Recovering the Unseen: Benchmarking the Generalization of Enhancement Methods to Real World Data in Heavy Fog

Mario Bijelic[1,2*] Paula Kysela[1] Tobias Gruber[1,2] Werner Ritter[1] Klaus Dietmayer[2]

[1] Daimler AG, Germany
[2] Ulm University, Germany

Abstract

Due to the ill-posed problem of collecting data within adverse weather scenarios, especially within fog, most approaches in the field of image de-hazing are based on synthetic datasets and standard metrics that mostly originate from general tasks as image denoising or deblurring. To be able to evaluate the performance of such a system, it is necessary to have real data and an adequate metric. We introduce a novel calibrated benchmark dataset recorded in real, well defined weather conditions. The aim is to give a possibility to test developed approaches on real fog data. Furthermore, we claim to be the first showing an investigation of heavy fog conditions up to a total degradation of the considered images. We present a newly developed metric providing more interpretable insights into the system behavior and show how it is superior to several current evaluation methods as PSNR and SSIM. For this purpose, we evaluate current state-of-the-art methods from the area of image defogging and verify the proposed dataset and our developed evaluation framework.

1. Introduction

Environment perception for autonomous driving and robotics relies on the sensor data of a large number of different sensors such as camera, radar and lidar. In clear weather conditions these sensors usually provide reliable sensor streams that can be fed into intelligent algorithms for object detection, depth estimation or semantic understanding. However, in real world scenarios including fog, haze, snow and rain, the performance of these sensors drop significantly [7, 6, 29]. While in camera images contrast degenerates, lidar pointclouds lose points and get dispersed due to atmospheric attenuation and scattering. Although a huge effort is spent on improving sensor hardware, there will always be disturbances in the sensor data because physical laws cannot be overcome. This impairment in the sensor

Figure 1: Example image dehazing results based on our introduced benchmark dataset. It enables evaluating the system performance based on calibrated reflective targets and assessing the contrast enhancement in depth for a challenging illumination setting in different fog densities.

stream tremendously increases the risk for a malfunction of the successive algorithms and can have fatal consequences, e.g. when the detection of pedestrians fails. This problem can be solved either by increasing the robustness of the

---

*Main correspondence is mario.bijelic@daimler.com

successive algorithms or by enhancing the disturbed signal such that the algorithms designed for undisturbed input data can be used for disturbed input data as well. Due to the versatility of disturbances in real-world scenarios, enhancing the sensor signal itself offers a large benefit as the disturbance can be decoupled from the actual algorithm which supports understanding the functioning of the whole system.

This work focuses on the enhancement of camera images, but could be further extended to other sensors such as lidar. Development and evaluation of algorithms for image enhancement require data from both clear and disturbed scenarios. However, most work has been done on synthetically disturbed images. For a meaningful evaluation of different image enhancement algorithms, real test data and an appropriate metric are inevitable. In this work we present a novel benchmark dataset recorded in a professional weather chamber under defined and adjustable weather conditions. This offers the possibility to test image enhancement methods on real data covering different fog levels.

Moreover, we present a novel metric that is adopted for the evaluation of de-fogging algorithms. Our metric provides the opportunity to evaluate the contrast improvement of the system in a more intuitive and interpretable way. In order to compare the novel metric with standard image enhancement metrics, we show the results for various state-of-the-art image enhancement methods. The implemented reference methods range from supervised approaches depending on pixelwise annotation to methods trained on unpaired image data.

Particularly, we make the following contributions:

- We provide a novel benchmark dataset recorded under defined and adjustable real weather conditions.

- We introduce a metric describing the performance of image enhancement methods in a more intuitive and interpretable way.

- We compare a variety of state-of-the-art image enhancement methods and metrics based on our novel benchmark dataset.

2. Previous Work

Image Quality Assessment The quality of images can be evaluated either based on ground truth data, i.e. full reference image quality assessment (FR-IQA) or by approaches independent of ground truth data, called no reference image quality assessment (NR-IQA). The mean squared error (MSE), the well-known signal-to-noise ratio (SNR) and the peak signal-to-noise ratio (PSNR) are part of the most often used FR-IQA measures [33, 60]. Hereby, the considered images are compared with their respective ground truth. Nevertheless, it was shown that these metrics do not correlate well with the perception of the human visual system (HVS) [22]. The structural similarity index (SSIM) [67] and the visual information fidelity (VIF) [61] are improved image quality assessment methods based on structural information such as contrast and luminance and natural scene statistics (NSS), respectively. For the learned perceptual image patch similarity (LPIPS) metric [77] both images are processed by one of three pretrained neural networks [63, 35, 31]. The result is the difference between selected network layers weighted by pretrained values based on human perceptual judgments [77]. A similar metric has been applied in [40] for several popular hazing algorithms.

In many domains ground truth is not always available as for example regarding real adverse weather conditions. Methods evaluating image quality without respective ground truth data have been applied to different types of noise [46, 69, 62, 45, 50, 48, 43, 40]. Regardless of the availability of reference, almost all presented metrics were developed and tested for rather standard image distortions such as Gaussian blur, white noise or JPEG compression. Image degeneration caused by weather-related phenomena, especially fog and haze, has different characteristics. Therefore, we see the need for a new evaluation method being able to evaluate the system performance on real data with real reference measurements in a quantitative manner.

Adverse Weather Simulation Methods Considering adverse weather scenarios, it is difficult to collect paired data to train supervised algorithms. Many approaches create synthetic data as for example rainy images by adding streaks to clear weather images [23, 74, 53, 73] or raindrops onto the windshield [65]. Recently, methods have been developed to adapt daytime images for nighttime scenes [58, 44]. For creating synthetic foggy images many existing methods are based on the well-known physical model of Koschmieder [34] which is defined as follows:

$$I(x) = R(x)t(x) + L(1 - t(x)). \tag{1}$$

Hereby, $I(x)$ is the foggy image at pixel x considered color channel wise and $R(x)$ is the respective value within the clear image. L is the atmospheric light, also denoted as airlight, which is the light being scattered by the atmosphere reaching an observer captured by a camera [68]. $t(x)$ is the transmission defined as the scene radiance captured by the camera and depends on the scene depth $l(x)$ which is the calculated distance of the considered pixel x to the camera [56]. Based on Eq. (1), synthetic foggy images are created by manipulating airlight and transmission of the clear weather image depending on the respective scene depth [41, 24, 57]. In order to capture training data under real conditions impressive work has been done in [29, 5, 4, 28].

Datasets Existing adverse weather datasets can be divided into synthetic and real weather scenarios. In [57, 64, 78,

41, 24, 40] datasets are created with synthesized fog based on Koschmieder's physical model [34]. In contrast to the synthetically generated datasets, real data is provided in [2, 5]. In O-HAZE [5] and I-HAZE [4] 35 indoor and 45 outdoor pairs of images were captured with a professional haze machine. Those datasets offer a good evaluation opportunity but are rather small. Furthermore, all of the investigated state-of-the-art approaches were tested by standard full reference image quality assessment metrics. Our dataset contains more image pairs and offers a novel metric enabling a quantitative measure and comparison of different approaches.

Enhancement Methods Adverse weather conditions as rain, fog or snow, as well as motion blur and any other type of noise can affect the performance of cameras. Because of its fundamental role, image denoising and reconstruction are big research subjects in signal processing and computer vision in general. Regarding image denoising, many traditional techniques based on spatial filtering and wavelet coefficients [13] have been developed [1, 19, 18, 11, 59, 21, 52]. The famous block-matching and 3-D filtering (BM3D) approach [16] takes advantage of methods used within video compression to remove Gaussian noise. More recently, there have been techniques introduced for general image denoising based on deep learning [75, 76, 38] with different levels of noise type generalization. Current best performing approaches are based on Generative Adversarial Networks (GANs) [27]. By formulating an adversarial loss function, two neural networks, called generator and discriminator, work against each other until the generator ideally creates images that are indistinguishable from real world images. Generative Adversarial Networks have started with simple tasks as generating digits, faces and simple classes [27] out of random vectors. Now, they have found their way to even more complex areas such as image manipulation [79], style transfer [25, 80], image inpainting [17], image to image translation [32] and super resolution [55, 37]. As an alternative to classic unsupervised data clustering or autoencoders, it was shown that GANs can be successfully used for representation learning [54, 9]. Regarding image denoising and deblurring, DeblurGAN [36] removes blur from an image and additionally shows how object detection can be improved by this procedure. Furthermore, several approaches have been presented in the area of image deraining [74, 23, 12, 53]. While [23] is working with a decomposition of images into low-frequency and high-frequency layers and a CNN, a GAN was successfully applied in [74, 53].

Image De-fogging and De-hazing For many image defogging and de-hazing algorithms, a physical model is used to describe foggy conditions [30, 57, 10, 39, 42]. Based upon Eq. (1), parameters as the atmospheric light and transmission are estimated locally. In [30] for example, the dark
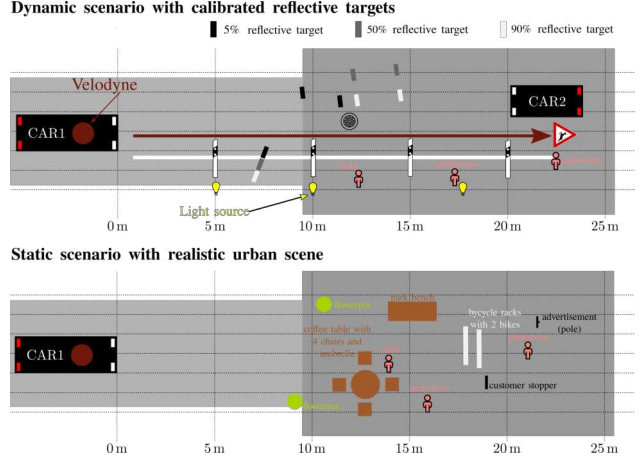

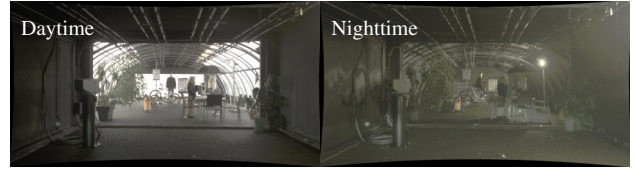Figure 2: Scene setup for the dynamic scenario (top) and the static scenario (bottom).

Figure 3: Photos of the static clear scenario for daytime (left) and nighttime (right).

channel prior is used to remove haze from images. In [57] the problem of semantic foggy scene understanding (SFSU) is investigated by adding simulated fog to the Cityscapes dataset [15]. The fog simulation is based on the standard optical model [34] for daytime fog and the semantic segmentation is performed with a dilated convolution network (DCN) [71]. Based on the usage of a physical model, further approaches were introduced in [10, 39]. DehazeNet [10] estimates the airlight and the transmission separately, while in [39] the parameters are estimated collectively to avoid amplified errors. A summary of further recent methods is presented in [3] where several algorithms tackling the image dehazing task are introduced, including GAN approaches such as CycleDehaze [20].

3. Dataset

In order to advance the evaluation of enhancement algorithms in difficult weather situations, we have collected a dataset in well defined weather conditions in a weather chamber [14]. The used chamber is able to emit real fog while the meteorological visibility is controlled by a visibility sensor. The meteorological viewing distance V is defined by

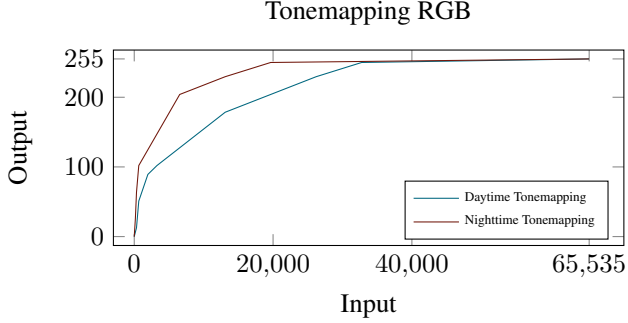$$V = -\frac{\ln(0.05)}{\beta}, \qquad (2)$$

Figure 4: RGB camera tonemapping from 16 bit to 8 bit images. The 12 bit images are first linearized to 16 bit according to the camera characteristic.

where β is the physical attenuation constant. At the beginning of the measurement, the whole chamber is flooded with fog down to a visibility of 5 m. Afterwards, the fog slowly dissipates during the recording while the current visibility is continuously tracked. The natural dissipation takes about 10-15 min from 10-125 m but highly depends on the current outside weather conditions. For dynamic measurements the visibility can be stabilized to a certain visibility range by gently adding fog when the visibility exceeds the desired threshold.

Sensor Setup For collecting the dataset, a test vehicle was equipped with state-of-the-art imaging, range and weather sensors. Regarding image recording, a passive monocular front-facing color camera (Aptina ar0230, 30 Hz, 1920 × 1024) and an active NIR gated camera (BrightwayVision, 1280×720) were selected. The color camera provides 12 bit high dynamic range images tonemapped to 8 bit LDR images following Figure 4. Due to its operating principle the gated camera avoids early back scatter and serves as a qualitative upper bound achieved by hardware optimizations [7]. For range data acquisition a Velodyne HDL64 S3D laser scanner is used. All sensors are time synchronized and egomotion corrected utilizing a proprietary inertial measurement unit (IMU). The system operates at an overall sampling rate of 10 Hz implemented in ROS. Thereby, the laser scanner provides the strongest and the last echo of a reflected signal.

Dynamic Scenarios In order to collect reflective targets for different distances, dynamic scenarios are conducted as illustrated in Figure 2. Large Zenith Polymer panels (0.5 m × 0.5 m) with diffuse reflectivities of 5 %, 50 % and 90 % are mounted at a height of 0.9 m and are carried from the front of the car to the end of the chamber with constant speed. The panels are moved away from the car to ensure that the fog is not compressed in front of the target. Due to the size and weight of the panels each target has to be moved separately. During one recording the fog density is kept constant within visibility ranges of 20-30 m, 30-40 m

and 40-50 m. After the recordings, the position of the targets are manually annotated while the distance is extracted semi-supervised following [6] based on the last echo of the lidar sensor. For evaluation we use all images from the color camera running at 30 Hz resulting in approximately 1500 images for each target sequence recorded in the three fog density ranges. Including clear reference data approximately 18000 annotated images are available for evaluation in total.

Static Scenarios For being able to perform full reference image quality assessment, the static urban scenario described in Figure 2 is setup for daytime and nighttime conditions. The scene contains several pedestrians, flower pots, tables with chairs, an advertisement pole, a customer stopper, a park bench and a bicycle rack. Hereby, sensor data is recorded during three iterations covering the continuous changes in visibility ranges. For reference, the scenario is recorded under clear weather conditions without any emitted fog. In total, approximately 12000 static images with different fog densities are available.

4. Metric

By using the measurements of the dynamic scenarios, we introduce a new metric based on the intensity of the calibrated reflective targets. We calculate the RMS contrast c_{rms}, given by

$$c_{\mathrm{rms}} = \sqrt{\frac{\left(I_{90} - I_{50}\right)^2}{2} + \frac{\left(I_{50} - I_5\right)^2}{2}}, \qquad (3)$$

where the raw intensities I_5, I_{50}, I_{90} for each target with reflectivities of 5 %, 50 % and 90 % respectively, are extracted by manually labeled bounding boxes. Figure 5 shows the extracted intensities over the distance of the reflective targets from the testing vehicle for clear and foggy conditions with a visibility range of 30-40 m. Thereby, the intensities are averaged over depth bins resulting in the shown error bars. Images with disturbances as windshield wipers are annotated and removed from evaluation. In Figure 5 three prominent maxima can be observed. The peaks highlighted with yellow background originate from the illumination sources mounted at the ceiling of the weather chamber. The respective positions are marked in the scene setup in Figure 2. Figure 6 illustrates the calculated RMS contrast from Eq. (3) within the three different visibility ranges. As expected, the contrast is impaired for decreasing visibility. Since the range of the depth reference is limited by the low performance of the lidar in fog, the contrast curves stop when the reflective targets vanish in the lidar pointcloud. Looking at the degeneration of the contrast for different distances, it can be seen that the chamber regions illuminated by the light sources are widened and lose rapidly
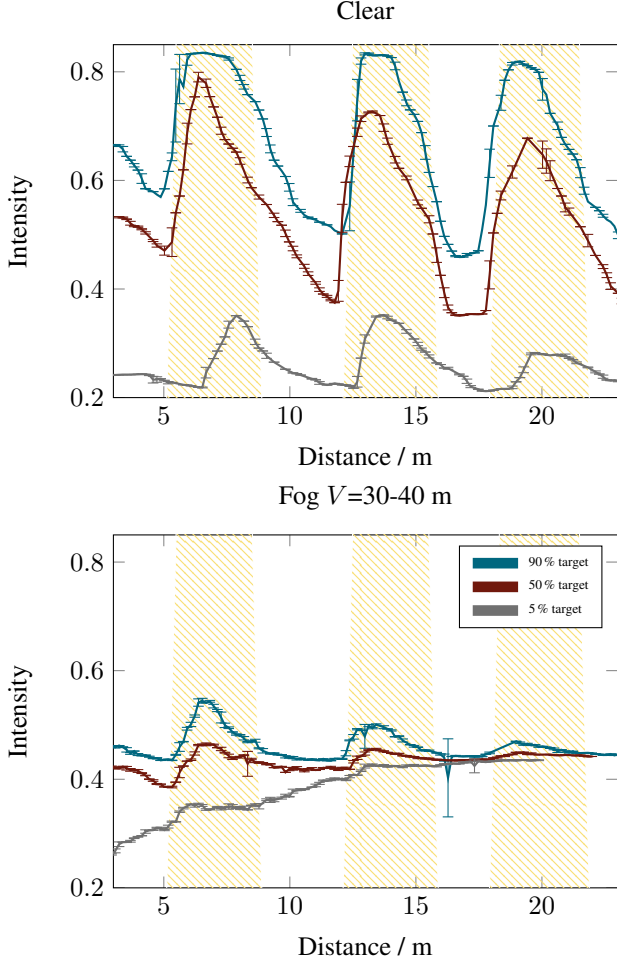
Clear

Fog V=30-40 m

Figure 5: Reference intensity measurements for clear conditions (top) and for a fog density of 30-40 m (bottom). The reference intensites for the three reflective targets with a reflectivity of 5%, 50% and 90% are illustrated relative to their distance. In the lower image it can be observed that for low reflective targets the intensity curves increase caused by the rising airlight per distance, while the intensity of high reflective targets is attenuated. The decrease in intensity between the targets results in a total loss of contrast which is shown in Figure 6.

in contrast. For the regions in-between, the illuminated areas lose the contrast almost completely making the image reconstruction very difficult. Therefore, we split the evaluation into dark and illuminated areas. Total scores are given as an average for each of the regions separately.

In contrast to the dynamic scenarios, the static scenario enables to quantify the amount of restored information in form of full reference based evaluation metrics such as LPIPS [77], MSE, PSNR, SSIM [67] and VIF [60]. Figure 7 visualizes LPIPS, SSIM and MSE as a function of meteorological visibility for the three independent measurement iterations, namely FogA, FogB and FogC.
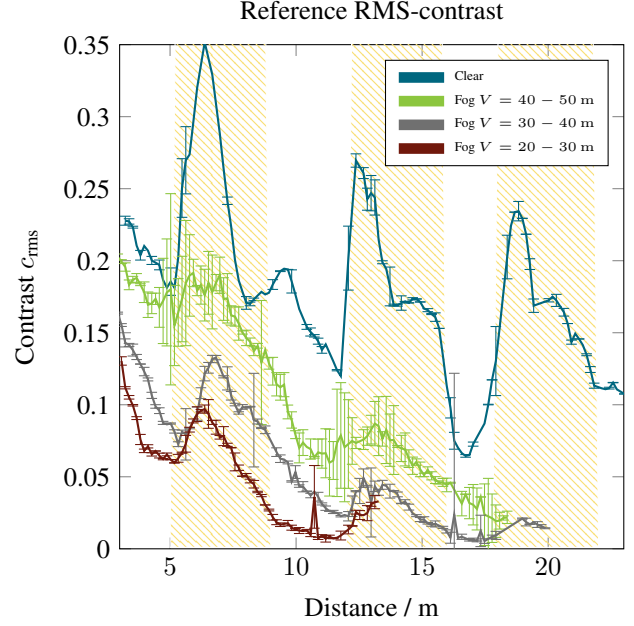

Reference RMS-contrast

Figure 6: Reference contrast without enhancement for different fog densities shown from clear to increased fog levels. The region of the chamber illuminated by light sources are highlighted in yellow.

Label	LPIPS[77]	PSNR	SSIM[67]	MSE	VIF[61]
PIX2PIXHD-BJ [8]	0.01379	29.7192	0.9565	247.2536	0.7488
PIX2PIXHD [66]	0.01280	29.4731	0.9529	258.2451	0.7408
PIX2PIXHD-CJ [8]	0.01684	29.0464	0.9539	252.4982	0.7147
FCN [49]	0.03913	23.7101	0.9001	634.1447	0.7424
FCN† [49]	0.07896	16.3522	0.7694	3591.658	0.7171
DCPDN [72]	0.05197	21.9583	0.8606	2182.039	0.6295
DCPDN† [72]	0.07581	16.0040	0.7823	3668.973	0.6134
PFF-NET [47]	0.04265	27.9973	0.9181	284.5090	0.5912
DEHAZENET [10]	0.09735	14.8190	0.7472	5003.525	0.5893
AOD-NET† [39]	0.08923	14.0910	0.7240	4410.643	0.5359
DEHAZENET† [10]	0.13308	13.6074	0.6676	5698.850	0.5047
CYCLEGAN† [20]	0.11873	13.8803	0.7164	3049.224	0.4199
PFF-NET† [47]	0.16579	14.7734	0.7194	5454.050	0.3976

Table 1: Quantitative enhancement results for the foggified KITTI dataset. All methods except the methods marked with † were finetuned on the KITTI dataset with additional fog densities.

5. Experiments

In order to validate our novel metric, we evaluate different enhancement methods in the described dense foggy conditions. For this purpose, we select several methods recently presented at the NTIRE challenge [3], namely PFF-Net [47], CycleGan [20] and FCN [49], but also established methods such as AOD-Net [39], DCPDN [72] and DehazeNet [10], as well as methods from other domains as Pix2PixHD [66] and Clahe [70]. Since most of the tested methods are only trained down to visibilities of 150 m, we fine tune the methods, if possible, on higher fog densities
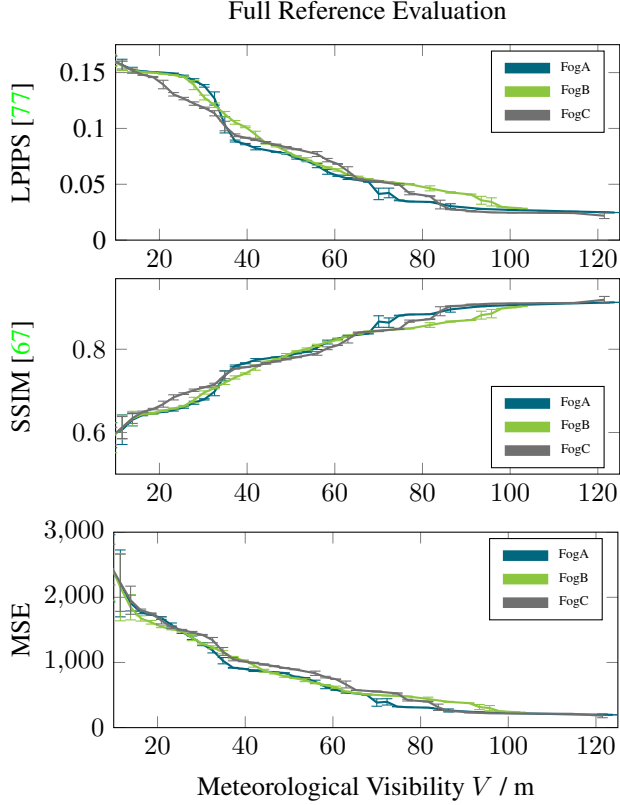
Figure 7: Three different full reference metrics for each independent fog run FogA, FogB, FogC with continuously changing fog density from V=10-125 m.

down to appprox. 20 m.[1] Methods without finetuning are kept as further reference and marked with †. To match the increased fog density we apply Eq. (1) to the KITTI dataset [26] following [57] in order to reach visibilities down to approximately 20 m. The results on the KITTI validation set are presented in Table 1. Nonetheless, the dataset itself shows only daytime scenarios without simulated light sources in late night or twilight conditions. The presented methods were trained on their default input image sizes, see Table 4, and applied to our dataset with an image size of 1920×1024.

All learned enhancement methods show different artifacts, especially close to the light sources or at the transition from darker to brighter areas. Figure 1 presents several qualitative examples for a visibility range of 30-40 m. The poor performance in these transition regions already indicates that better simulation performance in such critical regions will increase the overall enhancement performance in difficult scenarios.

To further analyze the generalization ability, two different derivatives of Pix2PixHD [66] with data aug-

---

[1]The training data split contains equally weighted fog densities from ∞ m - 20 m visibility.
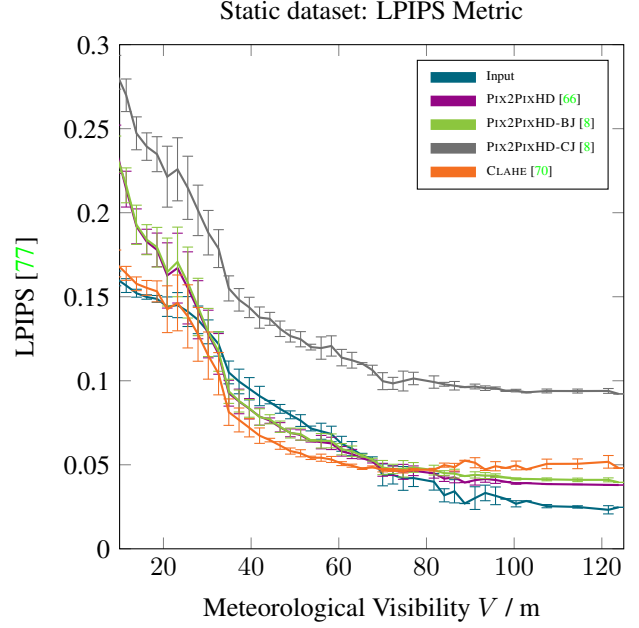

Figure 8: Visibility dependent full reference based evaluation for the LPIPS metric [77] and the four best performing methods based on contrast enhancement. Note that lower scores are better in terms of the LPIPS metric. Interestingly, the scores are higher for applying enhancement methods compared to no enhancement. This is not demonstrating the true enhancement performance as the reference clear image (Figure 3) neither represents the best possible enhancement result nor the training data distribution. It can be stated that the images are semantically identical but differ in color style leading to lower full reference based evaluation results.

METHOD	DEFAULT SIZE
PIX2PIXHD	1248×384
PIX2PIXHD-BJ	1248×384
PIX2PIXHD-CJ	1248×384
CYCLEGAN†	256×256
AOD-NET†	640×480
DEHAZE-NET	16×16
DEHAZE-NET†	16×16
PFF-NET	1248×384
PFF-NET†	520×520
DCPDN	512×512
DCPDN†	512×512
FCN	128×128

Table 4: Default input image sizes for the different methods.

mentation were introduced. We use color jitter transforms from PyTorch [51] with the following parameters: brightness = 0.125, saturation = 0.5, hue = 0.2 and contrast = 0.5. PIX2PIXHD-BJ only augments brightness and contrast while PIX2PIXHD-CJ additionally augments saturation and hue following [8]. Hence, the networks do

Daytime Scores

METRIC	LPIPS[77]			PSNR			SSIM[67]			MSE			VIF[61]		
VISIBILITY V	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m
DATASET	0.133	0.066	0.030	16.536	20.025	23.986	0.680	0.813	0.897	1485.548	675.181	268.602	0.759	0.741	0.675
FCN[49]	0.150	0.082	0.042	11.894	14.844	17.885	0.576	0.719	0.819	4266.039	2202.639	1073.552	0.787	0.716	0.643
AOD-NET†[39]	0.155	0.112	0.084	12.340	13.806	14.939	0.536	0.528	0.522	3814.738	2724.660	2088.643	0.631	0.624	0.577
DEHAZENET†[10]	0.120	0.056	0.039	17.293	18.787	19.600	0.675	0.818	0.865	1248.004	863.807	721.841	0.489	0.623	0.562
PFF-NET[47]	0.142	0.075	0.042	11.442	13.978	16.503	0.586	0.719	0.808	4717.768	2661.093	1468.997	0.443	0.542	0.535
FCN†[49]	0.150	0.077	0.049	16.299	19.160	21.554	0.681	0.747	0.756	1560.878	808.573	459.052	0.489	0.539	0.517
PFF-NET†[47]	0.130	0.065	0.041	16.043	20.794	25.025	0.661	0.838	0.907	1671.934	588.500	208.348	0.357	0.520	0.528
DCPDN[72]	0.130	0.066	0.034	15.794	17.534	20.218	0.674	0.755	0.797	1725.621	1187.259	625.004	0.476	0.495	0.481
CLAHE[70]	0.126	0.054	0.049	16.682	19.395	21.439	0.642	0.798	0.826	1438.851	758.978	469.573	0.349	0.452	0.441
DCPDN†[72]	0.135	0.062	0.035	16.618	20.286	22.546	0.658	0.757	0.783	1463.797	629.698	364.122	0.383	0.432	0.433
PIX2PIXHD[66]	0.146	0.061	0.041	14.244	17.898	20.162	0.595	0.772	0.822	2606.062	1078.579	632.362	0.217	0.421	0.490
PIX2PIXHD-BJ[8]	0.148	0.062	0.043	14.163	17.841	19.808	0.577	0.765	0.811	2702.028	1090.882	684.552	0.200	0.400	0.462
PIX2PIXHD-CJ[8]	0.205	0.117	0.096	13.877	17.426	19.272	0.539	0.743	0.784	2818.792	1193.507	773.717	0.173	0.358	0.411
CYCLEGAN[20]	0.152	0.091	0.066	14.258	14.966	15.577	0.571	0.668	0.707	2472.775	2076.760	1803.073	0.222	0.331	0.342

Nighttime scores

METRIC	LPIPS[77]			PSNR			SSIM[67]			MSE			VIF[61]		
VISIBILITY V	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m	10-40 m	40-75 m	75-125 m
DATASET	0.101	0.067	0.056	16.356	18.785	20.282	0.724	0.805	0.834	1530.795	872.267	611.921	0.529	0.638	0.661
FCN[49]	0.116	0.085	0.077	16.761	16.229	17.960	0.669	0.692	0.685	1373.256	1552.370	1720.260	0.439	0.554	0.580
AOD-NET†[39]	0.109	0.084	0.078	20.322	18.925	18.009	0.737	0.742	0.730	607.776	838.632	1029.609	0.363	0.463	0.491
DCPDN[72]	0.101	0.062	0.051	16.637	19.707	21.563	0.717	0.799	0.826	1450.630	711.037	456.233	0.356	0.430	0.445
DEHAZENET†[10]	0.113	0.073	0.061	10.823	11.680	12.771	0.620	0.711	0.750	5399.507	4447.837	3444.956	0.340	0.417	0.442
FCN†[49]	0.111	0.073	0.063	17.417	21.074	23.373	0.726	0.818	0.845	1218.643	524.433	302.125	0.317	0.397	0.416
PFF-NET[47]	0.106	0.072	0.064	16.571	17.630	17.270	0.686	0.748	0.750	1452.691	1123.258	1219.574	0.275	0.384	0.410
DCPDN†[72]	0.110	0.071	0.060	15.742	18.255	19.994	0.699	0.778	0.804	1760.583	989.804	654.898	0.284	0.339	0.355
PFF-NET†[47]	0.113	0.076	0.070	14.966	18.185	19.111	0.672	0.791	0.819	2143.134	995.687	802.029	0.184	0.322	0.350
CLAHE[70]	0.115	0.083	0.076	16.219	18.169	19.251	0.652	0.731	0.754	1575.560	998.911	774.356	0.235	0.308	0.326
CYCLEGAN†[20]	0.130	0.099	0.091	15.809	17.735	18.848	0.656	0.721	0.741	1719.124	1101.370	851.247	0.173	0.211	0.223
PIX2PIXHD[66]	0.173	0.158	0.160	13.604	14.246	14.528	0.593	0.635	0.632	2850.898	2446.887	2293.717	0.144	0.205	0.216
PIX2PIXHD-BJ[8]	0.168	0.169	0.174	12.741	14.122	14.740	0.554	0.592	0.585	3484.929	2522.743	2184.644	0.133	0.188	0.199
PIX2PIXHD-CJ[8]	0.227	0.227	0.226	12.509	13.215	13.202	0.512	0.541	0.534	3669.302	3101.872	3110.837	0.124	0.181	0.192

Table 2: Static evaluation scores for daytime (top) and nighttime illumination (bottom) settings. The table shows column wise different metrics averaged within three different fog visibility ranges. Note that for the metrics LPIPS [77] and MSE lower scores are better, while for SSIM [67], PSNR and VIF [61] higher scores are desired. The best and second best performing models are marked for each metric and fog visibility. Best models are marked **magenta** and second best models **blue**. The models are sorted according to the vif metric.

VISIBILITY	20-30 m			30-40 m			40-50 m			Clear		
DISTANCE	5-9 m	12-16 m	9-12 m	5-9 m	12-16 m	9-12 m	5-9 m	12-16 m	9-12 m	5-9 m	12-16 m	9-12 m
PIX2PIXHD-CJ[8]	0.146	0.064	0.087	0.346	0.145	0.192	0.438	0.369	0.286	0.440	0.426	0.424
PIX2PIXHD[66]	0.175	0.042	0.030	0.315	0.112	0.107	0.409	0.203	0.168	0.524	0.468	0.383
PIX2PIXHD-BJ[8]	0.142	0.030	0.052	0.252	0.091	0.074	0.363	0.151	0.158	0.473	0.485	0.389
CLAHE[70]	0.122	0.055	0.027	0.202	0.073	0.069	0.281	0.147	0.119	0.303	0.299	0.258
FCN†[49]	0.082	0.038	0.018	0.122	0.044	0.041	0.199	0.094	0.083	0.284	0.257	0.194
PFF-NET†[47]	0.096	0.047	0.018	0.125	0.050	0.035	0.190	0.092	0.069	0.267	0.213	0.164
DEHAZENET†[10]	0.098	0.036	0.020	0.137	0.048	0.040	0.204	0.091	0.078	0.272	0.202	0.128
DCPDN[72]	0.105	0.051	0.031	0.132	0.028	0.038	0.241	0.078	0.084	0.273	0.228	0.162
PFF-NET[47]	0.069	0.026	0.012	0.104	0.032	0.031	0.171	0.068	0.062	0.273	0.216	0.162
FCN[49]	0.063	0.026	0.014	0.094	0.030	0.027	0.183	0.067	0.066	0.269	0.205	0.151
DATASET	0.066	0.025	0.013	0.099	0.031	0.029	0.165	0.066	0.063	0.230	0.187	0.131
CYCLEGAN†[20]	0.053	0.033	0.023	0.104	0.043	0.053	0.213	0.061	0.090	0.271	0.258	0.240
AOD-NET†[39]	0.038	0.016	0.006	0.056	0.018	0.012	0.116	0.036	0.038	0.175	0.149	0.109
DCPDN†[72]	0.042	0.015	0.033	0.136	0.047	0.044	0.247	0.034	0.102	0.355	0.310	0.238

Table 3: Dynamic contrast evaluations for three regions in front of the testing vehicle within three different visibility ranges and reference clear conditions. The distance range 9-12 m lays in-between the illumination sources and therefore shows the lowest contrast. The best and second best performing models are marked for each visibility and region. Best models are marked **magenta** and second best models **blue**. The models are sorted following the contrast results for a visibility 40-50 m and the region 12-16 m.

not only learn to remove fog but also to revert disturbed colors which has a positive effect on generalizing to real adverse weather data.

Even though the two methods using data augmentation achieve the best visual enhancement (see Figure 1), the values for the full reference based evaluation metrics (Table 1 and Table 2) aggravate. The downside of these two methods is that the networks can not infer the present illumination during training correctly which is unfavorable when target and output image have to match precisely.
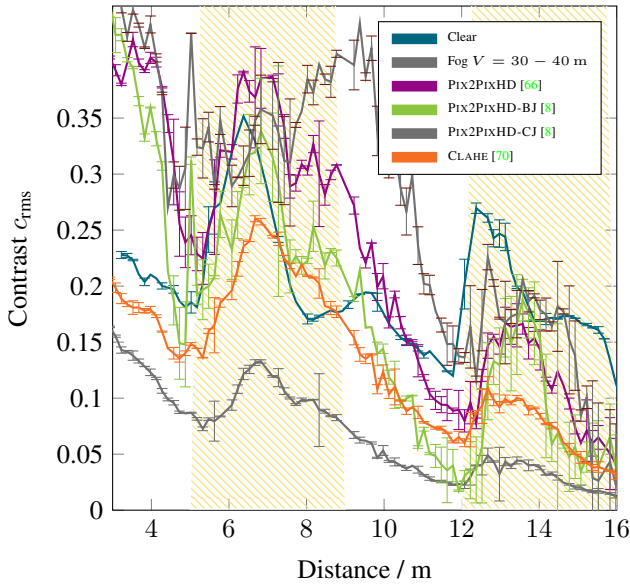
Figure 9: Distance dependent contrast results for the four best enhancement methods compared to the fog free reference input and clear conditions. In some regions the enhanced contrast overcomes the clear reference.

Full reference methods do a great job in assessing whether semantics are correctly transferred between images but are lacking in meaningfulness if two images differ only in image color distribution but match in semantics. This is interesting as almost all presented methods achieve lower scores compared to no enhancement based on full reference evaluation and Figure 8 as clear reference. Therefore, the static evaluations do not necessarily represent the real visual enhancement. This behavior can be explained by two reasons. The first reason is that the methods have been entirely trained on a different dataset with a different color distribution leading to a contrast enhancement but at the same time change the color distributions towards the training domain. Facing this shortcoming in full reference based evaluation methods, especially regarding adverse weather conditions, it is interesting to assess the dynamic reflective target curves. Table 3 and Figure 9 show the evaluations. Given those results it is possible to reproduce the performance ranking from the synthetic fine tunings in Table 1 to the extent that models with data augmentation achieve better contrast enhancements than without data augmentation.

Equipped with this knowledge the second reason becomes clear as well: Successful enhancement methods have achieved higher contrast values than the fog free reference. Hence, the network performed enhancement steps going beyond haze removal. Evidently, current de-hazing methods are also capable to achieve general image enhancements.

6. Conclusion

In this paper we have benchmarked recent enhancement methods in new challenging environment conditions. Especially in difficult illumination settings, e.g. in-between illumination sources, current methods fail. As the presented methods do not incorporate simulation and modeling of different levels of disturbance caused by differently mounted light sources, this can fuel the development of future enhancement techniques. Additionally, the presented benchmark dataset is large enough to enable elaborating signal disturbances caused by light sources. Our introduced benchmark data set does not only empower the development of new vision based simulation and enhancement methods but also offers the possibility for further signal improvement by providing sensor data from other sensor types as gated camera and lidar. Furthermore, it has been proven that full reference based image evaluation methods do not provide sufficient validity for an in depth enhancement assessment. Effects coming along with real world conditions, as contrast and illumination, should be taken into account. Otherwise, a well-founded comparison of different approaches can not be assured as superior methods might be punished because of the metrics lack in incorporating relevant information. We therefore developed a novel metric addressing these real world effects. Benchmarking state-of-the-art approaches shows how this novel metric leads to a meaningful and more intuitive way of evaluating enhancement methods in challenging environment conditions.

Acknowledgment

References

[1] M. Ahmad and D. Sundararajan. A fast algorithm for two dimensional median filtering. *IEEE Transactions on Circuits and Systems*, 34(11):1364–1374, 1987. 3

[2] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer. D-HAZY: A dataset to evaluate quantitatively dehazing algorithms. *IEEE International Conference on Image Processing (ICIP)*, pages 2226–2230, 2016. 3

[3] C. Ancuti, C. O. Ancuti, and R. Timofte. NTIRE 2018 Challenge on Image Dehazing: Methods and Results. *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 3, 5

[4] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer. I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images. *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 620–631, 2018. 2, 3

[5] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer. O-HAZE: A Dehazing Benchmark with Real Hazy and Haze-Free Outdoor Images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 867–8678, 2018. 2, 3

[6] M. Bijelic, T. Gruber, and W. Ritter. A Benchmark for Lidar Sensors in Fog: Is Detection Breaking Down? *IEEE Intelligent Vehicles Symposium (IV)*, pages 760–767, 2018. 1, 4

[7] M. Bijelic, T. Gruber, and W. Ritter. Benchmarking Image Sensors Under Adverse Weather Conditions for Autonomous Driving. *IEEE Intelligent Vehicles Symposium (IV)*, pages 1773–1779, 2018. 1, 4

[8] M. Bijelic, F. Mannan, T. Gruber, W. Ritter, K. Dietmayer, and F. Heide. Seeing Through Fog Without Seeing Fog: Deep Sensor Fusion in the Absence of Labeled Training Data. *CoRR*, abs/1902.08913, 2019. 1, 5, 6, 7, 8

[9] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. In *ICLR*, 2017. 3

[10] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. DehazeNet: An End-to-End System for Single Image Haze Removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 1, 3, 5, 7

[11] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546, 2000. 3

[12] J. Chen, C.-H. Tan, J. Hou, L.-P. Chau, and H. Li. Robust video content alignment and compensation for clear vision through the rain. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[13] A. Cohen and J. Kovacevic. Wavelets: The mathematical background. *Proceedings of the IEEE*, 84(4):514–522, 1996. 3

[14] M. Colomb, J. Dufour, M. Hirech, P. Lacôte, P. Morange, and J.-J. Boreux. Innovative artificial fog production device-A technical facility for research activities. 2004. 3

[15] 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 3

[16] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 3

[17] U. Demir and G. B. Ünal. Patch-based image inpainting with generative adversarial networks. *CoRR*, abs/1803.07422, 2018. 3

[18] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. 3

[19] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 3

[20] D. Engin, A. Genç, and H. Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 825–833, 2018. 1, 3, 5, 7

[21] G. Fan and X.-G. Xia. Image denoising using a local contextual hidden markov model in the wavelet domain. *IEEE Signal Processing Letters*, 8(5):125–128, 2001. 3

[22] X. Fei, L. Xiao, Y. Sun, and Z. Wei. Perceptual image quality assessment based on structural similarity and visual masking. *Signal Processing: Image Communication*, 27(7):772–783, 2012. 2

[23] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017. 2, 3

[24] A. Galdran. Image dehazing by artificial multiple-exposure image fusion. *Signal Processing*, 149:135–147, 2018. 2, 3

[25] L. A. Gatys, A. S. Ecker, and M. Bethge. A Neural Algorithm of Artistic Style. *CoRR*, abs/1508.06576, 2015. 3

[26] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 6

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 3

[28] S. Hasirlioglu, I. Doric, A. Kamann, and A. Riener. Reproducible fog simulation for testing automotive surround sensors. *IEEE 85th Vehicular Technology Conference (VTC Spring)*, pages 1–7, 2017. 2

[29] S. Hasirlioglu, A. Kamann, I. Doric, and T. Brandmeier. Test methodology for rain influence on automotive surround sensors. In *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2242–2247. IEEE, 2016. 1, 2

[30] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2011. 3

[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *CoRR*, abs/1602.07360, 2016. 2

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 3

[33] V. Kamble and K. Bhurchandi. No-reference image quality assessment algorithms: A survey. *Optik - International Journal for Light and Electron Optics*, 126:1090–1097, 2015. 2

[34] H. Koschmieder. *Theorie der horizontalen Sichtweite*. 1924. 2, 3

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1, 2012. 2

[36] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. DeblurGAN: Blind motion deblurring using conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8183–8192, 2018. 3

[37] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 3

[38] S. Lefkimmiatis. Universal denoising networks: a novel CNN architecture for image denoising. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2018. 3

[39] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. An all-in-one network for dehazing and beyond. *CoRR*, abs/1707.06543, 2017. 1, 3, 5, 7

[40] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, Jan 2019. 2, 3

[41] K. Li, Y. Li, S. You, and N. Barnes. Photo-realistic simulation of road scene for data-driven methods in bad weather. *Proceedings of the IEEE International Conference on Computer Vision*, pages 491–500, 2017. 2, 3

[42] Y. Li, R. T. Tan, and M. S. Brown. Nighttime haze removal with glow and multiple light colors. pages 226–234, 2015. 3

[43] L. Liu, B. Liu, H. Huang, and A. C. Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014. 2

[44] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised Image-to-Image Translation Networks. In *NIPS*, 2017. 2

[45] W. Lu, K. Zeng, D. Tao, Y. Yuan, and X. Gao. No-reference image quality assessment in contourlet domain. *Neurocomputing*, 73(4-6):784–794, 2010. 2

[46] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: application to JPEG2000. *Signal processing: Image Communication*, 19(2):163–172, 2004. 2

[47] K. Mei, A. Jiang, J. Li, and M. Wang. Progressive Feature Fusion Network for Realistic Image Dehazing. *CoRR*, abs/1810.02283, 2018. 1, 5, 7

[48] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 2

[49] R. Mondal, S. Santra, and B. Chanda. Image dehazing by joint estimation of transmittance and airlight using bi-directional consistency loss minimized FCN. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 920–928, 2018. 1, 5, 7

[50] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. 2

[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. De-Vito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. 2017. 6

[52] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain. In *ICIP*, 2001. 3

[53] J. Pu, X. Chen, L. Zhang, Q. Zhou, and Y. Zhao. Removing rain based on a cycle generative adversarial network. *13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 621–626, 2018. 2, 3

[54] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*, 2016. 3

[55] M. S. Sajjadi, B. Scholkopf, and M. Hirsch. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017. 3

[56] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018. 2

[57] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, pages 1–20, 2018. 2, 3, 6

[58] C. Sakaridis, D. Dai, and L. Van Gool. Semantic Nighttime Image Segmentation with Synthetic Stylized Data, Gradual Adaptation and Uncertainty-Aware Evaluation. *CoRR*, abs/1901.05946, 2019. 2

[59] L. Sendur and I. W. Selesnick. Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency. *IEEE Transactions on Signal Processing*, 50(11):2744–2756, 2002. 3

[60] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:iii–709, 2004. 2, 5

[61] H. R. Sheikh and A. C. Bovik. A visual information fidelity approach to video quality assessment. In *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, volume 7, 2005. 2, 5, 7

[62] H. R. Sheikh, A. C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005. 2

[63] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 2

[64] J.-P. Tarel, N. Hautiere, L. Caraffa, A. Cord, H. Halmaoui, and D. Gruyer. Vision enhancement in homogeneous and heterogeneous fog. *IEEE Intelligent Transportation Systems Magazine*, 4(2):6–20, 2012. 3

[65] A. von Bernuth, G. Volk, and O. Bringmann. Rendering Physically Correct Raindrops on Windshields for Robustness Verification of Camera-based Object Recognition. *IEEE Intelligent Vehicles Symposium (IV)*, pages 922–927, 2018. 2

[66] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic

manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 1, 5, 6, 7, 8

[67] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, volume 13, pages 600–612, 2004. 2, 5, 6, 7

[68] WMO. *Guide to Meteorological Instruments and Methods of observation*, volume I & II. 2008. 2

[69] S. Wu, W. Lin, S. Xie, Z. Lu, E. P. Ong, and S. Yao. Blind blur assessment for vision-based applications. *Journal of Visual Communication and Image Representation*, 20(4):231–241, 2009. 2

[70] G. Yadav, S. Maheshwari, and A. Agarwal. Contrast limited adaptive histogram equalization based enhancement for real time video system. In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2392–2397. IEEE, 2014. 1, 5, 6, 7, 8

[71] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[72] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3194–3203, 2018. 1, 5, 7

[73] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 695–704, 2018. 2

[74] H. Zhang, V. Sindagi, and V. M. Patel. Image De-raining Using a Conditional Generative Adversarial Network. *CoRR*, abs/1701.05957, 2017. 2, 3

[75] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions Image Processing*, 26(7):3142–3155, 2017. 3

[76] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning Deep CNN Denoiser Prior for Image Restoration. In *CVPR*, pages 2808–2817. IEEE Computer Society, 2017. 3

[77] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2, 5, 6, 7

[78] Y. Zhang, L. Ding, and G. Sharma. HazeRD: An outdoor scene dataset and benchmark for single image dehazing. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3205–3209, 2017. 3

[79] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. *European Conference on Computer Vision (ECCV)*, pages 597–613, 2016. 3

[80] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3