ZeroScatter: Domain Transfer for Long Distance Imaging and Vision through Scattering Media

Zheng Shi[1*] Ethan Tseng[1*] Mario Bijelic[2*] Werner Ritter[2] Felix Heide[1,3]

[1]Princeton University [2]Mercedes-Benz AG [3]Algolux

Abstract

Adverse weather conditions, including snow, rain, and fog, pose a major challenge for both human and computer vision. Handling these environmental conditions is essential for safe decision making, especially in autonomous vehicles, robotics, and drones. Most of today's supervised imaging and vision approaches, however, rely on training data collected in the real world that is biased towards good weather conditions, with dense fog, snow, and heavy rain as outliers in these datasets. Without training data, let alone paired data, existing autonomous vehicles often limit themselves to good conditions and stop when dense fog or snow is detected. In this work, we tackle the lack of supervised training data by combining synthetic and indirect supervision. We present ZeroScatter, a domain transfer method for converting RGB-only captures taken in adverse weather into clear daytime scenes. ZeroScatter exploits model-based, temporal, multi-view, multi-modal, and adversarial cues in a joint fashion, allowing us to train on unpaired, biased data. We assess the proposed method on in-the-wild captures, and the proposed method outperforms existing monocular descattering approaches by 2.8 dB PSNR on controlled fog chamber measurements.

1. Introduction

In the presence of a scattering medium, such as fog or snow, photons no longer propagate along a straight path but instead are redirected by particles, potentially many times, until arriving at the camera. This includes forward scattered light emitted from sources in the scene, e.g., an oncoming vehicle headlight, captured as a passive component by an RGB camera or human eye, and backward scattering observed when actively illuminating the scene, e.g., in automotive lidar or with the ego-vehicle headlights. While adverse weather conditions that include severe scattering are heavily underrepresented in existing training and evaluation datasets [45, 13, 9], these rare scenarios are a significant contributing factor for fatal automotive accidents [4], as a direct result of vision impairment for human drivers.

Supervised imaging and vision approaches are also fundamentally limited in adverse weather conditions. Adverse weather conditions follow a long-tail distribution where such environments are rarely encountered during day-today driving, making data collection, training, and evaluation challenging [37]. As a result, critical computer vision tasks such as object detection and tracking are often trained on clear day inputs and fail to generalize when the input scene is perturbed by adverse effects from scattering media. Even if adverse weather data is available, the scattering media would still affect the quality of human annotations used for supervision. Furthermore, supervised dehazing and defogging methods are restricted by the difficulty of acquiring paired perturbed and clear data, which is infeasible due to the dynamic nature of real-world automotive scenes. As such, supervised training on real-world data has been a fundamental challenge for imaging and vision in harsh weather conditions. To tackle this problem, existing approaches attempt to solve a domain transfer problem using simulated scattering media [35, 36, 42, 18]. However, these simulation models do not adequately simulate the effects that are observed in the wild. Unsupervised learning approaches have demonstrated impressive ability for image domain transfer but remain restricted to a single domain, e.g. faces, and small image resolutions [57, 24].

Researchers have also adopted alternative sensing modalities beyond conventional intensity imaging, e.g. lidar and radar, in robotic and automotive applications. However, they do not offer a solution in backscatter-limited weather scenarios. Specifically, pulsed lidar sensors that record the round-trip time of the first response fail to extract meaningful scene surfaces in severe snow and fog, fundamentally limited by backscatter [5], and indeed trail the performance of RGB stereo depth methods [16] in dense fog. While the mm-wavelengths of radar systems penetrate dense fog, existing radar systems are limited to low angular resolution, and hence do not allow for scene understanding tasks beyond the detection and tracking of objects with a large radar cross-section [27]. At the same time, RGB intensity cameras have become a ubiquitous sensor technology because of their low-cost and high spatial resolutions up to 250 MPix in modern commodity sensors [38], deployed across application domains from miniature smartphone cameras to automotive imaging systems. As such, in this work, we address the task of imaging through scattering media using conventional RGB cameras.

*indicates equal contribution.

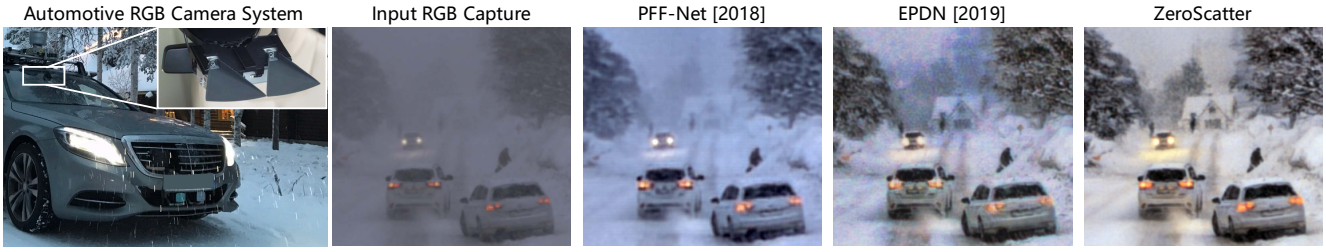

Figure 1: Scattering stemming from snow, rain, or fog significantly reduces the perceptible quality of RGB captures and impact downstream computer vision tasks such as object detection. The proposed method, which we dub "ZeroScatter", reliably removes these scattering effects for unseen automotive scenes.

We tackle this challenge by proposing ZeroScatter, a novel domain transfer method that converts RGB images corrupted by adverse weather effects into clear day scenes. To do this, we exploit a variety of training signals in order to achieve robust descattering performance on real-world examples. First, we employ a synthetic weather model using cycle consistency training. Second, we employ temporal and multi-view consistency to ensure stable model performance and to eliminate spurious adverse weather effects such as snowflakes, leveraging an adverse weather dataset [5]. Third, we employ multi-modal supervision using auxiliary data acquired by gated imagers [17]. Gated imaging is an emerging time-of-flight imaging technology that records photons with specific return times which allows it to image objects at select distances. This imaging modality is less susceptible to path lengths and provides higher contrast training signal for ZeroScatter. All of these training cues enable ZeroScatter to reliably reconstruct RGB captures that have been corrupted by adverse weather. For quantitative evaluation, we evaluated ZeroScatter on scenes with synthetically generated and laboratory generated adverse weather where we demonstrate $2.8\,\mathrm{dB}$ PSNR improvement over state-of-the-art methods.

Specifically, we make the following contributions:

- We propose a novel domain adaptation method which we call ZeroScatter for eliminating scattering media from conventional RGB captures, operating at real-time frame rates of 20 FPS.
- We employ a novel combination of synthetic and real-world data to train ZeroScatter with unpaired, biased datasets. To this end, we incorporate model-based cues jointly with multi-modal, multi-view, temporal and adversarial cues.
- In addition to qualitative improvements on real-world captures, we outperform state-of-the-art methods in controlled fog-chamber evaluation. Our method also outperforms state-of-the-art object detection in harsh weather at long distances.

2. Related Work

Descattering A variety of image descattering techniques have been proposed in recent years. Several works have been proposed for single dedicated tasks such as dehazing [8, 29, 35, 36, 40], removing rain [21, 11, 53, 39, 52, 56, 48], removing snow [34], and translating night to day [56].

Earlier descattering approaches that employed convolutional neural networks (CNNs) [8, 36] learned the scattering effects as a residual image by separately estimating the airlight and the transmission. However, this disjoint learning approach can amplify prediction errors. Li et al. [29] proposes to learn both parameters in an end-to-end fashion by inverting the image formation model. Similar approaches have been proposed for removing rain [11, 48]. These methods demonstrate strong performance through their explicit image formation models., but are difficult to apply to other adverse weather types. Recent methods [35, 21] directly learn the desired descattering without a prescribed image formation model. These methods are trained entirely using synthetically simulated weather conditions, and therefore struggle with real-world scenes.

Domain Adaptation The recent development of GAN architectures [51, 40] has demonstrated impressive results for image translation. However, most of these methods require paired simulated data consisting of full pixel-wise ground truth images for supervised training.

Methods that do not require paired ground truth [10, 55, 50] are based on CycleGAN [57]. While this allows for better training stability, it is difficult to learn both directions of the cycle, specifically the descattering and re-scattering processes. We alleviate these limitations for ZeroScatter by employing a novel cycle training approach where we train the descatterer but utilize a fixed adverse weather simulator for the reverse direction of the cycle. Furthermore, our use of temporal, multi-view, and multi-modal supervision improves ZeroScatter's generalization to real-world inputs over methods that do not utilize additional cues.

Weather Simulation and Datasets Adverse weather simulation techniques have been developed for snow-

fall [34], rainfall [19, 18], blur [28], fog [31, 12, 43], night driving [44, 32], and raindrops on the windshield [47]. Most datasets [43, 46, 1, 54, 31, 30] are based on Koschmieder's physical model [26]. These techniques overlay clear weather images with one type of adverse weather perturbation to create paired examples for supervised training. Very few datasets contain real-world adverse weather scenes [2, 3, 30, 16]. RESIDE [30] contains 4322 real foggy scenes obtained from the internet, along with their annotated object detection labels to enable task-driven dehazing. The O-HAZE [3] and I-HAZE [2] datasets contain real outdoor and indoor hazy scenes respectively which were generated with professional haze machines. However, the datasets are very small with only 45 outdoor and 35 indoor image pairs. Gruber et al. [16] provides a recent depth benchmark dataset with four scenes under different conditions that, as such, is too small for training purposes. In order to provide a variety of training cues for ZeroScatter, we utilize an adverse weather dataset containing real-world automotive captures from northern Europe [5]. In addition to RGB captures, the dataset consists of multi-modal data in the form of gated images [15], multi-view stereo data, and temporal sequences.

3. Domain Transfer with ZeroScatter

3.1. Formulation

To train a reconstruction network G without supervised training data available, we employ cues from adverse weather simulation, multi-modal cues that other sensors can provide, multi-view cues, and temporal consistency cues. Specifically, let X be the domain of raw RGB images, Y be the (unpaired) domain of processed daytime RGB images, and S be the (unpaired) domain of RGB images with scattering present. We train the mapping $G : X \cap S \rightarrow Y \setminus S$, which itself is composed of a translation block $G_\mathrm{T} : X \cap S \rightarrow Y \setminus S$ for image domain transfer and a consistency block $G_\mathrm{C} : Y \setminus S \rightarrow Y \setminus S$ for minimizing temporal and spatial jitter. As illustrated in Figure 2, we employ several auxiliary mapping functions to facilitate our learning scheme.

The model-based learning cycles utilize a user-defined ISP processing function $F_\mathrm{Proc} : X \rightarrow Y$ and an adverse weather simulator $F_\mathrm{Syn} : S^c \rightarrow S$. These mappings enable two training cycles, one involving clear daytime images, which we call "Clear to Scatter to Clear":

$$I_\mathrm{in} \rightarrow F_\mathrm{Syn}(I_\mathrm{in}) \rightarrow G_\mathrm{T}(F_\mathrm{Syn}(I_\mathrm{in})) \approx F_\mathrm{Proc}(I_\mathrm{in}), \quad (1)$$

where $I_\mathrm{in} \in X \setminus S$ is clear daytime images; and another involving scatter corrupted daytime images which we call "Scatter to Clear to Scatter":

$$I_\mathrm{in} \rightarrow G_\mathrm{T}(I_\mathrm{in}) \rightarrow F_\mathrm{Syn}(G_\mathrm{T}(I_\mathrm{in})) \approx F_\mathrm{Proc}(I_\mathrm{in}), \quad (2)$$

where $I_\mathrm{in} \in X \cap S$ is scatter corrupted daytime images.

Indirect supervision with multi-modal data is performed using gated images, as it is less affected by scatters. We pretrain a neural network $F_\mathrm{RGB2Gated} : Y \setminus S \rightarrow Z$ for inferring gated images Z from processed clear daytime scenes. We then use it with the real captured gated images I_gated:

$$I_\mathrm{in} \rightarrow F_\mathrm{RGB2Gated}(G_\mathrm{T}(I_\mathrm{in})) \approx I_\mathrm{gated} \quad (3)$$

where $I_\mathrm{in} \in X \cap S$ is scatter corrupted daytime images. Lastly, we utilize temporal and multi-view data as learning cues. This indirect supervision is facilitated by a temporal warper $F_\mathrm{TempWarp} : X^{(t+\epsilon)} \rightarrow X^{(t)}$ which warps temporally adjacent frames to the current frame and a stereo warper $F_\mathrm{StereoWarp} : X^{(r)} \rightarrow X^{(l)}$ which warps the right stereo image $X^{(r)} = X \cap R$ onto the left viewpoint $X^{(l)} = X \cap L$. We feed the warped images in addition to the current left capture through G_T and then we train G_C to complete the following training paths:

$$I_\mathrm{in}^{(l,t)} \rightarrow G_\mathrm{C}(G_\mathrm{T}(I_\mathrm{in}^{(l,t)})) \approx G_\mathrm{T}(F_\mathrm{TempWarp}(I_\mathrm{in}^{(l,t+\epsilon)})), \quad (4)$$

and

$$I_\mathrm{in}^{(l,t)} \rightarrow G_\mathrm{C}(G_\mathrm{T}(I_\mathrm{in}^{(l,t)})) \approx G_\mathrm{T}(F_\mathrm{StereoWarp}(I_\mathrm{in}^{(r,t)})). \quad (5)$$

In the following, we first describe each of these training components in more detail before discussing the generator architecture we employ for $G_\mathrm{C}(G_\mathrm{T}(\cdot))$.

3.2. Model-Based Synthetic Supervision

Our model-based training scheme has two training cycles called "Clear to Scatter to Clear" (C2C) and "Scatter to Clear to Scatter" (S2S). We employ a fixed adverse weather simulator $F_\mathrm{Syn} : S^c \rightarrow S$ that applies simulated adverse weather to RGB images, based on haze estimation following Koschmieder's model [26] with several modifications that promote generalization to real-world scenes. This ensures ZeroScatter is able to handle various-intensity and depth-dependent scatter effects. As many computer vision applications consume ISP processed images instead of raw camera captures, we also employ a post-processing function $F_\mathrm{Proc} : X \rightarrow Y$. This function can be arbitrarily defined by the user, for this work we define F_Proc to be a raw capture to daytime RGB mapping. These two functions are applied in a cyclic manner to the output of the generator translation block $G_\mathrm{T} : X \cap S \rightarrow Y \setminus S$ as shown in Figure 2. For more detail on F_Syn and F_Proc please refer to the Supplemental Document.

Our model-based supervision aims to minimize

$$\mathcal{L}_\mathrm{Model} = \mathcal{L}_\mathrm{C2C} + \mathcal{L}_\mathrm{S2S}. \quad (6)$$

For the C2C cycle we compute the loss using the input clear weather image $I_\mathrm{in} \in X \setminus S$:

$$\mathcal{L}_\mathrm{C2C} = (\mathcal{L}_1 + \mathcal{L}_\mathrm{perc} + \mathcal{L}_\mathrm{grad} + \mathcal{L}_\mathrm{adv})(I_\mathrm{T}, I_\mathrm{target}), \quad (7)$$

where $I_\mathrm{T} = G_\mathrm{T}(F_\mathrm{Syn}(I_\mathrm{in}))$ and $I_\mathrm{target} = F_\mathrm{Proc}(I_\mathrm{in})$ is the processed target image, $\mathcal{L}_1$ is the Mean Absolute Error loss,

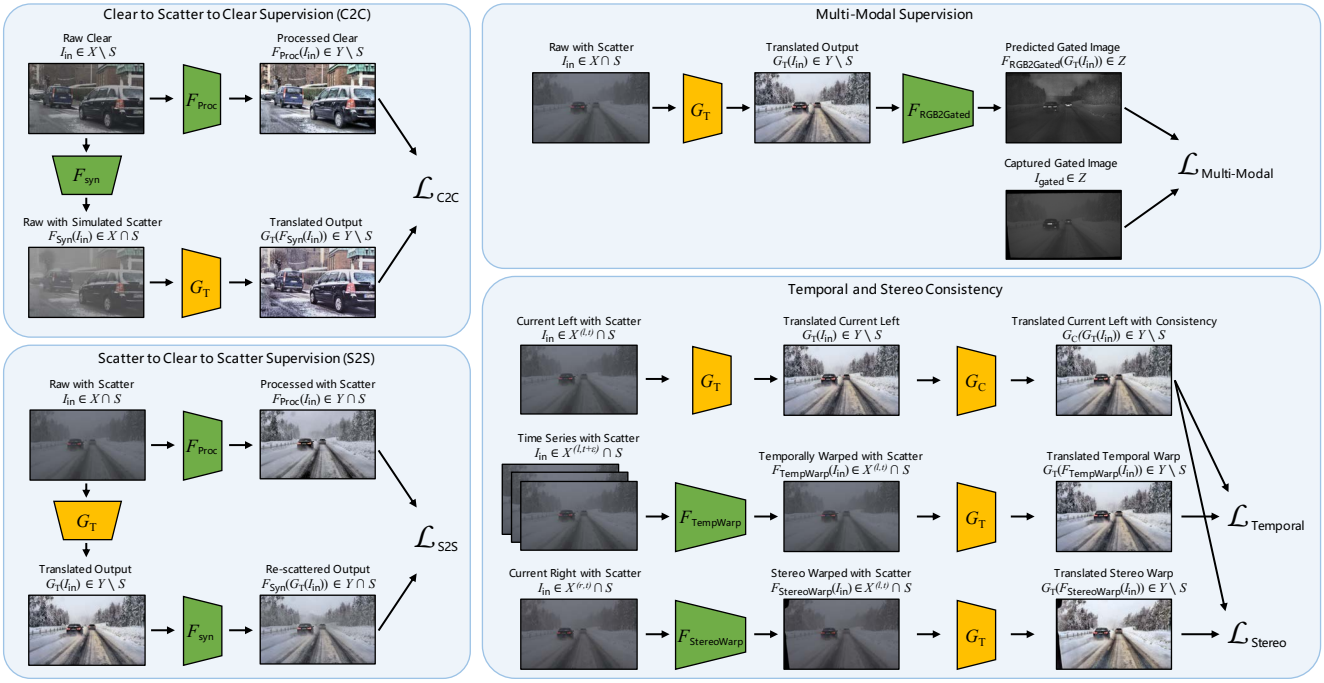

Figure 2: Overview of the proposed method. We train our generator using a novel combination of training cues that promote high-contrast, scatter-free, jitter-free results on unseen real-world scenes. We employ model-based supervision using cycle training which is facilitated by a robust adverse weather model, multi-modal supervision in the form of gated images for training on real heavy weather scenes, and consistency supervision in the form of temporal and stereo losses.

$\mathcal{L}_{\text{perc}}$ is a VGG-19 based perceptual loss [23], $\mathcal{L}_{\text{grad}}$ is an image gradient loss, and $\mathcal{L}_{\text{adv}}$ is a GAN based adversarial loss [14].

For the S2S cycle we compute the loss using the input adverse weather image $I_{\text{in}} \in X \cap S$ as

$$\mathcal{L}_{\text{S2S}} = \mathcal{L}_{\text{adv}}(F_{\text{Proc}}(I_{\text{in}}), F_{\text{Syn}}(G_{\text{T}}(I_{\text{in}}))). \tag{8}$$

Since there are a wide variety of plausible adverse scatter effects, we avoid using $\mathcal{L}_1$, $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{perc}}$ in the S2S cycle and instead use only an adversarial loss.

3.3. Multi-Modal Indirect Supervision

We employ a multi-modal indirect supervision approach to facilitate training on data captured in-the-wild, which makes use of emerging gated imagers [15, 17, 5] that uses active flash illumination to acquire high contrast images by temporally gating out scattering components.

As such, gated images are less affected by adverse weather than RGB cameras [6]. However, they cannot be directly used for training supervision due to the domain shift between gated images and RGB images, e.g. gated images lack color information, see Fig. 2. To overcome this domain shift, we train an RGB2Gated network $F_{\text{RGB2Gated}} : Y \setminus S \rightarrow Z$, where Z is the domain of gated images. This network predicts the gated image corresponding to a processed clear day RGB capture. By training our RGB2Gated network only on clear day images, we teach the network to predict the gated image in the absence of scattering media. We apply $F_{\text{RGB2Gated}}$ to the RGB output of G_{T}, which then allows us to compute a loss with respect to the actual gated image. As a result, our gated supervision loss encourages our generator to remove adverse weather effects to match the underlying image with scattering removed. For details on the RGB2Gated network architecture and training procedure please see the Supplemental Document.

During training we apply $F_{\text{RGB2Gated}}$ to $I_{\text{T}} = G_{\text{T}}(I_{\text{in}})$, $I_{\text{in}} \in X \cap S$, and compare the resulting image I'_{gated} to the corresponding real gated image I_{gated}. To filter out areas that contain insufficient information due to extreme long distance and overly strong reflections from retroreflectors, we apply a mask M_{ent} based on the local entropy of the real gated capture. The multi-modal supervision loss is expressed as

$$\mathcal{L}_{\text{Multi-Modal}} = \mathcal{L}_{\text{perc}}(M_{\text{ent}} \odot I_{\text{gated}}, M_{\text{ent}} \odot I'_{\text{gated}}), \tag{9}$$

where $\odot$ is point-wise multiplication.

We emphasize that we only use gated images for training supervision and that the generator only requires RGB inputs at test time. Our multi-modal loss provides a better training signal for the proposed ZeroScatter method but does not require the specialized gated imaging system at test time.

3.4. Temporal and Stereo Consistency

We employ an indirect consistency supervision to ensure a temporally and stereo consistent output. To do this, we align the multi-view and temporal outputs of our network with respect to the current left viewpoint. For stereo rectification, this is done by employing a depth-based warp $F_{\text{StereoWarp}} : X^{(r)} \rightarrow X^{(l)}$ which maps the right viewpoint images onto the left viewpoint images. For temporal alignment we apply an optical flow warp $F_{\text{TempWarp}} : X^{(t+\epsilon)} \rightarrow X^{(t)}$ to determine a warped current image from a temporally adjacent frame. For details on the warping procedures please refer to the Supplemental Document.

In addition to temporal and stereo consistency losses during training, we employ a consistency block $G_{\text{C}} : Y \setminus S \rightarrow Y \setminus S$ as a downstream network after the translation block to achieve high quality consistent outputs. Directly applying the consistency losses to a single-stage network produces inferior results as the single-stage network struggles to remove both fine scattering effects, such as haze and coarse scattering effects such as snowflakes, in addition to other jitters such as sensor noise. We train G_{C} using the consistency losses while G_{T} focuses on the other losses previously described. See our ablation comparison in Section 5.1 for the benefits of our two-stage sequential network.

Putting everything together, we train the consistency block G_{C} to minimize the following consistency loss:

$$\mathcal{L}_{\text{Consistency}} = \mathcal{L}_{\text{Temp}} + \mathcal{L}_{\text{Stereo}}. \tag{10}$$

The temporal loss component is computed as

$$\mathcal{L}_{\text{Temp}} = (\mathcal{L}_1 + \mathcal{L}_{\text{perc}})(G_{\text{C}}(G_{\text{T}}(I_{\text{in}})), G_{\text{T}}(I'_{\text{in}})), \tag{11}$$

where

$$I'_{\text{in}} = M_{\text{Temp}}(F_{\text{TempWarp}}(I_{\text{in}}^{(t-1)}), F_{\text{TempWarp}}(I_{\text{in}}^{(t+1)})) \tag{12}$$

is the warped current input computed from temporally adjacent frames $I_{\text{in}}^{(t-1)} \in X^{(t-1)}$ and $I_{\text{in}}^{(t+1)} \in X^{(t+1)}$, and M_{Temp} is a visibility mask that merges the two warped temporally adjacent frames by recovering out-of-view pixels and occlusions, see Supplemental Document for details.

The stereo loss component is computed as

$$\mathcal{L}_{\text{Stereo}} = M_{\text{Stereo}} \odot \mathcal{L}_1(G_{\text{C}}(G_{\text{T}}(I_{\text{in}})), G_{\text{T}}(I'_{\text{in}})), \tag{13}$$

where $I'_{\text{in}} = F_{\text{StereoWarp}}(I_{\text{in}}^{(r)})$ is the warped right stereo image, and $M_{\text{Stereo}} = \exp(-\alpha \mathcal{L}_1(I_{\text{in}}, I'_{\text{in}}))$ is a visibility mask calculated from the warping error between the left input and the warped right input, and we empirically set $\alpha = 10$.

3.5. Generator Architecture

Our ZeroScatter generator network is illustrated in Figure 3. The architecture consists of two sequential components: a translation block G_{T} that eliminates scattering and performs domain transfer from a raw RGB adverse weather capture into a clear daytime scene, and a consistency block

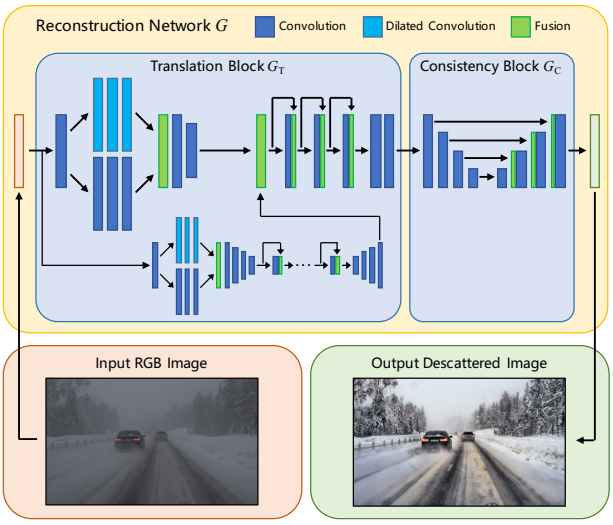

Figure 3: ZeroScatter generator network architecture. Our generator consists of a translation block that translates raw RGB captures into clear daytime scenes and a consistency block that removes erratic scattering media such as snowflakes.

G_{C} that further refines the translated output by removing stereo and temporal artifacts. Drawing inspiration from recent image translation networks [22, 49], our translation block architecture consists of two streams, one which operates at the full resolution and the other at a lower resolution. To allow the network to better recognize global features, we use an extended encoder with parallel feature extraction streams: one with 3×3 convolution layers to extract relative local context and one with 5×5 kernels with a dilation rate of 2 to allow the network to extract greater global context. Our consistency network consumes the output of the translation block and enforces consistency by removing distortions caused by adverse effects such as snowflakes and sensor noise. The architecture follows a U-Net [41] structure with 4 downsampling stages.

4. Unpaired Training Data and Setup

We train our model using a dataset from Bijelic et al. [5], who captured harsh weather scenarios in over $10\,000\,\text{km}$ of driving in northern Europe. Unlike previous works [5, 16, 7] we also leverage temporal sequences. The dataset we use consists of 12997 video sequences of length $0.5\,\text{s}$ and acquired at $20\,\text{Hz}$, resulting in a total of 120000 individual frames. The video sequences allow us to train for weather and sensor degradations that fluctuate over time, such as sensor noise and snowflakes. Please refer to the Supplemental Document for details on dataset distribution, split, and implementation details of the proposed approach.

We train ZeroScatter using Adam [25] with a learning rate of $5\text{e}{-5}$. After training, we implement the recon-

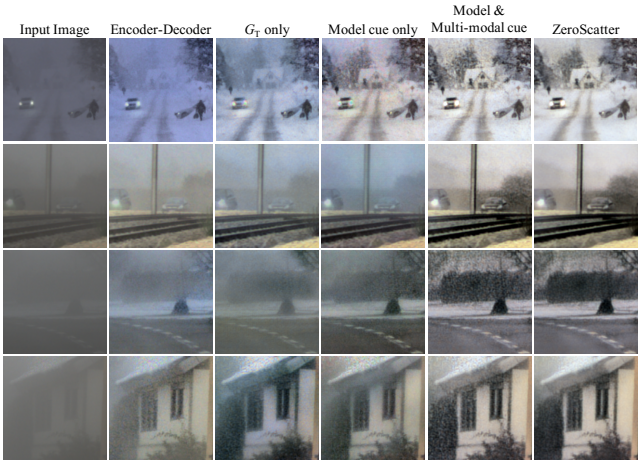

Figure 4: Ablation study qualitative results on unseen automotive RGB captures. Our sequential architecture and composite loss design enables enhanced contrast at long distances while minimizing snowflakes and sensor noise.

struction network for real-time inference at 20 FPS using fp16 precision for 768×1280 resolution images using an NVIDIA GeForce RTX 2080 Ti GPU. This allows for real-time vision and display applications in automotive systems.

5. Assessment

In this section, we validate the proposed method quantitatively and qualitatively. Our quantitative evaluation is performed on two test sets with paired clear reference data: fog chamber measurements, see Supplemental Material, that allow us to assess robustness to adverse weather in controlled fog scenarios, and a synthetic dataset where the scattering media is produced by F_{Syn}. For additional experimental details and qualitative results on the synthetic dataset, please refer to the Supplemental Document. Before reporting the performance of the proposed method compared to state-of-the-art image reconstruction approaches, we first validate model architecture choices in an ablation study.

5.1. Ablation Study

We conduct an ablation study to validate the effectiveness of our network architecture and the benefits from our novel combination of model-based, multi-modal, temporal, and multi-view supervision. The quantitative results for fog chamber measurements are shown in Table 1 together with the ablation configurations. Qualitative results are shown on unseen real-world data are shown in Figure 4.

We observed that relying solely on model-based training cues limits the performance on real-world data, as shown by the "Model cue only" configuration. The model outputs suffer from reduced contrast and this model is unable to adequately handle spurious sensor noise. "Model & Multi-Modal cue" illustrates how incorporating multi-modal indirect supervision improves performance with better removal of scattering components and increased contrast. Adding the consistency supervision grants us our proposed model ZeroScatter, which has the best descattering performance overall. Temporal and stereo consistency supervision enables effective removal of snowflakes and local fluctuations including sensor noise.

On the architecture side, our ablation study demonstrates the benefits of our sequential architecture. If we applied a standard encoder-decoder architecture [41] then minimal descattering is achieved, as shown by the "Encoder-Decoder" configuration. We attribute this to the limited receptive field which is unable to robustly recognize and remove adverse weather. Our translation block remedies this by using dilated convolutions to obtain a wider field of view and this results in better descattering as shown by the "G_{T} only" configuration. However, without the consistency block G_{C}, the translation block G_{T} falls into a local minimum where it avoids descattering. This is because the presence of some types of adverse weather such as haze can inadvertently increase temporal and stereo consistency by blurring out image details. As a result, our final network architecture that uses both G_{T} and G_{C} obtains the best performance across all variants compared in this work.

5.2. Controlled Experimental Evaluation

We compare our work against state-of-the-art image descattering networks [40, 35, 36, 8], image domain transfer networks [56, 20, 57], and traditional image refinement techniques [58]. Quantitative results are shown in Table 2 and qualitative results are reported in Figure 5. Please see the Supplemental Document for training details for these baselines and qualitative comparisons against CycleGAN, CyCADA, Bidirectional-FCN, and DehazeNet.

Traditional methods such as CLAHE [58] (shown as F_{Proc}) work well to stylize the image, but fail to remove severe fog and haze in the images. Image domain transfer networks, such as CycleGAN [57], CyCADA [20] , and ForkGAN [56] obtain better results, but are still unable to recover high-quality images from the degraded input images. Deep learning approaches designed for processing adverse weather such as EPDN [40] , PFF-Net [35] , DehazeNet [8], and Bidirectional-FCN [36] all perform well on the synthetic dataset, however, these methods are not robust to out of training distribution inputs and consequently fail to generalize to the real-world fog chamber measurements. We attribute this to the inability of these methods to incorporate real-world data into their training scheme. ZeroScatter remedies these limitations and as a result is able to achieve the highest image quality.

Table 1: Quantitative ablation study of different network structures and loss combinations on the fog chamber measurements.

	G_T	G_C	$\mathcal{L}_{\text{model}}$	$\mathcal{L}_{\text{multi-modal}}$	$\mathcal{L}_{\text{consistency}}$	1 - LPIPS	PSNR	SSIM
ZeroScatter	✓	✓	✓	✓	✓	**0.878**	**18.8**	**0.695**
Model & Multi-Modal cue	✓	-	✓	✓	-	0.875	18.5	0.685
Model cue only	✓	-	✓	-	-	0.870	17.4	0.658
G_T only	✓	-	✓	✓	✓	0.872	16.9	0.665
Encoder-Decoder [41]	-	-	✓	✓	✓	0.870	16.6	0.650

Table 2: Quantitative evaluation of image descattering methods. We evaluate descattering performance on the synthetic dataset and with controlled fog chamber measurements, see Supplemental Document. We also evaluate object detection performance after applying each descattering method.

	Fog Chamber Measurements			Synthetic Dataset			Object Detection		
	1 - LPIPS	PSNR	SSIM	1 - LPIPS	PSNR	SSIM	Easy mAP	Med mAP	Hard mAP
ZeroScatter	**0.878**	**18.8**	**0.695**	**0.873**	**19.2**	**0.750**	91.36	**90.11**	**82.71**
EPDN [40]	0.844	12.7	0.565	0.840	18.4	0.715	**91.60**	88.50	80.08
PFF-Net [35]	0.841	15.6	0.627	0.827	18.4	0.707	91.37	89.48	81.01
Bidirectional-FCN [36]	0.830	12.9	0.559	0.847	14.4	0.673	91.21	87.11	80.94
DehazeNet [8]	0.799	9.60	0.390	0.814	13.6	0.575	91.02	85.90	80.43
CyCADA [20]	0.819	13.1	0.506	0.808	14.3	0.572	90.97	88.18	80.46
CycleGAN [57]	0.779	11.7	0.505	0.794	13.7	0.578	90.99	85.56	80.15
ForkGAN [56]	0.718	11.6	0.374	0.720	13.8	0.383	87.81	84.53	78.71
F_{Proc} [58]	0.852	16.0	0.607	0.851	14.4	0.678	88.59	86.95	80.93
Input Image	0.812	14.3	0.517	0.753	13.4	0.492	90.50	86.50	80.91

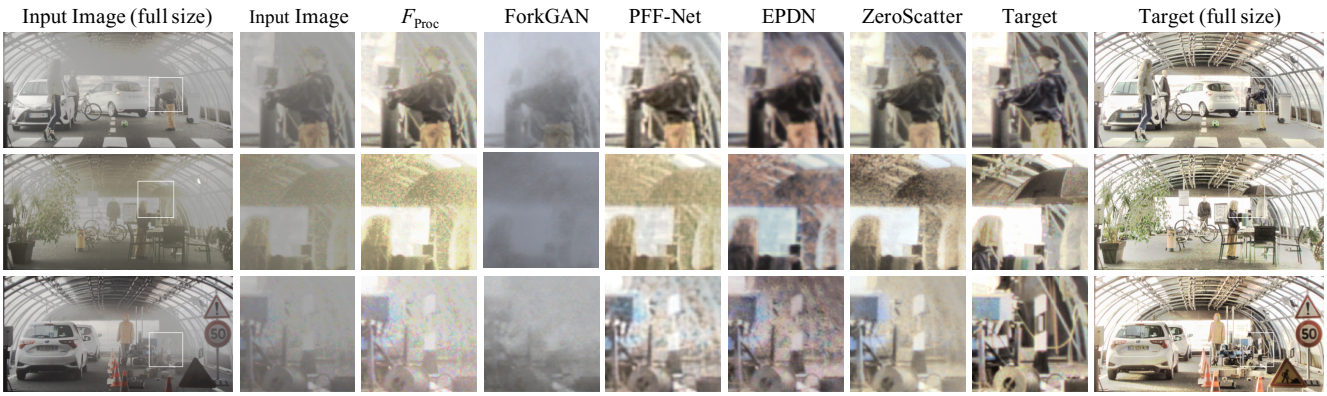

Figure 5: Qualitative performance comparison on controlled fog chamber measurements, see text. The proposed method significantly reduces scattering media present in the scene and most closely resembles the processed daytime target image.

5.3. In-the-Wild Experimental Evaluation

We showcase the performance of ZeroScatter and the baseline methods on real-world unseen measurements in Figures 1 and 6. Our high-quality reconstructions shown in these two figures as well as in the Supplemental Document validate the proposed method for diverse real-world scenes. Objects at long distances such as trees, houses, and cars, that have been obscured by adverse weather are revealed by the proposed method. Because the baseline methods do not utilize multi-modal information, their outputs suffer from residual noise and low contrast in the resulting images. Furthermore, without consistency supervision, their processed outputs accentuate sensor noise and fail to remove snowflakes.

5.4. Descattering for Object Detection

Furthermore we evaluate whether descattering improves 2D object detection in adverse weather. For this evaluation, we again use real-world adverse weather captures. Ground-truth annotations are performed manually, and difficulty levels are defined based on bounding box height, occlusion level and truncation following [13]. We employ SSD [33] object detectors with identical architecture that we finetune on the output of each descattering method for a fair comparison. Quantitative Average Precision (AP) scores

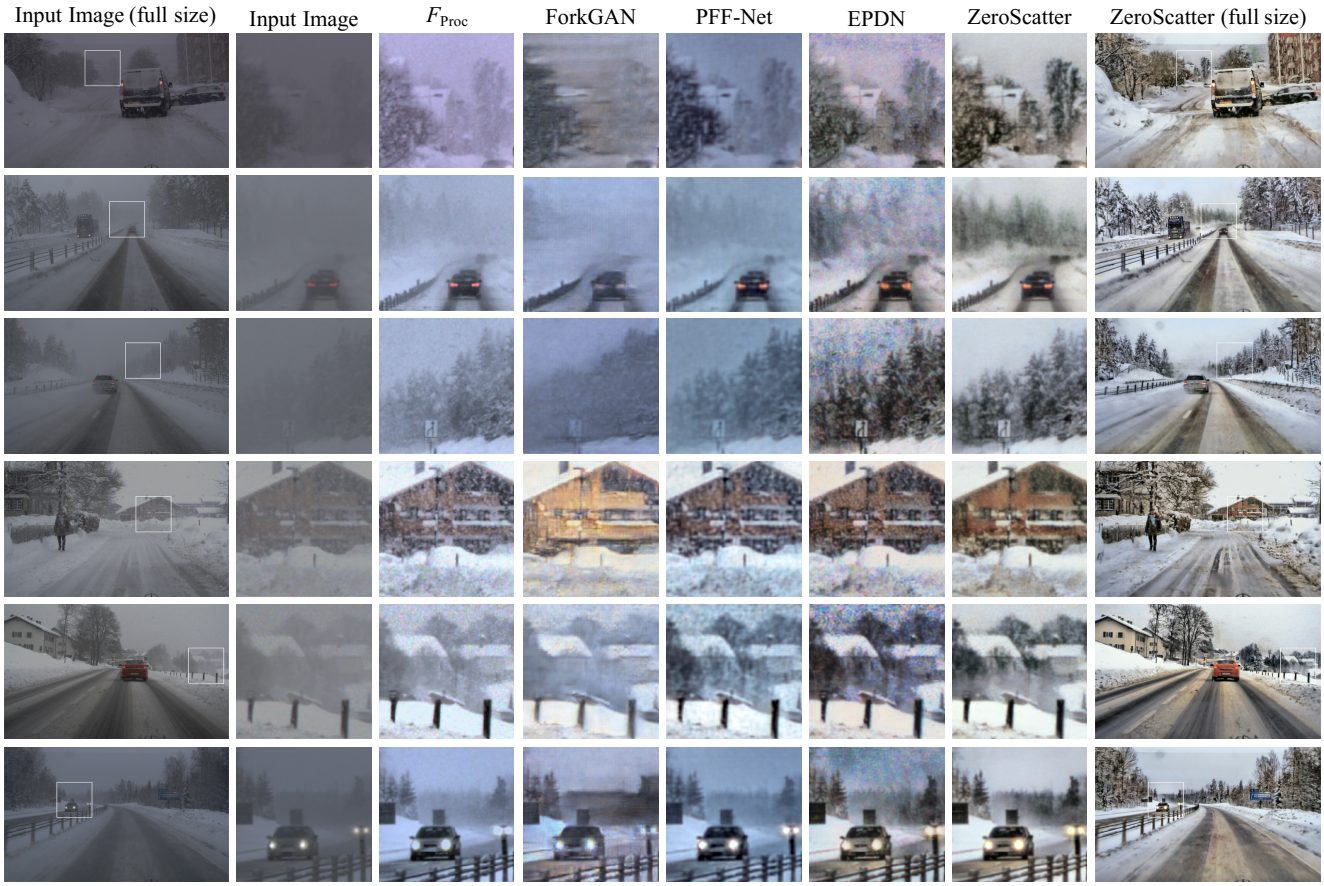

Figure 6: Real-world data qualitative comparisons. The proposed method significantly reduces scattering present in the scene and reveals object in long distance, such as the house and trees in the top two examples above. Compared to EPDN and PFF-Net, ZeroScatter is able to produce images with better contrast and less noise. ZeroScatter is able to remove snowflakes in the 3^{rd} and 4^{th} examples and sensor noise in the 5^{th} and 6^{th} examples.

are reported in Table 2, qualitative examples and training details are shown in the Supplemental Document. Among all descattering methods, ZeroScatter achieved the highest AP for the medium and hard settings while still maintaining near top performance on the easy setting. We attribute it to ZeroScatter's ability to remove scattering media in adverse conditions which in turn improves object detection through higher confidence detections and bounding box tightness, especially at long distances.

6. Conclusion

We introduce ZeroScatter, a novel domain transfer method that maps RGB images captured with strong scattering in adverse weather for removing scattering media from conventional RGB camera captures. We propose a combination of synthetic and real-world data by exploiting model-based, temporal, multi-view, multi-modal, and adversarial training cues. We validate the method by demonstrating that ZeroScatter significantly outperforms approaches both quantitatively in simulation and controlled experimental conditions, and on in-the-wild scenes. Moreover, we validate that removed scattering at long distances with ZeroScatter also enables state-of-the-art object detection results in harsh weather. In the future, we anticipate that ZeroScatter will not only allow human drivers and detectors to see in harsh weather but also assist human annotators for adverse weather scenes, overcoming the fundamental data bias in these scenarios. We envision the proposed training method as a basic building block for vision systems beyond imaging and object detection, especially for autonomous driving and robotics.

Acknowledgement

Felix Heide was supported by NSF CAREER Award (2047359) and a Sony Faculty Innovation Award. This research received funding from the Ministry for Economic Affairs and Energy within the project "VVM – Verification and Validation Methods for Automated Vehicles Level 4 and 5" (19A19002G). The authors would also like to acknowledge Klaus Dietmayer from University of Ulm.

References

[1] 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

[2] 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. 3

[3] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 867–8678. IEEE, 2018. 3

[4] W. S. Ashley, S. Strader, D. C. Dziubla, and A. Haberlie. Driving blind: Weather-related vision hazards and fatal motor vehicle crashes. *Bulletin of the American Meteorological Society*, 96(5):755–778, 06 2015. 1

[5] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 4, 5

[6] 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. 4

[7] M. Bijelic, P. Kysela, T. Gruber, W. Ritter, and K. Dietmayer. Recovering the unseen: Benchmarking the generalization of enhancement methods to real world data in heavy fog. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2019. 5

[8] 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. 2, 6, 7

[9] 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. 1

[10] 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. 2

[11] 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

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

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 1, 7

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 4

[15] Y. Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. 3, 4

[16] T. Gruber, M. Bijelic, F. Heide, W. Ritter, and K. Dietmayer. Pixel-accurate depth evaluation in realistic driving scenarios. In *International Conference on 3D Vision (3DV)*, pages 95–105, Sept. 2019. 1, 3, 5

[17] T. Gruber, F. Julca-Aguilar, M. Bijelic, and F. Heide. Gated2depth: Real-time dense lidar from gated images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1506–1516, 2019. 2, 4

[18] S. S. Halder, J.-F. Lalonde, and R. d. Charette. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10203–10212, 2019. 1, 3

[19] S. Hasirlioglu and A. Riener. A general approach for simulating rain effects on sensor data in real and virtual environments. *IEEE Transactions on Intelligent Vehicles*, pages 1–1, 2019. 3

[20] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998. PMLR, 2018. 6, 7

[21] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8014–8023. IEEE, 2019. 2

[22] A. Ignatov, L. V. Gool, and R. Timofte. Replacing mobile camera isp with a single deep learning model. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2275–2285, 2020. 5

[23] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 4

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE Computer Society, 2020. 1

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] H. Koschmieder. *Theorie der horizontalen Sichtweite*. 1924. 3

[27] I. Koshurinov. How startups are shaping the automotive radar of tomorrow. 1

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

[29] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[30] 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. 3

[31] K. Li, Y. Li, S. You, and N. Barnes. Photo-realistic simulation of road scene for data-driven methods in bad weather. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 491–500. IEEE, 2017. 3

[32] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 3

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016. 7

[34] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 2, 3

[35] K. Mei, A. Jiang, J. Li, and M. Wang. Progressive feature fusion network for realistic image dehazing. In *Asian Conference on Computer Vision (ACCV)*, 2018. 1, 2, 6, 7

[36] R. Mondal, S. Santra, and B. Chanda. Image dehazing by joint estimation of transmittance and airlight using bi-directional consistency loss minimized fcn. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1033–10338, 2018. 1, 2, 6, 7

[37] G. J. Oldenborgh, P. Yiou, and R. Vautard. On the roles of circulation and aerosols in the decline of mist and dense fog in europe over the last 30 years. *Atmospheric Chemistry and Physics*, 10:4597–4609, 2009. 1

[38] D. Petrov. Samsung working on a 250mp camera with sensor larger than the p40 pro. 1

[39] 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

[40] Y. Qu, Y. Chen, J. Huang, and Y. Xie. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8152–8160, 2019. 2, 6, 7

[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 5, 6, 7

[42] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 1

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

[44] C. Sakaridis, D. Dai, and L. Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7374–7383, 2019. 3

[45] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020. 1

[46] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmaoui. Improved visibility of road scene images under heterogeneous fog. In *2010 IEEE Intelligent Vehicles Symposium*, pages 478–485. IEEE, 2010. 3

[47] 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. 3

[48] H. Wang, Q. Xie, Q. Zhao, and D. Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[49] 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. IEEE Computer Society, 2018. 5

[50] W. Yang, R. T. Tan, S. Wang, and J. Liu. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[51] 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. 2

[52] 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

[53] H. Zhang, V. Sindagi, and V. M. Patel. Image De-raining Using a Conditional Generative Adversarial Network. *IEEE Transactions on Circuits and Systems for Video Technology*, abs/1701.05957, 2017. 2

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

[55] J. Zhao, J. Zhang, Z. Li, J.-N. Hwang, Y. Gao, Z. Fang, X. Jiang, and B. Huang. Dd-cyclegan: Unpaired image dehazing via double-discriminator cycle-consistent generative adversarial network. *Engineering Applications of Artificial Intelligence*, 82:263 – 271, 2019. 2

[56] Z. Zheng, Y. Wu, X. Han, and J. Shi. Forkgan: Seeing into the rainy night. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 2, 6, 7

[57] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 6, 7

[58] K. Zuiderveld. *Contrast Limited Adaptive Histogram Equalization*, page 474–485. Academic Press Professional, Inc., USA, 1994. 6, 7