in the second row of (b)∼(e), GMSD scores are 0.144, 0.145, **0.136** and 0.144. VSI scores are 0.912, 0.926, **0.927** and 0.920. FSIMc scores are 0.847, 0.841, **0.852** and 0.835. FSIM scores are 0.848, 0.852, **0.857** and 0.848. SSIM scores are 0.659, **0.752**, 0.731 and 0.728. PSNR scores are 13.70, **20.00**, 16.27 and 18.94. Note that the lower LPIPS, FADE or GMSD scores are, the better. For other metrics, the higher scores are, the better.

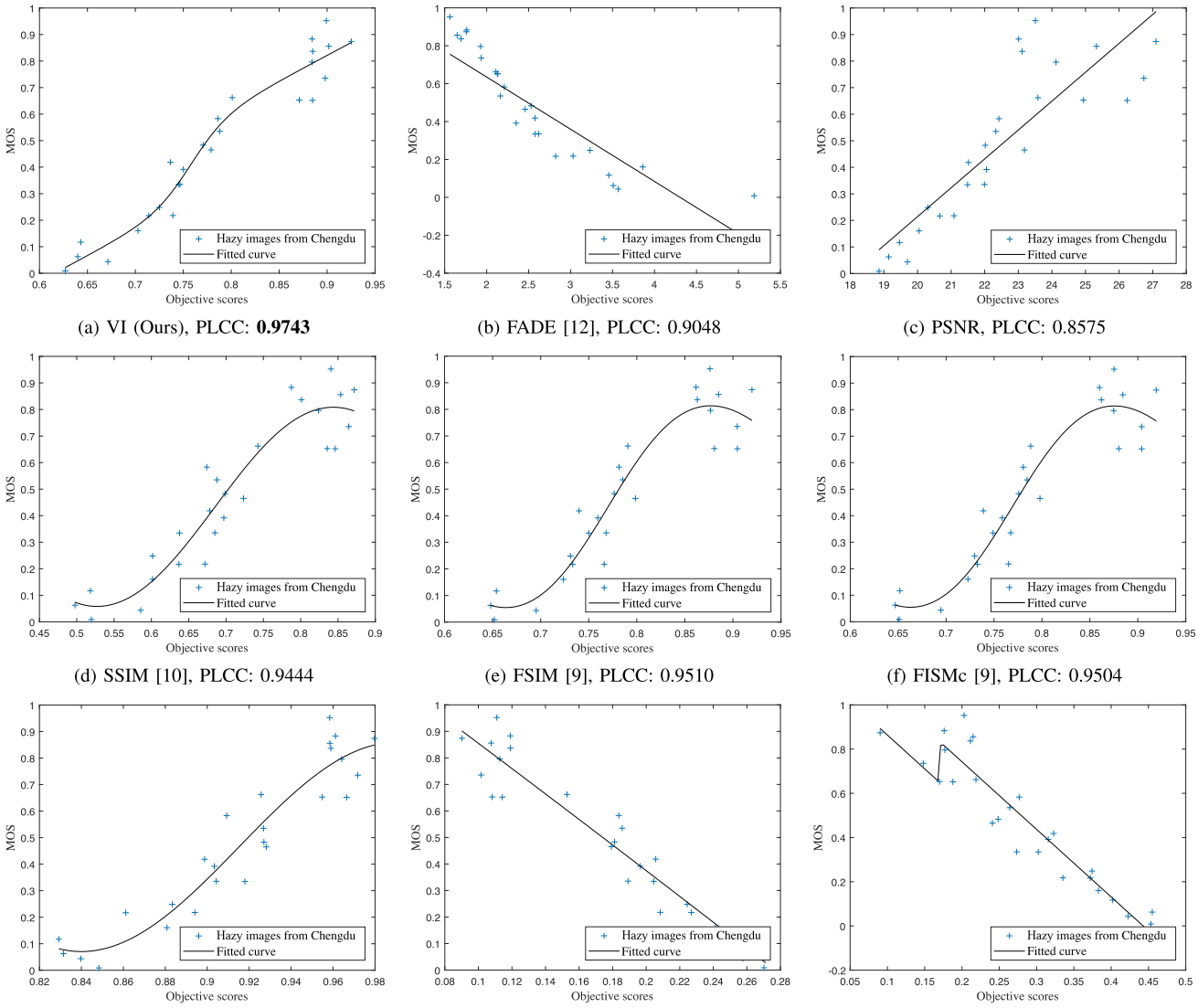
TABLE III

COMPARISONS OF DIFFERENT FEATURES IN DEHAZING GROUPS OF THE EXBEDDE. THE SRCC AND KRCC ARE CONSIDERED AS A WHOLE. THE TOP TWO PERFORMANCE VALUES IN TERMS OF THE SUM OF THE TWO METRICS ARE HIGHLIGHTED IN RED AND BLUE

City		Trans	GM	PC	VS	ChromIQ	ChromMN
Beijing	SRCC	-0.09873	0.60317	**0.82589**	**0.73700**	0.63314	0.67556
	KRCC	-0.08935	0.46612	**0.66497**	**0.57615**	0.49602	0.52265
Changsha	SRCC	**0.78144**	0.46798	**0.64400**	0.09155	0.63745	0.59172
	KRCC	**0.59767**	0.26774	**0.46881**	0.09430	0.47538	0.39091
Chengdu	SRCC	0.57495	0.35244	0.66903	**0.67118**	0.56856	**0.73533**
	KRCC	0.47147	0.30134	0.49169	**0.52477**	0.47102	**0.57332**
Hangzhou	SRCC	0.73799	0.66080	0.71870	0.76885	**0.82075**	**0.86323**
	KRCC	0.59025	0.48079	0.57817	0.59007	**0.67160**	**0.72431**
Hefei	SRCC	0.68413	0.62411	**0.75101**	0.28503	0.63223	**0.75514**
	KRCC	0.52408	0.47958	**0.60916**	0.21471	0.49434	**0.59934**
Hong Kong	SRCC	0.54934	0.43068	0.64932	0.32577	**0.66888**	**0.74030**
	KRCC	0.43600	0.33274	0.51202	0.28150	**0.50150**	**0.58724**
Lanzhou	SRCC	**0.74326**	**0.74326**	0.67402	0.66674	0.53921	0.73959
	KRCC	**0.58968**	**0.57183**	0.50914	0.52395	0.40187	0.56580
Nanchang	SRCC	0.56851	0.34872	**0.76050**	0.60635	0.33310	**0.66673**
	KRCC	0.44940	0.26482	**0.57592**	0.46624	0.21829	**0.52625**
Shanghai	SRCC	0.40050	0.55389	**0.77407**	0.56861	**0.76116**	0.67175
	KRCC	0.30706	0.43383	**0.57391**	0.43446	**0.61286**	0.51823
Shenyang	SRCC	0.60270	0.35769	**0.71788**	0.04526	0.34291	**0.60902**
	KRCC	0.45409	0.29087	**0.57099**	0.02349	0.23194	**0.46526**
Tianjin	SRCC	-0.16992	0.61655	**0.79211**	0.03000	0.60083	**0.65632**
	KRCC	-0.11418	0.43906	**0.69699**	0.00220	0.46212	**0.51258**
Wuhan	SRCC	0.71939	0.53183	**0.82381**	0.76862	**0.84373**	0.78290
	KRCC	0.58970	0.40321	**0.67125**	0.60220	**0.70390**	0.63873
Average	SRCC	0.50779	0.52426	**0.73336**	0.46375	0.61516	**0.70730**
	KRCC	0.40049	0.39433	**0.57692**	0.36117	0.47840	**0.55205**

FSIM [9], the FSIMc [9], the SSIM [10] and the PSNR. The no-reference metrics are FADE [12], e [11], $\overline{r}$ [11] and σ [11]. In our comparison, we applied masks for all the FR-IQA metrics so that they were not affected by inconsistent regions of image pairs. Regarding the NR-IQA metrics, we exploited the whole hazy image without a mask to maximize the information that they obtained so that they could achieve their best performance.

We report the average SRCC, KRCC, PLCC, and RMSE for the hazy and dehazing groups of the exBeDDE. Table IV



(a) VI (Ours), PLCC: **0.9743** (b) FADE [12], PLCC: 0.9048 (c) PSNR, PLCC: 0.8575

(d) SSIM [10], PLCC: 0.9444 (e) FSIM [9], PLCC: 0.9510 (f) FISMc [9], PLCC: 0.9504

(g) VSI [7], PLCC: 0.9486 (h) GMSD [8], PLCC: 0.9612 (i) LPIPS [20], PLCC: 0.9461

Fig. 11. Scatter plots of the MOSs (subjective scores) against the objective scores predicted by various IQA metrics on the hazy group of the city Chengdu. The PLCC score for each metric is provided under each figure.

provides the comparisons between our VI and other IQA metrics on hazy groups. Note that the results of e, $\overline{r}$, and σ do not appear in this table because they require both hazy and dehazed images but there are only hazy images in the hazy groups. Table V presents the results of our RI and other metrics on exBeDDE's dehazing groups. In Table VI, we also present the results of the realness evaluation on dehazing groups with different haze levels. Note that the haze level of a dehazing group is the same as the level of the source hazy image of this group. To illustrate the results, the samples and scores of different IQA metrics are exhibited in Fig. 10. In addition, Fig. 11 shows scatter plots of subjective scores against objective scores predicted by the VI and other IQA metrics on the hazy group of Chengdu.

Table IV and Table V show that our VI outperforms all the competitors by a large margin in terms of visibility evaluation, while the RI achieves the leading performance in the realness evaluation as well. Such results clearly suggest

TABLE IV

COMPARISONS BETWEEN OUR VISIBILITY INDEX AND OTHER IQA METRICS IN HAZY GROUPS OF THE EXBEDDE. THE TOP THREE PERFORMANCE VALUES ARE HIGHLIGHTED IN RED, BLUE AND BOLDFACE

Method	SRCC	KRCC	PLCC	RMSE
PSNR	0.6843	0.5625	0.7072	0.1593
SSIM [10]	0.7869	0.6654	**0.8629**	0.1108
FSIM [9]	**0.8054**	**0.6731**	**0.8634**	**0.1094**
FSIMc [9]	**0.7953**	0.6664	0.8600	0.1098
VSI [7]	0.7887	**0.6719**	0.8370	**0.1075**
GMSD [8]	**0.7953**	0.6586	0.8592	0.1113
LPIPS [20]	0.5468	0.4358	0.7045	0.1634
FADE [12]	0.7026	0.5539	0.6443	0.1622
VI (Ours)	**0.8805**	**0.7725**	**0.9296**	**0.0801**

that the proposed two criteria are more suitable for evaluating dehazing methods. In addition, LPIPS also achieves excellent performance on dehazing groups. In terms of the SRCC and KRCC, it even outperforms the RI. Considering these results, we regard this metric as a good alternative to our RI and

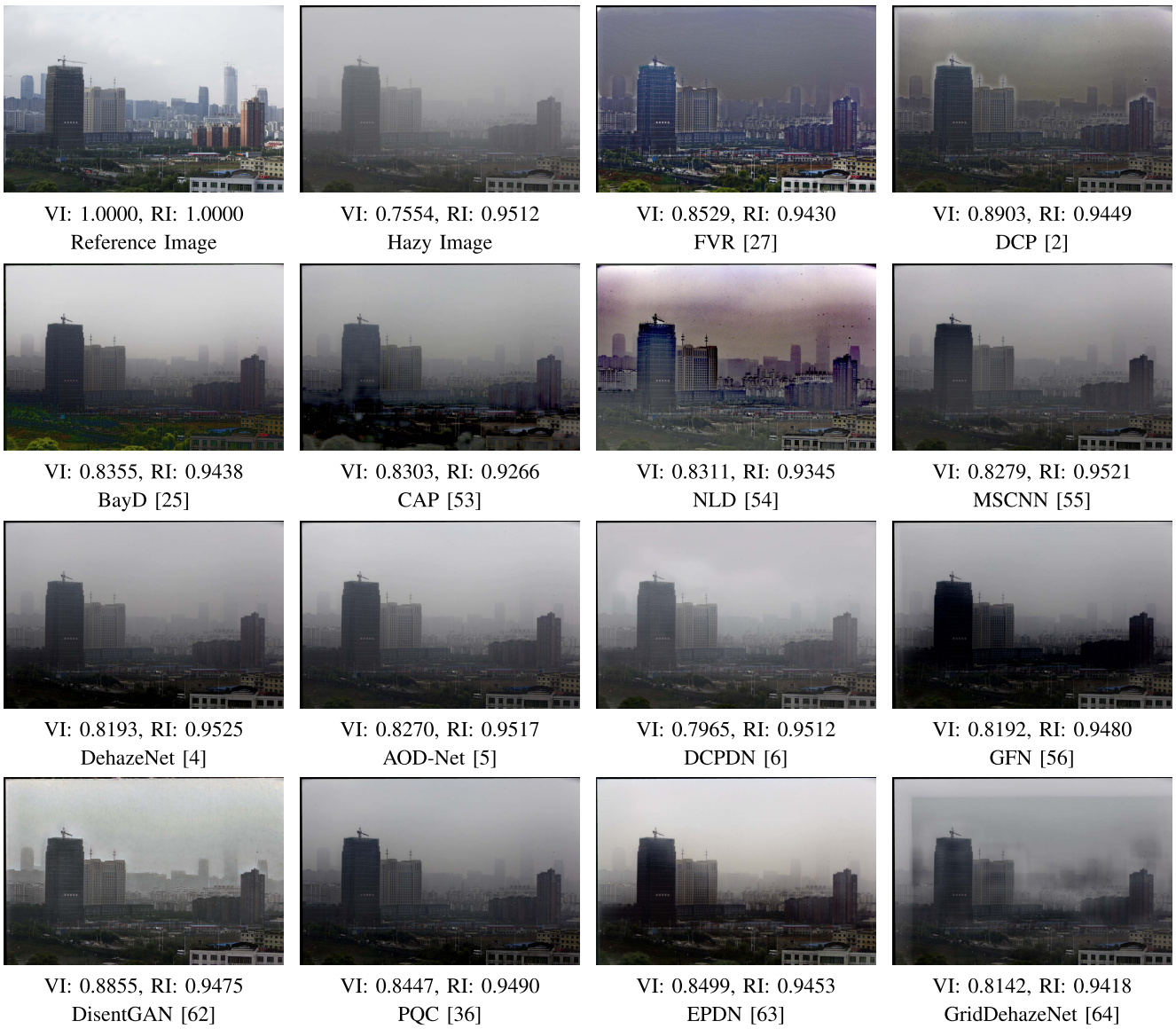


Fig. 12. Results of 14 dehazing methods for two BeDDE samples. The VI score and the RI score for each method are displayed below each image.

TABLE V

COMPARISONS BETWEEN OUR REALNESS INDEX AND OTHER IQA METRICS IN ALL THE DEHAZING GROUPS OF THE EXBEDDE. THE TOP THREE PERFORMANCE VALUES ARE HIGHLIGHTED IN RED, BLUE AND BOLDFACE

Method	SRCC	KRCC	PLCC	RMSE
PSNR	0.6276	0.4813	0.6217	0.1940
SSIM [10]	0.7167	0.5596	0.7602	0.1639
FSIM [9]	0.5848	0.4512	0.6954	0.1757
FSIMc [9]	0.6005	0.4636	0.6962	0.1750
VSI [7]	0.7026	0.5512	**0.8199**	**0.1346**
GMSD [8]	0.6247	0.4775	0.7239	0.1723
LPIPS [20]	**0.8358**	**0.6876**	**0.8676**	**0.1272**
e [11]	0.4715	0.3162	0.5997	0.2076
$\overline{r}$ [11]	0.5565	0.4005	0.7240	0.1756
σ [11]	0.048	0.0230	0.2751	0.2515
FADE [12]	**0.7675**	**0.6097**	0.8010	0.1557
RI (Ours)	**0.7743**	**0.6170**	**0.8731**	**0.1177**

report the LPIPS scores of different dehazing methods as dehazing baselines in Sect. V-D. In addition, although LPIPS [20] and FADE [12] have good performances for the realness evaluation, they fail to achieve similar performances for the visibility evaluation. This phenomenon further supports our hypothesis that dehazing evaluation should be considered from two separate aspects due to its dual characteristics.

As shown in Fig. 10, the images in the first row of (b)∼(e) are increasingly clear in visibility. Our VI is able to rank them correctly, while the majority of other metrics fail. Moreover, (b) contains natural images without degradations. The images of (c) and (d) bear slight degradations with some artifacts. Additionally, the images of (e) suffer serious degradation with obvious halos. Our RI can judge these differences and provide results consistent with human perceptions, whereas other metrics seem to be less plausible. In Fig. 11, our VI obtains a better PLCC value, and the predicted objective scores correlate better with the subjective scores (or MOSs). All these results demonstrate the superiority of the VI and RI for dehazing evaluation.

D. Dehazing Baselines

In this experiment, we provided the VI, RI and LPIPS [20] scores of 14 dehazing methods on the BeDDE as baselines for

TABLE VI

COMPARISONS BETWEEN OUR REALNESS INDEX AND OTHER IQA METRICS IN DEHAZING GROUPS OF DIFFERENT HAZE LEVELS. NOTE THAT THE HAZE LEVEL OF A DEHAZING GROUP IS THE SAME AS THE LEVEL OF THE SOURCE HAZY IMAGE OF THIS GROUP. THE TOP THREE PERFORMANCE VALUES ARE HIGHLIGHTED IN RED, BLUE AND BOLDFACE

		PSNR	SSIM [10]	FSIM [9]	FSIMc [9]	VSI [7]	GMSD [8]	LPIPS [20]	e [11]	$\overline{r}$ [11]	σ [11]	FADE [12]	RI (Ours)
Light Groups	SRCC	0.6461	0.6922	0.6859	0.6882	0.7972	0.7546	**0.8696**	0.4629	0.5842	0.0229	**0.7818**	**0.8138**
	KRCC	0.5009	0.5367	0.5421	0.5438	**0.6497**	0.5927	**0.7324**	0.3022	0.3867	0.0020	0.6289	**0.6684**
	PLCC	0.6143	0.7353	0.8001	0.8017	**0.8631**	0.7898	**0.8982**	0.5621	0.7635	0.2236	0.7984	**0.8833**
	RMSE	0.1963	0.1733	0.1533	0.1524	**0.1216**	0.1588	**0.1146**	0.2191	0.1641	0.2592	0.1576	**0.1159**
Medium Groups	SRCC	0.6622	0.7272	0.5647	0.5888	0.7203	0.6295	**0.8321**	0.4461	0.5420	0.0933	**0.7753**	**0.7847**
	KRCC	0.5075	0.5722	0.4384	0.4565	0.5641	0.4777	**0.6830**	0.3021	0.3966	0.0603	**0.6046**	**0.6113**
	PLCC	0.6741	0.7673	0.1816	0.6729	**0.8334**	0.7225	**0.8667**	0.6060	0.7046	0.2871	0.8122	**0.8875**
	RMSE	0.1843	0.1614	0.6678	0.1811	**0.1334**	0.1724	**0.1279**	0.2043	0.1805	0.2488	0.1511	**0.1143**
Heavy Groups	SRCC	0.5176	0.6296	0.1994	0.2334	0.4063	0.0326	**0.7732**	0.4168	0.4297	0.1419	**0.6732**	**0.6289**
	KRCC	0.3816	0.4729	0.1401	0.1645	0.2878	0.0343	**0.6035**	0.2707	0.3424	0.0926	**0.5238**	**0.4957**
	PLCC	0.5418	0.7287	0.4095	0.3560	0.6202	0.5159	**0.8001**	0.5538	0.5982	0.3882	**0.7614**	**0.7812**
	RMSE	0.2112	0.1657	0.2259	0.2267	0.1714	0.2032	**0.1536**	0.2118	0.2054	0.2361	**0.1628**	**0.1356**

TABLE VII

QUANTITATIVE COMPARISONS OF DEHAZING METHODS ON IMAGES OF DIFFERENT HAZE LEVELS. TOP THREE PERFORMANCE VALUES ARE HIGHLIGHTED IN RED, BLUE AND BOLDFACE. NOTE THAT THE MSCNN, DEHAZENET, AOD-NET, DCPDN AND GFN ARE CNN-BASED METHODS. "HAZE" REFERS TO THE ORIGINAL HAZY IMAGES. "ALL IMAGES" REFERS TO ALL HAZY IMAGES

Method	Light			Medium			Heavy			All Images		
	VI	RI	LPIPS [20]	VI	RI	LPIPS [20]	VI	RI	LPIPS [20]	VI	RI	LPIPS [20]
FVR [27]	0.8079	0.9572	0.4766	0.8001	0.9500	0.5039	**0.8188**	0.9327	0.5195	0.8054	0.9511	0.4976
DCP [2]	**0.9354**	0.9740	0.4155	**0.9027**	0.9638	0.4502	**0.8543**	0.9401	0.5059	**0.9111**	0.9654	0.4469
BayD [25]	0.8432	0.9323	0.5006	0.8288	0.9367	0.4510	0.7795	0.9286	0.4855	0.8294	0.9340	0.4707
CAP [53]	0.8607	0.9496	0.3231	0.8536	0.9492	0.3152	0.8005	0.9385	0.3586	0.8507	0.9482	0.3233
NLD [54]	0.8364	0.9601	0.4758	0.8192	0.9559	0.5135	**0.8321**	0.9381	0.5688	0.8278	0.9557	0.5093
MSCNN [55]	0.9240	0.9775	0.2700	0.8842	0.9687	0.2877	0.8042	0.9492	0.3611	0.8920	0.9702	0.2920
DehazeNet [4]	0.9304	**0.9793**	0.2575	0.8772	**0.9700**	0.2610	0.7935	**0.9501**	**0.3315**	0.8902	**0.9718**	0.2692
AOD-Net [5]	**0.9312**	0.9778	**0.2495**	**0.8868**	0.9686	**0.2559**	0.8032	0.9489	0.3332	**0.8961**	0.9703	**0.2641**
DCPDN [6]	**0.9326**	**0.9802**	**0.2061**	0.8833	**0.9688**	**0.2308**	0.7931	**0.9520**	**0.3060**	0.8940	**0.9717**	**0.2332**
GFN [56]	0.9076	0.9750	0.3135	0.8493	0.9609	0.3581	0.7793	0.9459	0.4268	0.8659	0.9651	0.3535
DisentGAN [62]	0.8803	0.9696	0.4032	0.8620	0.9579	0.4292	0.8445	0.9363	0.4746	0.8678	0.9604	0.4186
PQC [36]	0.9274	0.9765	0.2598	0.8821	0.9677	0.2720	0.8030	0.9497	0.3345	0.8923	0.9694	0.2738
EPDN [63]	0.9270	0.9744	0.2651	**0.8869**	0.9608	0.2951	0.8177	0.9459	0.3655	**0.8960**	0.9649	0.2903
GridDehazeNet [64]	0.9281	0.9788	0.2629	0.8671	0.9647	0.2995	0.7868	0.9474	0.3751	0.8837	0.9687	0.2926
Haze	0.9138	**0.9808**	**0.1928**	0.8240	**0.9703**	**0.2267**	0.7361	**0.9511**	**0.3270**	0.8518	**0.9726**	**0.2295**

related studies. These methods are FVR [27], DCP [2], BayD [25], CAP [53], NLD [54], MSCNN [55], DehazeNet [4], AOD-Net [5], DCPDN [6], GFN [56], the disentangled dehazing network (DisentGAN) [62], the patch quality comparator (PQC) [36], the Enhanced Pix2pix Dehazing Network (EPDN) [63], and GridDehazeNet [64]. In addition, the evaluation results of original hazy images are provided. Their average scores on hazy images of different haze levels are presented in Table VII as quantitative comparisons. The dehazing results of two samples from the BeDDE along with their scores are shown in Fig. 12 as qualitative comparisons. From these results, we have several interesting findings.

First, non-CNN methods, such as FVR [27] and NLD [54], demonstrated in Fig. 12, are more likely to over-enhance the contrast of hazy images. Thus, they produce many artifacts that seriously degrade the realness of the restored images. In contrast, CNN-based methods are able to produce results that are close to natural images. Therefore, almost all CNN-based methods outperform the non-CNN-based methods in terms of the RI.

Second, the recently proposed CNN-based GFN [56] performs worse than the other CNN-based dehazing methods in terms of both the VI and the RI. There are several potential causes. On the one hand, the GFN fuses three traditional image enhancement techniques (i.e., white balance, contrast enhancement, and gramma correction) together to generate the dehazed image based on the weights provided by CNNs. However, traditional enhancement techniques are not suitable for dehazing because they are unable to effectively handle the degradation caused by haze which is highly correlated with the depth of the scene. On the other hand, the training set and the test set of the GFN are all simulated using the indoor images of NYU2. Therefore, GFN may perform well on its own test set due to overfitting, but it fails to handle real images of the BeDDE.

Third, DCP [2] achieves the best performance in terms of the VI, even though CNN-based methods seem to produce more appealing images. This result is still reasonable because we have been accustomed to seeing natural phenomena such as fog and haze and our eyes are more sensitive to distortions in images. In regard to judging the image quality without special requirements, we are more likely to notice artifacts than haze. However, if we focus on what we can see clearly in the scene, the output images of DCP might be better.

VI. CONCLUSION

In this paper, we focus on how to evaluate the performance of dehazing algorithms and establish two benchmark datasets, the BeDDE and exBeDDE. The BeDDE is the first dataset in this field to consist of real-world hazy images and the corresponding clear references. As an extension of the BeDDE, the exBeDDE provides subjective scores for both hazy and

dehazed images and is designed to assess dehazing evaluation metrics. Being aware of the dual characteristics of dehazing, we propose two criteria, the VI and the RI, to evaluate dehazing methods in visibility and realness, respectively. Through extensive experiments, the effectiveness of the VI and RI is verified and guaranteed. Moreover, the evaluation results of 14 dehazing methods using our criteria are provided as baselines for relevant studies. In the future, we will enlarge the BeDDE with more hazy and haze-free pairs. Moreover, we will consider more advanced ways to evaluate dehazing methods, such as merging the VI and RI into a robust holistic index.

REFERENCES

[1] G. J. van Oldenborgh, P. Yiou, and R. Vautard, "On the roles of circulation and aerosols in the decline of mist and dense fog in europe over the last 30 years," *Atmos. Chem. Phys.*, vol. 10, no. 10, pp. 4597–4609, 2010.

[2] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[3] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 1–14, Dec. 2014.

[4] 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 Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[5] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-net: All-in-One dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4780–4788.

[6] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.

[7] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[8] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[9] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[11] N. Hautière, J.-P. Tarel, D. Aubert, and É. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Anal. Stereol.*, vol. 27, no. 2, pp. 87–95, 2011.

[12] L. Kwon Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.

[13] W. E. K. Middleton, "Vision through the atmosphere," in *Geophysik II/Geophysics II*, J. Bartels, Ed. Berlin, Germany: Springer, 1957.

[14] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 754–762.

[15] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Berlin, Germany: Springer, 2018, pp. 620–631.

[16] M. Bijelic, P. Kysela, T. Gruber, W. Ritter, and K. Dietmayer, "Recovering the unseen: Benchmarking the generalization of enhancement methods to real world data in heavy fog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 11–21.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[18] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[19] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[21] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, p. 72, 2008.

[22] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[23] J. Kopf *et al.*, "Deep photo: Model-based photograph enhancement and viewing," *ACM Trans. Graph.*, vol. 27, no. 5, p. 116, 2008.

[24] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, Dec. 2009, pp. 1956–1963.

[25] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 263–278, Jul. 2012.

[26] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 617–624.

[27] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2201–2208.

[28] C. O. Ancuti and C. Ancuti, "Single image dehazing by multi-scale fusion," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3271–3282, Aug. 2013.

[29] L. He, J. Zhao, N. Zheng, and D. Bi, "Haze removal using the difference-structure-preservation prior," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1063–1075, Mar. 2017.

[30] A. Galdran, A. Bria, A. Alvarez-Gila, J. Vazquez-Corral, and M. Bertalmio, "On the duality between retinex and image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8212–8221.

[31] K. Ma, W. Liu, and Z. Wang, "Perceptual evaluation of single image dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3600–3604.

[32] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, "Consistent depth maps recovery from a video sequence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 974–988, Jun. 2009.

[33] S. Salazar-Colores, E. Cabal-Yepez, J. M. Ramos-Arreguin, G. Botella, L. M. Ledesma-Carrillo, and S. Ledesma, "A fast image dehazing algorithm using morphological reconstruction," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2357–2366, May 2019.

[34] Q. Liu, X. Gao, L. He, and W. Lu, "Single image dehazing with depth-aware non-local total variation regularization," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5178–5191, Oct. 2018.

[35] T. M. Bui and W. Kim, "Single image dehazing using color ellipsoid prior," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 999–1009, Feb. 2018.

[36] S. Santra, R. Mondal, and B. Chanda, "Learning a patch quality comparator for single image dehazing," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4598–4607, Sep. 2018.

[37] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8202–8211.

[38] A. Wang, W. Wang, J. Liu, and N. Gu, "AIPNet: Image-to-Image single image dehazing with atmospheric illumination prior," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 381–393, Jan. 2019.

[39] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760.

[40] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[41] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.* Berlin, Germany: Springer, 2014, pp. 31–42.

[42] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 478–485.

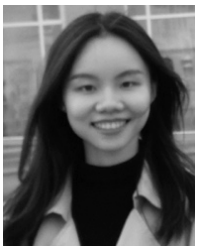
[43] J.-P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer, "Vision enhancement in homogeneous and heterogeneous fog," *IEEE Intell. Transp. Syst. Mag.*, vol. 4, no. 2, pp. 6–20, Summer 2012.

[44] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-HAZY: A dataset to evaluate quantitatively dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2226–2230.

[45] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.
[46] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.
[47] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.
[48] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[49] S. Zhao, L. Zhang, S. Huang, Y. Shen, S. Zhao, and Y. Yang, "Evaluation of defogging: A real-world benchmark dataset, a new criterion and baselines," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1840–1845.
[50] M. Colomb, J. Dufour, M. Hirech, P. Lacôte, P. Morange, and J.-J. Boreux, "Innovative artificial fog production device-A technical facility for research activities," in *Proc. Int. Conf. Fog, Fog Collection Dew*, 2004, pp. 1–5.
[51] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[52] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.
[53] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.
[54] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.
[55] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 154–169.
[56] W. Ren *et al.*, "Gated fusion network for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.
[57] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.
[58] M. C. Morrone, J. Ross, D. C. Burr, and R. Owens, "Mach bands are phase dependent," *Nature*, vol. 324, no. 6094, pp. 250–253, Nov. 1986.
[59] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.
[60] J.-M. Geusebroek, R. Van Den Boomgaard, A. W. M. Smeulders, and A. Dev, "Color and scale: The spatial structure of color images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2000, pp. 331–341.
[61] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[62] X. Yang, Z. Xu, and J. Luo, "Towards perceptual image dehazing by physics-based disentanglement and adversarial training," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[63] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced Pix2pix dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8160–8168.
[64] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.

Shiyu Zhao received the B.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2017, where he is currently pursuing the master's degree. His research interests are visibility enhancement for bad weather images, scene understanding, and machine learning.

Lin Zhang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University. In August 2011, he joined the School of Software Engineering, Tongji University, Shanghai, where he is currently a Full Professor. His current research interests include the environment perception of intelligent vehicles, pattern recognition, computer vision, and perceptual image/video quality assessment.



Shuaiyi Huang received the B.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2017. She is currently pursuing the master's degree with the School of Information Science and Technology, ShanghaiTech University. Her research interests include image visual correspondence, scene understanding, and machine learning.

Ying Shen (Member, IEEE) received the B.S. and M.S. degrees from the School of Software Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2012. In 2013, she joined the School of Software Engineering, Tongji University, Shanghai, where she is currently an Associate Professor. Her research interests include bioinformatics and pattern recognition.

Shengjie Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.S. degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004. He is currently a Professor with the School of Software Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research at Lucent Technologies, Whippany, NJ, USA, and China Aerospace Science and Industry Corporation, Beijing. His research interests include big data, wireless communications, image processing, and signal processing. He is a fellow of the Thousand Talents Program of China.