EVALUATION OF DEFOGGING: A REAL-WORLD BENCHMARK DATASET, A NEW CRITERION AND BASELINES

Shiyu Zhao[1], *Lin Zhang*[1,*], *Shuaiyi Huang*[2], *Ying Shen*[1,*], *Shengjie Zhao*[1], *Yukai Yang*[3]

[1]School of Software Engineering, Tongji University, Shanghai, China
[2]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[3]Department of Statistics, Uppsala University, Uppsala, Sweden

ABSTRACT

Modern defogging methods are able to achieve very comparable results whose differences are too subtle for people to qualitatively judge. On the other hand, existing quantitative evaluation methods are also not convincing due to a lack of proper datasets. In this work, we attempt to address these issues and establish a long-term lacking benchmark dataset, namely BeDDE (BEnchmark Dataset for Defogging Evaluation), for evaluating the performance of defogging algorithms. To our knowledge, **BeDDE is the first real-world dataset comprising foggy images with their registered clear counterparts**. Using BeDDE, we set up a new criterion for evaluating defogging methods where VSI, a full reference image quality assessment metric, is calculated and averaged on registered ROIs of all image pairs. The evaluation results of the proposed criterion correlate well with human judgements. 10 state-of-the-art defogging methods are evaluated as baselines on BeDDE. BeDDE is available online[1].

***Index Terms*—** Benchmark dataset for defogging evaluation, FR-IQA, metric for defogging evaluation

1. INTRODUCTION

Under foggy conditions, image quality is seriously impaired due to the scattering of atmospheric aerosol particles, leading to performance degradation of the related vision algorithms or systems. Consequently, researchers show great enthusiasm for defogging and have presented a great number of defogging approaches [1, 2, 3, 4, 5]. However, due to a lack of benchmark datasets comprising foggy images with their ground-truth clear versions, how to evaluate the performance of those methods remains an open issue.

Widely adopted evaluation schemes can be categorized into three classes. The first class relies on subjective judgements of readers on defogged images. The second class adopts no reference IQA metrics [6, 7] which are specially designed for evaluating defogging methods. The third class simulates foggy images from clear images according to Koschmieder's law [8] and then employs FR-IQA (full reference image quality assessment) metrics, such as PSNR and SSIM [9], to evaluate defogging algorithms. However, all those schemes own their drawbacks, which will be discussed in Sect. 2.1. In this work, we attempt to address the issue of evaluating the performance of defogging algorithms objectively and reasonably.

2. RELATED WORK AND OUR CONTRIBUTIONS

2.1. Related work

Considering the concerns of this work, we briefly introduce several most well-known IQA metrics and then review current evaluation schemes for defogging.

Recent advances in FR-IQA. The pixel-based metrics, e.g. MSE and PSNR, correlate poorly with human perceptions, and thus human visual system (HVS) based IQA metrics were exploited. The most well-known one is the structure similarity (SSIM) index [9]. It considered that HVS is highly adapted to extract the structural information from the visual scene and thus leveraged the luminance, contrast and structural information to calculate the similarity. Different from SSIM, Zhang *et al.* [10] held the view that HVS understands an image mainly according to its low-level features and proposed two feature similarity indices, FSIM and FSIMc, which involve the phase congruency and image gradient magnitude features. In their later work, they replaced the phase congruency features with saliency maps and proposed a new metric named VSI [11]. In other studies, gradients were further investigated, and related metrics, such as GMSD [12], were proposed.

Current evaluation schemes for defogging. As aforementioned, there are three classes of evaluation schemes for defogging. The first class encourages an article to present defogged images or other intermediate outputs (e.g. transmission maps) generated by different algorithms and resorted to subjective judgments of readers only. In [2], Fattal provided a benchmark with 23 foggy images for subjective judgments. However, those images own similar visibility conditions and

∗Corresponding author. Email: {cslinzhang, yingshen}@tongji.edu.cn
[1]https://github.com/xiaofeng94/BeDDE-for-defogging

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00316

evaluating on them only may lead to a preference to handle foggy images with certain conditions. Additionally, some defogged images or outputs are too similar for people to judge.

The second class makes use of specially designed no reference IQA metrics. In [6], Hautière *et al.* considered that defogging methods should be able to restore the contrast of foggy images and proposed three indicators, i.e., e, $\overline{r}$, σ, to describe changes in edges, norms and intensities of images before and after the restoration. Later, Choi *et al.* [7] proposed another no reference assessment method, called FADE (Fog Aware Density Evaluator), based on natural scene statistics (NSS) and fog aware statistical features. However, no reference IQA is still an open issue and less reliable than FR-IQA. By contrast, the third class explores FR-IQA metrics to evaluate defogging methods.

Due to a lack of real-world image pairs, the third class usually simulated foggy images from clear ones according to Koschmieder's law. In [2], Fattal made use of 11 clear images with depth maps provided by [13] to simulate foggy images with ground-truth transmission maps and calculated mean L_1 distance between estimated transmission maps and their ground-truths. More works [14, 3, 15, 4, 5, 16] employed existing indoor datasets with depth maps, such as NYU2 [17] and Middlebury datasets[18], to handle the lack of depth information, which is required in simulation but not easy to acquire in outdoor scenes, and adopted FR-IQA metrics, such as MSE, PSNR and SSIM to evaluate defogging methods on pairs of clear images and the restored ones. However, such a strategy is questionable. First, indoor scenes actually dissatisfy the premise on which Koschmieder's law is established. Second, there is a certain gap between real foggy images and simulated ones. Third, different researchers used different images, making the comparison results less convincing.

To handle issues in exploiting FR-IQA metrics for evaluating defogging methods, there are a few works considering establishing proper datasets for defogging. Using SiVIC^{TM} software, Tarel *et al.* constructed two synthetic outdoor datasets, namely, FRIDA (Foggy Road Image DAtabase) [19] and FRIDA2 [20], for testing defogging methods. Those two datasets contain 90 synthetic images of 18 urban road scenes and 330 synthetic images of 66 diverse scenes, respectively, and provide both homogeneous and heterogeneous fogs. However, their images are in low resolutions and not realistic-looking. In [21], Li *et al.* established a dataset named RESIDE (REalistic Single-Image DEhazing) which provides indoor and outdoor images with simulated fogs for training and testing defogging models. The indoor images also came from NYU2 and Middlebury datasets, and thus suffered the same problem as other works did. The outdoor images were collected from the Internet and their depth maps were estimated from those monocular images. However, depth estimation from a monocular image is highly ill-posed and thus the acquired depth maps are unreliable, leading to a poor simulation quality. But inspiringly, they proposed a task-driven evaluation scheme. In their scheme, state-of-the-art object detection algorithms are used to detect the objects of interests on defogged images which are generated by different defogging methods from real foggy images and then mean Average Precisions (mAP) of detection algorithms are calculated as scores of defogging methods. In [22], Sakaridis *et al.* added synthetic fogs to images from Cityscapes [23] and established a dataset named Foggy Cityscapes. However, the depth information in Cityscapes is not complete and thus the quality of simulated foggy images cannot be guaranteed. Therefore, it is also improper to evaluate on Foggy Cityscapes.

2.2. Our motivations and contributions

Having investigated the literature, we find that existing studies for the evaluation of defogging methods have limitations at least in two respects.

First, in order to effectively evaluate the effect of defogging algorithms, for each foggy image we need to have its corresponding clear reference image. In existing datasets of this field, clear reference images or foggy images are synthesized rather than collected from the real physical world. In other words, a real-world benchmark dataset comprising foggy images with aligned clear references does not exist in the literature. For this reason, it is difficult for us to faithfully evaluate the performance of defogging algorithms on the restoration of real-world foggy images.

Second, there is not a widely accepted criterion for assessing the quality of defogging results. The majority of defogging studies adopt their own test samples and metrics, which makes the comparisons unfair and less convincing.

In this work, we attempt to fill the aforementioned research gaps and our major contributions are listed as follows.

(1) We overcome the difficulty of collecting real-world images under different weather conditions and provide a long-term lacking benchmark dataset, called BeDDE (BEnchmark Dataset for Defogging Evaluation), for evaluating defogging algorithms. BeDDE contains 208 pairs of foggy images and well aligned clear references. Its raw images were collected from 23 provincial capital cities of China. For each raw image pair, the foggy image and the corresponding clear one were roughly registered. Due to slight changes in viewpoints and contents during data collections, all raw image pairs are aligned and then their common ROIs are delineated by manually labeled masks. Registered ROIs between foggy images and their clear references make it possible to explore FR-IQA metrics to assess the quality of defogging results. To our knowledge, **as a benchmark dataset for evaluating the performance of defogging algorithms, BeDDE is the first one whose foggy images and their clear references are all collected from the real physical world**.

(2) It is widely accepted that state-of-the-art FR-IQA methods are very reliable in predicting the quality of an image given its high-quality reference. In our case, registered ROIs

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

Fig. 1. The clear image and foggy ones with different visibility conditions taken from Chengdu, the capital city of Sichuan province, China. (a) is the clear image. (b)∼(e) are foggy images whose visibility conditions become worse sequentially.

between foggy images and their clear references are available. Thus, it is a natural idea to assess the quality of defogging results by applying an FR-IQA metric on registered ROIs. Inspired by this motivation, a new criterion to evaluate defogging methods is proposed. In this criterion, a state-of-the-art FR-IQA metric, namely VSI [11], is adopted. To evaluate a defogging method, the mVSI (mean score of VSI over ROIs of all image pairs in BeDDE) is calculated as the score of this method. For convenience of related studies, scores of 10 state-of-the-art defogging methods are provided as baselines on BeDDE. The evaluation results by using our VSI-based criterion correlate well with the human perception.

3. BEDDE: A BENCHMARK DATASET COLLECTED FROM THE REAL-WORLD

BeDDE is the first dataset containing real-world foggy images and their corresponding clear references for defogging studies. In this section, we will present an overview of BeDDE and the pipeline of establishing it.

3.1. Dataset overview

BeDDE contains 208 image pairs collected from 23 provincial capital cities of China. For each city, one clear image and several foggy images of the same places are provided. For each image pair, the foggy image is well aligned to the corresponding clear image via a 2-D projective transformation and a manually labeled mask is provided. This mask is used to delineate regions with the same contents in those two images, which we call the ROI of this pair, and will be used to calculate scores of defogging methods.

Besides clear reference images and masks, another outstanding merit of BeDDE is its diversity of visibility conditions. To illustrate this merit, several images of a city (Chengdu) are provided in Fig. 1. In this figure, the leftmost image is a clear one while the others are foggy ones taken under different visibility conditions.

3.2. Pipeline of establishing BeDDE

There are four steps, namely, data acquisition, image registration, data cleaning and mask labeling, in the pipeline of establishing BeDDE.

Fig. 2. An overlaid image for a badly aligned pair. Badly aligned edges are highlighted by a red box and a blue box.

Data acquisition. In this step, an image of the same place was collected at a time between 8:00 to 9:00 in each day of 40 days. Such collections were conducted simultaneously at 34 provincial capitals of China in one year and the representative scenes of those cities were chosen as the collection sites. Thanks to the 44 photographers, we acquired 1269 high resolution images eventually.

Image registration. Although images of a city were taken in the same place, slight changes in viewpoints are inevitable. Therefore, for each city, we choose one image as the reference image, which is in overcast weather and provides a good visibility, and align all the other images to this image. If there is no proper reference image for a city, we simply drop all images of this city. Afterwards, we follow a standard image registration procedure which is composed of keypoint detection, feature extraction and matching, transformation matrix estimation, and transformation application. Since there are only slight changes in the viewpoints, we adopt a 2-D projective matrix as the transformation model which can be formulated as,

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

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

Data cleaning. In this step, we need to filter out undesired images including other fog-free ones and badly aligned ones. To better visualize the registration quality, we adopt an image overlay technique where the gray scale version of the reference image and that of the foggy image are assigned to different channels of a black image (zero values in RGB

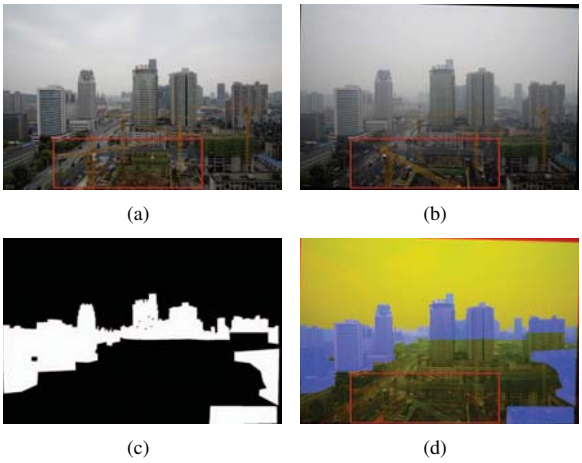

Fig. 3. Examples to illustrate the different contents in spite of a good alignment. (a) and (b) are the clear image and the foggy one, respectively, in a pair of BeDDE. (c) is the mask of this pair. (d) is an overlaid image for this pair and the blue region is the ROI of this pair. The red box highlights the major differences between (a) and (b).

channels) and thus badly aligned edges become salient. An overlaid image for a badly aligned pair is displayed in Fig. 2. As we can see, the registration quality is quite easy to judge with this technique.

Mask labeling. In spite that images of a city have been well aligned, there are still contents which can be different in them, such as vehicles, pedestrians, trees and water. Examples of such differences can be seen in Fig. 3. To handle this problem, we manually label a mask to delineate regions with the same contents between two images in a pair, namely, the ROI of this pair. In the evaluation phase, we only calculate the score over ROIs to rank defogging methods.

4. A NEW CRITERION FOR EVALUATING DEFOGGING RESULTS

The selected metric, VSI, was originally proposed by Zhang *et al.* [11] and formulated as,

$$VSI = \frac{\sum_{\mathbf{x}\in\Omega} S(\mathbf{x}) \cdot VS_m(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} VS_m(\mathbf{x})} \tag{2}$$

Here, Ω means the whole spatial domain, and $\mathbf{x}$ represents a point in Ω. $S(\cdot)$ refers to feature maps composed of VS (visual saliency), GM (gradient modulus) and chrominance features. $VS_m(\cdot)$ refers to the maximum map of VS maps of the distorted image and its reference. Such a formula can be applied in our case easily, as long as we replace Ω with the ROI delineated by a mask. Furthermore, we calculate the mVSI, namely, mean score of VSI over ROIs of all image pairs in BeDDE, for various defogging methods and rank those methods according to their scores.

Table 1. Comparison of time costs. Time-CPU represents the time cost with CPU only and Time-GPU refers to the time cost with GPU acceleration. '-' means that no implementation is provided.

Method	Time-CPU (s)	Time-GPU (s)
FVR [24]	25.48	-
DCP [1]	1.66	-
BayD [25]	77.81	-
CAP [14]	7.90	-
NLD [26]	20.63	-
MSCNN [15]	20.91	11.669
DeN [3]	14.61	-
AOD-Net [4]	3.20	0.047
DCPDN [5]	15.19	0.034
GFN [27]	33.49	0.399

Using BeDDE and mVSI, 10 representative or state-of-the-art defogging methods were evaluated, including Fast Visibility Restoration (FVR) [24], Dark Channel Prior (DCP) [1], Bayesian Defogging (BayD) [25], Color Attenuation Prior (CAP) [14], Non-Local image Dehazing (NLD) [26], MSCNN [15], DehazeNet (DeN) [3], AOD-Net [4], DCPDN [5], GFN [27]. Official implementations for those methods were used. Furthermore, all parameters were set to default and pre-trained models for CNN-based methods (the last 5 evaluated methods) came from original authors. The time costs of them using CPU only and those of 4 CNN-based methods with GPU acceleration are provided in Table 1. The mVSI scores of those methods are presented in Table 2 for quantitative comparisons. Results of two samples selected from BeDDE are shown in Fig. 4 for qualitative comparisons. From those results, several interesting conclusions can be drawn.

First, non-CNN methods, such as FVR [24] and NLD [26] demonstrated in Fig. 4, are more likely to over-enhance the contrast of foggy images and to produce artifacts that seriously degrade the quality of defogged images. By contrast, CNN-based methods are able to preserve the quality of images and restore the visibility simultaneously. Therefore, it is not weird that almost all CNN-based methods outperform the non-CNN ones on BeDDE.

Second, GFN [27], a CNN-based defogging approach proposed quite recently, performs worse than the other CNN-based methods on BeDDE. There are several potential causes. On one hand, in GFN, three traditional image enhancement techniques, namely, white balance, contrast enhancement and gramma correction, are weighted to generate the defogged image and the weights are learned by CNNs. However, traditional enhancement techniques are not suitable for defogging, since the degradation in fog is highly correlated with the depth of the scene and those techniques cannot handle such correlations. On the other hand, the training set and test set of GFN

Table 2. Quantitative comparisons for 10 representative state-of-the-art defogging methods. Red, green and blue texts highlight the champion, the runner-up and the second runner-up, respectively. Note that MSCNN, DeN, AOD-Net, DCPDN and GFN are CNN-based methods.

Method	FVR [24]	DCP [1]	BayD [25]	CAP [14]	NLD [26]
Score	0.8858	0.9462	0.9079	0.9156	0.8959
Method	MSCNN [15]	DeN [3]	AOD-Net [4]	DCPDN [5]	GFN [27]
Score	0.9471	0.9522	0.9540	0.9541	0.9389

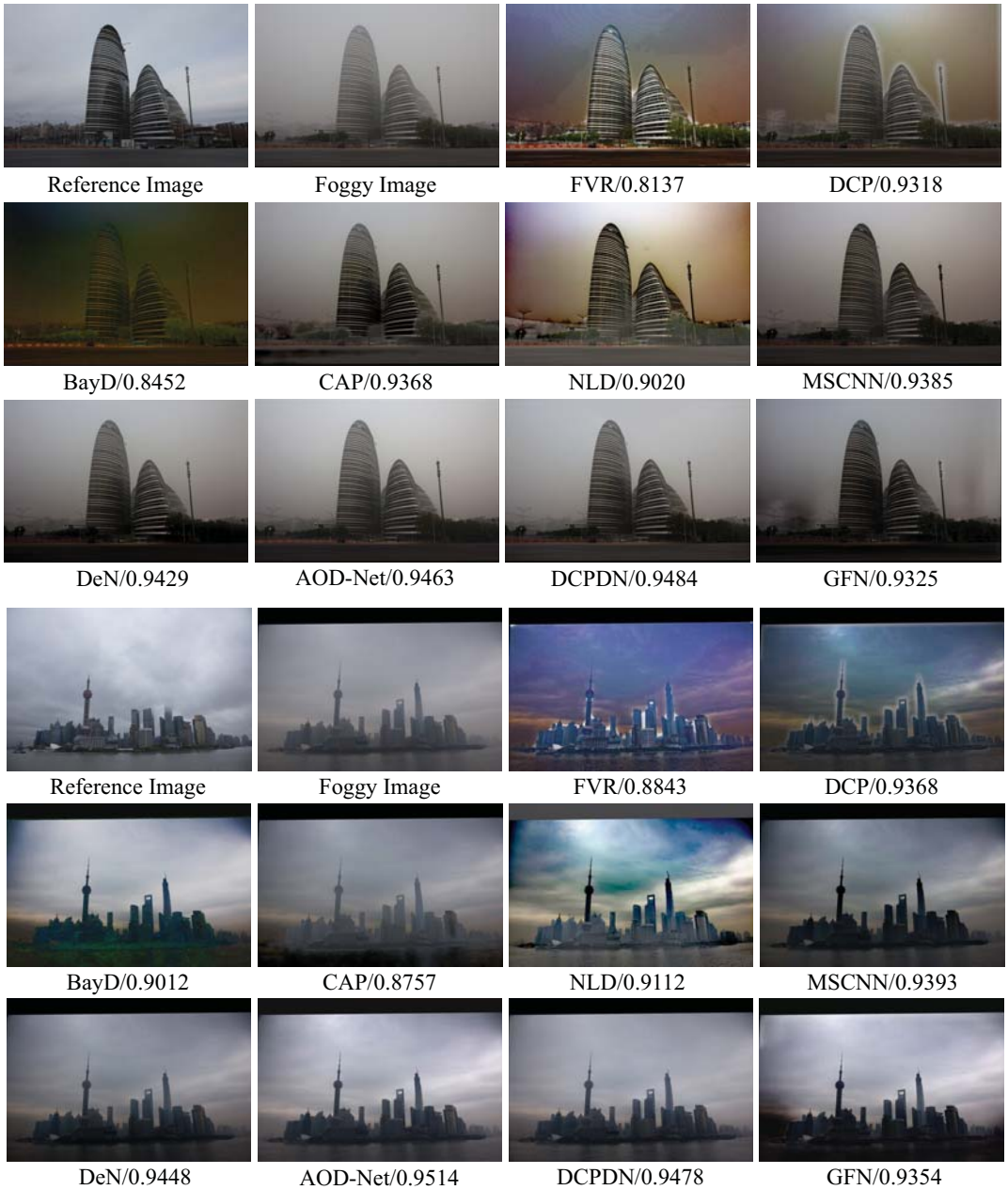

Fig. 4. Results of 10 defogging methods on two samples in BeDDE. The VSI score of each defogged image is provided at the bottom of this image in the form of "method name/VSI score".

are all simulated from indoor images of NYU2. Therefore, GFN may perform well on its own test set due to overfitting but fails to handle real foggy images in BeDDE.

5. CONCLUSION

In this paper, we focus on how to evaluate the performance of defogging algorithms and established a benchmark dataset, namely BeDDE, whose images were all collected from the real physical world. It is the first dataset containing real-world foggy images and the corresponding clear references in this field. Furthermore, by exploiting advances in the field of FR-IQA, we proposed a new criterion for evaluating defogging methods by calculating VSI scores on registered ROIs. In the future, we will continuously enlarge BeDDE to include more real-world samples.

6. ACKNOWLEDGMENT

This research was funded in part by the Natural Science Foundation of China under Grant No. 61672380, in part by the National Key Research and Development Project under Grant No. 2017YFE0119300, and in part by the Fundamental Research Funds for the Central Universities under Grant No. 2100219068.

7. REFERENCES

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

[2] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graphics*, vol. 34, no. 1, pp. 13, 2014.

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

[4] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *ICCV*, 2017, pp. 4780–4788.

[5] H. Zhang and V. M Patel, "Densely connected pyramid dehazing network," in *CVPR*, 2018.

[6] N. Hautière, J. Tarel, D. Aubert, and E. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Anal. & Stereology*, vol. 27, no. 2, pp. 87–95, 2011.

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

[8] W. E. K. Middleton, "Vision through the atmosphere," 1957, in: Geophysik II/Geophysics II (Edited by J. Bartels).

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

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

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

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

[13] G. Zhang, J. Jia, T. Wong, and H. Bao, "Consistent depth maps recovery from a video sequence," *IEEE Trans. PAMI*, vol. 31, no. 6, pp. 974–988, 2009.

[14] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. IP*, vol. 24, no. 11, pp. 3522–3533, 2015.

[15] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *ECCV*, 2016, pp. 154–169.

[16] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," in *CVPR*, 2018.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *ECCV*, 2012, pp. 746–760.

[18] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Ger. Conf. Pattern Recog.*, 2014, pp. 31–42.

[19] J. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in *IEEE Intel. Vehic. Symposium*, 2010, pp. 478–485.

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

[21] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Trans. IP*, vol. 28, no. 1, pp. 492–505, 2019.

[22] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int'l J. Comp. Vis.*, pp. 1–20, 2018.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[24] J. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in *ICCV*, 2009, pp. 2201–2208.

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

[26] D. Berman and S. Avidan, "Non-local image dehazing," in *CVPR*, 2016, pp. 1674–1682.

[27] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M. Yang, "Gated fusion network for single image dehazing," in *CVPR*, 2018.