Image Dehazing Assessment: A Real-World Dataset and a Haze Density-Aware Criteria

Jiyou Chen, Gaobo Yang, Shengchun Wang, Dewang Wang, and Xin Liao

***Abstract*—Full-reference image dehazing quality assessment (FR-IDQA) evaluates the visual quality of a dehazed image by measuring its differences with a clear reference. The existing FR-IDQA methods are not convincing due to the lack of well-aligned datasets of hazy and clear image pairs and the limited hand-crafted features make it difficult to simulate the complicated perception by the human visual system (HVS). In this work, we build a real-world image dataset, namely RW-Haze, which comprises natural hazy images and their well-aligned clear references. Each clear image is paired with several hazy images with diverse haze levels from slight to heavy. Meanwhile, the existing FR-IDQA works evaluate the dehazed image quality in a global manner, without considering local haze distributions in the original hazy image. Actually, the perceived haze in a natural hazy image is not uniformly distributed, and the haze density varies with scene depth. Based on this priori observation, we design a haze density-aware convolutional neural network (CNN), namely DehIQA, for FR-IDQA. It adopts transfer learning to alleviate the issue of lacking sufficient labeled data. Specifically, we divide image dehazing assessment into two tasks. The source task is to classify unpaired clear and hazy images, which enforces the deep network to learn haze-related features. The target task is image quality assessment, which is achieved by transferring the trained model for the source task to the target task. Considering the fact that the perceived distortion in a dehazed image is also not uniform, we present a haze density-aware mechanism into DehIQA, which assigns different weights for different local regions in a dehazed image in terms of the dark channel of the original hazy image. Extensive experimental results show that DehIQA outperforms the state-of-the-art (SOTA) works on the benchmark dataset and achieves better consistency with human perceptions.**

***Index Terms*—Dehazing assessment metrics, transfer learning, benchmark dataset.**

Manuscript received 25 April 2023; revised 25 November 2023; accepted 12 December 2023. Date of publication 20 December 2023; date of current version 4 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62372164, 61972143, and 62272160, in part by the 14th Five-Year Plan" Key Disciplines and Application-oriented Special Disciplines of Hunan Province (Xiangjiaotong [2022] 351), and in part by the Science Foundation of Hengyang Normal University of China under Grant 2023QD14. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Lamberto Ballan. *(Correspondence author: Gaobo Yang.)*

Jiyou Chen is with the School of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China, and also with the School of Computer Science and Technology, Hengyang Normal University, Hengyang 421008, China (e-mail: jiyouchen@hnu.edu.cn).

Gaobo Yang, Dewang Wang, and Xin Liao are with the School of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: yanggaobo@hnu.edu.cn; dewang_wang@126.com; xinliao@hnu.edu.cn).

Shengchun Wang is with the College of Information Science and Electronic Engineering, Hunan Normal University, Changsha 410081, China (e-mail: scwang@hunnu.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3345141

I. Introduction

IMAGE dehazing has attracted wide research attention, and many works have achieved impressive results. However, image dehazing assessment still falls behind. A reliable full-reference image dehazing quality assessment (FR-IDQA) metric promotes image dehazing research by making comparisons among various image dehazing works. Meanwhile, due to the lack of real-world datasets that contain sufficient hazy and clear image pairs, most dehazing works were only evaluated on synthetic hazy images [1], [2], [3]. They achieve desirable results on synthetic hazy images, but behave poorly for real-world hazy images. Moreover, the existing metrics, which include PSNR, SSIM, visibility index (VI), and realness index (RI), have been shown to be not always consistent with human perception when they are used to measure the visual quality of a dehazed image [4]. As shown in Fig. 1, the dehazed images in the left column are much better than the ones in the right, whereas their PSNRs are relatively lower.

In general, FR-IDQA should be conducted by comparing a dehazed image from a real-world hazy image with its clear reference. Users can select the best image dehazing work for practice dehazing tasks. However, it is extremely difficult or almost impossible to collect an image dataset of real-world hazy images and their well-aligned clear references. It is well-known that the real weather condition for a scene at a specific moment is either hazy or not, but the concept of well-aligned implies that a hazy image and its clear one share the same spatial alignment and background content. To ensure this, a camera should be fixed when capturing them, and there are no other changes except for weather conditions. However, it is easy to ensure hazy and clear image pairs share the same spatial alignment by using a fixed camera, but not trivial to keep the same background content. As we know, a hazy image and its clear one are usually captured at different moments or even different days, easily leading to some changes in the image background due to falling or fallen leaves and moving objects. Especially, if they are captured over long time intervals, the background changes will be noticeable. Though there are a few real-world datasets [5], [6] collected for image dehazing and its evaluation, there still exists the so-called misalignment issue between hazy images and their clear references. Apparently, this is not acceptable for the existing pixel-based metrics such as PSNR and RMSE.

In the literature, there exist several FR-IDQA approaches [3], [5], [6], [7]. They share the similarity that some hand-crafted features are designed by exploiting some HVS-related prior knowledge. Nevertheless, since human visual perception is a

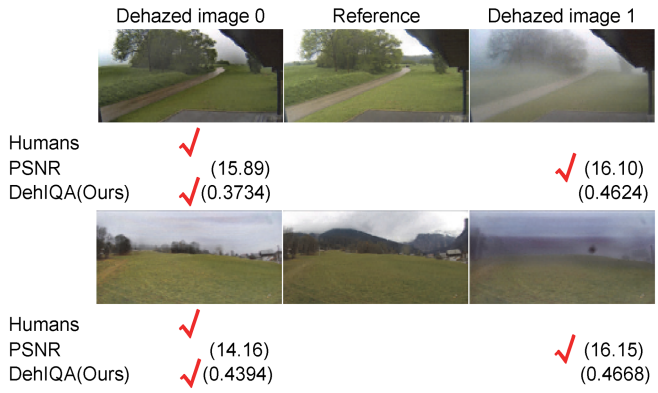

Fig. 1. Which dehazed image (left or right) is "closer" to the reference in two examples? Apparently, the traditional metric PSNR is not consistent with human perceptions, whereas the proposed DehIQA is completely consistent. Note that a high PSNR indicates better subjective image quality, and the opposite is true for DehIQA.

complicated process, limited features make it difficult to simulate HVS well. Moreover, they compute the distance of the features computed from a dehazed image and its clear reference in a global manner, without considering the original hazy image. However, since the dehazed image is restored from its original hazy image, haze distribution has a direct impact on possible artifacts or left haze residues in the local regions of a dehazed image. Fig. 2 shows several hazy images and their dehazed ones. We can observe that if a local region of the hazy image has dense haze, there are more artifacts or haze residues left in the corresponding local region of the dehazed image. In recent years, several CNN-based works have been proposed for full-reference image quality assessment (FR-IQA) [8], [9], [10], which are widely-used in many image applications such as lossy compression. Nevertheless, there is still no CNN-based FR-IDQA approach. The reason behind this is that training a CNN-based model for FR-IDQA requires a large-scale real-world dataset with pairwise labeled data and their mean opinion score (MOS), which is both time-consuming and cumbersome to collect.

In this work, we establish a real-world benchmark dataset, namely RW-Haze, and propose a haze density-aware CNN, namely DehIQA, for FR-IDQA. The RW-Haze dataset is made up of clear images and their well-aligned hazy ones with different haze densities, and their spatial resolutions are 1440×2560. Each image pair of a hazy image and its clear reference was collected by a fixed camera from a weather station to ensure the same spatial alignment. Moreover, each image pair was captured within a short interval, so that there were almost no background changes except subtle environmental conditions. Note that our previous conference version of this dataset [11] is extended and incorporated into this work to explain the necessity of developing DehIHA and its practical effectiveness for evaluating a dehazed image from a real-world hazy image. To alleviate the issue of lacking sufficient image pairs for training DehIHA, we introduce transfer learning to exploit unpaired data. The whole task is split into the source task and the target task. The source task is image classification, which trains the model to classify unpaired clear and hazy images that are easily collected from natural scenes. In essence, it is to enforce the model to learn haze-relevant features. The target task is image quality assessment, which is achieved by transferring the trained model for the source task to image dehazing evaluation. As claimed earlier, for the region with thick haze in a hazy image, more noticeable artifacts or haze residues are left in the dehazed image. Due to the perceived non-uniform distortions in a dehazed image, we present a haze density-aware module for DehIQA, which assigns different weights for different local regions in a dehazed image in terms of the dark channel of the original hazy image. Specifically, the proposed DehIQA model is a triple-stream network. Each stream inherits the structure and parameters of the classifier network. The dark channel map extracted from the hazy image, the reference clear image and the dehazed image are fed into DehIQA simultaneously to learn their features, respectively. Then, the features of the dark channel map are multiplied by the features of the dehazed images and reference images to calculate the final score for FR-IDQA. The main contributions of this work are three-fold.

- A real-world dataset, namely RW-Haze, is established for FR-IDQA by collecting clear and hazy images from six cities in China. For each scene, a clear image and multiple hazy images with different haze densities from slight to heavy were captured from the same scene with almost no activity, which ensures image pairs are well-aligned.
- A haze density-aware CNN-based DehIQA is presented for FR-IDQA by comprehensively considering the correlation between the distorted regions of the dehazed image and the haze distribution in the raw hazy image. The haze density-aware mechanism enables DehIQA to pay more attention to the local regions with more distortions or haze residues in a dehazed image.
- A simple yet effective transfer learning strategy is introduced for DehIQA, which addresses the issue of insufficient training samples with labels. Extensive experiments show that DehIQA outperforms the state-of-the-art (SOTA) FR-IDQA works.

II. Related Works

In this section, we briefly introduce the existing hazy image datasets for image dehazing and its assessment, and summarize the existing IDQA methods and the representative CNN-based IQA works.

A. Hazy Image Datasets

The existing hazy image datasets can be divided into two categories.

1) *Haze generated by haze machine:* Ancuti et al. built four hazy image datasets including O-HAZE [12], I-HAZE [13], Dense-haze [14] and NH-HAZE [15] for image dehazing and its assessment. Fig. 3 shows some examples from these four datasets. However, the haze was simulated by a professional haze machine, which is still far away from the natural haze. As we know, natural haze is a complex compound of various small particles such

Fig. 2. Illustrate the possible distortion or left haze residue in the dehazed and its relationship with the original hazy image. The denser haze in the local regions of the hazy image, the more distortion or haze residue in the corresponding regions of the dehazed image.

Fig. 3. Example images from O-HAZE [12], I-HAZE [13], Dense-haze [14], and NH-HAZE [15]. (a) is the reference, and (b) is the corresponding hazy images. The hazy images and corresponding references are almost components aligned with spatial and background. However, since the haze is machine-generated, the haze of hazy images looks unnatural and more like smoke.

as vapor, dust and aerosol in the air. Because different particles absorb and scatter light in different ways, it is not easy to simulate natural haze with a professional haze machine.

2) *Natural haze images:* There are a few works [5], [6], [11] dedicated to collecting image pairs from real-world scenes. The BeDDE [5] dataset was constructed by collecting 208 real-world image pairs in one year for image dehazing evaluation. Due to a slight change in camera viewpoint, raw image pairs are not well spatially aligned and have no uniform background. Thus, image registration is used to align partial image regions, and the masks are manually labeled to delineate their common regions of interest, which are used for FR-IDQA. The MRFID dataset [6] has paired hazy and clear images from 200 outdoor scenes. Each clear image has 4 hazy images, which were picked out from the AMOS dataset [16] that was collected by static webcams from 2006 to 2017. The hazy images are spatially aligned with their clear references, but there were long intervals when capturing them, leading to background content changes. Figs. 4 and 5 show some hazy images and clear references from BeDDE and MRFID, respectively.

TABLE I
COMPARISONS OF RW-HAZE AND THE EXISTING REAL-WORLD HAZY IMAGE DATASETS

Dataset	clear scene	Haze	Quantity	Aligned
O-HAZE [12]	Real-world	Artificial	45	Yes
I-HAZE(Indoor) [13]	Real-world	Artificial	35	Yes
Dense-haze [14]	Real-world	Artificial	33	Yes
NH-HAZE [15]	Real-world	Artificial	55	Yes
BeDDE [5]	Real-world	Natural	208	Manual
MRFID [6]	Real-world	Natural	800	Yes
RW-Haze(Ours)	Real-world	Natural	210	Yes

Table I compares the proposed RW-Haze and the existing real-world hazy image datasets. There is still rich room for improving the real-world image datasets for dehazing and its assessment. The hazy images and clear references are almost completely aligned in O-HAZE [12], I-HAZE [13], Dense-haze [14] and NH-HAZE [15], but the haze is artificial. While the haze is

Fig. 4. Example images from BeDDE [5]. (a) is the reference, and (b) is the corresponding hazy images. The hazy images are misaligned with their reference due to changes in viewpoints and contents during data collection, the yellow and red boxes highlight the misalignment in the physical space and background, respectively.

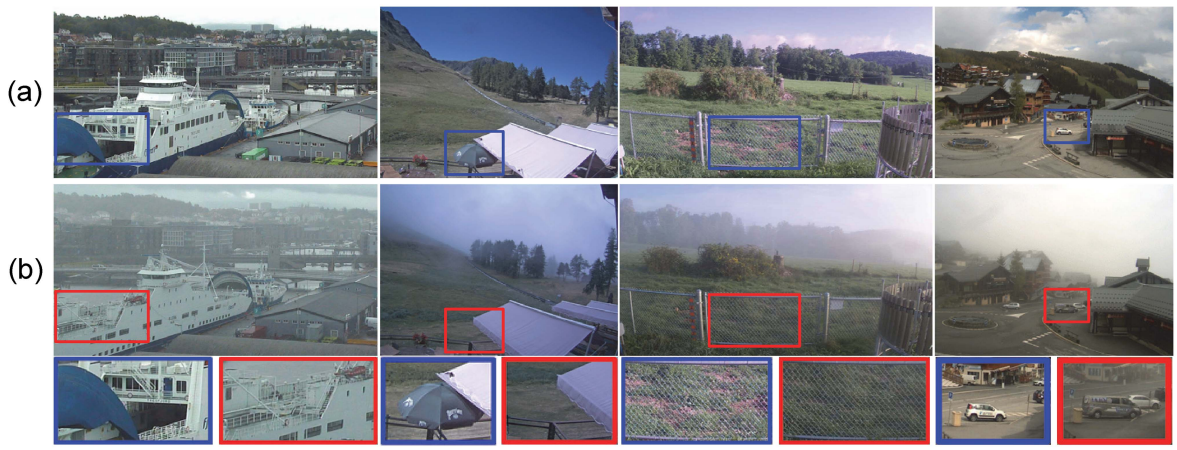

Fig. 5. Example images from MRFID [6]. (a) is the reference, and (b) is the corresponding hazy images. The hazy images are well aligned with their reference in physical space, while the background changes due to the long acquisition interval. The obvious changes between the hazy image and the corresponding reference are highlighted with red and blue boxes, respectively.

natural in BeDDE [5] and MRFID [6], their image pairs are not well-aligned in either spatial or background. The proposed RW-Haze is similar or smaller in scale to the existing datasets, but it was collected from natural scenes, and the hazy images are still well-aligned with their clear references. In fact, a desirable benchmark dataset with well-aligned image pairs is a prerequisite for FR-IDQA. In contrast, the quantity of image pairs is not so much required.

B. Existing Metrics for Image Dehazing Evaluation

The widely-used IQA metrics including PSNR and SSIM are simple to understand, but consider not the special requirements of dehazing [17]. Choi et al. [18] proposed a no-reference IQA metric by predicting the density of the residual haze in a dehazed image, which was named the Fog-Aware Density Evaluator (FADE). Nevertheless, FADE still provides a high score for a dehazed image with severe color distortion due to over-dehazing. Guo et al. [7] proposed an objective dehazing assessment method by using a synthetic hazy image as the reference. Min et al. [19] designed three groups of hand-crafted features, namely haze-removing features, structure-preserving features and over-enhancement features, for FR-IDQA. These features capture the most key aspects related to dehazing. However, it is conducted between a dehazed image and its reference hazy image. Zhao et al. [5] addressed IDQA from the restoration perspectives of visibility and realness, and presented the visibility index (VI) and the realness index (RI). Moreover, Liu et al. [6] developed a Fog-relevant Feature-based SIMilarity index (FRFSIM) for FR-IDQA. Specifically, both dark channel features and mean subtracted contrast normalized features are defined to measure the changes in haze density, whereas gradient similarity and chroma similarity are defined to measure texture distortions and color distortions.

C. Learning-Based FR-IQA Models

Benefiting from the developments of deep learning, many learning-based FR-IQA models were presented, which exploit deep models to learn features from training data without prior knowledge. Nevertheless, to the best of our knowledge, there

is still no learning-based FR-IDQA work. Since a dehazed image is still a restored image, IDQA can be treated as a special type of IQA. Thus, we briefly summarize the learning-based models for IQA. Gao et al. [20] proposed a DeepSim model for FR-IQA by measuring the local similarities between the learned features from the test and reference images. The overall quality score is estimated by gradually pooling local quality indices. Zhang et al. [4] investigated the effectiveness of deep features across different architectures and tasks, and claimed that deep features as a perceptual metric outperform all classic metrics such as PSNR by large margins. Then, a Learned Perceptual Image Patch Similarity (LPIPS) was presented. Ding et al. [8] presented a Deep Image Structure and Texture Similarity (DISTS), which combines the correlations of spatial averages (texture similarity) with the correlations of feature maps (structure similarity). Liao et al. [21] proposed to address image quality degradation in perceptual space from a statistical distribution perspective, where the quality is measured based on the Wasserstein distance of the learned deep features. In addition, some works addressed the learning-based IQA from the perspective of HVS. Kim et al. [22] proposed a DeepQA model, in which human visual sensitivity is learned from the underlying data distribution of IQA databases. Prashnani et al. [23] presented a perceptual image error assessment via pairwise preference (PieAPP), which is pairwise-learning framework to predict the preference of one distorted image over the other. Cao et al. [24] proposed a dual-branch network including reference and distortion branches, which combines semi-supervised and positive-unlabeled learning to address the distribution inconsistency between labeled and unlabeled data.

Besides, Guan et al. [25] proposed a no-reference IDQA work, which learns both visibility and distortion-aware features and maps the learned features into the quality scores by support vector regression. Lv et al. [26] also proposed a blind dehazed image quality assessment (BDQM) model, where the effect of inhomogeneous distortion from the dehazing procedure is attenuated via a specifically designed patch attention mechanism.

D. Transfer Learning

Transfer learning refers to the reuse of a pre-trained model on a new problem. That is, an already trained model is introduced into a different yet related problem, which is dedicated to addressing the issue of lacking large amounts of labeled data to train a complex deep model [27]. The definition of transfer learning is unified as follows. Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, transfer learning improves the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$. Transfer learning is usually divided into three classes, namely inductive transfer learning, transductive transfer learning, and unsupervised transfer learning [28]. As claimed earlier, there is no large-scale real-world paired image data for FR-IDQA, but unlabeled clear images and hazy images can be easily collected from the Internet. Thus, transfer learning is introduced for FR-IDQA. Specifically, the source task $\mathcal{T}_S$ is the image classification that judges whether an input image is a clear image or a hazy image, and the trained model is exploited for the target task, namely FR-IDQA. Since both source and target domains are in the real world, namely $\mathcal{D}_S = \mathcal{T}_S$, and the source and target tasks are different, namely $\mathcal{T}_S \neq \mathcal{T}_T$, we actually exploit inductive transfer learning for FR-IDQA.

III. Proposed Method

In this section, we introduce the establishment of the RW-Haze dataset and present the proposed DehIQA model for FR-IDQA.

A. Building the RW-Haze Dataset

Data acquisition: The RW-Haze dataset was recorded from outdoor scenes of six cities in China during two haze-frequent seasons, namely autumn and winter. The captured images are in PNG format with a resolution of 2560×1440 and a depth of 24 bits. Since each image pair was collected by a fixed camera for weather monitoring, they are spatially aligned well. For each scene, the camera position and its parameters were kept fixed during the collection. To ensure almost the same background for image pairs, we also made the following three considerations when collecting the dataset. Firstly, since there are more hazy days in autumn and winter, we collected hazy images and clear references in the cloudy days of these two seasons. This ensures to a great extent that image pairs share consistent backgrounds, especially almost the same sky as the image background. In contrast, if clear reference images have a blue sky or white clouds, it may bring side effects to IDQA. Secondly, the scenes from which we collected this dataset are on the top of high mountains. Due to few people and cars, there is almost no scenery change in a short period. Thirdly, hazy images and their clear references were captured within short intervals, even within an hour. This also ensures almost no background change, except for subtle weather conditions. With the above measures, we ensure to the maximum extent that raw hazy images and their clear references have desirable spatial alignment and background contents.

Data clearing: The raw data captured by the camera are digital videos. They are first converted into video frames. Then, we remove some undesired images, such as those with serious blur or degraded images by dark night. That is, for hazy images and their references, we only choose those images in which human eyes find it difficult to perceive any differences in their environmental conditions, except for their haze densities. Further, we invite five persons to make individual judgments on image pairs and pick out only those image pairs that are regarded as desirable samples. Moreover, a clear reference is carefully selected with multiple hazy images with distinct haze levels. These image pairs are grouped into three categories, e.g., slight, moderate, and heavy haze. Finally, we obtained 210 image pairs from more than 3000 images to build the RW-Haze benchmark dataset, which will be opened for academic use soon.

Fig. 6 shows some example image pairs in the RW-Haze dataset. The record time of each image is also displayed in the upper left corner. For the convenience of viewing, the recorded time is zoomed in and displayed at the top. We can observe that

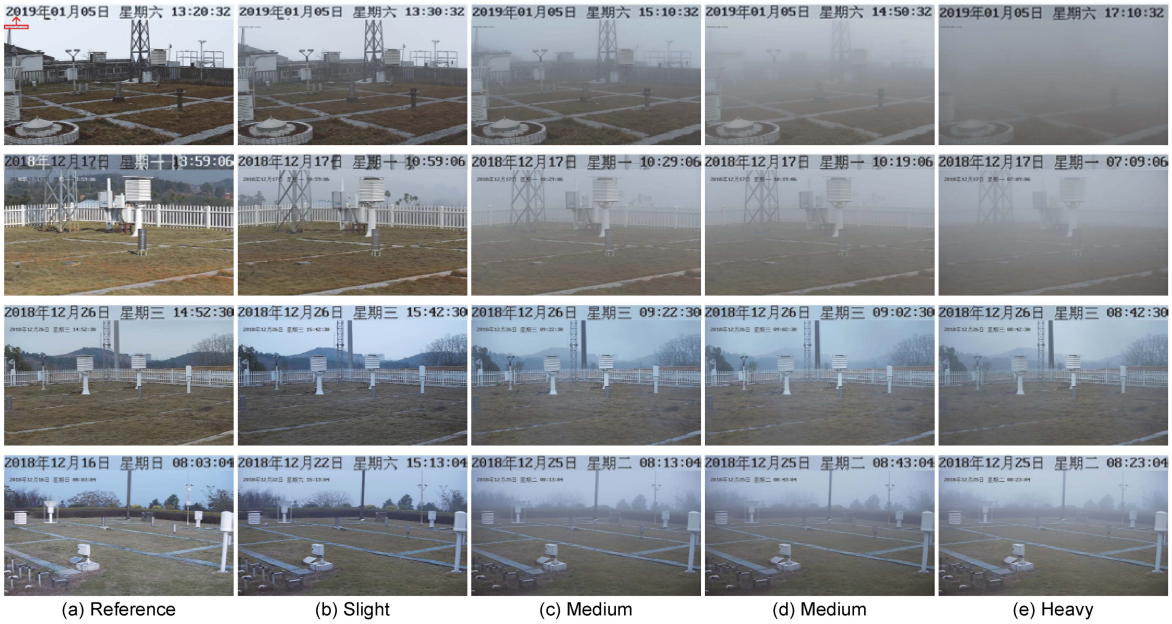

Fig. 6. Example images from the proposed RW-Haze dataset. (a) is the reference images. (b)–(e) are well-aligned real-world hazy images with increasing haze levels, and their labels are presented below. Each scene is captured by a fixed camera, ensuring the spatial alignment of images. Furthermore, the hazy images and corresponding references are captured at short intervals, some within an hour. The pristine recording time is displayed in the upper left corner and zoomed in on the top.

each clear reference is well-aligned with its hazy images with different haze densities.

B. Proposed DehIQA Model for FR-IDHA

Problem Formulation: Let $x = (I_{Ref}, I_{Hazy}, I_{Dehazed})$ be a triplet of the the primitive-quality reference image I_{Ref}, the hazy image I_{Hazy}, and the dehazed image $I_{Dehazed}$, and y be the ground-truth MOS of $I_{Dehazed}$. The learning-based FR-IDQA is to find a mapping $f(x)$ parameterized by θ^f to predict the quality score $\hat{y}$, which should be as close to the ground-truth y as possible. However, collecting massive MOS annotations is challenging and time-consuming. There is also no open dataset for training deep models. In this work, we exploit the transfer learning strategy to address these issues in a feasible way. Specifically, a binary classification network is trained to differentiate clear references and hazy images, which is trained with publicly available datasets. Then, only the feature extraction module of the classification network is retained and the parameters of this network are frozen, which is then transferred to the IDQA task.

Source Task for Classification: The source task is to classify each input image as a hazy image or a clear reference. In fact, image classification is often used as a source task in transfer learning [27] or a pretext task in self-supervised learning [29]. For the image classification tasks, there are many deep models including AlexNet [30], VGG16 [31], SqueezeNet [32] and ResNet18 [33]. Moreover, the learned features of VGG trained on ImageNet [34] for image classification have been proved to be prominently useful as a perceptual loss for image style transfer [35], image synthesis [36], and video synthesis [37]. Thus, we also exploit several typical classification networks including AlexNet, VGG16, SqueezeNet, and ResNet18 for the source task. Specifically, we label clear images and hazy images with 0 and 1 for training, respectively. In essence, the trained classification network can learn well the haze-related features from input images, which can be used for the IDQA task. Actually, the existing classic IDQA methods also depend seriously on hand-crafted features to describe the haze distortions in a dehazed image [5], [6], whereas we exploit deep models to learn haze-related features for FR-IDQA.

Target Task for FR-IDQA: Fig. 7 is the proposed DehIQA for FR-IDQA, which is a triple-branch structure. Each branch shares the same structure and weight. It accepts the reference clear image, the hazy image, and the dehazed image as inputs. Specifically, the reference clear image and the dehazed image are fed into the feature extraction network, which is transferred from the classification network, to obtain the reference feature f^s_{Ref} and the dehazed feature f^s_{Deh} $(s = 1, 2, \ldots, m)$, respectively. To focus on the regions with different haze levels, a haze attention strategy was used to integrate into the network. Finally, a score is obtained to evaluate the visual quality of the dehazed image by comparing the reference feature f^s_{Ref} and the dehazed feature f^s_{Deh}.

The IQA of a dehazed image is usually achieved by making a comparison with its clear reference, without considering the primitive hazy image from which the dehazed image is restored [5], [6]. Apparently, this ignores the important information that can be provided by the primitive hazy image. As we know, the haze is not uniformly distributed in a hazy image. There are usually more haze residues or artifacts in the regions with thicker haze than those regions with lighter haze. When evaluating the dehazed image, more attention should be paid to the regions with thicker haze in its primitive hazy image. However, the clear reference can not provide any information about

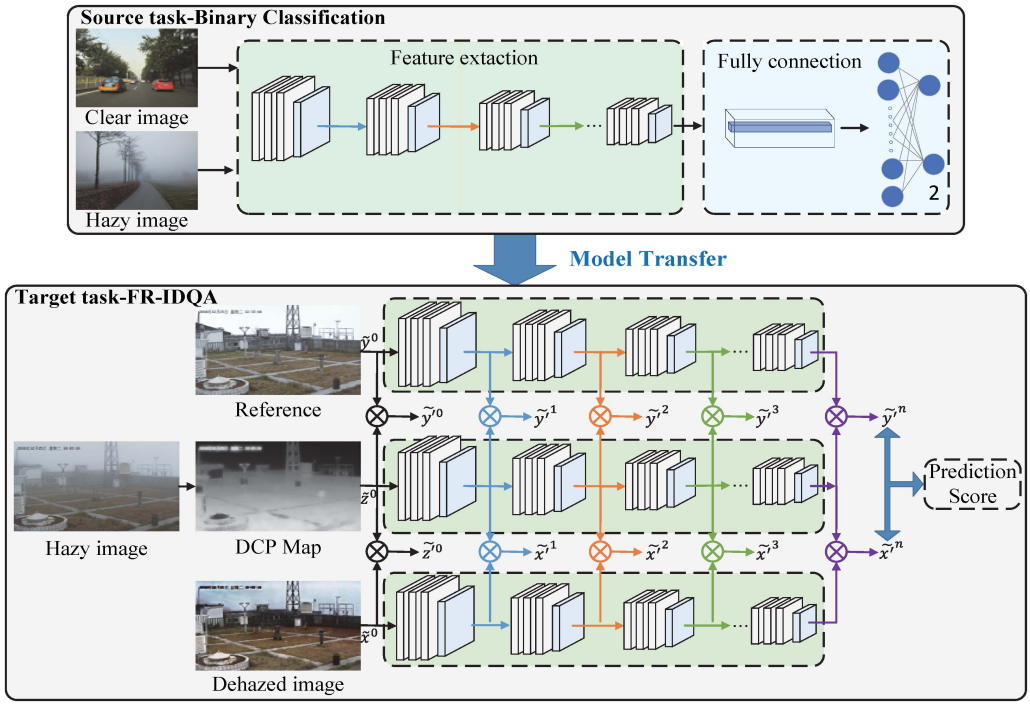

Fig. 7. Proposed FR-IDQA. It adopts a triple-branch structure, i.e., one for the reference image, one for the DCP map, and the other for the dehazed image. The feature extraction network works at three scales, namely the three input images are downsampled by a factor of 1, 2, and 4, respectively.

the haze distribution. Luckily, the importance of local regions can be obtained from the Dark Channel Map (DCP) of the primitive hazy image [38]. Thus, we design a simple yet effective haze-aware mechanism based on the DCP features. Specifically, the DCP features are multiplied by the features learned from the dehazed image and the clear reference, respectively. In this way, those regions with more haze are allocated with higher weights for IDQA, and vice versa.

To ensure an injective mapping, the features of the reference images are represented as

$$f_{ref}(x) = \tilde{x}_j^{(i)}; i = 0, \ldots, m; j = 1, \ldots, n_i, \tag{1}$$

where m is the number of the convolution layers, and n_i is the number of the feature maps in the ith convolution layer. Similarly, we define the features of the dehazed image and the dark channel maps as

$$f_{deh}(y) = \tilde{y}_j^{(i)}; i = 0, \ldots, m; j = 1, \ldots, n_i, \tag{2}$$

$$f_{dak}(z) = \tilde{z}_j^{(i)}; i = 0, \ldots, m; j = 1, \ldots, n_i, \tag{3}$$

After the attention module, the features of the reference image and the dehazed image are

$$f'_{ref}(x') = \tilde{x}_j^{(i)} \times \tilde{z}_j^{(i)}; i = 0, \ldots, m;\ j = 1, \ldots, n_i, \tag{4}$$

$$f'_{deh}(y') = \tilde{y}_j^{(i)} \times \tilde{z}_j^{(i)}; i = 0, \ldots, m;\ j = 1, \ldots, n_i, \tag{5}$$

where x' and y' can be denoted as the new features for the reference image and dehazed image, respectively. Similar to the definition of SSIM [39], we define the following formula to measure the differences by comparing the distributions of the feature maps.

$$l(\tilde{x'}_j^{(i)}, \tilde{y'}_j^{(i)}) = \frac{2\mu_{\tilde{x'}}^{(i)}\mu_{\tilde{y'}}^{(i)} + c_1}{(\mu_{\tilde{x'}}^{(i)})^2 + (\mu_{\tilde{y'}}^{(i)})^2 + c_1}, \tag{6}$$

$$s(\tilde{x'}_j^{(i)}, \tilde{y'}_j^{(i)}) = \frac{2\sigma_{\tilde{x'}\tilde{y'}}^{(i)} + c_2}{(\sigma_{\tilde{x'}}^{(i)})^2 + (\sigma_{\tilde{y'}}^{(i)})^2 + c_2}, \tag{7}$$

where $\mu_{\tilde{x'}}^{(i)}$, $\mu_{\tilde{y'}}^{(i)}$ are the means of $\tilde{x'}^{(i)}$ and $\tilde{y'}^{(i)}$, respectively. $\sigma_{\tilde{x'}\tilde{y'}}^{(i)}$, $(\sigma_{\tilde{x'}}^{(i)})^2$, and $(\sigma_{\tilde{y'}}^{(i)})^2$ are the covariances. c_1 and c_2 are small constants to avoid the denominators be close to zero.

The quality measurements for one scale image is

$$D(x', y') = 1 - \sum_{i=0}^{m}\sum_{j=1}^{n_i}(l(\tilde{x'}_j^{(i)}, \tilde{y'}_j^{(i)}) + s(\tilde{x'}_j^{(i)}, \tilde{y'}_j^{(i)})), \tag{8}$$

Finally, the scores at all scales are averaged to generate the final score,

$$D_{DehIQA} = \sum_{i=0}^{k}(D(x', y')). \tag{9}$$

where the k is empirically set with 3. That is, the input image is downsampled by a factor of 2 and 4.

IV. Experiments

A. Experiment Settings

Training Datasets: As claimed earlier, the classification network is to classify hazy-free and hazy images. RTTS [40] and

RSCM [41] are two datasets for single image dehazing and weather classification, respectively. They have large amounts of real-world haze-free and hazy images. We randomly selected 12,000 haze-free images and hazy images, respectively. These images are partitioned into the training, validation, and testing sets with the ratio of $8:1:1$ for training.

Testing Datasets: A desirable FR-IDQA approach should obtain an evaluation score that is consistent with the subjective MOS. For image dehazing assessment, the reference images, the hazy images and the dehazed images should be provided with their MOSs. Luckily, the RW-Haze dataset provides hazy images with different haze levels and their references, and image pairs are well-aligned. The dehazed images are obtained by the existing open-source image dehazing works. We select 10 representative dehazing works, which include one prior-based work and nine deep learning-based works to obtain the dehazed images. Specifically, they are the dark channel prior (DCP) based work [38], and the recent domain adaptation dehazing (DAD) [42], unsupervised dehazing (DCPLoss) [43], semi-supervised dehazing (SEMI) [44], ultra-high-definition dehazing (4kDehazing) [45], principled synthetic-to-real dehazing (PSD) [46], unsupervised and untrained dehazing (YOLY) [47], weakly supervised dehazing (RefineDNet) [48], self-augmented unpaired dehazing (D4) [49], and unsupervised single image dehazing (USID) [50].

In addition, we selected 30 hazy images with different haze levels. By using the 10 image dehazing works, we obtain 10 dehazed images for each hazy image, which are treated as a group. There is a total of 300 dehazed images. Then, we invite 10 volunteers to rank the dehazed images in each group in terms of their subjective qualities, namely the amounts of left haze residues and artifacts/distortions. The paired comparison and sorting strategy is adopted for making comparisons. That is, two dehazed images from the same group are simultaneously shown to a volunteer to rank them. Finally, we obtain the overall orders of the dehazed images in each group, which are converted into the MOSs as follows.

$$Score = \frac{1}{M}\sum_{i=0}^{M}\left(1-\frac{n_i}{N}\right). \tag{10}$$

where M and N are the number of volunteers and the number of images in each group, respectively. n_i is the order of the image in a group given by the ith volunteer. The higher the score, the better quality of a dehazed image.

Baselines: To prove the superiority of the proposed DehIQA for FR-IDQA, eleven existing metrics are used for comparison. Among them, VI [5], RI [5], and FRFSIM [6] are specially designed for image dehazing assessment, which exploits hand-crafted features. The other eight metrics are for general IQA. Specifically, they include seven FR-IQA metrics, namely PSNR, SSIM [39], FSIM [51], VSI [52], LPIPS [4], DISTS [8] and PieAPP [23], and one no-reference IQA (NR-IQA) metric, namely VDA-DQA [25]. Note that LPIPS, DISTS, and PieAPP are deep learning-based metrics.

Evaluation Criteria: Four commonly-used criteria are used to evaluate the proposed DehIQA metric. The Spearman Rank-order Correlation Coefficient (SRCC) and Kendall Rank-order correlation Coefficient (KRCC) are used to measure prediction monotonicity, whereas the Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE) are used to measure prediction accuracy. To calculate the PLCC and RMSE, a non-linear regression is needed to fit the subjective scores and the objective scores. Following a previous study [5], the non-linear regression function is defined for mapping.

$$f(x) = \beta_1\left(\frac{1}{2}-\frac{1}{1+e^{\beta_2(x-\beta_3)}}\right)+\beta_4 x+\beta_5. \tag{11}$$

where $\beta_i (i = 1, \ldots, 5)$ are the parameters to fit, and their initial values are set with 1. x and $f(x)$ denote the objective scores and the subjective scores, respectively. For SRCC, KRCC, and PLCC, a higher value means a better IQA metric and a lower value of RMSE means a better IQA metric.

Training strategy and parameters. All the images are resized to 256×256. The classification network is implemented by Pytorch 1.3.0 and Python 3.6.9 on a PC with two NVidia RTX 2080 TI GPUs. The Adam optimizer is adopted with the exponential decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train the network. The learning rate is set to 1×10^{-4}, which is multiplied by 0.1 after 10 epochs. The batch size is set to 64. The classification network is based on the pre-trained ResNet18 [33].

B. Comparison With SOTA Methods

The proposed DehIQA is compared with seven common IQA metrics and four specially designed FR-IDQA metrics on three levels of dehazed groups by calculating the SRCC, KRCC, PLCC, and RMSE. Note that PSNR, SSIM [39], FSIM [51] and VSI [52] are traditional IQA metrics, whereas LPIPS [4], DISTS [8] and PieAPP [23] are the CNN-based general IQA metrics. In addition, there are four specially-designed IDQA metrics, including RI [5], VI [5], FRFSIM [6] for FR-IDQA, and VDA-DQA [25] for NR-IDQA.

We report the average SRCC, KRCC, PLCC, and RMSE for the dehazed groups of the RW-HAZE dataset. Table II provides the comparison between DehIQA and other IQA metrics on the dehazed groups. We observe that DehIQA outperforms all the competitors by a larger margin, including three specially designed FR-IDQA metrics. In addition, we have the following observations.

Firstly, the proposed DehIQA achieves the best average performance for three groups of dehazed images and outperforms all the competitors by a large margin (SRCC = 0.7512, KRCC = 0.6197, PLCC = 0.7641, and RMSE = 0.1621), which shows that our DehIQA provides consistent results with human perceptions. In addition, thanks to the proposed haze density-aware mechanism, there are obvious advantages in the heavy dehazed group with more distortions or haze residues, which benefits from the strong representation abilities of the learned deep features.

Secondly, the existing models or metrics have inconsistent performances for those dehazed images from hazy images with

TABLE II
PERFORMANCE COMPARISON FOR DIFFERENT IQA METHODS. NOTE THAT THE HIGHER SRCC, KRCC, OR PLCC SCORES ARE, THE BETTER, WHILE LOWER SCORES FOR RMSE ARE BETTER. THE BEST TWO VALUES ARE MARKED WITH RED AND GREEN

Type		PSNR	SSIM [51]	FSIM [52]	VSI [53]	LPIPS [4]	DISTS [8]	PieAPP [23]	VDA-DQA [25]	RI [5]	VI [5]	FRFSIM [6]	Ours
Light	SRCC	0.2848	0.4454	0.8091	0.6303	0.5939	0.7818	0.5545	0.2333	0.7101	0.7727	0.1575	0.8273
	KRCC	0.1777	0.3000	0.6666	0.4889	0.4666	0.6000	0.4000	0.1444	0.5250	0.6111	0.1444	0.6888
	PLCC	0.4080	0.5953	0.8658	0.7698	0.6404	0.8074	0.6611	0.2771	0.7843	0.8322	0.2390	0.7971
	RMSE	0.2340	0.2038	0.1236	0.1578	0.1993	0.1498	0.1942	0.2468	0.1553	0.1383	0.2507	0.1520
Medium	SRCC	0.1743	0.1848	0.6515	0.6515	0.6151	0.7515	0.5605	0.2545	0.6939	0.5381	0.2818	0.7454
	KRCC	0.1170	0.1444	0.5111	0.5111	0.4889	0.6222	0.4444	0.1889	0.5333	0.4060	0.2111	0.6000
	PLCC	0.3745	0.5104	0.6862	0.7019	0.6772	0.7910	0.5828	0.3328	0.7133	0.6927	0.4043	0.7536
	RMSE	0.2304	0.2155	0.1780	0.1779	0.1838	0.1449	0.1922	0.2291	0.1756	0.1940	0.2266	0.1626
Heavy	SRCC	0.2567	0.2444	0.4141	0.5959	0.3740	0.4141	0.2121	0.2566	0.5393	0.4909	0.1919	0.6808
	KRCC	0.1336	0.0963	0.2888	0.4814	0.2972	0.3185	0.1703	0.1852	0.3778	0.3777	0.1259	0.5704
	PLCC	0.4679	0.5491	0.6044	0.6847	0.5818	0.5672	0.2937	0.3408	0.5824	0.6426	0.1601	0.7416
	RMSE	0.2248	0.2141	0.1921	0.1855	0.2079	0.2042	0.2374	0.2335	0.2063	0.1927	0.2471	0.1717
All Images	SRCC	0.2385	0.2916	0.6249	0.6259	0.5277	0.6491	0.4424	0.2481	0.6478	0.6005	0.2104	0.7512
	KRCC	0.1428	0.1802	0.4888	0.4938	0.4176	0.5135	0.3382	0.1728	0.4787	0.4649	0.1604	0.6197
	PLCC	0.4168	0.3532	0.7188	0.7189	0.6332	0.7219	0.5125	0.3169	0.6934	0.7225	0.2678	0.7641
	RMSE	0.2297	0.2111	0.1646	0.1737	0.1970	0.1663	0.2079	0.2364	0.1791	0.1668	0.2414	0.1621

distinct haze levels. Note that FSIM [51] and VI [5] handle well the dehazed images from hazy images with light haze, but they behave poorly for those dehazed images from hazy images with medium or heavy haze. Though RI [5] and VI [5] are specially designed to evaluate dehazed images in terms of realness and visibility, a single RI or VI does not provide an overall evaluation of a dehazed image, despite they might be good individually. In addition, FRFSIM [6] seems to be unsatisfactory, since its performance is even worse than some general IQA metrics.

Third, the learning-based IQA models are more consistent with human perception, even if they are better than the specially designed FR-IDQA metrics. Nevertheless, the pioneer CNN-based IQA models, including LPIPS [4], DISTS [8], and PieAPP [23], still have limited performances. DISTS is seen to perform the best for groups of moderately dehazed images and outperform the proposed metrics (SRCC $=$ 0.7515, KRCC $=$ 0.6222, PLCC $=$ 0.7910, and RMSE $=$ 0.1449). However, it does not work well for dehazed images with severe distortion because it ignores the effect of haze on IQA. Actually, they do not consider the effect of haze on image visibility. If an image dehazing method generates a dehazed image with almost no noise or distortion, but there remains a large amount of haze. In this case, the general IQA metrics may treat this as desirable dehazing, but actually, it is the opposite.

Fourth, VDA-DQA [25] is an NR-IDQA metric that might be considered to be not appropriate for making fair comparisons. However, since NR-IQA is a future trend, we still believe that making comparisons between DehIQA and VDA-DQA is useful. Moreover, compared with the well-known PSNR, VDA-DQA is much better for IDQA simply because it is specially designed for dehazing assessment.

C. Dehazing Baselines

To further prove the effectiveness of IDQA when evaluating the existing dehazing works, we provide the RI, VI, FRFSIM and DehIQA scores of 10 dehazing methods on the RW-HAZE as baselines. Fig. 8 shows an example from the RW-Haze dataset, in which their scores of different IQA metrics are reported. Note that the higher the score of RI, VI, and FRFSIM, the better the dehazed images, while the opposite is true for DehIQA. Table III reports the scores of RI [5], VI [5], FRFSIM [6], and DehIQA when evaluating the dehazed images from different hazy images groups. From Fig. 8 and Table III, we have the following observations.

First, there is still no single dehazing work that simultaneously achieves SOTA results for hazy images with three haze levels including light, moderate, and heavy. The average score of the dehazed image obtained by D4 [49] is the highest (DehIQA $=$ 0.4344), but for dense hazy images, it is worse than the other works. In contrast, PSD [46] ranks first (DehIQA $=$ 0.4851) for dense hazy images but performs poorly for mild and moderate hazy images.

Second, most existing image dehazing works are dedicated to the design of new deep models and improving the training strategy, but few works address the benchmark datasets, including real-world hazy image datasets and more reliable synthetic hazy image datasets. As claimed earlier, the most popular and recent RESIDE [40] dataset still has rich room for improvement to promote image dehazing. As shown in Fig. 8, D4 [49] attempts to synthesize more photo-realistic hazy images to better train deep dehazing network and improve its generalization ability, it obtains the best score in terms of RI, VI and DehIQA.

Third, the traditional DCP [38] achieves stable dehazing results for various hazy images in terms of RI, VI, and DehIQA. However, the dehazed images by DCP seem to be over-enhanced. The CNN-based dehazing works are quite different. Some of them behave well for light hazy images, but there are a lot of distortions for moderate or heavy hazy images. In fact, this is caused by the lack of large-scale real-world datasets for training. Fourth, the unsupervised dehazing works such as DCPLoss [43], YOLY [47], and USID [50] can overcome the tedious process of data collection, but they usually achieve

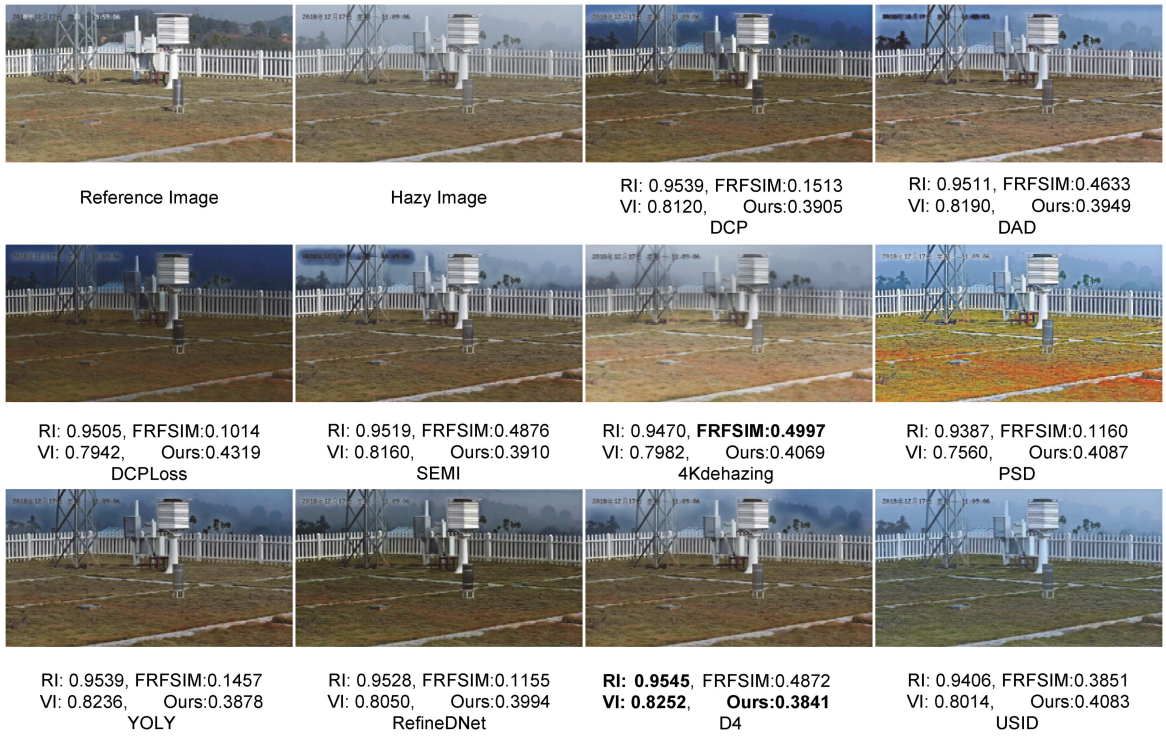

Fig. 8. Qualitatively and quantitatively comparison of the dehazed results generated by 10 different dehazing methods.

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT DEHAZING METHODS BY DIFFERENT IQA METRICS ON THE RA-HAZE DATASET. THE TOP TWO DEHAZING PERFORMANCES ARE INDICATED WITH RED AND GREEN

Method	Light				Medium				Heavy				All Images			
	RI	VI	FRFSIM	Ours	RI	VI	FRFSIM	Ours	RI	VI	FRFSIM	Ours	RI	VI	FRFSIM	Ours
DCP [38]	0.9582	0.8284	0.1820	0.3837	0.9431	0.7822	0.1295	0.4319	0.9171	0.7101	0.0862	0.4992	0.9395	0.7736	0.1325	0.4383
DAD [42]	0.9559	0.8365	0.4803	0.3853	0.9358	0.7809	0.4162	0.4363	0.9141	0.7074	0.2253	0.4894	0.9361	0.7749	0.3739	0.4370
DCPLoss [43]	0.9542	0.8026	0.1137	0.4278	0.9383	0.7601	0.0885	0.4684	0.9088	0.6942	0.0663	0.5316	0.9337	0.7523	0.0895	0.4759
SEMI [44]	0.9526	0.8353	0.3506	0.3815	0.9372	0.7592	0.1386	0.4377	0.9093	0.6758	0.1002	0.4977	0.9342	0.7568	0.1964	0.4390
4kDehazing [45]	0.9517	0.8124	0.5152	0.4001	0.9317	0.7334	0.4161	0.4475	0.9053	0.6374	0.2315	0.5052	0.9295	0.7277	0.3876	0.4509
PSD [46]	0.9415	0.7625	0.1369	0.3988	0.9323	0.7518	0.1271	0.4409	0.9095	0.6997	0.3221	0.4851	0.9278	0.7380	0.1953	0.4416
YOLY [47]	0.9553	0.8316	0.1250	0.3909	0.9334	0.7467	0.1960	0.4500	0.9167	0.6462	0.1481	0.4921	0.9351	0.7415	0.1563	0.4443
RefineDNet [48]	0.9575	0.8236	0.1351	0.3912	0.9436	0.7681	0.1235	0.4372	0.9179	0.6976	0.1158	0.4885	0.9397	0.7631	0.1248	0.4389
D4 [49]	0.9590	0.8427	0.3467	0.3771	0.9429	0.7613	0.2503	0.4320	0.9213	0.6723	0.2661	0.4940	0.9411	0.7588	0.2877	0.4344
USID [50]	0.9452	0.8230	0.3893	0.3984	0.9317	0.7550	0.3456	0.4466	0.9114	0.6786	0.2330	0.4959	0.9294	0.7522	0.3226	0.4470

relatively much poorer results than those supervised methods. Nevertheless, we believe that unsupervised dehazing is promising and may surpass supervised works shortly.

Fifth, with the network growing wider and deeper, the expansion of the network makes the improvement of the dehazing performance limited and less cost-effective. In turn, high-quality datasets with precise annotations can maximize the vitality of the network and improve the dehazing performance to a greater extent.

D. Evaluation on Other Datasets

To further evaluate image dehazing works on other datasets, we select several relatively high-quality image pairs from the MRFID dataset [6] for experiments. First, we use 10 dehazing works to obtain the dehazed images and calculate their scores by the proposed metrics. Second, a single dehazing work [49] is used to process hazy images with different haze levels from the same scene, and the scores are calculated for the dehazed images. Figs. 9 and 10 show the evaluation results of the dehazed images and their scores for the above two scenarios, respectively. We can observe that the proposed metric can accurately evaluate the dehazed images, which further demonstrates the robustness of the proposed metric.

E. Comparisons of Different Classification Networks

To verify the effectiveness of the proposed DehIQA, different baseline networks including AlexNet [30], Vgg16 [31], SqueezeNet [32], and ResNet18 [33] are used for comparisons. Specifically, they are used as the binary classification network for training to achieve clear and hazy image classification. Then, the feature extraction network is embedded into DehIQA for FR-IDAQ. Table IV reports the average scores of SRCC, KRCC,

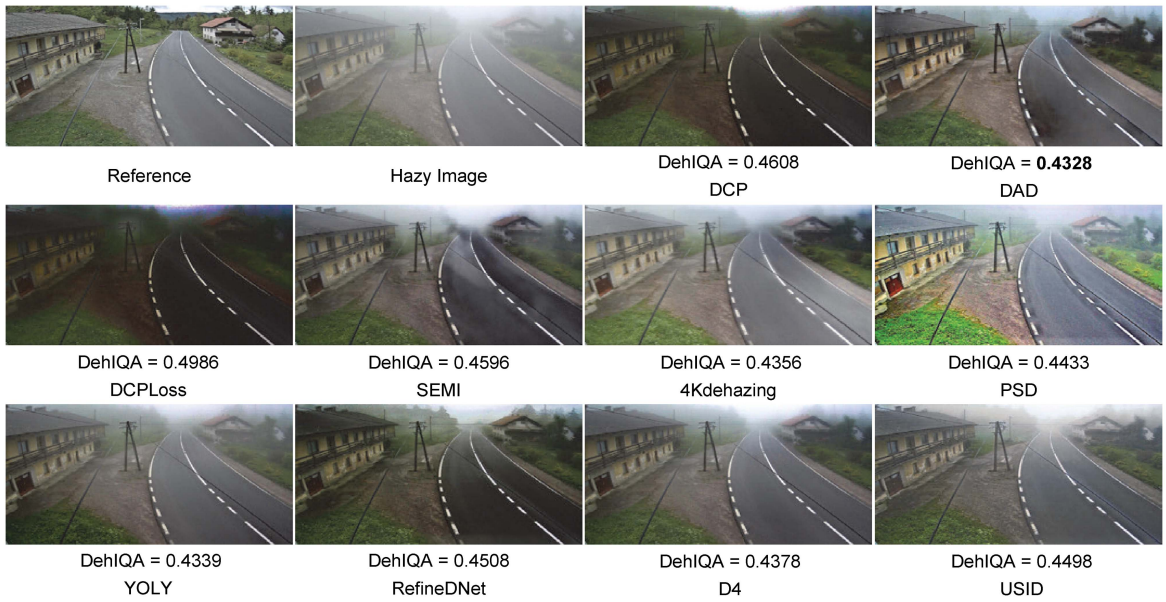

Fig. 9. Qualitatively and quantitatively evaluation of the dehazed images generated from 10 different dehazing methods by DehIQA, where the hazy images are selected from the MRFID dataset [6]. Lower scores are better.

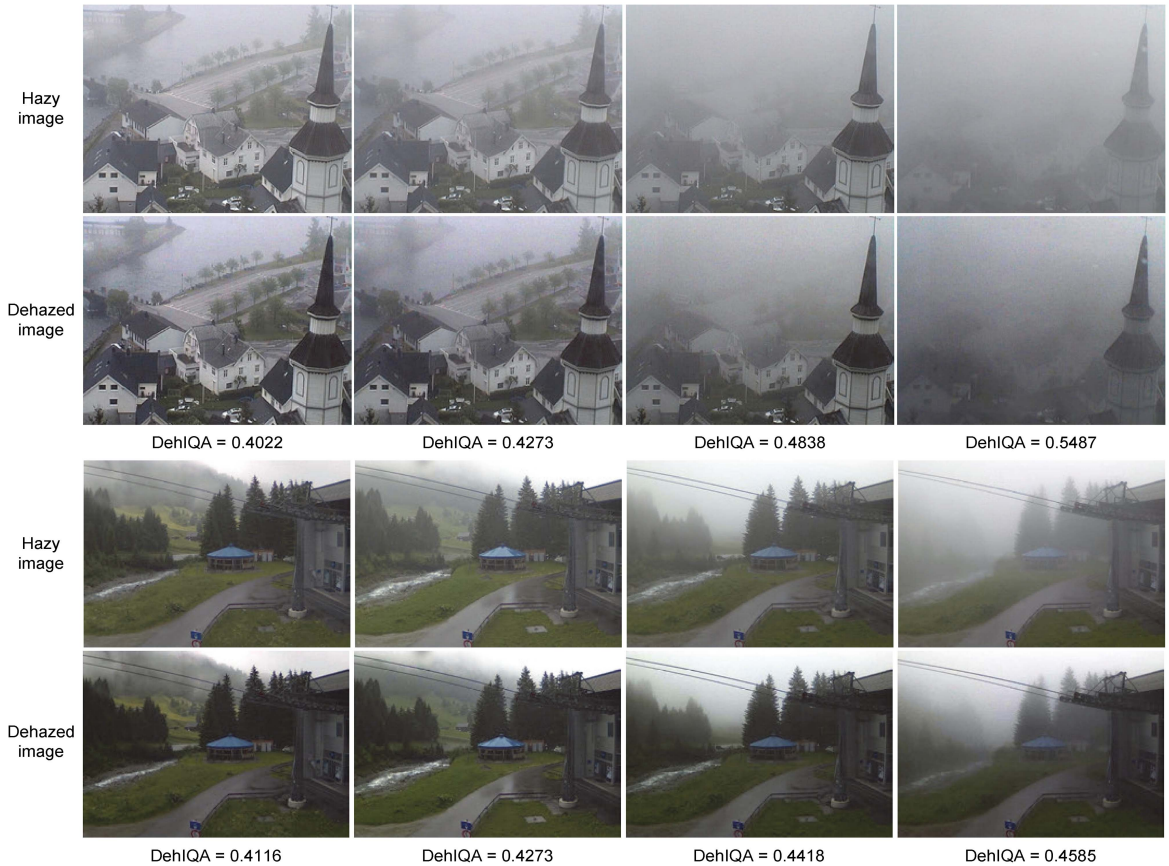

Fig. 10. Qualitatively and quantitatively evaluation of the dehazed results generated by D4 [49], where the hazy images come from the same scene with different haze densities in the MRFID dataset.

TABLE IV
QUANTITATIVE COMPARISONS OF FOUR DIFFERENT BASELINE NETWORKS. THE TOP SCORES ARE INDICATED IN BOLD

Model	SRCC	KRCC	PLCC	RMSE
DehIQA(AlexNet)	0.6939	0.5419	**0.7729**	**0.1615**
DehIQA(Vgg16)	0.7131	0.5715	0.7466	0.1679
DehIQA(SqueezeNet)	0.7117	0.5691	0.7231	0.1697
DehIQA(ResNet18)	**0.7512**	**0.6197**	0.7641	0.1621

TABLE V
ABLATION STUDY OF THE PROPOSED DEHIQA

Model	SRCC	KRCC	PLCC	RMSE
Baseline	0.7212	0.5556	0.7456	0.1819
Baseline+attention (DehIQA)	**0.7512**	**0.6197**	**0.7641**	**0.1621**

The top scores are indicated in bold.

PLCC, and RMSE for four different baselines. We can observe that Resnet18 obtains the best performance. The other three networks also obtain comparable performances, which proves that the learned deep features have been remarkably useful for image dehazing assessment.

F. Ablation Study

To verify the effectiveness of the proposed haze attention module, we remove it from the proposed DehIQA and keep others unchanged for experiments. The classification model is ResNet18. Table V reports the results. We can observe that the haze attention module improves performance. That is, learning more features from the regions with more haze in the original hazy images benefits IDQA.

V. LIMITATION

The RW-Haze dataset is made up of clear images and well-aligned hazy images with distinct haze levels from mist to dense haze, which makes it suitable for promoting image dehazing and its evaluation. However, collecting large-scale real-world hazy and clear image datasets is very challenging, and the cleaning and annotation process also requires extensive human interventions. Thus, the collected RW-Haze dataset is relatively small and from limited scenes. In the future, more real-world clear and hazy image pairs will be collected from diverse scenes to extend the RW-Haze dataset, which will further promote the robustness of both image dehazing and its assessment.

VI. CONCLUSION

In this work, we established a new RW-Haze dataset, which comprises real-world hazy images and their well-aligned clear references. Note that they were collected from natural outdoor scenes, and the hazy images have distinct haze levels. We also presented a haze density-aware DehIQA, which can serve as an objective FR-IDQA metric. Transfer learning was introduced to alleviate the lack of sufficient labeled datasets. Motivated by the observation that there are more visual artifacts or haze residues near those regions with thicker haze in a dehazed image, we designed a haze density-aware mechanism to enforce DehIQA to learn more from the regions with more haze in the primitive hazy image for IDQA. The experimental results demonstrate that the proposed DehIQA is more effective than the existing metrics such as VI and RI for FR-IDQA.

REFERENCES

[1] 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.*, 2016, pp. 2226–2230.

[2] Y. Zhang, L. Ding, and G. Sharma, "HazeRD: An outdoor scene dataset and benchmark for single image dehazing," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3205–3209.

[3] X. Min et al., "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, vol. 21, pp. 2319–2333, 2019.

[4] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[5] S. Zhao, L. Zhang, S. Huang, Y. Shen, and S. Zhao, "Dehazing evaluation: Real-world benchmark datasets, criteria, and baselines," *IEEE Trans. Image Process.*, vol. 29, pp. 6947–6962, 2020.

[6] W. Liu, F. Zhou, T. Lu, J. Duan, and G. Qiu, "Image defogging quality assessment: Real-world database and method," *IEEE Trans. Image Process.*, vol. 30, pp. 176–190, 2021.

[7] F. Guo, J. Tang, and Z.-X. Cai, "Objective measurement for image defogging algorithms," *J. Central South Univ.*, vol. 21, no. 1, pp. 272–286, 2014.

[8] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[9] G. Yin et al., "Content-variant reference image quality assessment via knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3134–3142.

[10] A. Ghildyal and F. Liu, "Shift-tolerant perceptual similarity metric," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 91–107.

[11] J. Chen, S. Wang, X. Liu, and G. Yang, "RW-haze: A real-world benchmark dataset to evaluate quantitatively dehazing algorithms," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 11–15.

[12] 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 Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 754–762.

[13] 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. Adv. Concepts Intell. Vis. Syst.*, 2018, pp. 620–631.

[14] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1014–1018.

[15] C. O. Ancuti, C. Ancuti, and R. Timofte, "NH-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 444–445.

[16] N. Jacobs, N. Roman, and R. Pless, "Consistent temporal variations in many outdoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[17] S. Wei, H. Zhou, Y. Fu, X. Shang, and H. Jing, "A novel image quality assessment method for dehazed image," in *Proc. IEEE 7th Int. Conf. Comput. Commun. Syst.*, 2022, pp. 331–337.

[18] L. K. 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.

[19] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective quality evaluation of dehazed images," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2879–2892, Aug. 2019.

[20] F. Gao et al., "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017.

[21] X. Liao et al., "DeepWSD: Projecting degradations in perceptual space to wasserstein distance in deep feature space," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 970–978.

[22] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1676–1684.

[23] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1808–1817.
[24] Y. Cao, Z. Wan, D. Ren, Z. Yan, and W. Zuo, "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5851–5861.
[25] T. Guan et al., "Visibility and distortion measurement for no-reference dehazed image quality assessment via complex contourlet transform," *IEEE Trans. Multimedia*, vol. 25, pp. 3934–3949, 2023.
[26] X. Lv, T. Xiang, Y. Yang, and H. Liu, "Blind dehazed image quality assessment: A deep CNN-based approach," *IEEE Trans. Multimedia*, vol. 25, pp. 9410–9424, 2023.
[27] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[29] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[32] F. N. Iandola et al., "SqueezeNet: AlexNet-level accuracy with $50\times$ fewer parameters and $<$0.5 mb model size," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[34] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.
[36] T. Wang et al., "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 8798–8807.
[37] T.-C. Wang et al., "Video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1152–1164.
[38] 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.
[39] 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.
[40] B. Li et al., "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.
[41] D. Lin, C. Lu, H. Huang, and J. Jia, "RSCM: Region selection and concurrency model for multi-class weather recognition," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4154–4167, Sep. 2017.
[42] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2805–2814.
[43] A. Golts, D. Freedman, and M. Elad, "Unsupervised single image dehazing using dark channel prior loss," *IEEE Trans. Image Process.*, vol. 29, pp. 2692–2701, 2020.
[44] L. Li et al., "Semi-supervised image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 2766–2779, 2020.
[45] Z. Zheng et al., "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16180–16189.
[46] Z. Chen, Y. Wang, Y. Yang, and D. Liu, "PSD: Principled synthetic-to-real dehazing guided by physical priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7180–7189.
[47] B. Li et al., "You only look yourself: Unsupervised and untrained single image dehazing neural network," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1754–1767, 2021.
[48] S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, "RefineDNet: A weakly supervised refinement framework for single image dehazing," *IEEE Trans. Image Process.*, vol. 30, pp. 3391–3404, 2021.
[49] Y. Yang et al., "Self-augmented unpaired image dehazing via density and depth decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2027–2036.
[50] J. Li, Y. Li, L. Zhuo, L. Kuang, and T. Yu, "USID-Net: Unsupervised single image dehazing network via disentangled representations," *IEEE Trans. Multimedia*, vol. 25, pp. 3587–3601, 2023.
[51] 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.
[52] 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.