Image Defogging Quality Assessment: Real-World Database and Method

Wei Liu, Fei Zhou, Tao Lu, *Member, IEEE*, Jiang Duan, and Guoping Qiu

***Abstract*— Fog removal from an image is an active research topic in computer vision. However, current literature is weak in the following two areas which in many ways are hindering progress for developing defogging algorithms. First, there is no true real-world and naturally occurring foggy image datasets suitable for developing defogging models. Second, there is no suitable mathematically simple and easy to use image quality assessment (IQA) methods for evaluating the visual quality of defogged images. We address these two aspects in this paper. We first introduce a new foggy image dataset called multiple real-world foggy image dataset (MRFID). MRFID contains foggy and clear images of 200 outdoor scenes. For each scene, one clear image and 4 foggy images of different densities defined as slightly foggy, moderately foggy, highly foggy, and extremely foggy, are manually selected from images taken from these scenes over the course of one calendar year. We then process the foggy images of MRFID using 16 defogging methods to obtain 12,800 defogged images (DFIs) and perform a comprehensive subjective evaluation of the visual quality of the DFIs. Through collecting the mean opinion score (MOS) of 120 subjects and evaluating a variety of fog-relevant image features, we have developed a new Fog-relevant Feature based SIMilarity index (FRFSIM) for assessing the visual quality of DFIs. We present extensive experimental results to show that our new visual quality assessment measure, the FRFSIM, is more consistent with the MOS than other IQA methods and is therefore more suitable for evaluating defogged images than other state-of-the-art IQA methods. Our dataset and relevant code are available at http://www.vistalab.ac.cn/MRFID-for-defogging/.**

***Index Terms*— Foggy image dataset, fog density, image defog, image quality assessment.**

Manuscript received February 2, 2020; revised August 17, 2020 and October 9, 2020; accepted October 15, 2020. Date of publication October 29, 2020; date of current version November 19, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 62001334 and Grant 62072350, in part by the Science Research Foundation of Wuhan Institute of Technology under Grant K202034, in part by the Science and Technology Research Project of Hubei Provincial Department of Education under Grant Q20201507, in part by the Hubei Technology Innovation Project under Grant 2019AAA045, and in part by the Education Department of Guangdong Province, China, under Project 2019KZDZX1028. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel L. Lau. *(Corresponding author: Guoping Qiu.)*

Wei Liu and Tao Lu are with the Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China.

Fei Zhou is with the Guangdong Key Laboratory for Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Laboratory of Artificial intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial intelligence and Robotics for Society, Shenzhen 518172, China.

Jiang Duan is with the School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 610074, China, and also with Chengdu Everimaging Ltd., Chengdu 610041, China.

Guoping Qiu is with the Guangdong Key Laboratory for Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Laboratory of Artificial intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China, also with the Shenzhen Institute of Artificial intelligence and Robotics for Society, Shenzhen 518172, China, and also with the School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K. (e-mail: guoping.qiu@nottingham.ac.uk).

Digital Object Identifier 10.1109/TIP.2020.3033402

I. Introduction

IMAGE defogging is an important research issue in the field of computer vision. It is widely used in many applications, such as urban traffic monitoring, remote sensing, military reconnaissance, etc. Its purpose is to restore degraded images caused by fog. Although defogging methods have made great progress in the last few decades, how to evaluate the visual quality of defogged images (DFIs) is understudied and an unsolved problem. The lack of a good quality measure of DFIs makes it very hard to compare the relative merits of defogging techniques and is hindering technological advancement. In this paper, we fill a gap in the literature by first presenting an image database suitable for evaluating defogging algorithms and then an image quality assessment (IQA) method specifically designed for evaluating defogging methods.

A. Defogging Techniques

Atmospheric degradation model [30], [34], [35] is widely used to describe the formation of foggy images, which can be formally expressed as:

$$I(x) = J(x)T(x) + A[1 - T(x)] \tag{1}$$

where $J(x)$ is the fog-free image and $I(x)$ is the observed foggy image, $T(x)(0 \leq T(x) \leq 1)$ denotes the transmission map, and A denotes the atmospheric light. If we want to recover J from I, we need to first determine the parameters A and T.

There are various methods developed for the fog removal problem in the literature. These methods are classified into prior-based methods, fusion-based methods, and learning-based methods. Based on the atmospheric degradation model (1), both prior-based methods or learning-based methods are pursuing the goal of accurately estimating the transmission map $T(x)$ and atmospheric light A.

1) Prior-Based Methods: These methods are also called hand-crafted fog removal methods, such as dark-channel Prior [18], color attenuation prior [61], contrast color-lines [13], hue disparity prior [3], haze-line prior [6], etc. They are all based on the atmospheric degradation model. Despite the remarkable defogging performance by these methods, hand-crafted features (such as textural, contrast,

etc.) have difficulty in achieving good performance in some cases since the assumptions are predetermined. For instance, it is assumed in dark-channel prior that at least one color channel of the clear image has some pixels whose intensity is very low and even close to zero except the sky area. Artifacts may occur in the defogged result by using this method when the scene objects are similar to the atmospheric light, such as the sky or white buildings.

2) Fusion-Based Methods: The characteristics of these methods are recovering the foggy image without using the atmospheric degradation model, and the performance is on par with the prior-based methods. Ancuti and Ancuti [2] provided a multi-scale fusion method for single image dehazing. They first extract three weight maps (luminance map, chromatic map and saliency map) from two enhancement inputs of a foggy image. Then they combined the Gaussian and Laplacian pyramids to fuse the weight maps with inputs to remove fog in a single image. Galdran [14] introduced an artificial multiple-exposure image fusion strategy to recover a foggy image. Several over-exposed images are first obtained from the original hazy image, and then use a Laplacian pyramid decomposition method to merge these images into a haze-free result. This method can recover a fog-free image, which has good contrast and vivid color information. In [16], Gao *et.al.* proposed a self-constructing image fusion method to recover a single foggy image based on a scale-invariant feature transform flow. However, all these methods need to obtain several derived inputs from the foggy image, which means they require a significant amount of work in the early. Moreover, the derived inputs do not well reflect the depth information of the scene, resulting in poor defogging performance when the fog density is high.

3) Learning-Based Methods: Recently, deep learning-based defogging methods have captured much attention. These methods [7], [38], [48] first estimate the transmission map by convolution neural networks, and then estimate the atmospheric light by the prior-based methods to recover the clear image with the atmospheric degradation model. Others use cGAN [24] or network based on encoder-decoder structure [22], [36], [39] to directly recover the image without using the atmospheric degradation model. Either way, both of these methods can overcome the disadvantage of prior-based methods. However, learning-based methods are data-driven. The training data of these methods almost always rely on synthesized foggy images, and the results usually contain artifacts, color distortions and even fail to process real foggy images. To solve this problem, methods based on training unpaired samples have been developed, such as Cycle-dehazing [11], PBGAN [28], etc. These methods do not require the preparation of synthetic foggy images which is time-consuming and laborious. They can better reflect the distributions of real-world foggy images and can generate good defogged results in terms of sharpness and brightness.

B. Foggy Image Datasets

Today, there are still very few foggy image datasets available. Foggy images in all defogging datasets are synthesized by the atmospheric degradation model (1) with known depth information or a professional haze machine. Most works first select the clean images and their corresponding depth maps from the NYU dataset [47], and then use them with the model (1) to synthesize the foggy images. However, these depth maps are captured indoors and models built using indoor scenes are necessarily suitable for processing outdoor scenes. Moreover, there exist very few real world naturally occurring fog datasets that contain paired fog and fog-free ground-truth images for evaluating and comparing defogging methods. In [4] and [5], the authors stated that the foggy images are real, however, the fog is generated by a professional machine, therefore the foggy images are still synthesized. In [57], HAZERD contains fifteen real outdoor scenes, and for each of which five different weather conditions are synthesized. There are five subsets in RESIDE [23] including both indoor and outdoor images. However, the foggy images are also synthesized from their clear versions. In [1], Ancuti *et.al.* introduce a dataset that contains 1400 foggy images. These images are derived from the Middlebury [41] and NYU [47] depth datasets. The depth map associated with each clear image was used to yield a synthesized foggy image based on the optical model (1). In [59], the authors stated that they proposed the first real-world foggy dataset. This dataset contains 208 pairs of nature images, which are collected from 23 provincial capital cities in China. However, for each clear reference image, there are serious image matching problems in the corresponding foggy image.

From the above discussion, it is clear that most datasets used in the defogging literature are synthesized. The use of synthetic data is due to the fact that it is difficult to capture paired clear and foggy images of the same nature scene. In particular, obtaining a large number of such paired images sufficient for developing defogging models is very challenging. Another reason for using synthetic data is that it is more convenient to use traditional IQA methods (such as PSNR, SSIM.) to evaluate the defogging results because synthesized clear and foggy images have identical scene illumination. Although it is relatively easy to synthesize a foggy image if the depth map is known, it is difficult to generate a foggy image for scenes that have long-distance views. For instance, we can not synthesis fog for a distant mountain. Moreover, the fog in a synthetic foggy image is uniformly distributed, whereas the atmosphere is heterogeneous. Therefore, to facilitate research in defogging technology, there is an urgent need for a true real-world database containing naturally occurring paired fog and fog-free images of nature scenes.

C. Defogging Quality Assessment

In the field of IQA, full-reference (FR) IQA and non-reference (NR) IQA have attracted extensive research attention. The distinction between these two type methods is that the former has reference images and the latter does not, which means the 'perfect' quality image is available in FR-IQA. Thus, the FR-IQA metric is correlated better with human perceptions than the NR-IQA metric [29]. Moreover, for the defogging task, there is usually some kinds of distortions in defogged result, such as color artifacts, distorted contrast, etc. The NR-IQA metric is not very sensitive to

such information. As a result, this method is not reliable in evaluating the defogged image. Therefore, in this work, we focus on the FR-IQA metric for evaluating the defogged result.

A suitable FR-IQA method should calculate quality degradation that is consistent with subjective evaluation, which herein is generally referred to as the Mean Opinion Score (MOS). In most single image defogging literature, many authors use mean squared error (MSR), peak signal-to-noise ratio (PSNR), or structural similarity (SSIM) [52] to evaluate the defogged image quality. Although these methods are mathematically simple and easy to use, they can't truly reflect the quality of defogged images very well. For example, for a foggy image, part of the textural and structural information is lost due to the presence of fog. This results in the edge information of the image being badly damaged. The SSIM will fail to predict the quality of the image since it underestimates the effect of this information [26]. In order to improve SSIM, many IQA methods have been designed to evaluate distorted images, such as multi-scale SSIM (MS-SSIM) [54], information content weighted SSIM (IW-SSIM) [53], information fidelity criterion (IFC) [45], visual information fidelity (VIF) [44], etc. However, these methods are still unable to consistently reflect the MOS of subjective evaluation when used to measure the defogged images. We will discuss this in section III. Therefore, it is important to design a suitable method specifically for evaluating defogged images. Unfortunately, there are few evaluation methods specifically for this type of degraded images.

To our best knowledge, there are only four IQA methods designed for evaluating defogged images, including three NR-IQA based methods [9], [17], [20] and one FR-IQA based method [59]. In [17], a blind contrast enhancement assessment method is proposed to evaluate defogged images. However, this method is based on the visible edges of the image. There is usually some remaining fog in the defogged image, which will cause the method to miscalculate the fog as visible edges. In [9], a model called Fog Aware Density Evaluator (FADE) is proposed to predict the fog density to evaluate the defogged image by calculating the deviations of twelve fog aware statistical features (such as low contrast, sharpness, image entropy, colorfulness, etc.) from the foggy image and the statistical characteristics observed in natural foggy and fog-free images. However, it is not clear why these twelve fog aware features are selected. Besides, the deviations range from 0 to $+\infty$, which means the level of fog density cannot be determined due to the lack of upper bound. In [20], a surrogate model based method is proposed to estimate the fog density by learning a polynomial regression for the depth map with seven fog-relevant features. However, this model can not accurately estimate the fog density from the depth map due to its fog-relevant features are extracted from synthesized foggy images. In [59], two independent criteria are proposed to evaluate the defogged result, which are the visibility index (VI) and the realness index (RI). In their work, VI is used to evaluate the haze level of the foggy image or defogged result, and RI is used to evaluate the distortions of the defogged result. However, in practice, we still cannot comprehensively and intuitively judge the quality of the defogging result through these two indicators due to they usually cannot give a unified opinion.

From the above discussions, it is very clear that current defogging research literature is weak in the following two areas which in many ways are hindering progress for the defogging research. First, there is no true real-world and naturally occurring foggy image datasets suitable for developing defogging models. Second, there are no suitable IQA methods for evaluating defogged images that are mathematically simple and easy to use. Around these two aspects, this paper makes the following contributions:

- Firstly, we introduce a new foggy image dataset called multiple real-world foggy image defogging dataset (MRFID). MRFID contains foggy and clear images of 200 outdoor scenes. For each scene, one clear image and 4 foggy images of different densities defined as slightly foggy, moderately foggy, highly foggy, and extremely foggy, are manually selected from images taken from these scenes over the course of one calendar year.
- Secondly, we present the first and comprehensive study of fog-relevant features of naturally occurring real foggy images. Moreover, we establish a defogging benchmark database, which contains 200 clear images as the reference and 12,800 DFIs created using 16 methods on 800 naturally occurring real foggy images. We have also designed a specific software system to subjectively evaluate the defogging visual quality, and 120 individuals participate in a psychovisual experiment procedure to obtain reliable MOS.
- Thirdly, based on the observation that the main factors affecting defogging quality are fog density and artifact distortion, we propose a new method to assess the visual quality of DFIs by separately considering these two aspects. For evaluating fog density, we proposed two similarities of dark channel feature and Mean subtracted contrast normalized (MSCN) feature to measure the fog density changes between reference and defogged image. For evaluating the artifact, gradient similarity and $Chroma_{HSV}$ similarity measures are calculated to measure the artifacts of texture distortion and color distortion respectively. We first calculate the individual similarity components to obtain the respective scores, and then fuse these scores into a final single evaluation score through a pooling procedure.
- Finally, extensive experiments conducted on MRFID and other fog database show that our new visual quality assessment method is more suitable for evaluating defogged images than other state-of-the-art IQA methods and significantly outperforms them.

II. A Database for Defogging Quality Assessment

In this section, we will introduce the detials of MRFID and the DFIs database, including the perparation of defogged images, subjective evaluation and data analysis.

A. Preparation of Defogged Images

In MRFID, the foggy images are manually picked from an image dataset called the Archive of Many Outdoor

(a) Clear (b) Slightly (c) Moderately (d) Highly (e) Extremely

Fig. 1. Example of four fog density images correspond to clear images in MRFID. (a) Clear image. (b)∼(e) Four different fog density images corresponding to (a).

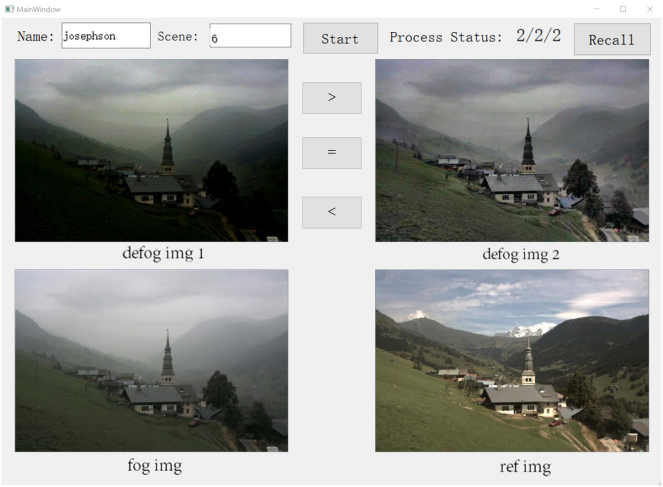


Fig. 2. Screenshot of the subjective evaluation software interface.

Scene (AMOS) [19], in which images were captured by 29,945 static webcams located around the world, and contains 1,128,087,180 images from 2006 to 2017. In MRFID, images of each scene were manually selected from images taken by a fixed camera over the course of one year. The image sizes range from 640×480 to $2,284 \times 914$. In this work, we have collected images from 200 natural outdoor scenes. Different from the real-world dataset BeDDE [59], each clear image has multiple fog density foggy images. And our image pairs do not have the image registration problem. The reason is that, first, our images come from fixed webcams, second, we have a large number of image pairs for the same scene. Thus, we can simply apply the image registration algorithm to filtrate the foggy images inconsistent with a clear image. Moreover, for each scene in our dataset, there are one well-aligned clear image and four corresponding images with different fog densities. The 200 clear images served as the reference images, and the scenes mainly include mountain, village, city, harbor, etc. Some examples are shown in Fig. 1. This dataset will be made available to researchers.

To obtain the defogged images, we use 16 defogging methods. These methods include 9 prior-based algorithms, 2 fusion-based algorithms, and 5 learning-based algorithms.

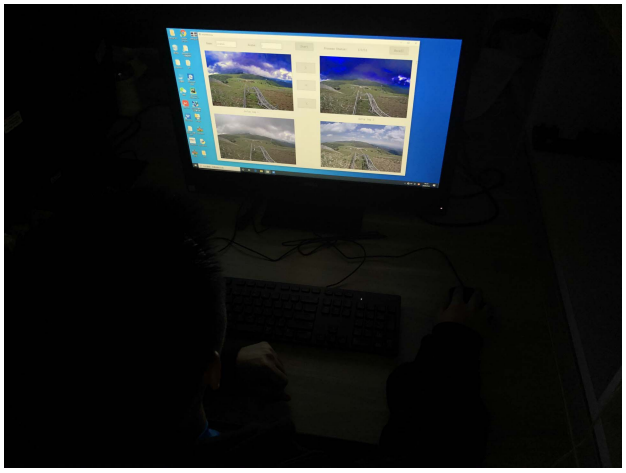


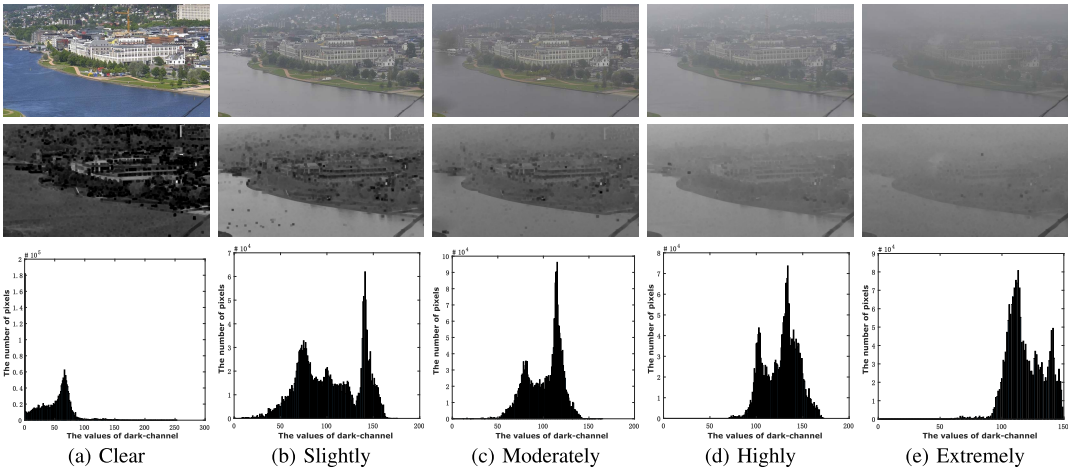
Fig. 3. Subjective evaluation environment.

The prior-based methods are dark channel prior (DCP) [18], non-local dehazing (NL) [6], DehRet [15], color attenuation prior (CAP) [61], Multi-scale optimal fusion dehazing (MOF) [58], fast visibility resrotation (FVR) [50], single image dehazing(SID) [12], fog density perception defogging (FDPD) [25] and boundary constraint and contextual regularization dehazing (BCCR) [31]. The fusion-based methods are multi-scale fusion dehazing (MSF) [2] and artificial multiple-exposure image fusion defogging (AMEF) [14]. The learning-based methods are Dehazenet [7], MSCNN [38], PBGAN [28], gated fusion network (GFN) [39] and enhanced Pix2Pix network (EPDN) [36]. Therefore, based on those methods, we can obtain 12,800 defogged results from MRFID, we called them DMRFIs. For each reference image, the number of defogged images (DFIs) is 64.

B. Subjective Evaluation and Data Analysis

To obtain the MOS data of subjective evaluation of the DFIs, we recruited 120 students to conduct the psychovisual experiments using a software interface as shown in Fig. 2. The experiment was performed in a dark indoor room. The subjective software was conducted on a 23.8-inch liquid crystal display monitor with spatial and temporal resolutions of 1440 by 900 pixels and 60Hz. Moreover, as suggested in [21], subjects were required to sit at least twice the screen height away from the screen to maintain the best evaluation effect. During the whole subjective evaluation, the other configurations of the monitor remained in the default state and unchanged, such as color calibration, brightness, and contrast. The evaluation environment is shown in Fig. 3.

As shown in Fig. 2, there are four image windows in the software interface. The top row is two defogged images to be evaluated, they are generated by different defogging methods from the same foggy image shown in the bottom-left window, and share the same reference image shown in the bottom-right window. The bottom-left window displays the foggy image, and the bottom-right displays the reference image (the clear image). Note that, in this subjcetive evaluation experiment, we add the original foggy image to the evaluation process and find some constructive results. We will discuss these results



(a) Clear (b) Slightly (c) Moderately (d) Highly (e) Extremely

Fig. 4. The dark-channel features of a clear image and different density foggy images. First row: the clear image and different density foggy images. Second row: the dark-channel feature maps of the first row. Third row: the histogram statistics of the second row.

in IV-D. When a subject uses this software to evaluate the DFIs, he/she first needs to fill in the "name" and "scene" then click the "start" button to start the evaluation journey. Besides, the numbers 2/2/2 in the interface represent progress status, where the first two numbers mean the evaluated image is the moderately foggy image of the second scene. The last number "2" refers to the number of comparisons, 17 in total. During the evaluation, subjects were instructed to click the button ">", "<", or "=" on the interface to give their judgments that which defogged image has the better quality. Especially, the button "=" is selected by the subject that he/she is hard to distinguish which one is better or consider they have equal quality. If the subject considers himself/herself gave a wrong rating or accidentally clicks the wrong button, he/she can cancel it by clicking the button "recall". Moreover, in this psychovisual evaluation, we employed the pair comparison sorting scheme [49], [60] to score the quality of the defogged images. For each reference image, sixteen defogged images are presented to the subject for comparative rating. Thus, we will obtain a rank ordered numbers range from 1 to 17. A higher number means the best quality.

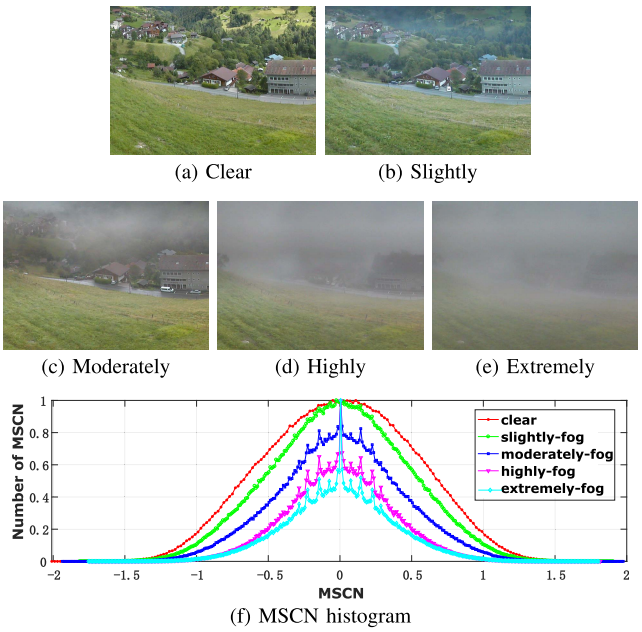
To obtain the MOS for each DFI, we first normalize the individual scores to the range of (0,10), as follows:

$$S_N(i) = 10 \times \frac{max(S_n(i)) - S_n(i)}{max(S_n(i)) - min(S_n(i))} \tag{2}$$

where $S_N(i)$ is the normalized result of the i^{th} DFIs ($i = 1, 2, \ldots, 12800$), $S_n(i)$ is the score of subject n ($n = 1, 2, \ldots, 120$) assigned to the i^{th} DFIs. The higher the quality score the best the quality rating. Then, we use the method in [46] to remove the outlier to ensure the reliability of the evaluation scores. Finally, the remaining valid scores are averaged as the MOS of each DFI.

III. Fog-Relevant Feature Based SIMilarity Index (FRFSIM) for DFIs

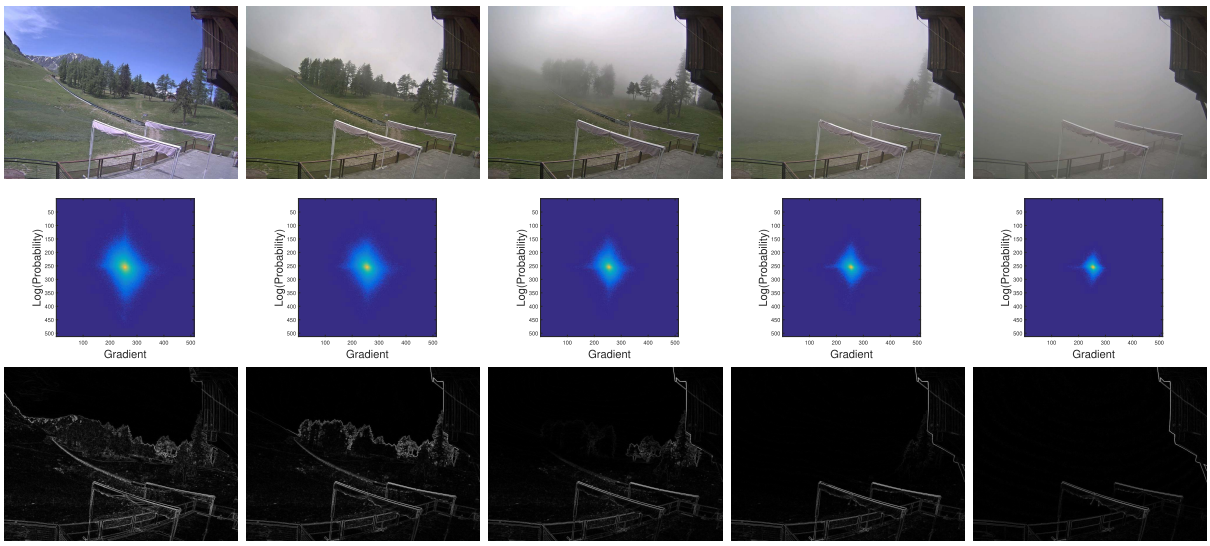
Through a large number of experiments, we observe that fog density and artificial distortion are two major factors affecting the defogging quality. Especially for the artificial distortion,



(a) Clear (b) Slightly (c) Moderately (d) Highly (e) Extremely (f) MSCN histogram

Fig. 5. The MSCN feature extracted from the clear image and different fog density images. (a)-(e) The clear image and four fog density images. (f) The histogram of MSCN.

the color distortion and loss of the texture information are a frequent occurrence in the defogged results. Thus, the strategy for evaluating the DFIs should be based on both fog density and artificial distortion. In this section, four fog-relevant features are used to evaluate the DFIs, including dark channel feature [18], MSCN feature [32], gradient features and $Chroma_{HSV}$ [51] features. In these features, dark channel and MSCN are used to measure the fog density, and the changes of the artificial distortion between the reference and defogged image are measured with the gradient(which related to texture changes) and the $Chroma_{HSV}$(which related to color distortion) features. Note that when we analyze these features of the images with different fog density (as shown in Fig. 4, Fig. 5, Fig. 6, and Fig. 7), we choose the foggy images of the same scene in the natural environment in our database MRFID.



(a) Clear (b) Slightly (c) Moderately (d) Highly (e) Extremely

Fig. 6. The gradient features of the clear image and different density foggy images. Top row: clear image and different density foggy images. Second row: the distribution of log gradient for the clear image and different density foggy images. Last row: the gradient map of the clear image and different density foggy images.

A. Dark Channel Feature

Dark channel prior is proposed by He *et al.* [18] and is a baseline approach for most defogging works and is mathematical simple and effective. The theory of Dark channel prior is that in most local areas (excluding sky regions) of outdoor haze-free images, at least one color channel will have some pixels whose intensity is very low and even close to zero. For an arbitrary image, its dark channel can be defined as follows [27]:

$$D(x) = \min_{i\in\Omega(x)} \left(\min_{c\in\{r,g,b\}} I^c(i) \right) \tag{3}$$

where I^c is a color channel of an image, $\Omega(x)$ is a square area centered at x. The dark channel of an image is obtained by performing the minimum operation twice: $\min_{c\in\{r,g,b\}}$ is used to process each pixel for each RGB channel, $\min_{i\in\Omega(x)}$ is a minimum value filter. For a fog-free image, hence, the intensity of its dark channel is very low and approaches zero ($d(x) \to 0$) except the sky region. Fig. 4 shows the dark-channel features of a clear image and four different fog density images of the same scene. As we can see that the structures in the dark-channel feature maps gradually blur and fade away as fog density increases. This phenomenon is due to the fact that higher fog density makes the pixels have larger intensity values, resulting in a brighter dark-channel map. The histograms of the dark channels of the clear images and their foggy images of different densities indicate that the dark-channel feature can approximately represent fog density in an image. Similar to SSIM [52], the dark channel similarity(DS) for the i-th pixel $FS_1(i)$ is defined as:

$$FS_1(i) = \frac{2D_r(i)D_d(i) + C_1}{D_r^2(i) + D_d^2(i) + C_1} \tag{4}$$

where the subscript r and d represent the reference image and the defogged image respectively. D_r denotes the dark channel feature of r. D_d denotes the dark channel feature of d. C_1 is a positive constant to prevent the denominator from being zero for dark channel similarity.

B. MSCN Feature

Mean subtracted contrast normalized (MSCN) coefficients were proposed by Mittal *et al.* [32] to evaluate natural image distortions. It is a regular natural scene statistical model that has been well established in the spatial domain [40]. We explore how this model mabe used to characterize different fog densities. The MSCN coefficients are formulated as [8], [20]:

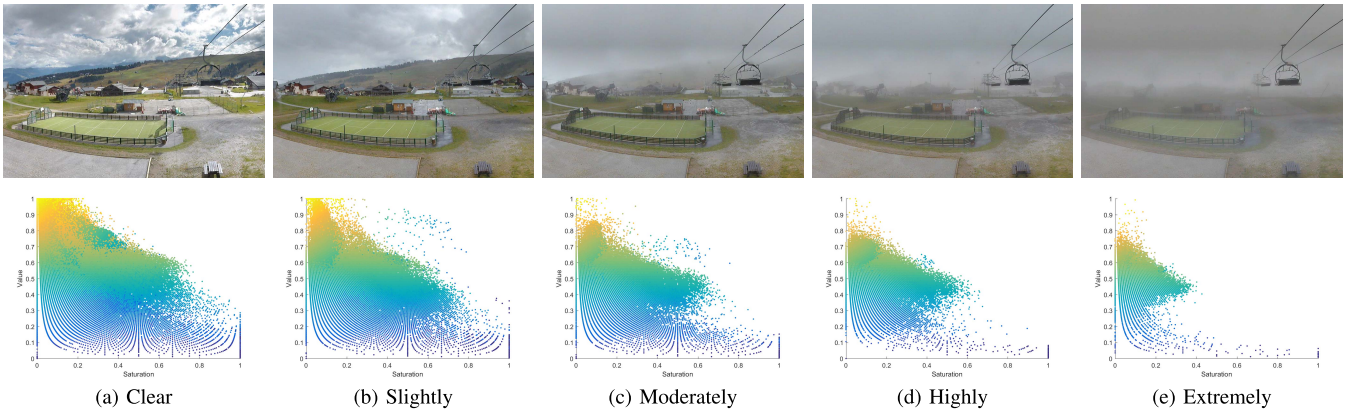
$$M(i,j) = \frac{I_{gray}(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{5}$$

$$\mu(i,j) = \sum_{u=-U}^{U} \sum_{v=-V}^{V} \eta_{u,v} I_{gray}(i+u, j+v) \tag{6}$$

$$\sigma^2(i,j) = \sum_{u=-U}^{U} \sum_{v=-V}^{V} \eta_{u,v} [I_{gray}(i+u, j+v) - \mu(i,j)]^2 \tag{7}$$

where I_{gray} is the grayscale of the foggy image, i and j are pixel indices ($i \in [1, M]$, $j \in [1, N]$, M and N are the image dimensions), and η is a 2D circularly symmetric Gaussian weighting function. μ and σ^2 denote the local mean values and the local normalized variance of a foggy image respectively. Statistics of MSCN feature from a clear image and its corresponding foggy images of different fog densities are shown in Fig. 5. In Fig. 5(f), we can see that the MSCN values decrease as the fog density increases. The MSCN values of the clear image are larger than that of the other images, and the extremely foggy image has the smallest MSCN values. This indicates that the MSCN features of a foggy image can quantitatively describe the fog density. Similar to (4), the MSCN similarity(MS) FS_2 is measured as:

$$FS_2(i,j) = \frac{2M_r(i,j)M_d(i,j) + C_2}{M_r^2(i,j) + M_d^2(i,j) + C_2} \tag{8}$$



(a) Clear (b) Slightly (c) Moderately (d) Highly (e) Extremely

Fig. 7. The $Chroma_{HSV}$ features of a clear image and different density foggy images. First row: the clear image and different density foggy images. Second row: the $Chroma_{HSV}$ feature maps of the first row.

where M_r and M_d denote the MSCN feature of the reference image and the defogged image respectively. C_2 is a positive constant to prevent the denominator from being zero for MSCN similarity.

C. Gradient Feature

It is well known that texture is an important feature of the spatial structure relationship between pixels in an image. It reflects the grayscale variation between image pixels. Moreover, the gradient of the image represents the rate of change of image grayscale, which is widely used to assess the structure information [10], [26], [60]. However, fog can change the structure information of an image greatly. Therefore, we chose the gradient feature to measure the texture changes between the defogged image and its clear reference image. For an image $f(x, y)$, the gradient of f is defined as:

$$\nabla f \equiv grad(f) \equiv \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{9}$$

where (x, y) represents the position of a pixel in the image, G_x and G_y denote the first-order derivative of f in x and y directions, respectively. Then, the gradient magnitude is calculated as:

$$G(i) = \sqrt{G_x^2(i) + G_y^2(i)} \tag{10}$$

where i denotes the i-th pixel. As shown in Fig. 6, the gradient decreases as the fog density increases. This indicates that the gradient features can correlate with texture information between foggy image and fog-free image. Based on this feature, the gradient similarity(GS) FS_3 is measured as:

$$FS_3(i) = \frac{2G_r(i)G_d(i) + C_3}{G_r^2(i) + G_d^2(i) + C_3} \tag{11}$$

where G_r and G_d represent the gradient feature of the reference image and the defogged image respectively. C_3 is a positive constant to prevent the denominator from being zero for gradient similarity.

D. $Chroma_{HSV}$ Feature

In computer vision and image processing, HSV (Hue, Saturation and Value) color space is widely used to extract the color information of an image. In [51], Javier *et. al.* calculated the histogram of the difference between a pair of the fog image and fog-free image in HSV space, and found that Saturation (S) component and Value (V) component are closely related to color changes. Thus, in this paper, we choose the S and V to measure the color distortion, which is defined as:

$$Chroma_{HSV}(i) = S(i) \times V(i) \tag{12}$$

where i stands for the i-th pixel. We can observe in Fig. 7 that the amount of the colorfulness in $Chroma_{HSV}$ decreases with increasing fog density in the image. As a result, the $Chroma_{HSV}$ feature can measure the color distortion in a foggy image or defogged result. Thus, similar to the above feature similarity, the color similarity(CS) FS_4 is measured as:

$$FS_4(i) = \frac{2Chroma_r(i)Chroma_d(i) + C_4}{Chroma_r^2(i) + Chroma_d^2(i) + C_4} \tag{13}$$

Note that, in (13), *chroma* denotes the $Chroma_{HSV}$. $Chroma_r$ and $Chroma_d$ represent the color feature of the reference image and the defogged image respectively. C_4 is a positive constant to prevent the denominator from being zero for color similarity.

E. Fog-Relevant Feature Similarity Index

To measure the final quality between a defogged image and its corresponding reference image, we need to pool the above similarity maps into a single score. Note that, in the above similarity maps, DS and MS are related to the fog, GS and CS are related to the artefacts. Thus, in our pooling strategy, we first pool the four similarity maps into four scores, and then combine the scores of DS and MS to measure the fog density (which is called the FD similarity) and combine the scores of GS and CS to measure the artefacts (which is called the AD similarity). Finally, we fuse the FD and AD into one. Each map score is defined as:

$$S_n = \frac{1}{N} \sum_{i=1}^{N} FS_n(i) \tag{14}$$

where N is the number of pixels in the image, S_n denotes the score of each similarity, $n \in \{1, 2, 3, 4\}$. Then the *FD* score and *AD* score are defined as:

$$S_{FD} = S_1 \cdot S_2$$
$$S_{AD} = S_3 \cdot S_4 \tag{15}$$

By fusing the above scores, the final single score *FRFSIM* is defined as:

$$FRFSIM = (S_{FD})^{\beta_1} \cdot (S_{AD})^{\beta_2} \tag{16}$$

where β_1 and β_2 are the positive parameters used to adjust the impact of different similarities, $\beta_1 + \beta_2 = 1$, $FRSFIM \in (0, 1)$. β_1, β_2 represents the impact factor of *FD*, *AD*, respectively. The higher the *FRFSIM*, the better the defogging quality. The appropriate values of β_1 and β_2 is determined empirically through analysing our data. We observed that they should be image dependant and should be related to dark channel similarity score, S_1 (see equation (11)). For a small S_1, the fog related similarity S_{FD} should be given more importance, β_1 should be small and β_2 should be large. For a large S_1, the artefact related similarity S_{AD} should be given more importance and β_2 should be small and β_1 should be large. We set β_1 and β_2 as follows:

$$\begin{cases} \beta_1 = \alpha, \beta_2 = 1 - \alpha; & 0 < S_1 < T \\ \beta_1 = 1 - \alpha, \beta_2 = \alpha; & T \leq S_1 < 1 \end{cases} \tag{17}$$

where α is a positive constant ($\alpha \in [0, 1]$), and T is a threshold($T \in (0, 1)$).

Fig. 8 shows a side-by-side comparison of different similarity maps and the *FRFSIM* maps for different fog density images and defogging methods. For the *DS* and *MS* maps in the third row and fourth row of Fig. 8, it can be seen that the edge information of them is not blurred with the increase of the fog density. However, the contrast between them is decreasing gradually. In comparison, as shown in Fig. 8 (f) - (h), these two similarity maps have changed significantly in defogged results. Especially for *MS* feature maps, as shown in the fourth row of Fig. 8 (g), the edge information is almost lost. This is because *DS* and *MS* are more sensitive to the light affected by fog than to the color distortion of the image. For the *CS* and *GS* maps in the fifth row and sixth row of Fig. 8, we can see that there are significant changes in the edge information of the house in both the foggy images and the defogged images. It is demonstrated that *CS* and *GS* can capture color distortion or structural distortion information caused by fog or defogging methods, respectively. Thus, as shown in the seventh row of Fig. 8, our fusion *FRFSIM* maps can effectively reflect the quality of fog map or defogging image. These are also reflected in the values of *FRFSIM*. As shown in Fig. 8 (b) - (e), with the increase of fog density, the value of *FRFSIM* gradually decreases, which means that the image quality becomes worse and worse. Moreover, it can be seen in Fig. 8 (f) - (h), compared with the original foggy image($FRFSIM = 0.2904$) in Fig. 8 (d), DehazeNet [7] has the best defog quality ($FRFSIM = 0.5202$), fellowed by DCP [18] ($FRFSIM = 0.5105$), and MSF [2] has the worst defog quality ($FRFSIM = 0.1333$).

IV. Experimental Results

In this section, we compare the proposed Fog-relevant feature based IQA method for DFIs with other state-of-the-art IQA methods to demonstrate its effectiveness by conducting experiments on our true real fog image database MRFID. 10 representative full-reference (FR) and 3 no-reference (NR) IQA methods are used for comparison. The FR-IQA metrics are PSNR, SSIM [52], GSM [26], IFC [45], VSI [55], DeltaE [43], Haar [37], VI [59], RI [59] and LPIPS [56]. The NR-IQA metrics are FADE [9], NIQE [33], BRISQUE [32]. Moreover, we qualitatively and quantitatively evaluate the defogging methods by using our IQA metric and other IQA metrics.

A. Evaluation Criteria and Parameter Setting

Four criteria are employed to evaluate the performance of FRFSIM: Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank order correlation coefficient (KROCC), and root mean squared error (RMSE). These metrics can be directly calculated between the predicted scores of the IQA methods and the subjective mean opinion scores (MOS) in Section II. Here, we employ SROCC and KROCC to measure the prediction monotonicity of IQA methods and use PLCC and RMSE to measure prediction accuracy. However, to calculate the PLCC and RSME, a non-linear regression is required to fit the objective model scores to the subjective scores by using a logistic function. According to the literature [42], we use the following logistic regression function:

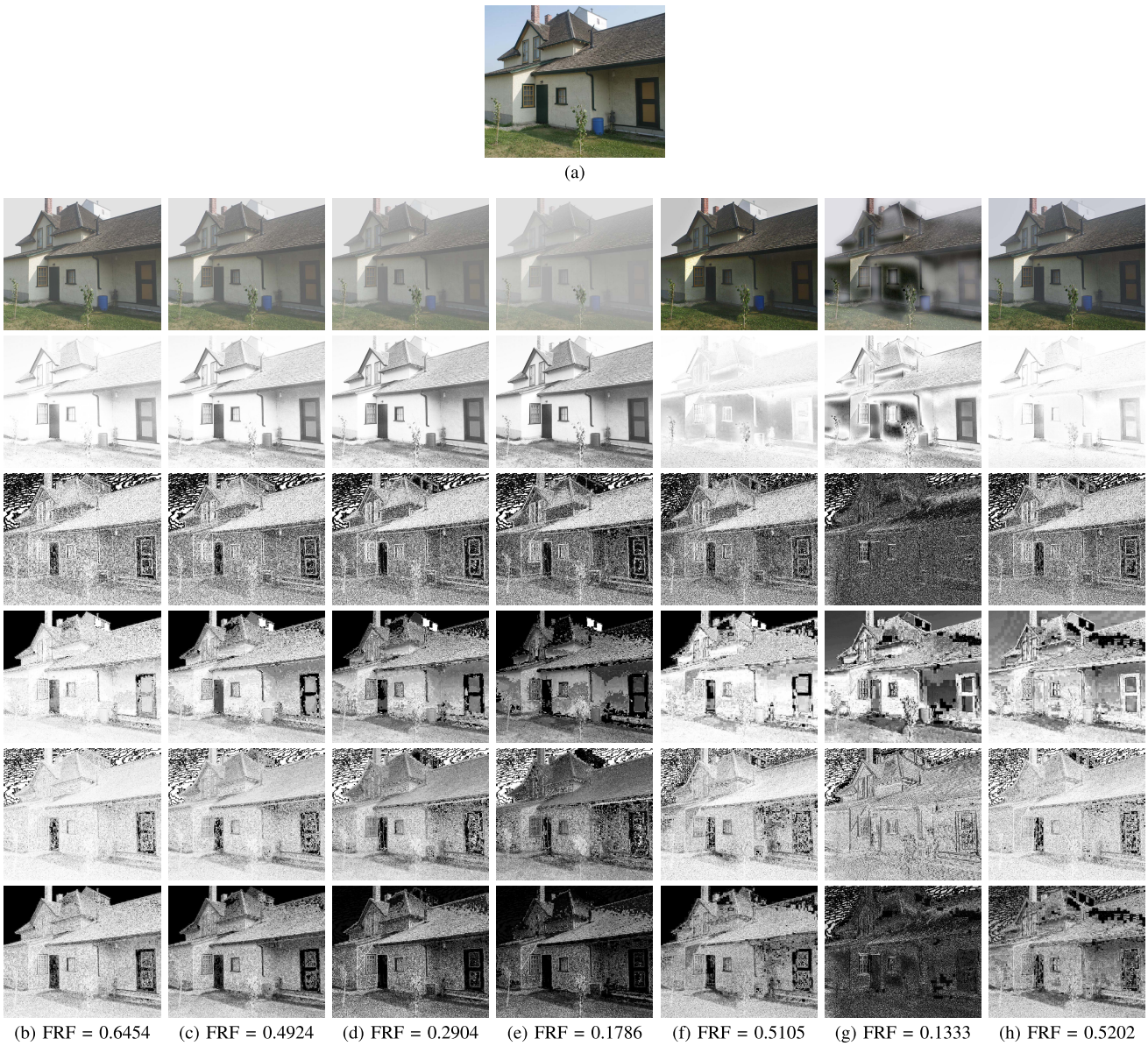
$$Q(x) = \lambda_2 + \frac{\lambda_1 - \lambda_2}{1 + e^{-(x - \lambda_3/|\lambda_4|)}} \tag{18}$$

where x denotes the output of the IQA methods, λ_1, λ_2, λ_3 and λ_4 are the optimal parameters that we use the Matlab function "nlinfit" to fit between the subjective scores and the objective scores, $Q(x)$ denotes the regression values of x. A better IQA metric will have higher SROCC, KROCC, and PLCC and lower RMSE.

In our proposed IQA model, several parameters are required to be determined. Similar to SSIM [52], C_i is assigned by $C_i = (K_i L)^2, i \in \{1, 2, 3, 4\}$, where L is the dynamic range of the pixel values (255 for 8-bit gray scale images), and $K \ll 1$. Through extensive experiment, we set these four positive parameters to following values: $K_1 = 0.0001$, $K_2 = 0.00005$, $K_3 = 0.00045$, and $K_4 = 0.0009$. Moreover, to determine the values of β_1 and β_2 in equation 17, we do the following: if S_1 is under the threshold of T = 0.85, $\beta_1 = 0.2$, $\beta_2 = 0.8$, and over that threshold $\beta_1 = 0.8$, $\beta_2 = 0.2$. This means that, for a small value of the dark channel (little fog), the index relies less on S_1 and S_2 (fog density related similarity maps), and vice-versa. These values T and β are obtained empirically.

B. Metric Performance Comparison

In this experiment, we compare the experimental results of different IQA methods on DFIs to demonstrate the superiority of our proposed criteria for defogging evaluation. As shown



(a)

(b) FRF = 0.6454 (c) FRF = 0.4924 (d) FRF = 0.2904 (e) FRF = 0.1786 (f) FRF = 0.5105 (g) FRF = 0.1333 (h) FRF = 0.5202

Fig. 8. Comparison of different similarity maps for different fog density images and defog methods. (a) Reference image. (b)-(e) Different fog density images with their corresponding similarity maps. The fog density is increasing from left to right. (f)-(h) The defogged results of (d) with their corresponding similarity maps produced by the approaches of DCP [18], MSF [2], and DehazeNet [7]. Second row: The original foggy images and defogged results. Third row: The dark channel similarity maps. Fourth row: The MSCN similarity maps. Fifth row: The $Chroma_{HSV}$ similarity maps. Sixth row: The gradient similarity maps. Seventh row: The *FRFSIM* maps. Note that FRF means our defogging quality score *FRFSIM*.

in Table I, for each criterion in different fog density groups, the best result is highlighted in boldface. It can be observed that the values of our method for PLCC, SROCC, and KROCC are larger than the compared methods, and for RMSE are smaller than that of the others. This means that, for evaluating the DFIs, these metrics are not well correlated with subjective evaluation scores. To give a few examples, the performance for PSNR and NIQE [33] is rather poor. For the DFIs, the widely-used evaluation criteria SSIM [52] also performs poorly, although it has better values than PSNR. Similar to SSIM [52], the performance of the methods based on image color and gradients are also unsatisfactory, e.g., GSM and DeltaE. Although the color or gradient information is an important factor in the evaluation of the defogging images, the reason for the above results is that none of these methods takes into account the fog density in the image. On the other hand, the methods specially designed for evaluating the defogging quality, VI [59], RI [59] and FADE [9], take fog density into account, but neglect the other information of the image. In particular, for VI [59], it uses the transmission map and gradient map to predict the visibility of the image, but ignores the change of color information in the defogging image. For RI [59], the phase congruency and Chrominance information are used to predict the artifacts from the defogged images, but the information about the fog density in the images is ignored. Thus, it can be seen from Table I that their performances are poor. It is interesting to note that the performance of LPIPS [56] performed very well and was second only to our

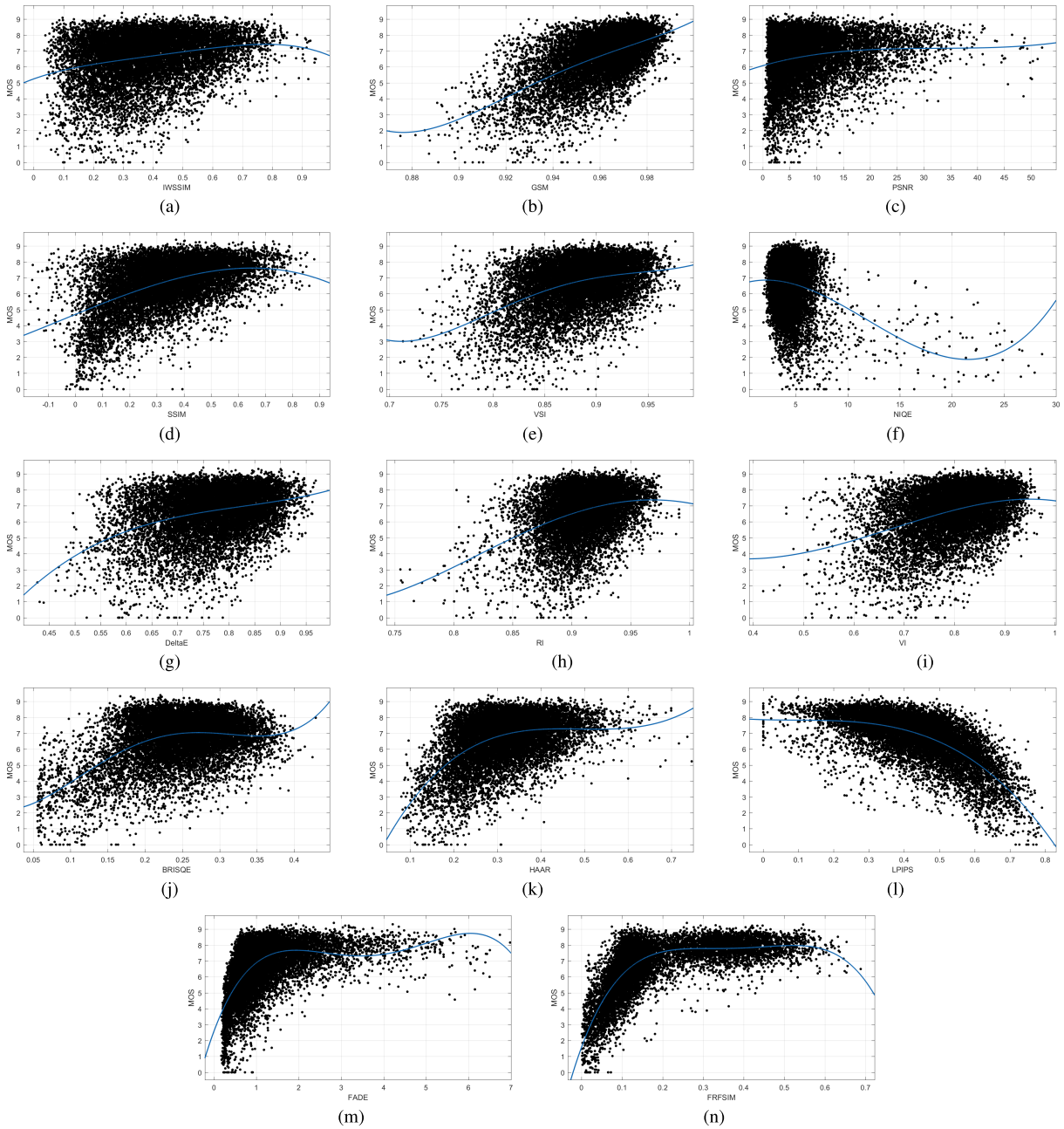


Fig. 9. Scatter plots of objective IQA scores versus MOS for all defogged images from our MRFID database by using 16 defog methods. (a)IW-SSIM [53]. (b) GSM [26]. (c) PSNR. (d) SSIM [52]. (e) VSI [55]. (f) NIQE [33]. (g) DeltaE [43]. (h) RI [59]. (i) VI [59]. (j) BRISQUE [32]. (k) HAAR [37]. (l)LPIPS [56]. (m) FADE [9]. (n) FRFSIM.

method. LPIPS is based on CNN features which means that convolutional neural networks can extract image features for effectively evaluate defogging image quality. These are also reflected in Fig. 9.

Fig. 9 shows the scatter plots of objective scores versus MOS obtained on DFIs for fourteen IQA models. From it, we can see that the points obtained by our proposed IQA method are more tightly distributed on the fitted curve than other methods. In contrast, the points obtained by other IQA models are scattered loosely around the fitted curve. It indicates that our method has a better consistency with the subjective evaluation scores, and it is more suitable for evaluating DFIs.

C. Qualitative and Quantitative Comparison on DFIs

In this section, we qualitatively and quantitatively compare the DFIs by using our proposed IQA method FRFSIM and other methods, including LPIPS [56], SSIM [52], and RI [59]. The defogging methods include DCP [18], CAP [61], Dehazenet [7], NL [6], DehRet [15], MOF [58], FVR [50], SID [12], FDPD [25], BCCR [31], MSF [2], AMEF [14], MSCNN [38], PBGAN [28], GFN [39], EPDN [36].

Qualitatively, as shown in Fig. 10, we compare the defogged results from different fog density images by using five representative defogging methods. And quantitatively, we evaluate these results with the metrics of FRFSIM, SSIM and LPIPS.

TABLE I

PERFORMANCE COMPARISON FOR DIFFERENT IQA METHODS ON DFIS

		IQA mtehods													
		PSNR	SSIM [52]	GSM [26]	IWSSIM [53]	DeltaE [43]	VSI [55]	Haar [37]	LPIPS [56]	VI [59]	RI [59]	FADE [9]	NIQE [33]	BRISQUE [32]	FRFSIM
Slightly	PLCC	0.2536	0.5187	0.5988	0.4486	0.5345	0.4321	0.5215	0.7192	0.5449	0.4682	0.6428	0.3270	0.5101	**0.8775**
	SROCC	0.2442	0.4565	0.5336	0.4211	0.4780	0.3589	0.4735	0.6833	0.5114	0.4032	0.5240	0.1841	0.4381	**0.8027**
	KROCC	0.1965	0.3179	0.3762	0.3488	0.3878	0.4458	0.4556	0.5019	0.4819	0.3944	0.4672	0.1551	0.3591	**0.7162**
	RMSE	0.5177	0.3131	0.2756	0.4877	0.4475	0.3762	0.3105	0.2177	0.4416	0.4288	0.2365	0.4915	0.3011	**0.1688**
Moderately	PLCC	0.1778	0.5073	0.5833	0.4218	0.4485	0.4493	0.5031	0.6925	0.4727	0.4076	0.5726	0.2503	0.5011	**0.8172**
	SROCC	0.1495	0.4254	0.5293	0.3168	0.3254	0.3642	0.3685	0.6635	0.4483	0.3984	0.5442	0.1213	0.4291	**0.7837**
	KROCC	0.1041	0.3959	0.4717	0.2893	0.3020	0.3294	0.3420	0.4837	0.3688	0.3025	0.4824	0.1062	0.3588	**0.6862**
	RMSE	0.4966	0.3105	0.2309	0.4677	0.4299	0.3586	0.3145	0.1971	0.3113	0.3143	0.2516	0.4725	0.3112	**0.1576**
Highly	PLCC	0.2175	0.5355	0.6204	0.2985	0.3630	0.4543	0.5277	0.7015	0.3868	0.3969	0.4981	0.2527	0.5130	**0.7881**
	SROCC	0.1859	0.4836	0.5696	0.2828	0.3242	0.4026	0.4281	0.6673	0.3671	0.3441	0.5454	0.1628	0.3537	**0.7353**
	KROCC	0.1457	0.3379	0.4013	0.1912	0.2195	0.2763	0.3956	0.4864	0.3507	0.2946	0.4807	0.1412	0.2694	**0.5451**
	RMSE	0.4836	0.2837	0.1921	0.4507	0.4163	0.3541	0.2915	0.1832	0.4017	0.3952	0.3220	0.4907	0.3047	**0.1357**
Extremely	PLCC	0.1705	0.5179	0.6188	0.2381	0.3553	0.4462	0.5247	0.7016	0.4015	0.3601	0.5079	0.2851	0.4991	**0.7293**
	SROCC	0.1376	0.4445	0.5392	0.2139	0.2569	0.3393	0.3758	0.6502	0.3418	0.2874	0.5523	0.0655	0.2278	**0.6252**
	KROCC	0.0926	0.3114	0.3789	0.1437	0.1725	0.2314	0.2569	0.4721	0.2329	0.1939	0.3870	0.0434	0.1512	**0.5381**
	RMSE	0.5115	0.3122	0.2049	0.4898	0.4339	0.3727	0.3059	0.1931	0.4052	0.4313	0.3263	0.5703	0.3292	**0.1879**

TABLE II

QUANTITATIVE COMPARISON OF DIFFERENT DEFOGGING METHODS WITH DIFFERENT IQA METHODS ON OUR MRFID. THE TOP THREE PERFORMANCE VALUES ARE HIGHLIGHTED IN RED, BLUE AND GREEN

Method	Slightly			Moderately			Highly			Extremely		
	FRFSIM	SSIM [52]	RI [59]	FRFSIM	SSIM [52]	RI [59]	FRFSIM	SSIM [52]	RI [59]	FRFSIM	SSIM [52]	RI [59]
DCP [18]	0.3179	0.4390	0.9249	0.2307	0.3777	0.9186	0.2052	0.3447	0.9124	0.1866	0.3266	0.9071
CAP [61]	0.3354	0.3858	0.9206	0.2798	0.3350	0.9173	0.2523	0.3192	0.9127	0.1675	0.2918	0.9081
Dehret [15]	0.1536	0.1415	0.8982	0.1511	0.1318	0.8957	0.0932	0.1231	0.8900	0.0509	0.1031	0.8853
SID [12]	0.1585	0.2020	0.8974	0.0962	0.1960	0.8912	0.0733	0.1930	0.8805	0.0636	0.1835	0.8693
FVR [50]	0.4093	0.4593	0.9208	0.3724	0.4046	0.9133	0.2823	0.3580	0.9066	0.2625	0.3245	0.9011
FDPD [25]	0.5242	0.4527	0.9203	0.3266	0.4035	0.9130	0.2523	0.3192	0.9127	0.2305	0.3347	0.9015
BCCR [31]	0.3900	0.4576	0.9208	0.2948	0.3901	0.9140	0.1978	0.3503	0.9084	0.1688	0.2977	0.8977
MOF [58]	0.4013	0.4320	0.9200	0.3869	0.3705	0.9126	0.2603	0.3197	0.9038	0.1297	0.2513	0.8944
NL [6]	0.3564	0.3453	0.9169	0.3182	0.2986	0.9132	0.2344	0.2914	0.9093	0.1487	0.2430	0.9040
AMEF [14]	0.4752	0.4587	0.9298	0.4515	0.4077	0.9223	0.3282	0.3753	0.9154	0.2757	0.3435	0.9073
MSF [2]	0.3883	0.3678	0.9077	0.2840	0.3293	0.8997	0.2642	0.2918	0.8854	0.1587	0.2715	0.8749
Dehazenet [7]	0.4771	0.4235	0.9312	0.3674	0.3581	0.9249	0.2571	0.3646	0.9175	0.2638	0.2958	0.9116
EPDN [36]	0.2561	0.4082	0.9269	0.2056	0.3670	0.9027	0.1389	0.3226	0.9127	0.1252	0.2847	0.9063
GFN [39]	0.3511	0.4512	0.9330	0.3071	0.3691	0.9254	0.2541	0.2904	0.9163	0.2149	0.2526	0.9078
PBGAN [28]	0.4965	0.4688	0.9314	0.3885	0.3914	0.9138	0.3590	0.3463	0.9077	0.2667	0.2996	0.9019
MSCNN [38]	0.3598	0.4686	0.9226	0.3186	0.4150	0.9256	0.2252	0.3805	0.9188	0.1648	0.3360	0.9117

In addition, we also use these metrics to evaluate the original foggy images. It can be seen from left to right in the first row of Fig. 10, the fog density is increasing. Accordingly, our FRFSIM is gradually decreasing, and LPIPS is gradually increasing. However, in contrast, SSIM does not show similar monotonicity. The SSIM of the extremely foggy image (as shown in Fig. 10 (e), SSIM = 0.342) is larger than the moderately (SSIM = 0.316) and slightly (SSIM = 0.282) foggy image(as shown in Fig.10(c) and (b) respectively). It means that SSIM is not appropriate for evaluating the foggy images. Similarly, as shown in Fig. 10 (f) - (j), the SSIM criterion is not suitable for evaluating the defogging image. From top to bottom in Fig. 10 (f) - (j), as the fog density increases, SSIM is ineffective in evaluating the defogged results. For instance, as shown in the third row of Fig. 10 (g) and (h), our FRFSIM and LPIPS are consistent in determining that DCP (FRFSIM = 0.404, LPIPS = 0.183) has a better defogging quality than MSF (FRFSIM = 0.373, LPIPS = 0.223), but SSIM has the opposite result (DCP is 0.279, MSF is 0.281). This phenomenon also occurs in the third row and fourth row of Fig. 10 (g), the defogged result from moderately foggy image (FRFSIM = 0.373, LPIPS = 0.233) has a better image quality on FRFSIM and LPIPS than the ones from highly foggy image (FRFSIM = 0.307, LPIPS = 0.356), but SSIM indicates the opposite (moderately is 0.281, highly is 0.332). The reason for these results is that SSIM does not consider the influence of the fog density characteristics on image edge information. In comparison, our proposed IQA method can better measure the quality of defogged images and none of the above phenomena occurred. As can be seen from Fig. 10 (f) - (j), the higher the fog density in foggy images or defogged images, the lower the magnitude of the FRFSIM index. Especially for the defogged result without fog, e.g., the NL's result in the third row (the result from slightly foggy image) and fourth row (the result from moderately foggy image) of Fig. 10 (i), our FRFSIM can also evaluate accurately and effectively. In these results, we can see that their FRFSIM is smaller than the others, although there is no fog in the image. The reason is that the NL's defogged results always have lower contrast and severe color distortion.

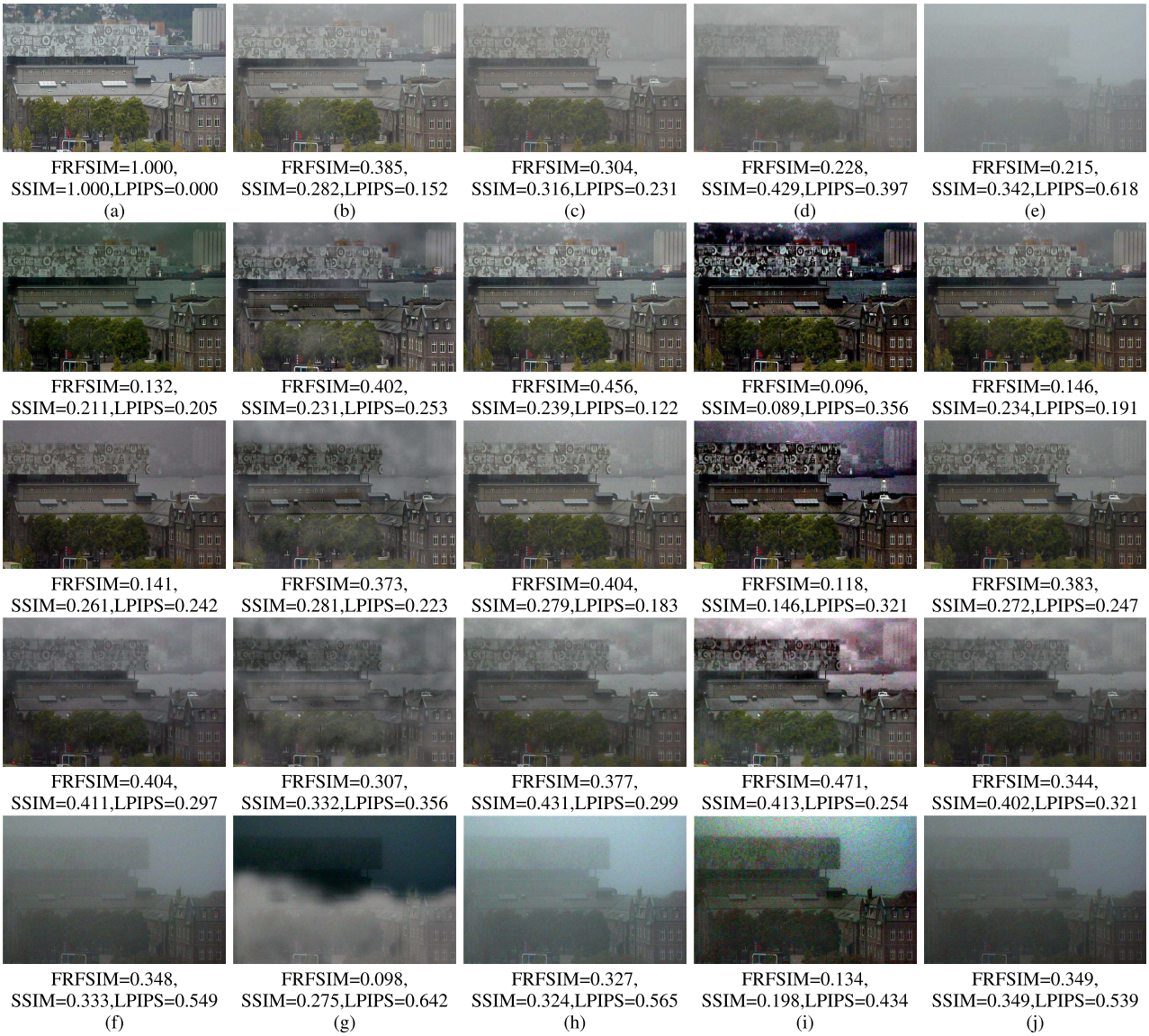


Fig. 10. Qualitatively and quantitatively comparison of the defogged results which are from the same scene with different fog densities.(a) Reference. (b) Slightly. (c) Moderately. (d) Highly. (e) Extremely. (f)-(j) The defogged results are generated by DehazeNet [7], MSF [2], DCP [18], NL [6], CAP [61]. Second row: the defogged results of (b). Third row: the defogged results of (c). Fourth row: the defogged results of (d). Fifth row: the defogged results of (e). Additionally, the FRFSIM, SSIM [52] and LPIPS [56] scores are below each image.

It further indicates that our method is more suitable for evaluating DFIs due to the FRFSIM is not only able to effectively predict the fog density, but also very sensitive to color distortion.

Moreover, we provide the scores of the FRFSIM, SSIM [52], and RI [59] of 16 defogging methods on the MRFID as baselines for related studies. As shown in Table II, the top three performance values are highlighted in red, blue and green. It can be seen that the learning-based methods (Dehazenet, EPDN, GFN, PBGAN, and MSCNN) have the better performance than the fusion-based methods (AMEF and MSF) and the prior-based methods (DCP, CAP, Dehret, SID, FVR, FDPD, BCCR, MOF, and NL). Especially, among the learning-based method, MSCNN has the best performance on FRFSIM, SSIM and RI, followed by the PBGAN, Dehazenet, GFN and EPDN. For the fusion-based method, AMEF has a better performance than MSF. For the prior-based methods, FVR, FDPD and MOF are better than the other methods in terms of overall performance. Dehret has the worst performance among all methods, followed by SID.

To further evaluate the defogging methods by using our FRFSIM, Fig.11 shows the FRFSIM and its similarity map of different defogging methods on other foggy images from I-HAZE [4], O-HAZE [5] and BeDDE [59] dataset. From Fig. 11, we can see that our proposed metric FRFSIM can still effectively evaluate the defogged results. To name just a few, the results of MOF [58] (as shown in Fig. 11 (e)) have the worst FRFSIM among other methods. It appears to be better at removing the fog than other methods, however, this method creates serious artifacts, such as over enhancement, color cast, etc. In contrast, as shown in Fig.11 (c), the FRFSIM of AMEF [14] are better than MOF, although there is some remaining fog in the results. It further demonstrates that our method can capture both of the fog density and artifact

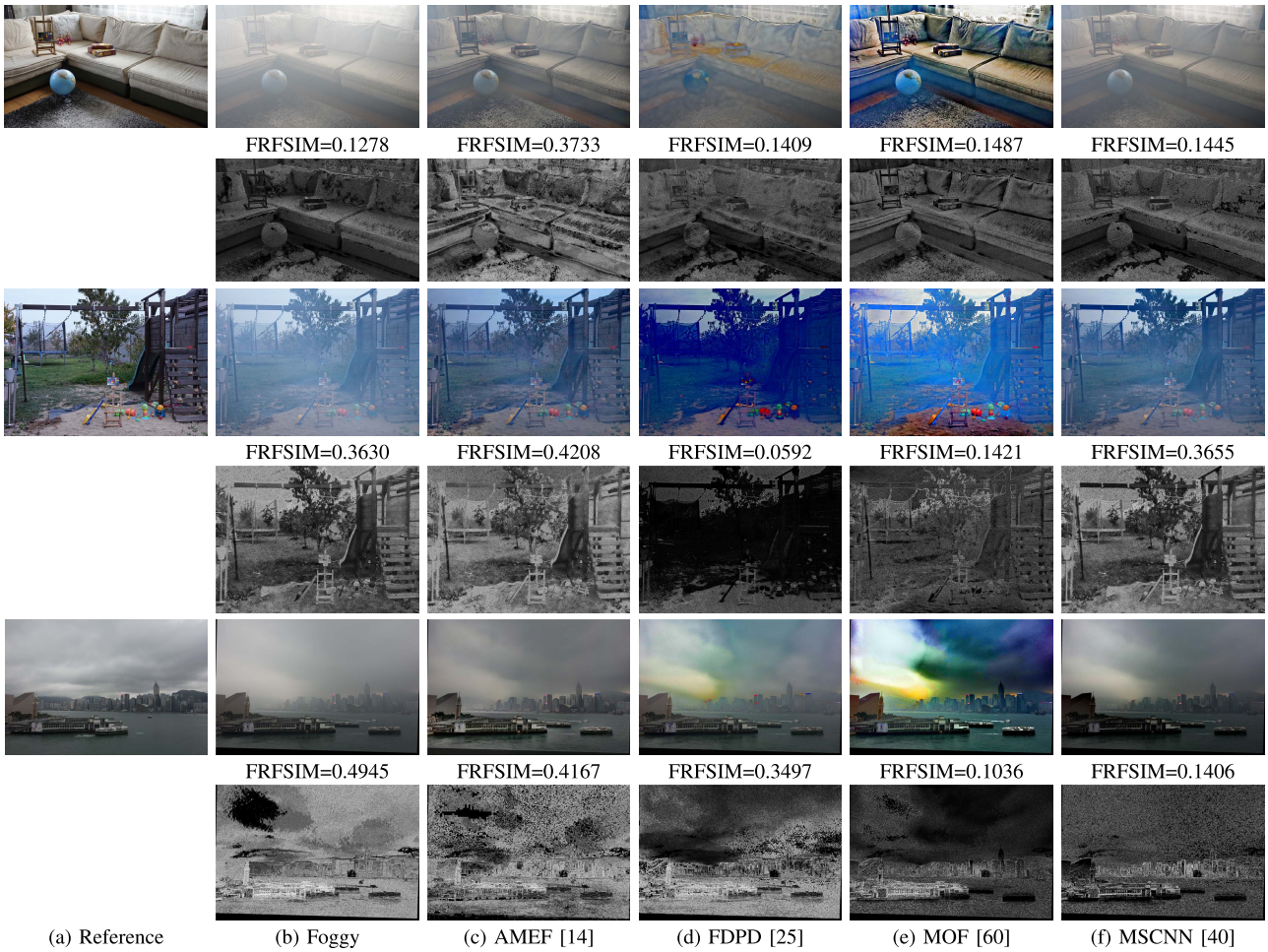


Fig. 11. Defogged results and their FRFSIM maps for the foggy images from other dataset [4], [5], [59].

information in defogged results, and then effectively evaluate the defogging quality through a single score.

D. Discussion

As mentioned in II-B, in our subjective evaluation experiment, we have added the original foggy image to the evaluation process. Results from analysing all MOS shown that amongst the 800 foggy images in our dataset, the MOS values of 408 original foggy images (51 %, a slight majority) have higher MOS values than any of the corresponding DFIs. This means that many subjects preferred the original foggy image to the DFIs. This is consistent with the finding of [51]. Based on such subjective criterion, all 16 defogging algorithms studied in this paper have failed in more than half of the cases, i.e., the DFI's have lower MOS's than the original foggy images.

This has raised questions if current defogging solutions are useful or necessary if the goal of defogging is to improve the subjective visual quality of the foggy images. At the same time, it also raises the question if subjective evaluation or MOS is the suitable criterion for assessing the performances of defogging algorithms if defogging is for other image analysis purposes such as object recognition. This therefore suggests that different IQA criterions should be developed for assessing defogging algorithms depending on the goal of defogging.

It is worth noting that no individual feature of FRFSIM seems to emerge to give a clear indication if a defogging algorithm has succeeded or failed. We have performed a series of data analysis to see if any of the four fog relevant features, D, M, G and C or four feature similarities DS, MS, GS, and CS, can indicate if a defogging algorithm has succeeded (the DFI has a higher MOS than the original foggy image) or failed (the foggy image has a higher MOS than the DFI). In one analysis, we computed the average values of the fog-relevant features for the failed group and succeeded group respectively. And in another, we compared the PLCC performances of the individual feature similarity maps of different defogging methods for the failed group and the succeeded group. From the results of these experiments, unfortunately, we found that no obvious pattern is emerging which will allow us to use any one of the fog-relevant features or fog-relevant feature similarities to indicate the success or failure of a defogging algorithm. These analyses further demonstrate that the quality of a DFI is influenced by multiple factors and a quality index should consider them together as in FRFSIM.

Current IQA metrics including FRFSIM are developed for the purpose of assessing the subjective visual quality of the DFIs. However, they cannot tell us if an algorithm has succeeded, i.e., if the subject would prefer the DFI to the original foggy image. A method for automatically determining

if a DFI has a higher visual quality than the original fogy image would be very useful. Additionally, if the purpose of defogging is for image analysis tasks such as improving object recognition or other image analysis tasks, specific IQA metrics should be developed for assessing such algorithms.

In defogging quality assessment, there are two types of images will receive low subjective opinions: scenes with heavy fog and images with defogging artifacts. Therefore, how to design an objective evaluation method to fully consider these two situations is particularly important. FRFSIM has attempted to measure both fog-related information and defogging artifact related features together. The fact that it has shown very good match with the MOS's suggests that this is a right strategy.

V. Concluding Remarks

In this study, a benchmark natural fog dataset MRFID is presented, which includes 200 clear images and each with four corresponding foggy images of different densities. Based on these foggy images, 16 defogging methods are implemented to build a defogged image dataset, named as DMRFIs. Meanwhile, we employed 120 subjects for evaluating the DMRFIs and obtained the MOS of the visual quality of defogged images. In search for an IQA method that is consistent with subjective evaluation results, we have developed the FRFSIM quality index. Comparisons between our FRFSIM and 13 state-of-the-art IQA methods on DMRFIs demonstrate that the proposed new IQA method is more suitable for evaluating DFIs and that FRFSIM is potentially a useful quality metric for evaluating defogging methods.

Acknowledgment

The authors would like to express their gratitude to the anonymous reviewers for their insightful suggestions which have lead to significantly improvements of the paper.

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. (ICIP)*, Sep. 2016, pp. 2226–2230.

[2] 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.

[3] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert, "A fast semi-inverse approach to detect and remove the haze from a single image," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 501–514.

[4] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images," 2018, *arXiv:1804.05091*. [Online]. Available: http://arxiv.org/abs/1804.05091

[5] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," 2018, *arXiv:1804.05101*. [Online]. Available: http://arxiv.org/abs/1804.05101

[6] 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.

[7] 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.

[8] L. K. Choi, J. You, and A. C. Bovik, "Referenceless perceptual fog density prediction model," *Proc. SPIE*, vol. 9014, Feb. 2014, Art. no. 90140H.

[9] 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.

[10] L. Ding, H. Huang, and Y. Zang, "Image quality assessment using directional anisotropy structure measurement," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1799–1809, Apr. 2017.

[11] D. Engin, A. Genc, and H. K. Ekenel, "Cycle-dehaze: Enhanced CycleGAN for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 825–833.

[12] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, 2008.

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

[14] A. Galdran, "Image dehazing by artificial multiple-exposure image fusion," *Signal Process.*, vol. 149, pp. 135–147, Aug. 2018.

[15] 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.

[16] Y. Gao, Y. Su, Q. Li, H. Li, and J. Li, "Single image dehazing via self-constructing image fusion," *Signal Process.*, vol. 167, Feb. 2020, Art. no. 107284.

[17] N. Hautière, J.-P. Tarel, D. Aubert, and É. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Anal. Stereol. J.*, vol. 27, no. 2, pp. 87–95, Jun. 2008.

[18] 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.

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

[20] Y. Jiang, C. Sun, Y. Zhao, and L. Yang, "Fog density estimation and image defogging based on surrogate modeling for optical depth," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3397–3409, Jul. 2017.

[21] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4725–4740, Oct. 2017.

[22] 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. 4770–4778.

[23] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[24] R. Li *et al.*, "Single image dehazing via conditional generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8202–8211.

[25] Z. Ling, J. Gong, G. Fan, and X. Lu, "Optimal transmission estimation via fog density perception for efficient single image defogging," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1699–1711, Jul. 2018.

[26] 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.

[27] W. Liu, X. Chen, X. Chu, Y. Wu, and J. Lv, "Haze removal for a single inland waterway image using sky segmentation and dark channel prior," *IET Image Process.*, vol. 10, no. 12, pp. 996–1006, Dec. 2016.

[28] W. Liu, R. Yao, and G. Qiu, "A physics based generative adversarial network for single image defogging," *Image Vis. Comput.*, vol. 92, Dec. 2019, Art. no. 103815.

[29] 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.

[30] E. J. McCartney, *Optics of the Atmosphere: Scattering by Molecules and Particles*, vol. 421. New York, NY, USA: Wiley, 1976, p. 1976.

[31] 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.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[34] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[35] S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 820–827.

[36] 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.

[37] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Process., Image Commun.*, vol. 61, pp. 33–43, Feb. 2018.
[38] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 154–169.
[39] 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.
[40] D. L. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.
[41] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 31–42.
[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[43] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, 2005.
[44] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[45] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[46] 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.
[47] 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.
[48] Y. Song, J. Li, X. Wang, and X. Chen, "Single image dehazing using ranking convolutional neural network," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1548–1560, Jun. 2018.
[49] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.
[50] 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.
[51] J. Vazquez-Corral, A. Galdran, P. Cyriac, and M. Bertalmío, "A fast image dehazing method that does not introduce color artifacts," *J. Real-Time Image Process.*, vol. 17, pp. 607–622, Aug. 2018.
[52] 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.
[53] 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.
[54] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[55] 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.
[56] 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.
[57] 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. (ICIP)*, Sep. 2017, pp. 3205–3209.
[58] D. Zhao, L. Xu, Y. Yan, J. Chen, and L.-Y. Duan, "Multi-scale optimal fusion model for single image dehazing," *Signal Process., Image Commun.*, vol. 74, pp. 253–265, May 2019.
[59] 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.
[60] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3528–3541, Jul. 2019.
[61] 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.

Wei Liu received the M.Eng. degree in control science and engineering and the Ph.D. degree in computer science and technology from the Wuhan University of Technology in 2013 and 2017, respectively. From 2018 to 2020, he was a Postdoctoral Researcher with Shenzhen University, China. He is currently an Assistant Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology. His research interests are in computer vision, deep learning, and image processing.

Fei Zhou received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, in 2013. From 2013 to 2016, he was a Postdoctoral Fellow with the Graduate School at Shenzhen, Tsinghua University. From 2017 to 2018, he was a Visiting Scholar with the Department of Statistical Science, University College London. He is currently an Assistant Professor with the College of Electronic and Information Engineering, Shenzhen University. He has authored more than 50 articles internationally. His research interests include image super-resolution, image decomposition, and image quality assessment. He is a reviewer of many well-known journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and *Information Sciences*. He also serves as a Guest Editor for *Neurocomputing* and *Signal Processing: Image Communication*.

Tao Lu (Member, IEEE) received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China, in 2003 and in 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, Wuhan University, in 2013. He is currently a Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology, and a Research Member with the Hubei Provincial Key Laboratory of Intelligent Robot. He was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Texas A&M University, from 2015 to 2017. His research interests include image/video processing, computer vision, and artificial intelligence.

Jiang Duan received the B.S. degree in mechanical engineering from Southwest Jiaotong University, the M.S. degree from the University of Derby in 2002, and the Ph.D. degree from the University of Nottingham in 2006. He is currently a Professor with the Southwestern University of Finance and Economics, China. He has published more than 30 academic articles, won more than ten international and national patents, and his researches have been funded by about 20 national and provincial funds. He is also an Expert of the National Thousand Talents Plan, the winner of the first Sichuan Outstanding Talent Award (the highest talent award of Sichuan province), the winner of the Sichuan Youth Science and Technology Award, the Standing Committee Member of Sichuan Association for Science and Technology, and the Vice Chairman of Sichuan Science and Youth Federation.

Guoping Qiu is currently a Distinguished Professor of Information Engineering, the Director of the Shenzhen University Intelligent Robotics Centre, Shenzhen University, China, and a Chair Professor of Visual Information Processing with the University of Nottingham, Nottingham, U.K. He has taught in universities in U.K., and Hong Kong, and also consulted for multinational companies in Europe, Hong Kong, and China. His research interests include image processing, pattern recognition, and machine learning. He is particularly known for his pioneering research in high-dynamic range imaging and machine learning-based image processing technologies. He has published widely and holds several European and U.S. patents. Technologies developed in his lab have laid the cornerstone for successful spinout companies who are developing advanced digital photography software enjoyed by tens of millions of global users.