# Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather (Supplemental Material)

Mario Bijelic[1,3] Tobias Gruber[1,3] Fahim Mannan[2] Florian Kraus[1,3] Werner Ritter[1] Klaus Dietmayer[3] Felix Heide[2,4]

[1]Mercedes-Benz AG [2]Algolux [3]Ulm University [4]Princeton University

This supplemental document provides additional information on the proposed dataset, additional method details, and additional results and comparisons.

## 1. Additional Dataset Details

The proposed dataset has been described in Section 3 of the main document. In this section, we present additional details, including the preselection and multimodal annotation processes, as well as the controlled weather capture setup.

We illustrate the diversity of our dataset in Figures 8, 9, 10, and 11.

### 1.1. Data Preselection Process

Before annotation, we preselect images as many of them are not relevant due to low scene variance, sensor failures, wipers, or no objects, see Figure 4. Frames with low scene variance usually contain scenes when waiting at a traffic light or following the same car at a long road. Sensor failures are caused either by technical problems or by sensors covered with snow or dirt.

Specifically, images at a frame rate of 0.1 Hz were uniformly sampled from the dataset, delivering a total of 17,799 images. These images were annotated with the scene weather and the semantic content (discard/dispensable/appropriate/very interesting). In addition, we tagged images with *interpolate* if interesting content was found close to a sampled frame. 44.66 % of the selected images were annotated with *discard* or *dispensable*. To increase the number of sequences with interesting semantic content, we additionally exported sequences that contained frames with *interpolate* annotations at a frame rate of 1 Hz, leading to additional 4,561 samples. After this process, since the resulting subset was biased towards good weather data, we additionally exported sequences in adverse weather with at least one *very interesting* tag at a frame rate of 1 Hz and obtained additional 6,444 frames. In total, 28,804 frames were annotated with scene weather and semantic content classification. From these frames, we filter out frames where the weather annotations changed quickly along the recordings – we consider these annotations as noisy or ambiguous. Finally, we chose the most interesting scenes annotated at least with *appropriate* scene content, leading to 12,000 annotated frames for bounding box labeling.

### 1.2. Data Annotation Process

The object annotation process is depicted in Figure 1. The process starts with a 3D bounding box annotation in a joint lidar and radar point-cloud. We annotated 3D boxes up to 80 m distance. Due to the sparsity of the point-cloud, the resulting 3D bounding boxes are visualized in the RGB camera frame for adjusting position and dimension. We use the RGB image visualizations to add 2D bounding boxes for objects missed during lidar/radar annotation. Furthermore, for each 3D box, additional 2D boxes were added by projecting the 3D boxes as a 2D box into the RGB camera frame and tightening them to fit the 2D object shape in the camera frame. If there are too few points for annotating a 3D box, visible objects are annotated with 2D boxes in both camera streams up to an object height of 30 pixels. Finally, the annotations are transferred into the gated frame.

We label objects according to the following types/classes: Pedestrian, Truck, Car, Cyclist, and DontCare. The total data distribution can be found in Figure 3. For each drawn object, we also included tags that indicate the occlusion level (no occlusion, $>10\,\%$, $>40\,\%$ and $>80\,\%$ occlusion) and the object visibility in each sensor stream. In addition, we include the following scene tags: image daytime (day, night, dawn), illumination (low dynamic range, high dynamic range, best computer vision weather, overall dark), weather (clear, rain, light fog, dense fog, snow), derivable path conditions (dry, wet, slushy, full snow coverage) and scene-setting (downtown, suburban, highway).

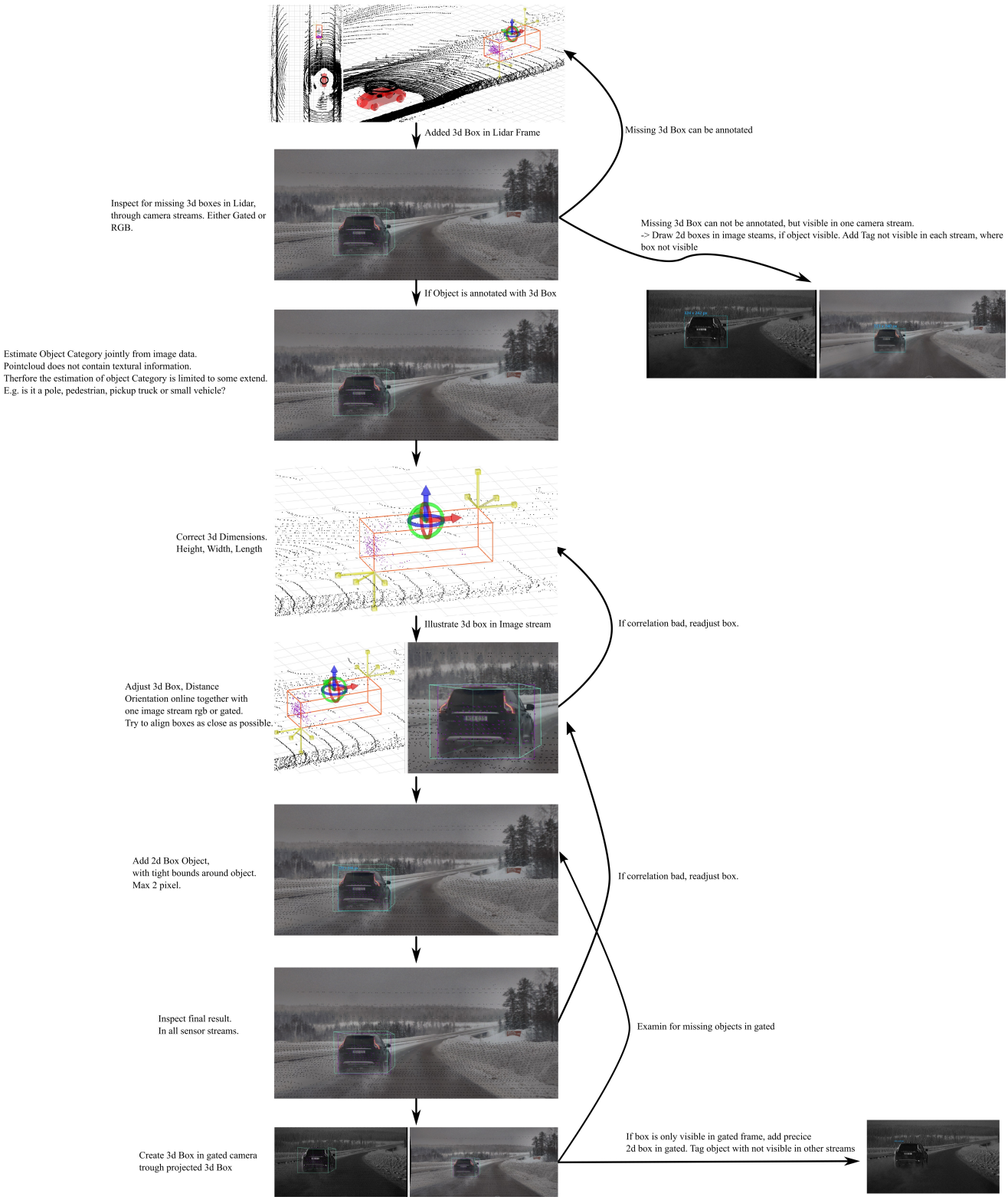


Figure 1: This figure illustrates the data annotation process, which is started in a combined lidar and radar frame using 3D bounding boxes. By visualizing the 3D boxes in the camera streams, they are corrected in order to fit the object dimension and location in every sensor stream. In addition, 2D bounding boxes are annotated in the RGB camera and the gated camera. For RGB data, human annotators tightened projected 3D boxes to the 2D object appearance. If there are any missing objects visible in a sensor stream, additional 2D boxes are added.

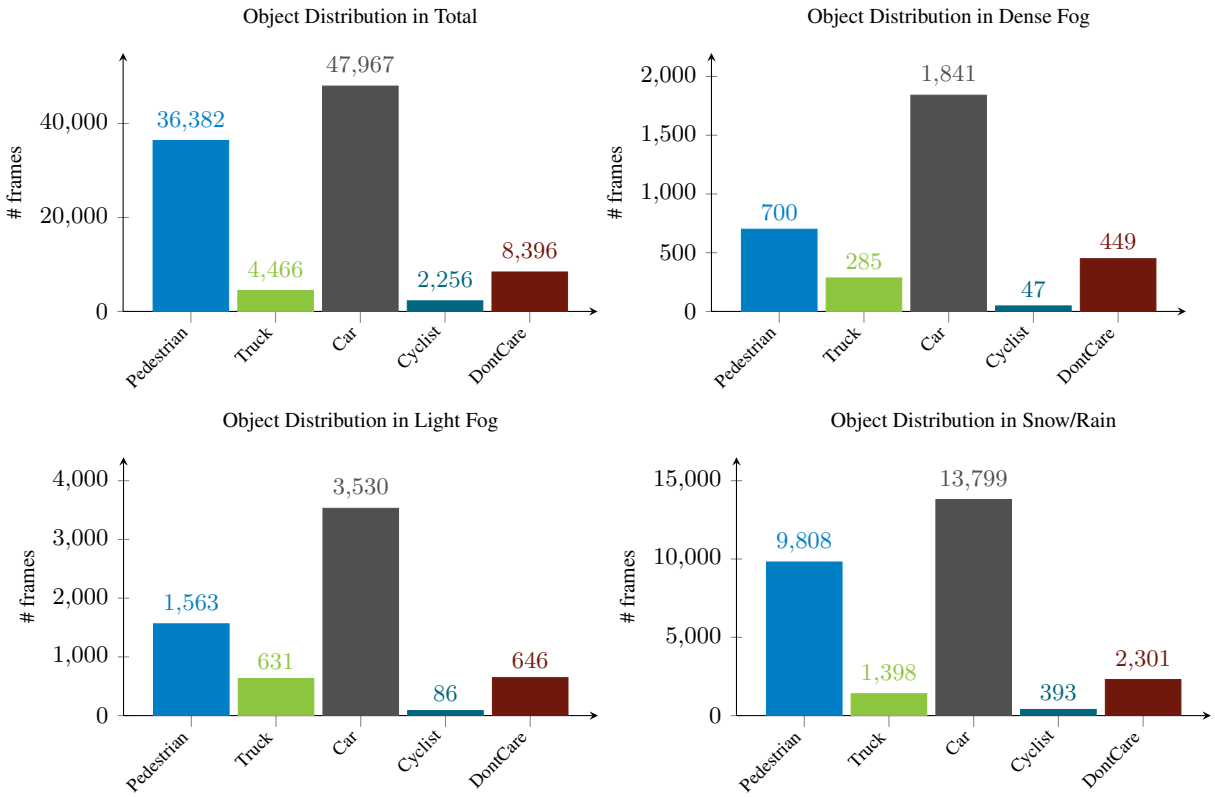


Figure 2: Object distribution in total and different weather conditions. Note how the object distribution changes from total to different weather condition.

### 1.3. Controlled Weather Dataset

We have recorded typical street scenarios in a weather chamber in controlled adverse weather conditions. We have designed six dynamic scenarios, as illustrated in Figure 6, namely crossing pedestrian, pedestrian on the sidewalk, cyclist on the street, crossing pedestrian and a cyclist on the street, oncoming car, crossing pedestrian and a cyclist on the street and an oncoming car. All of these dynamic scenarios have been performed in eight different lighting conditions and in three different fog densities. Lighting conditions can be changed by opening and closing the greenhouse part of the chamber, by varying the direction of the interacting car (oncoming/driving away) and by switching headlamps and streetlights on and off. In addition to three different fog levels at approximately 30 m, 40 m and 50 m visibility, some scenarios have been recorded in different rain intensities. However, these scenarios are limited due to a limited amount of available water. For the annotation process, we randomly selected 1,500 over all scenarios and lighting conditions. Since annotation of 3D bounding boxes is challenging due to the bad lidar performance in fog and rain, we annotated only 2D bounding boxes in the rgb and the gated camera. Figure 7 shows image example of the controlled weather data set.

## 2. Additional Qualitative Detection Results

We show additional qualitative detection results in Figures 13 and 16. In particular, we show a variety of distortions unseen during training, including those due to snowfall, spray, and incorrect exposure control. In all examples, the proposed method robustly handles the asymmetrical multimodal distortions, validating that the method generalizes well to challenging unseen conditions.

## 3. Additional Training Details

### 3.1. Anchor Boxes

To optimally represent the training data distribution, we use the K-means clustering algorithm with the intersection over union distance metric adopted from [14]. The training bounding boxes are clustered with respect to their width and height.

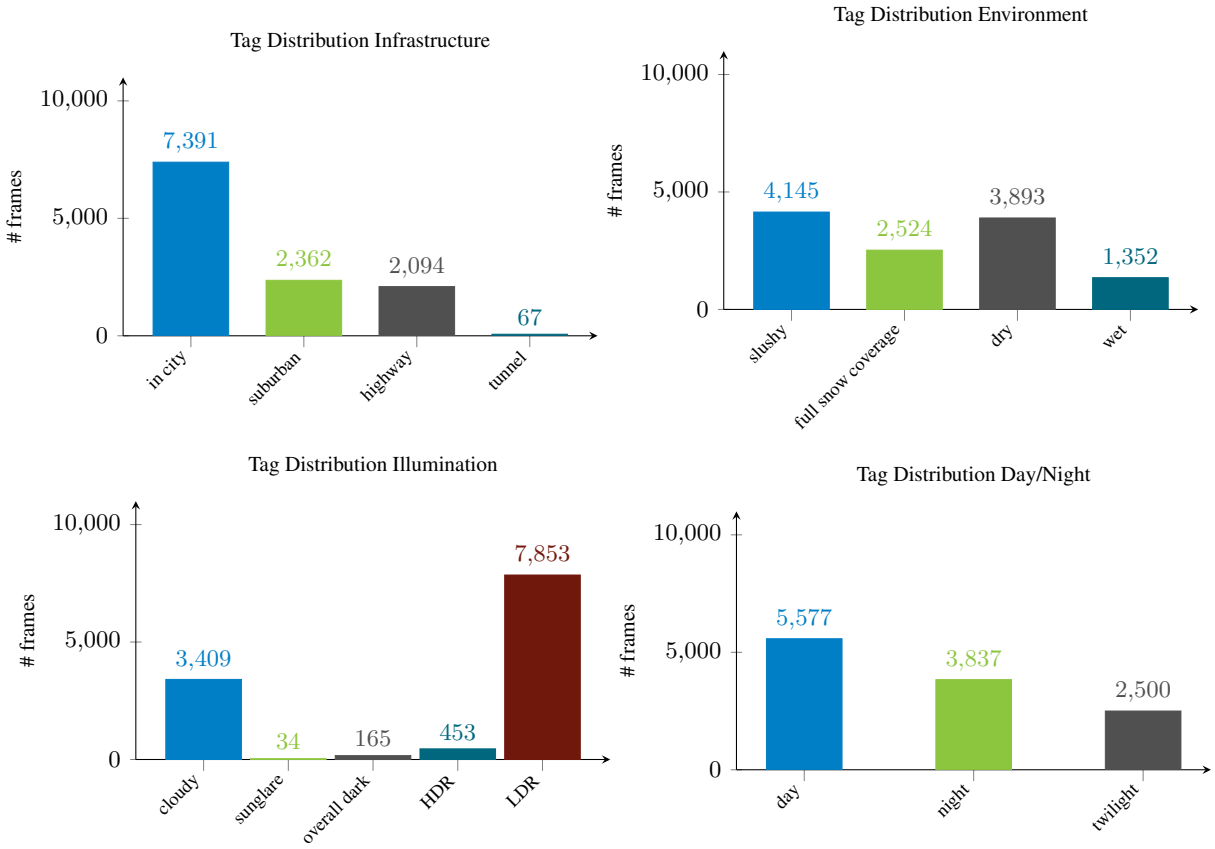


Figure 3: The tag distribution for the subsampled dataset is presented. Additional tags are provided for infrastructure, environment, illumination, and day/night condition.

| WEATHER | clear | | | light fog | | | dense fog | | | snow | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DIFFICULTY | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard |
| DEEP ENTROPY FUSION (THIS WORK) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -RADAR | -0.23% | -0.40% | -0.77% | -0.07% | -0.98% | -0.69% | -1.18% | -2.26% | -2.96% | -0.46% | -1.05% | -1.31% |
| -GATED | -0.54% | -1.94% | -2.91% | -1.28% | -2.26% | -2.68% | -3.80% | -6.26% | -8.65% | -1.32% | -4.13% | -3.98% |
| -LIDAR | -0.96% | -0.89% | -2.01% | -1.55% | -0.18% | -0.84% | 3.78% | 6.12% | 6.29% | -0.30% | -1.30% | -1.44% |

Table 1: Ablation study for dropping single sensors from the full fusion model.

A total number of 21 anchors were chosen. These anchor boxes are finally adjusted based on the resolution of each feature map, with earlier feature maps having smaller anchor boxes than the later ones.

### 3.2. Image Homography

To map the gated images to their corresponding RGB images we utilize a planar homography between both sensors. The planar homography is calibrated by a static scenario depicted in Figure 17. For each image, 170 corresponding points were labeled by a human annotator. The mapping was optimized using the RANSAC [5] optimization.

### 3.3. Contribution of Individual Sensors

To evaluate the per-sensor contribution we drop each sensor from the full fusion model. The results are described in Table 1. Please note the evaluation was based on the full RGB camera opening angle. Here only the RGB camera is covering the full view, while gated camera and radar provide smaller and lidar larger opening angles.

### 3.4. Lidar Input Representation

To test the best possible lidar input representation, we have tested different lidar input configurations. Per default, we have projected lidar points into the camera coordinate system and used distance, height, and intensity as inputs. This representation facilitates camera-lidar fusion in our model. The inputs are zero centered and scaled between $[-127.5, 127.5]$. In addition

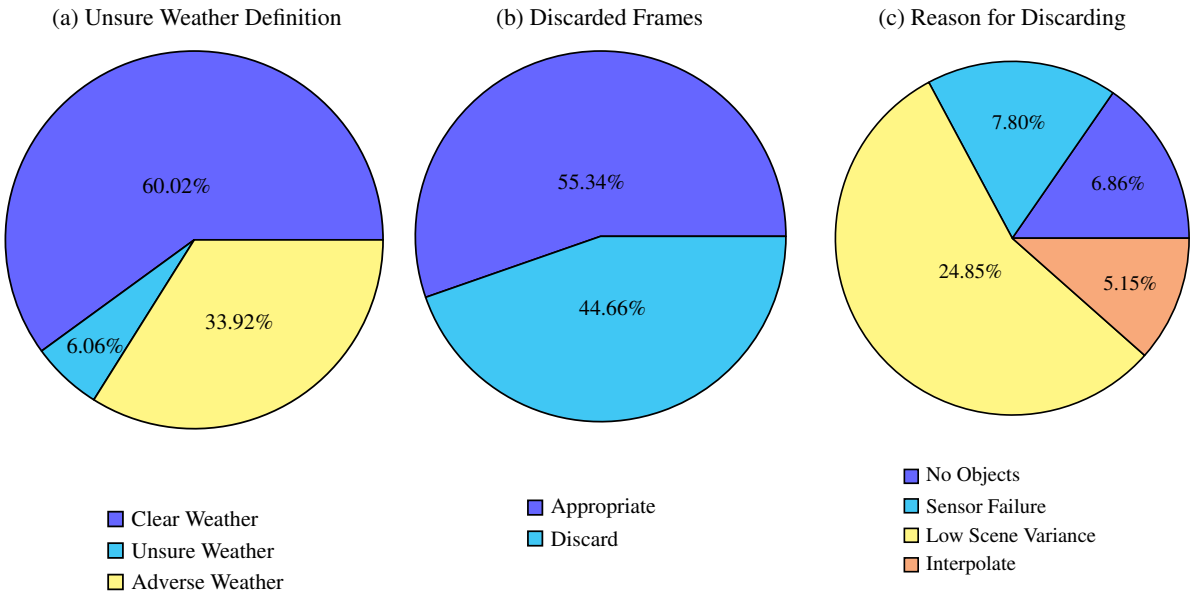


Figure 4: The recorded data has been preselected before annotation. Therefore, the whole dataset was exported at a frequency of 0.1 Hz. (a) To ensure well-defined weather conditions, we neglect samples where the weather annotations change quickly. (b) We neglect 44.66% of all selected frames due to low scene variance, sensor failure, or no objects within the scene. The distribution of the reasons for discarding is shown in (c). Frames annotated with *Interpolate* show a low scene variance, but the following frames could be interesting. Therefore, those sequences have been upsampled at a higher frame rate of 1 Hz and interesting frames have been selected.

| WEATHER | clear | | |
|---|---|---|---|
| DIFFICULTY | easy | mod. | hard |
| DEFAULT | 0 | 0 | 0 |
| UNIT SCALED | 0.58% | -0.75% | -0.95% |
| SHIFTED | -0.33% | -0.27% | -0.29% |

Table 2: Ablation study for different lidar data normalization approaches.

to this proposed representation, we test an shifted $[0, 255]$ input representation (shifted), and a normalization between [0,1] (unit scaled). In Table 2, we list quantitative results which indicate that all normalization approaches perform on-par.

### 3.5. Runtime Evaluations

Our unoptimized Tensor RT implementation runs at 22.6 Hz on our prototype inference platform using four Nvidia V100 GPUs, each GPU processing one sensor feature stack. This throughput performance is comparable to recent real-time camera-lidar detection methods, including AVOD operating at 12 Hz and PIXOR at 10.75 Hz, and PointFusion at 0.8 Hz. Note that our network is implemented in 32bit floating point precision and we leave integer quantization, commonly used in production deployments, for future work.

## 4. Additional Domain Adaption Results

For completeness, we validate the proposed approach against domain adaptation. Specifically, we compare against feature adaptation and dataset adaptation. For feature adaptation, the weights of a clear weather model are adapted to adverse weather conditions. For dataset adaptation, we learn an image to image mapping from clear weather conditions to our adverse weather scenes such that a clear weather training dataset can be converted to an adverse weather dataset with the same labels. This could reduce the amount of data collection campaigns in adverse weather scenes as the adverse weather style could be transferred to interesting clear weather scenes.

However, domain adaptation does *not model distortions* that can appear independently of the style. For example, fog can appear in summer or winter, which may potentially alter scene semantics and scene objects may be removed entirely due to fog. Moreover, these methods require large datasets of (unannotated) image data – which is not a trivial problem


Figure 5: Example image distortions in the proposed dataset. We annotate the distortion type as caption for each subfigure.

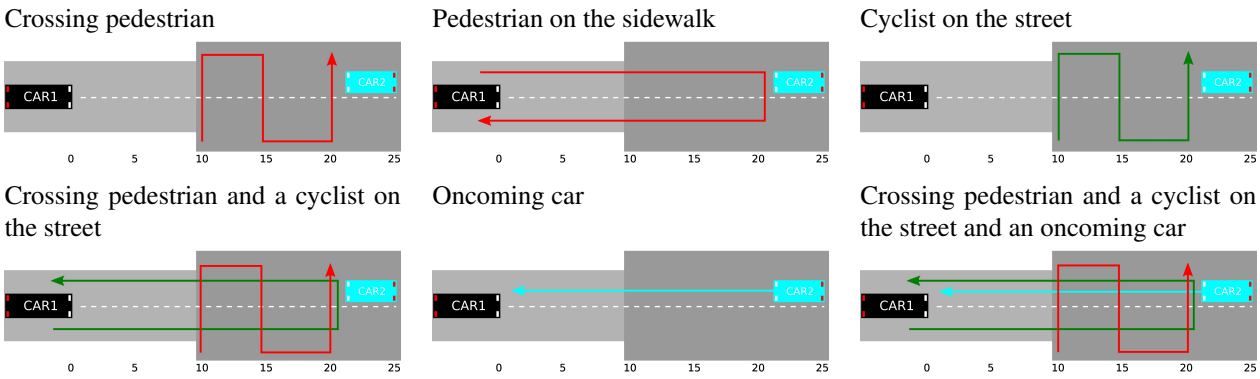


Figure 6: Illustration of six different scenarios for recordings under controlled weather conditions.

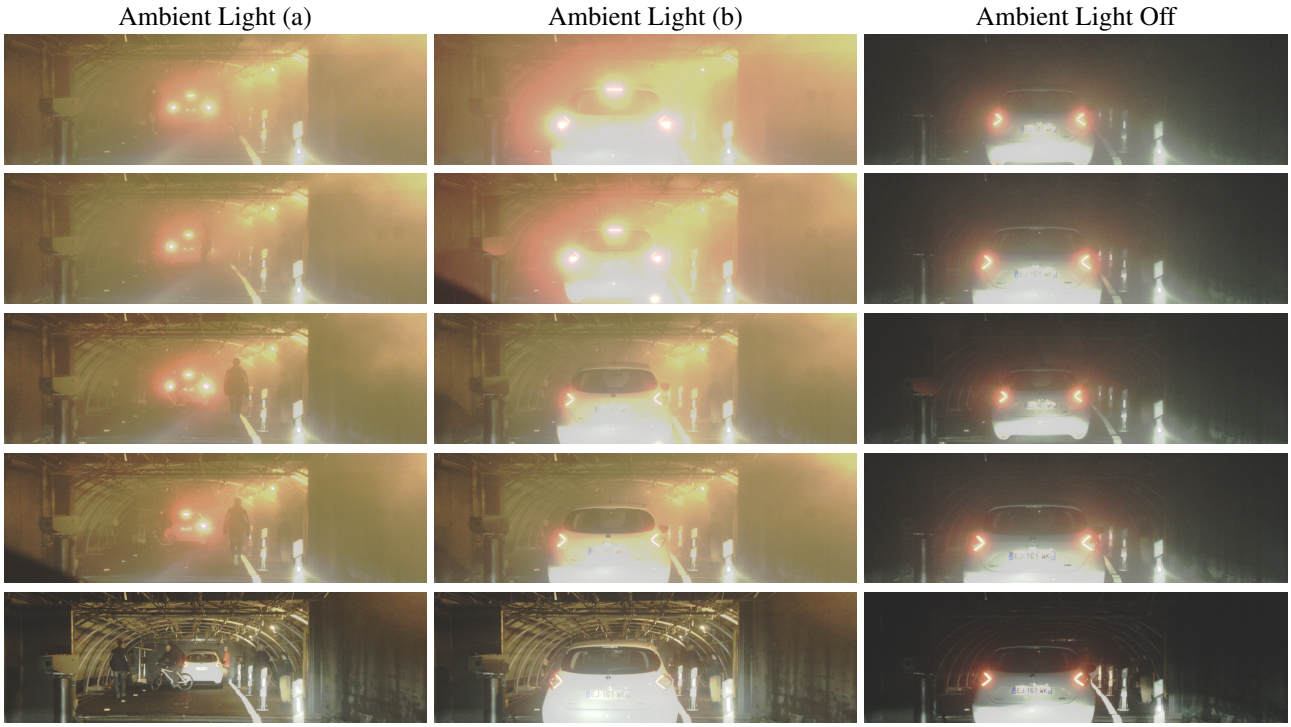


Figure 7: Example fog chamber images in two scenarios and two different illumination setting (a,b). The left column shows a dynamic scenario with moving pedestrians and cyclist. The middle and right columns show the same scenario with a moving car in front of the research car in different ambient light.

because adverse weather occurs rarely and changes quickly. Hence, domain adaptation does currently not offer a solution to overcome the bias in existing driving datasets in theory. We validate this with additional empirical experiments. Note that these experiments are *unfair in the sense that the compared domain-adapted methods have seen adverse weather data from the validation set*.

### 4.1. Additional Feature Adaptation Results

Adapting features from clear weather to adverse weather conditions could help reduce the amount of data required to enable the training of models in adverse weather conditions. Even though it requires unpaired pairs of adverse weather and clear samples, it does not need annotated 2D boxes. In our case, we use *Image-only SSD* as the pretrained model on clear weather conditions and cut the feature extraction part after the block4 of the VGG backbone. This feature extractor is

adapted to target images from sampled adverse weather conditions by training ADDA [16] with a batch size of 5, a learning rate of 0.0001, and an adversary network of three fully connected layers with 1024 hidden units, 2048 hidden units, and the adversarial discriminator output. Due to the size of the images, we could not fit more than five images per batch into the GPU. Even though the feature extractor has been adapted, the results decrease, as presented in Table 2 of the main manuscript. We assume that the decrease of performance is caused by different adverse weather appearances ranging from illumination changes to different environmental conditions, i.e. the snow coverage on the road as well as the disturbance patterns visible in the air (water droplet or snow flake). Nevertheless, also the changing number of road traffic participants changes the activation pattern. This can also be observed for an image to image mapping.

### 4.2. Additional Dataset Adaptation Results

Image to image mappings learn a mapping in-between stylistically different (e.g. summer/winter, or day/night) but semantically identical scenes. We use the the recent CyCADA [9] method to learn a mapping from clear large scaled KITTI data to our experimental scenes. Both dataset contain a large corpus of samples. The resulting model transfers data well between clear scenes, as shown in Figure 18 and in the first two rows of Figure 19. However, training a model from clear KITTI data to our full experimental training data set (including good and bad weather) fails, as shown in Figure 18 and in the last four rows of Figure 19. Qualitatively, the network only changes the image style, e.g. converts black roads to white roads or paints all green trees to gray/black. However, especially for creating fog and rain realistically, the scene semantics must be changed. This leads to substantial artifacts in the resulting domain-transformed adverse weather images.

To reduce domain differences, we map our clear images from our proposed dataset to our adverse weather scenes. In Figure 20, we can observe that the network does not learn to map difficult disturbances. The network only applies a style transfer from clear weather to winter scenes. The suggested simplification turned out to be a probable drawback as the discriminative force of the GAN discriminator may be too weak for this task and learns only simple features as "is there a snow-covered road?". Besides, it can be observed that different illumination conditions are not correctly transferred to adverse weather scenes.

### 4.3. Additional Semantic Adaptation Results

To test the semantic adaptation capabilities, we adapt from clear weather RGB images in Cityscapes [2] to our adverse weather RGB images utilizing DADA [17]. Please note that an adaptation was successful but the task implies more than changes in viewing angle and textures. In adverse weather, the semantics can change fundamentally, such as snow-covered roads leading to vanishing borders between roads and sidewalks or different degrees of degenerated information in adverse weather recordings, e.g. completely blackout areas in nighttime driving or halos through oncoming cars blacking out the vehicle contours and sky information mixing with objects in foggy driving. Representative examples are presented in Figure 21.

## 5. Additional Simulation Results

In this Section, we provide additional results adding simulated data to the clear training data as an alternative approach to tackle the rare harsh weather conditions. As fog measurements are rare in the proposed dataset, we opt to model measurement distortions introduced by fog, which is the weather condition where established fusion techniques drop the most. We address these foggy conditions through the recently proposed data augmentation techniques by Sakaridis et al. [15]. Note that the proposed forward models also help to explain observed distortions.

### 5.1. Intensity Imaging in Fog

In foggy conditions, light is scattered by the suspended water droplets before falling into the image sensor. This scattering phenomenon has two primary effects. First, the chief ray is attenuated before falling into the sensor, and second, a signal floor of scattering light is present. Both effects reduce contrast, and the observed foggy image can be modeled by [15]

$$I_{\text{foggy}}(x) = t(x) I_{\text{clear}}(x) + (1 - t(x))L, \tag{1}$$

where $I_{\text{clear}}$ is the latent clear image, depth-dependent transmissivity $t$, the global ambient component $L$. The transmission coefficient is $t(x) = \exp(-\beta d(x))$, where $\beta$ is the fog density (or attenuation) coefficient, and $d(x)$ is the scene depth at a pixel. The exponentially decaying model is consistent with controlled fog-chamber measurements, which we have validated in Figure 25. To this end, we capture measurements of the same targets as in Figure 23 at varying distances. We averaged the intensities on the target, which have been labeled through human annotators, and the corresponding depth has been

precisely measured by hand. Each curve corresponds to a different reflective target with values from 5 %, 50 % and 90 %. Interestingly, a peak at 8 m can be observed. This peak can be explained through the scene illumination. All surrounding illumination sources were turned off. The scene was only illuminated through the vehicle's high beams. From 9 m onwards, the targets were fully within the headlamp illumination cone. Given the calibrated intensity curves, it was possible to fit the model from Eq. 1 for a distance $d > 9\,\mathrm{m}$, i.e.,

$$I(d) = I_{\text{clear}} \exp(-\beta d) + L(1 - \exp(-\beta d)). \tag{2}$$

$I(d)$ denotes the average intensity at a distance $d$, $\beta$ corresponds to the measured fog density, $I_{\text{clear}}$ describes to the target baseline reflectivity, and $L$ is the airlight.

### 5.2. Pulsed Lidar in Fog

Scanning lidar systems actively illuminate the scene with focused high peak-power pulses, simplifying the measurement model to

$$L_{\text{foggy}}(x) = t(x) L_{\text{clear}}(x), \tag{3}$$

where $L_{\text{clear}}(x)$ is the emitted laser beam intensity, measured for a given repetition rate, and $L_{\text{foggy}}(x)$ is the received laser intensity. Heed that we assume that the beam divergence is not affected by the fog. In this model, a returned pulse echo is always registered as long as the received laser intensity is larger than the effective noise floor. However, severe back-scatter from fog may lead to direct back-scatter from points within the scattering fog volume, which is quantified by the transmissivity $t(x)$ from Eq. (1). Modern scanning lidar systems implement adaptive laser gain $g$ to increase the signal for a given noise floor, see also [7], yielding the maximum distance of

$$d_{\text{max}} = \ln\left(\frac{n}{L_{\text{clear}} + g}\right) / (2\beta), \tag{4}$$

with $n$ being the detectable noise floor. The detectable distance decreases logarithmically with the sum of the reciprocal of the received laser intensity from Eq. (3) and gain. Hence in fog, lidar measurements do not suffer only from loss in peak intensity but also from back-scattering, which results in a peak-shift inside the fog volume, and thus all information on the target scene point is lost. Figure 3 in the main paper shows a camera-lidar measurement in dense fog. We use Algorithm 1 to simulate lidar measurements that are distorted by fog. This forward model is based on Eq. (3), Eq. (4), and on additional fog chamber measurements validating the chosen hyperparameters. Note that the beam divergence in fog has been neglected, and we assume that a constant additive gain $g$ and noise-floor $n$ accurately describe the lidar depth measurement process. We also assume here that the intensities from detected objects decay exponentially following the attenuation model in Eq. (3).

We calibrate the laser scanner's gain and noise floor to achieve realistic fog disturbances in different fog densities $\beta$. Based on our quantitative fog-chamber measurements, qualitative real world measurements and the maximal viewing distance for the Velodyne HDL64 S2 lidar used in the of KITTI dataset [6], we calibrate the gain and noise-floor for the Velodyne S2 to be $g = 0.35$ and $n = 0.05$ and for the Velodyne HDLS3D to be $g = 0.45$ and $n = 0.04$. The maximal measured viewing distances are visualized in Figure 23 and Figure 24. We evaluate our model in two different fog types (advection and radiation fog defined in [4, 1]). The maximum viewing distance is the distance where all points on the targets are lost. This distance is estimated using calibrated diffusive Zenith Polymer targets with a reflectivity of 90 %, and moving the targets from the test vehicle's position to the farthest distance until they become invisible.

In addition to modeling the incident intensity, we also model distance distortions of the backscattered points to match typical bad weather performance. To match a backscattered point's distance observed in our prototype system, we assume that a point is backscattered by fog if the object distance is larger than a maximal distance achievable for a given fog level, which we calibrate. The point is then either lost with the attenuation probability $p_{\text{lost}} = \exp(-\beta \cdot d_{\text{max}})$, backscattered by the fog if the intensity in fog is half as large as the emitted intensity. Therefore, the backscattered point is at distance $d_{\text{new}} = -\ln(0.5)/\beta$. Furthermore, we model random distortions with with probability $p_{\text{random}}$ and a threshold of $\text{random}_{\text{thresh}} = 0.1$.

To achieve typical point cloud "wobbling" induced by fog inhomogeneities, which can also be observed at low ambient light induced through exhaust gases [8], we simulate this behavior using the heuristic model presented in Algorithm 2. Specifically, we add a set of sinusoids to the base fog density $\beta$ along the azimuth angle and height. The frequencies are chosen randomly in an interval of [0,2] for the azimuth direction and [0,5] for the height. Over time all functions are updated, creating characteristic point cloud wobbling effects due to inhomogeneous fog.

**Algorithm 1** Pseudo code for lidar image formation in fog, time $= t$, pointcloud $= Pc$, visibility $= \beta$, probabilities $= p$, distances $= d$, disturbed pointcloud $= Pc_{\text{dist}}$, gain $= g$, noiselevel $= n$

**procedure** HAZE_POINTCLOUD($t, Pc, b$)
  $Pc_{\text{dist}} = [\ ]$
  **for** p **in** Pc **do** ▷ calculate current distance
    $d = \sqrt{p.x^2 + p.y^2 + p.z^2}$ ▷ disturb $\beta$ periodically
    $\beta_{\text{dist}} = \text{func}_{\text{dist}}(t, \beta, p)$ ▷ calculate max. $d$ with $g = 0.35$ and $n = 0.05$
    $d_{\text{max}} = \ln\left(\frac{0.05}{p.I+0.35}\right)/2/b_{\text{dist}}$ ▷ calculate $p$ for loosing point
    $p_{\text{lost}} = 1 - \exp(-\beta_{\text{dist}} \cdot d_{\text{max}})$ ▷ calculate $d$ for self reflection
    $d_{\text{new}} = -\ln(0.5)/\beta_{\text{dist}}$
    **if** $d_{\text{max}} < d$ **then** ▷ point is scattered
      $p_{\text{lost}} = \exp(-\beta_{\text{dist}} \cdot d_{\text{max}})$
      **if** $p_{\text{lost}} >$ **then** ▷ point is lost, do nothing
        **pass**
      **else if** $d_{\text{new}} < d$ **then** ▷ reflection from fog
        $s = d_{\text{new}}/d$
        $I = \exp(-\beta_{\text{dist}} \cdot d_{\text{new}})$
        $Pc_{\text{dist}}$.append$([p.x \cdot s, p.y \cdot s, p.z \cdot s, I])$
      **else if** $p_{\text{random}} > \text{random}_{\text{thresh}}$ **then**
        $d_{\text{rand}} =$ append.append$(d_{\text{max}})$
        $I = p.I \cdot \exp(-\beta_{\text{dist}} \cdot d_{\text{rand}})$
        $s = d_{\text{rand}}/d$
        $Pc_{\text{dist}}$.append $([p.x \cdot s, p.y \cdot s, p.z \cdot s, I])$
      **end if**
    **else** ▷ only point intensity is attenuated
      $I = p.I \cdot \exp(-\beta_{\text{dist}} \cdot d)$
      $Pc_{\text{dist}}$.append $([p.x, p.y, p.Z, I])$
    **end if**
  **end for**
  **return** $Pc_{dist}$ ▷ returns disturbed pointcloud
**end procedure**

### 5.3. Gated Imaging in Fog

We model gated intensity imaging according to the intensity imaging model applied to only the narrow gating range considered. As most backscatter is eliminated with the first gate, the transmission function $\beta$ reduces local contrast significantly less compared to conventional intensity imaging.

### 5.4. Radar Measurements in Fog

We model radar as being unaffected by foggy conditions.

### 5.5. Additional Simulation-Augmented Detection Results

Here, we add additional synthetic fog images to the training corpus. Specifically, we extract dense stereo depth for our clear weather data and augment our existing dataset with synthetic fog images. We use a clear-data pretrained model and finetune for 10 epochs and sample the fog density uniform in discrete steps 0.0, 0.005, 0.1, 0.02, 0.04, 0.06, 0.08. For image-only models, we can observe a slight improvement in Table 3. For lidar-only models, we can only observe a slight improvement in dense fog conditions, while other results drop. This behavior is due to the sparsity of lidar point clouds and especially due to the loss of many points during augmentation. Therefore, the network learns that objects can also exist in regions without points, which leads to an increased number of false positives and a loss in performance. Furthermore, the lidar model is only validated for dense foggy conditions up to visibilities of 50 m, while light fog spans up to 1000 m. Applying the joint augmentation method to our *Deep Entropy Fusion* model leads to higher results in dense foggy conditions, while other categories drop. We attribute this behavior to imbalanced sensor stream tradeoffs that the network is making during

**Algorithm 2** Pseudo code for FUNC_DIST$(t, \beta, p)$, time $= t$, Point $= p$. Based on periodically sampled $\sin$ functions. A random set of functions is saved in params containing a frequency for angle and height as well as the phase shift.

**procedure** FUNC_DIST$(t, \beta, p)$
  **if** params exist **then**
    params = loadParams ▷ load existing params
  **else**
    params = sampleParams ▷ sample new params
  **end if**
  a = calculateAngle(p)
  h = calculateHeight(p)
  **for** curve **in** params **do**
    betadiff = 0
    betadiff += curve.amplitudeAngle$\times \sin$(curve.frequencyAngle$\times$a+curve.phaseAnlge)/curve.frequency)
    betadiff += curve.amplitudeAngle$\times \sin$(curve.frequencyAngle$\times$a+curve.frequencyHeight$\times$h+curve.phaseHeight)
    $\beta$ += abs(betadiff)
  **end for**
  **if** p.isLastPoint **then** ▷ if last point has been disturbed, update params for next pointcloud in timestep
    curve.phaseAnlge += curve.frequencyTimeAngle$\times$t.Timestep
    curve.phaseHeight += curve.frequencyTimeHeight$\times$t.Timestep ▷ propagate params in time
  **end if**
  **return** $\beta$ ▷ returns periodically augmented beta
**end procedure**

| WEATHER | clear | | | light fog | | | dense fog | | | snow | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DIFFICULTY | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard |
| **DEEP ENTROPY FUSION (THIS WORK) + AUG** | 89.47 | 83.85 | 78.71 | 90.12 | 86.30 | 83.32 | 88.41 | **84.69** | **81.15** | 88.86 | 82.59 | 77.33 |
| **DEEP ENTROPY FUSION (THIS WORK)** | 89.84 | **85.57** | **79.46** | **90.54** | **87.99** | **84.90** | 87.68 | 81.49 | 76.69 | **88.99** | **83.71** | **77.85** |
| IMAGE-ONLY SSD + AUG [15] | 85.59 | 75.25 | 67.90 | 88.24 | 78.12 | 73.73 | **88.82** | 79.00 | 76.50 | 85.13 | 73.81 | 66.84 |
| IMAGE-ONLY SSD | 85.43 | 75.75 | 67.79 | 87.76 | 78.52 | 70.43 | 87.89 | 78.25 | 74.96 | 84.33 | 74.38 | 67.01 |
| LIDAR-ONLY SSD + AUG | 64.74 | 49.97 | 46.06 | 59.44 | 45.55 | 42.78 | 31.54 | 26.06 | 25.09 | 58.44 | 44.31 | 41.24 |
| LIDAR-ONLY SSD | 73.46 | 57.32 | 54.62 | 68.43 | 54.82 | 51.91 | 28.98 | 25.24 | 24.56 | 67.50 | 52.26 | 46.83 |

Table 3: The influence of additional synthetically distorted data added to the training dataset (AUG), generated by simulating foggy observations from clean captured data. Quantitative detection AP on real unseen weather-affected data from dataset split across weather and difficulties easy/moderate/hard following [6]. The augmented model uses clear pretrained models from Table 5 in the main paper. The models were finetuned for 10 epochs on the augmented data sampling uniformly across different fog densities from $\beta = 0.0\ \frac{1}{\text{m}}$ to $\beta = 0.08\ \frac{1}{\text{m}}$ dimension.

training. If the network has learned through an extensive augmentation that lidar data is unreliable, it shifts towards other sensors sacrificing the performance on other weather types where lidar data performs well. We also observed this behavior with added adverse weather samples.

| WEATHER | clear | | | light fog | | | dense fog | | | snow | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DIFFICULTY | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard | easy | mod. | hard |
| DEEP ENTROPY FUSION (THIS WORK) | **89.84** | **85.57** | **79.46** | **90.54** | **87.99** | **84.90** | 87.68 | **81.49** | 76.69 | 88.99 | **83.71** | **77.85** |
| IMAGE-ONLY SSD + FOG REMOVAL [18] | 87.21 | 76.95 | 68.85 | 88.44 | 78.80 | 70.74 | **88.03** | 78.98 | **76.73** | 83.76 | 69.29 | 66.79 |
| IMAGE-ONLY SSD | 85.43 | 75.75 | 67.79 | 87.76 | 78.52 | 70.43 | 87.89 | 78.25 | 74.96 | 84.33 | 74.38 | 67.01 |

Table 4: The influence of additional image reconstruction as a preprocessing step on real unseen weather-affected data from dataset split across weather and difficulties easy/moderate/hard following [6]. The proposed model is trained solely on clean data without weather distortions. The best model is highlighted in bold. Here we compare our *Deep Entropy Fusion* network to *Single Image* and a *Single Image* with image preprocessing step removing haze.

## 6. Additional Image-Only Detection Results

As a further baseline, we evaluate existing fog removal methods and propose variations of popular image-to-image translation techniques for fog-removal prior to object detection. Specifically, we adopt the recent Pix2PixHD [18] method, which is a generative adversarial network that transforms images between domains while preserving the scene semantics. We extend this model with color and brightness jitter data augmentation (Pix2PixHD-CJ), and the K-matrix estimation proposed in AODNet [11] (Pix2Pix2HD-AOD). We trained these models using the simulated fog dataset from the previous section, which provides corresponding image pairs. Figure 26 demonstrates that the proposed augmentation scheme suffers from fewer artifacts and achieves relatively stable fog removal and contrast enhancement.

In particular, we propose to add color jitter during training to improve the robustness to measured data. We use PyTorch [12] transforms with the ColorJitter implementation and the following parameters: brightness $0.125$, saturation $0.5$, hue $0.2$ and contrast $0.5$. The undisturbed target images were kept as they are. Hence, during training, the network also learns to revert the colors if they are disturbed, which is beneficial as data augmentation step.

Qualitative results on real data are shown in Figure 26. Note that the vanilla AODNet [11] and Pix2PixHD with this modification, which we dub Pix2PixHD AOD, do not generalize to real data Figure 26.

Although we evaluate image enhancement methods in this work, note that this signal enhancement stage requires a suitable training dataset containing a large amount of clear and disturbed image patches. Those can be effectively created using the previously described simulation techniques without the need of real adverse weather data for training. But this also limits the performance to disturbance types, which can be simulated. Consequently, the image enhancement methods only improve detection results on real data slightly Table 4, because in the real world dataset fog often appears in combination with other disturbances as snowfall or a dirty windshield, see Figure 27. As our image-only fog removal model is only trained on synthetic data, it does not generalize to those cases. Currently, to our best knowledge, there exist no suitable simulation frameworks that faithfully model these distortions such that the resulting model generalizes to in-the-wild captures.

## References


[1] 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. In *Atmospheric Research*, 2004. 9

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 8

[3] A. Dudhane and S. Murala. C2msnet: A novel approach for single image haze removal. Jan 2018. 30

[4] P. Duthon, F. Bernardin, F. Chausse, and M. Colomb. Methodology used to evaluate computer vision algorithms in adverse weather conditions. *Transportation Research Procedia*, 14:2178–2187, 2016. 9

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 4

[6] 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*, pages 3354–3361, 2012. 9, 11, 12

[7] D. Hall. High definition lidar system, 2007. 9

[8] S. Hasirlioglu, A. Riener, W. Ruber, and P. Wintersberger. Effects of exhaust gases on laser scanner data quality at low ambient temperatures. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1708–1713, June 2017. 9

[9] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2017. 8, 23, 24, 25

[10] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–8. IEEE, 2018. 19, 20

[11] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng. Aod-net: All-in-one dehazing network. In *International Conference on Computer Vision (ICCV)*, pages 4780–4788, Oct 2017. 12, 30

[12] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems*, 2017. 12

[13] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018. 19, 20

[14] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2017. 3

[15] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, pages 1–20, 2018. 8, 11

[16] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. 8

[17] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 8, 26

[18] 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*, pages 8798–8807, 2018. 12, 30

Figure 8: Random sampled RGB images from the adverse weather dataset to illustrate the diversity of scenes, illumination and weather conditions. The images are cropped from 1920×1024 to 1248×384.

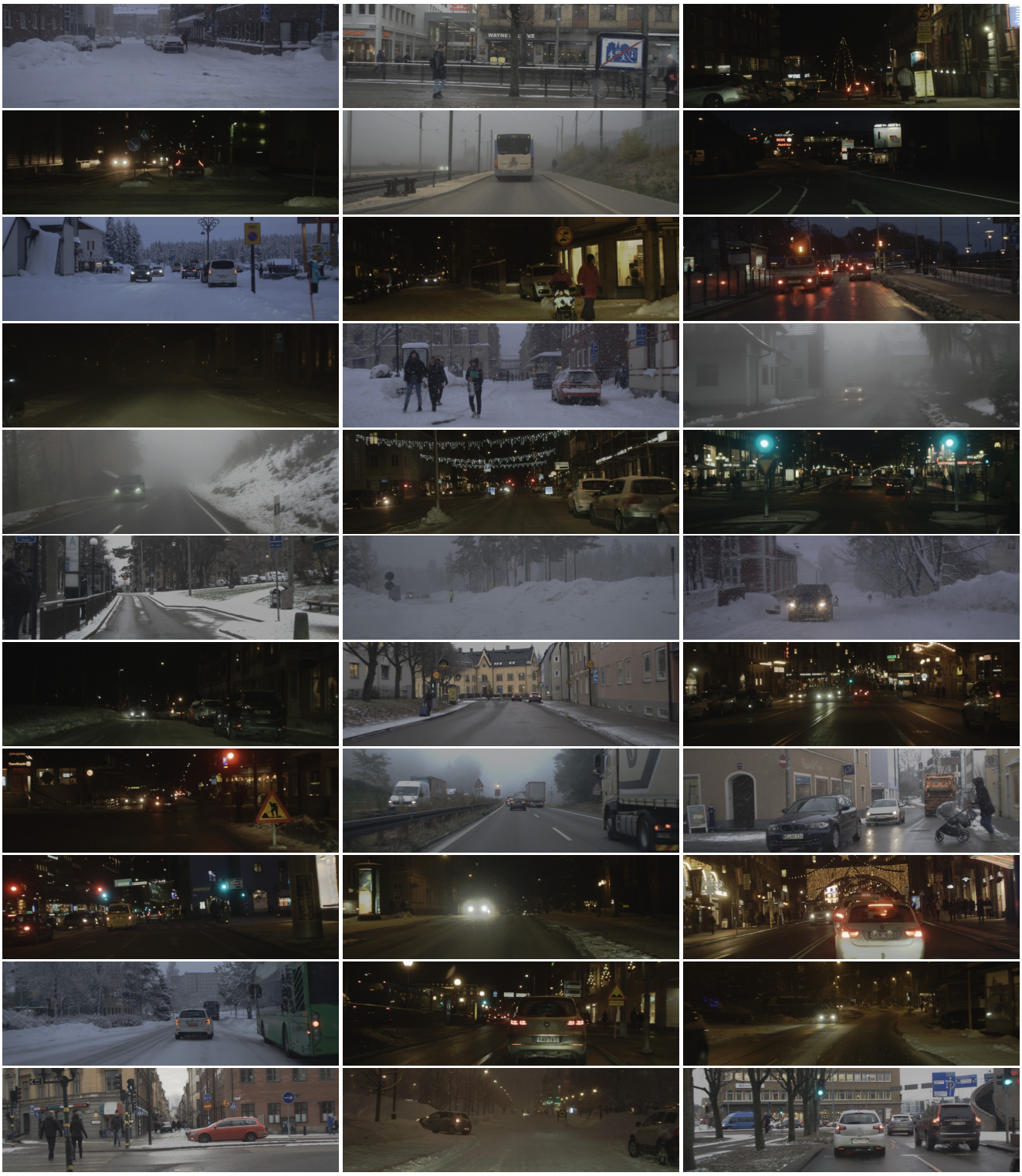

Figure 9: Random sampled RGB images from the adverse weather dataset to illustrate the diversity of scenes, illumination and weather conditions. The images are cropped from 1920×1024 to 1248×384.

Figure 10: Random sampled RGB images from the adverse weather dataset to illustrate the diversity of scenes, illumination and weather conditions. The images are cropped from 1920×1024 to 1248×384.

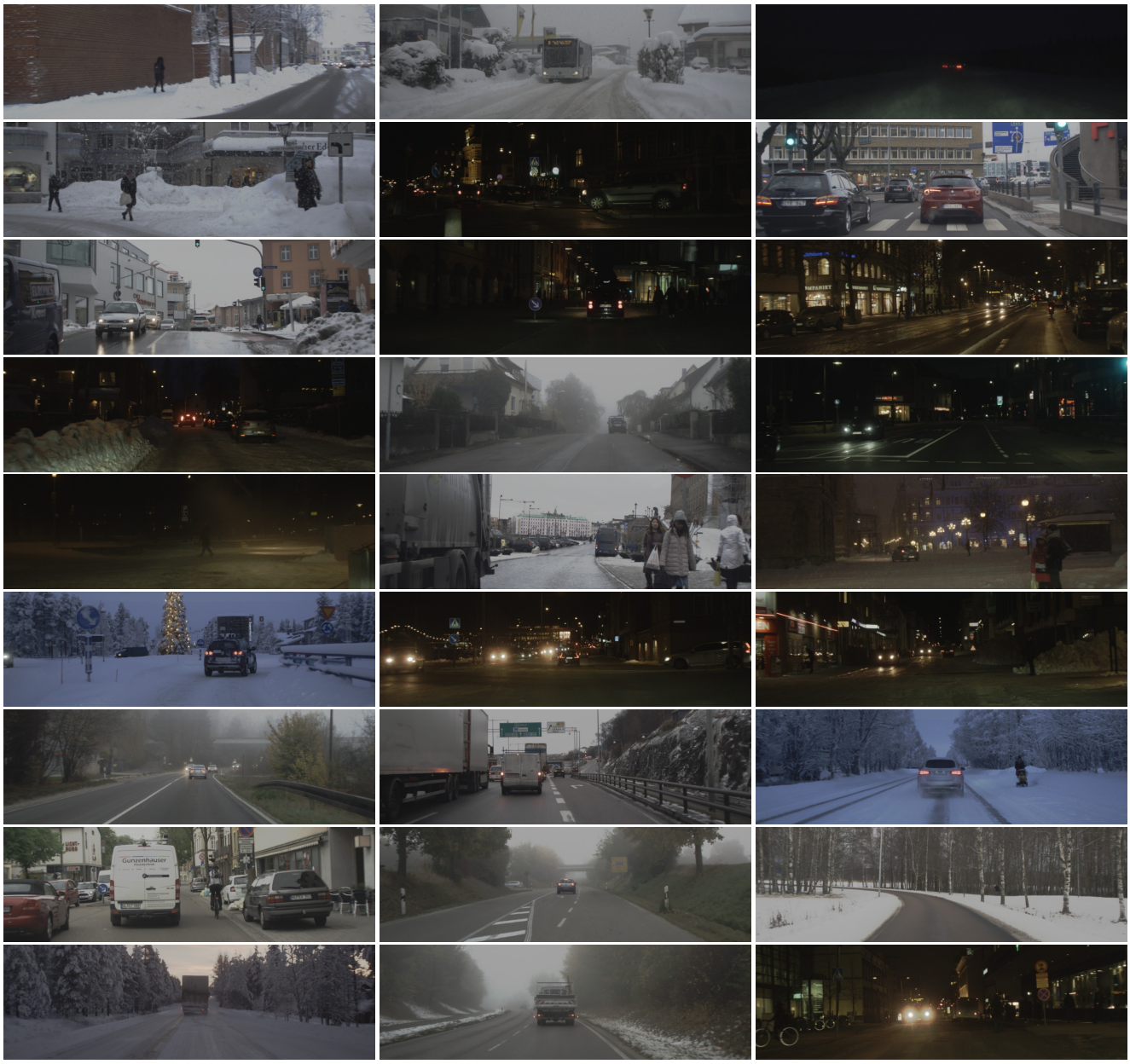

Figure 11: Random sampled RGB images from the adverse weather dataset to illustrate the diversity of scenes, illumination and weather conditions. The images are cropped from 1920×1024 to 1248×384.

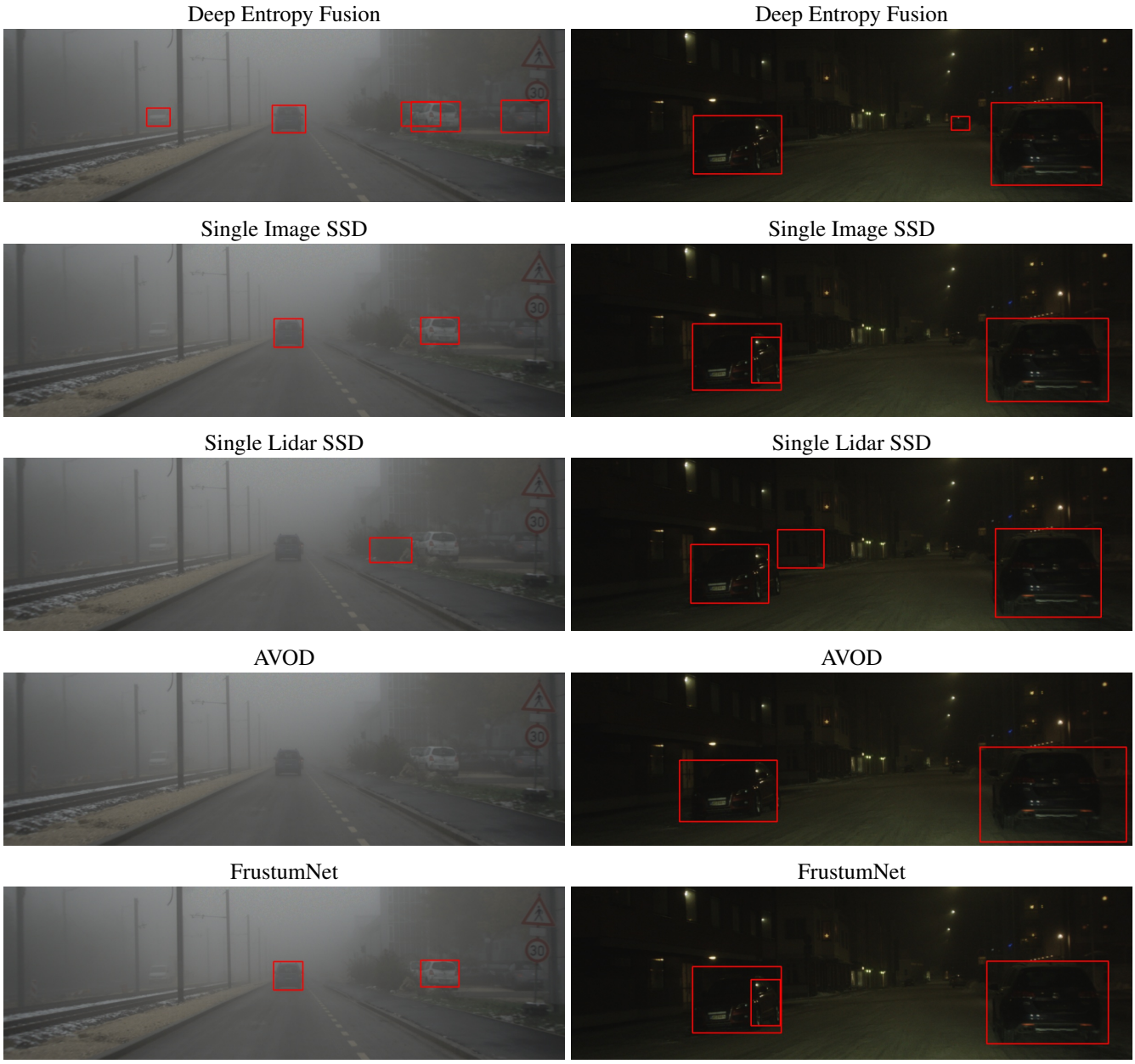


Figure 12: Additional qualitative detection results for in-the-wild measurement distortions that have not been seen during training. The left column shows disturbances through dense fog in an urban scenario, the right column show detections in nighttime conditions.


Figure 13: Additional qualitative detection results for in-the-wild measurement distortions that have not been seen during training. The left and right columns show disturbances in dense fog in different illumination settings on a suburban road.

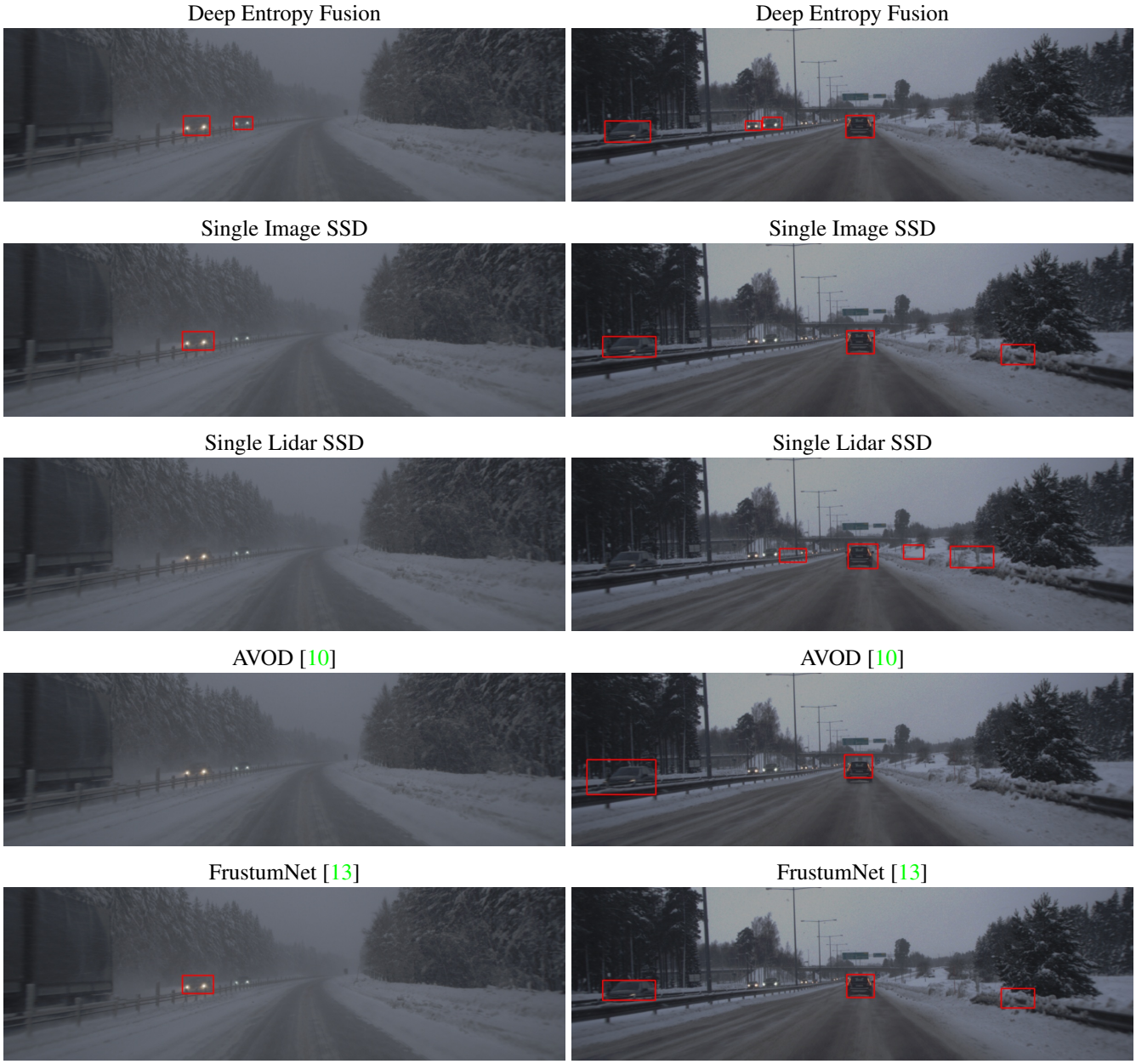


Figure 14: Additional qualitative detection results for in-the-wild measurement distortions that have not been seen during training. The left and right column show a suburban road with subarctic climate, with (left) and without (right) fog.

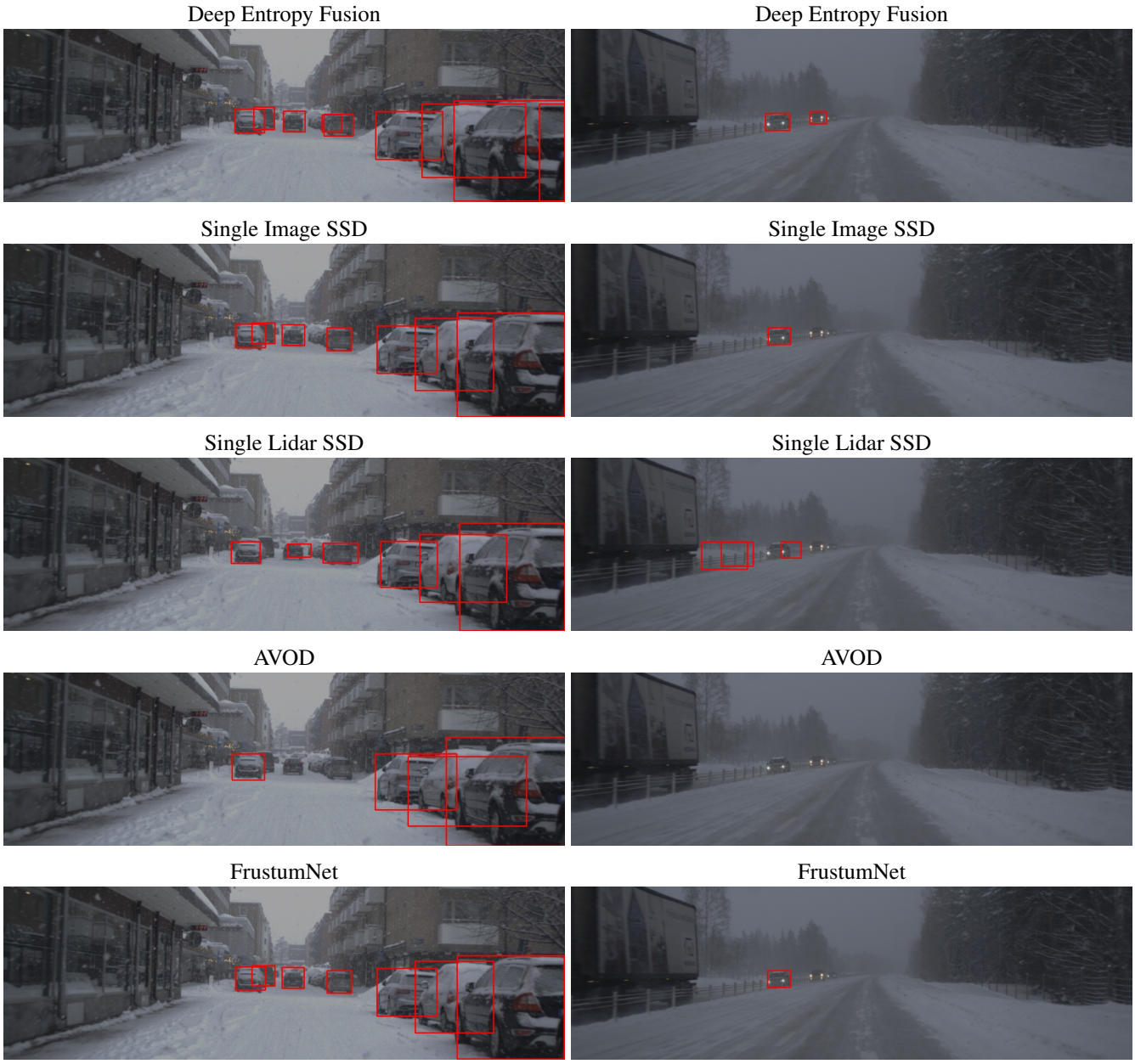


Figure 15: Additional qualitative detection results for in-the-wild measurement distortions that have not been seen during training. The left column shows an urban scenario during snowfall and the right column show detection performance on highway roads in fog.

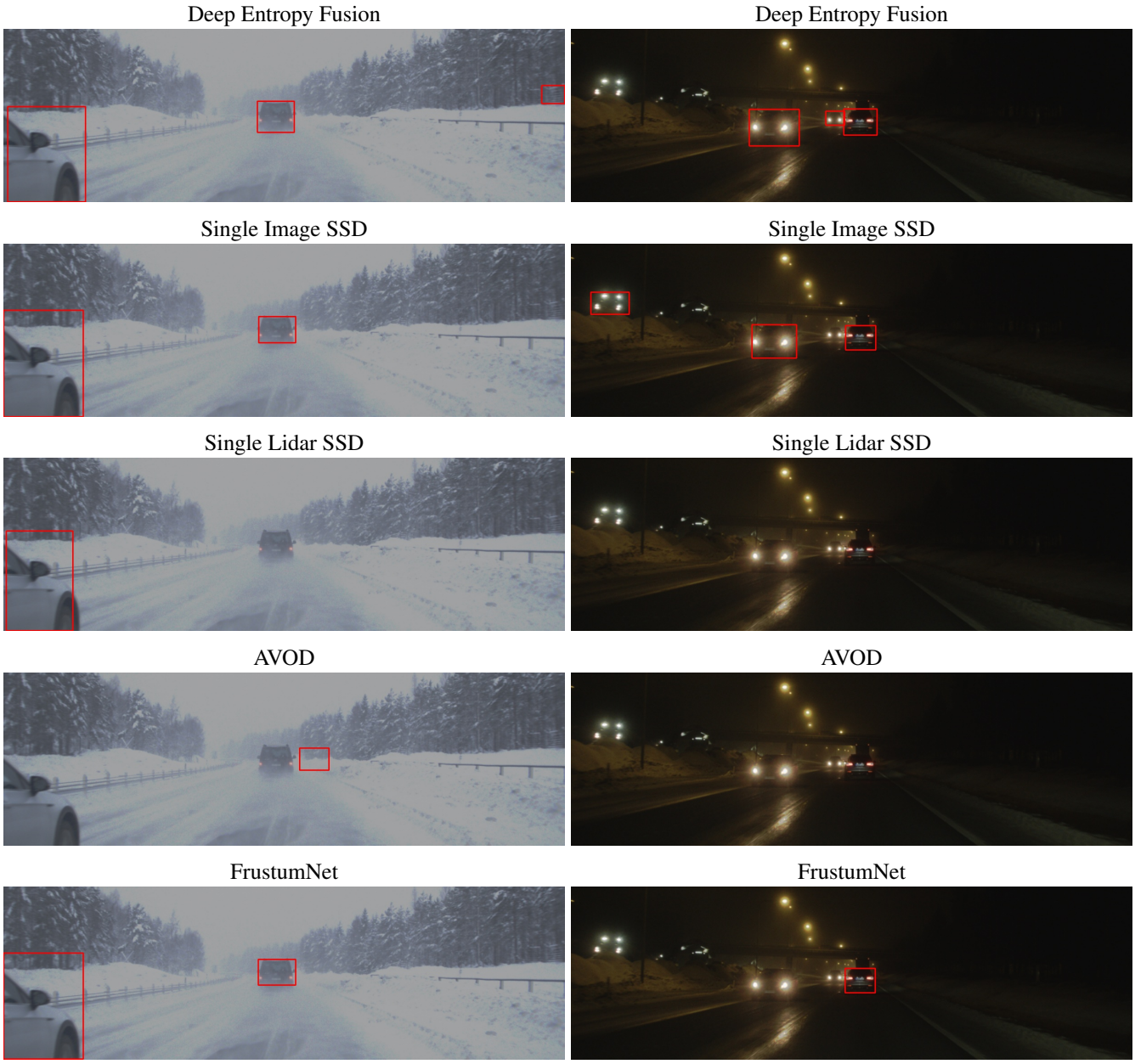


Figure 16: Additional qualitative detection results for in-the-wild measurement distortions that have not been seen during training. The left column shows disturbances through spray and an incorrect auto exposure, the right column show detections in during fog and nighttime.

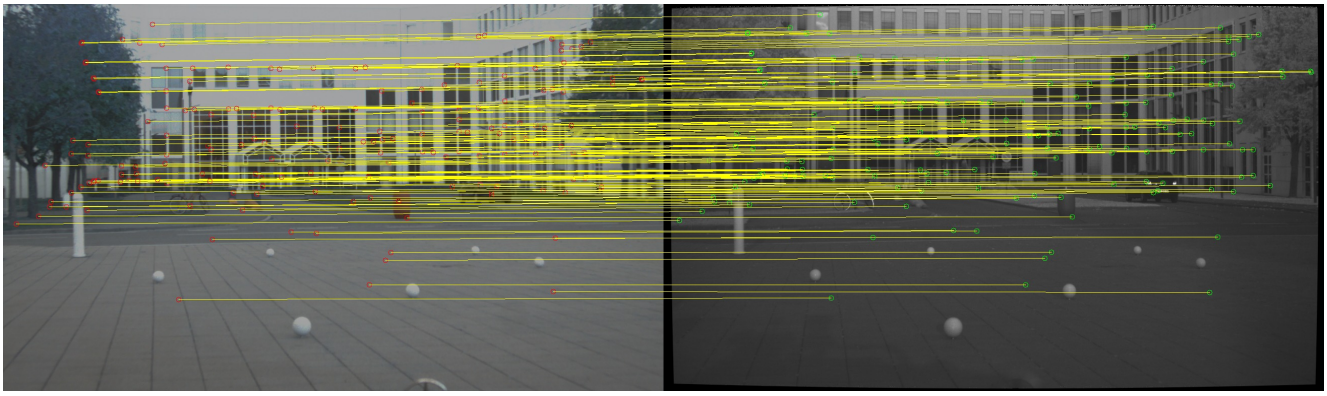

Figure 17: Image homography used for the image warping of gated images into the RGB image frame. Note the rgb image is cropped to the same field of view as the gated camera.


Figure 18: Examples for domain adaptation using CyCADA[9]. Adapting KITTI (left) to our experimental data adapts to winter scenes with snow, but does not properly model fog distortions (middle). For completeness, (right) shows a reverse transfer. The reverse transfer does not recover enough information to model a clear scene correctly.


Figure 19: Examples for domain adaptation from KITTI to our proposed dataset. The first two rows show a mapping from clear KITTI to clear experimental data using CyCADA [9], which contains only few artifacts. Snow on the sidewalks and a transfer from green trees to gray trees without leaves is correctly learned. In contrast, the last four rows show results from clear KITTI to adverse weather data using the same method which fails. The third row shows red shining backlights randomly placed in the scene. The fourth row shows random skit marks on the road. Row six shows wrongly interpreted shadows completely blacking out the scene.


Figure 20: Examples for domain adaptation from clear winter captures to adverse weather scenes. The first two rows show a mapping from clear images to clear winter captures with style transfer using CyCADA [9], which contains only a few artifacts, but note the illumination settings are not correctly changed. The last two rows show the mapping to foggy/snowy captures, which does not change the images at all because the appearance of the clear input images is already winter-like but without any adverse weather disturbances.

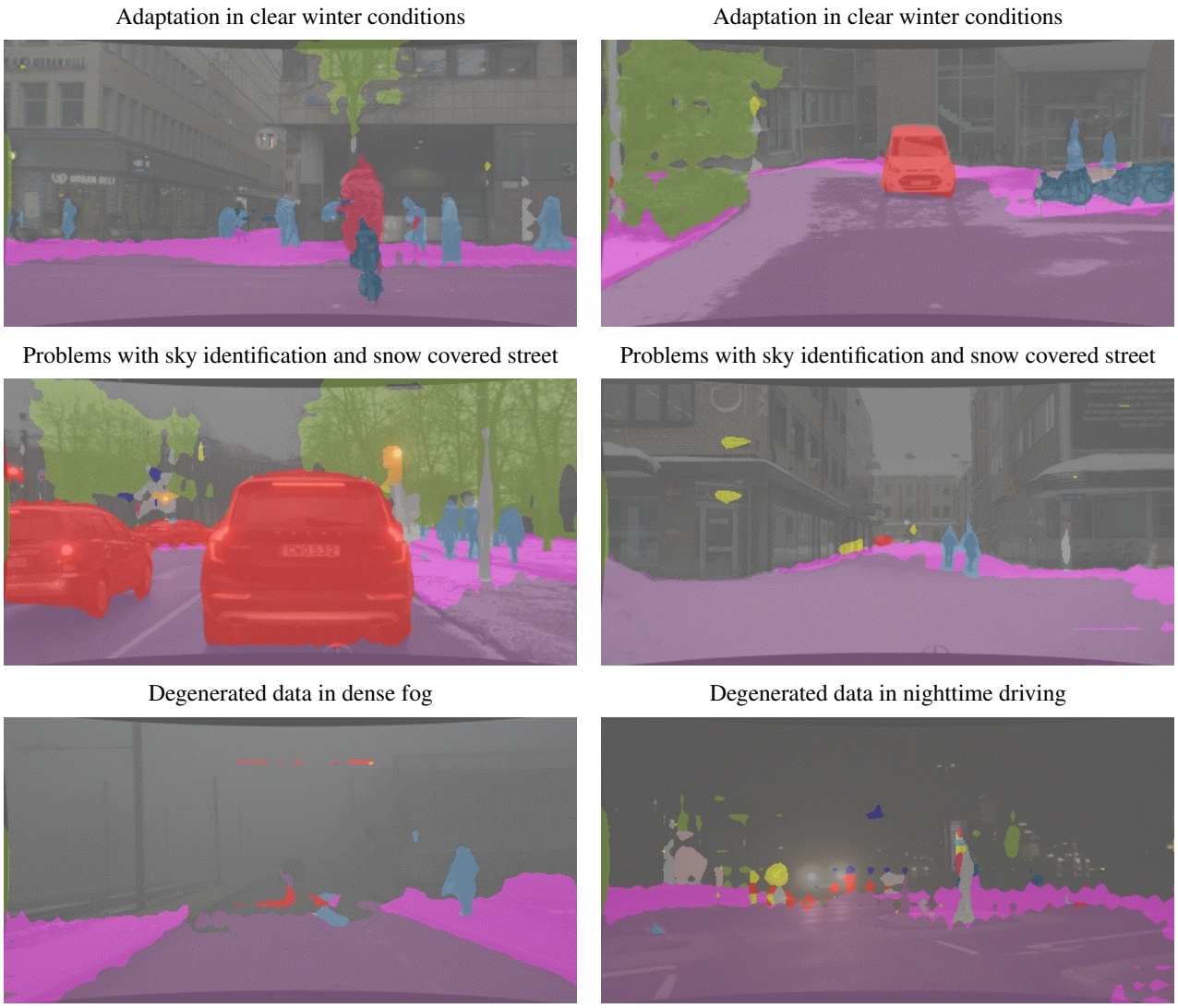


Figure 21: Examples for domain adaptation from a semantic segmentation model trained on clear Cityscapes mapped to our adverse weather scenes. The first row shows a transfer from clear Cityscapes images to clear winter captures using DADA [17], which contains only a few artifacts. The second row shows incorrect mapping to snowy winter captures. Here the sidewalks/roads covered with snow and the sky are incorrect. The last row shows an adaptation to dense fog and nighttime scenes. The image information has fully degenerated, and the adaptation fails.

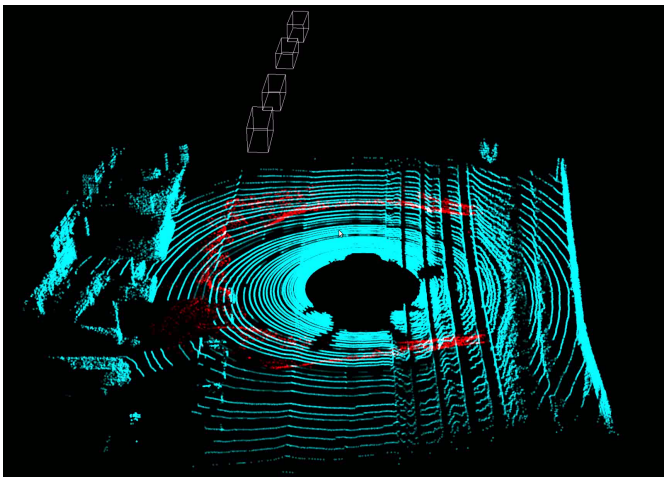

Figure 22: Synthetic example of a fog point cloud disturbed in fog ($\beta = 0.04$) with objects missing. Valid backscattered objects are marked in blue, backscatter from fog is marked red.

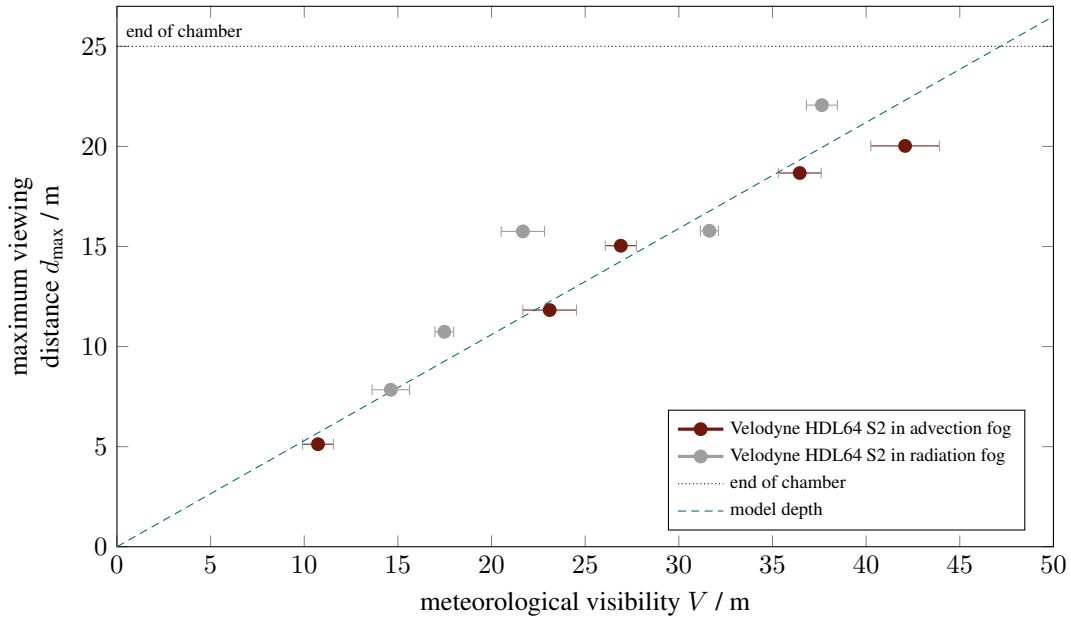


Figure 23: Here, the maximum viewing distance for the Velodyne HDL64 S2 lidar sensors in advection fog in a controlled fog chamber at different fog densities, fog types, and reflective targets with 90% reflectivity compared to our model prediction in the dashed line is presented.

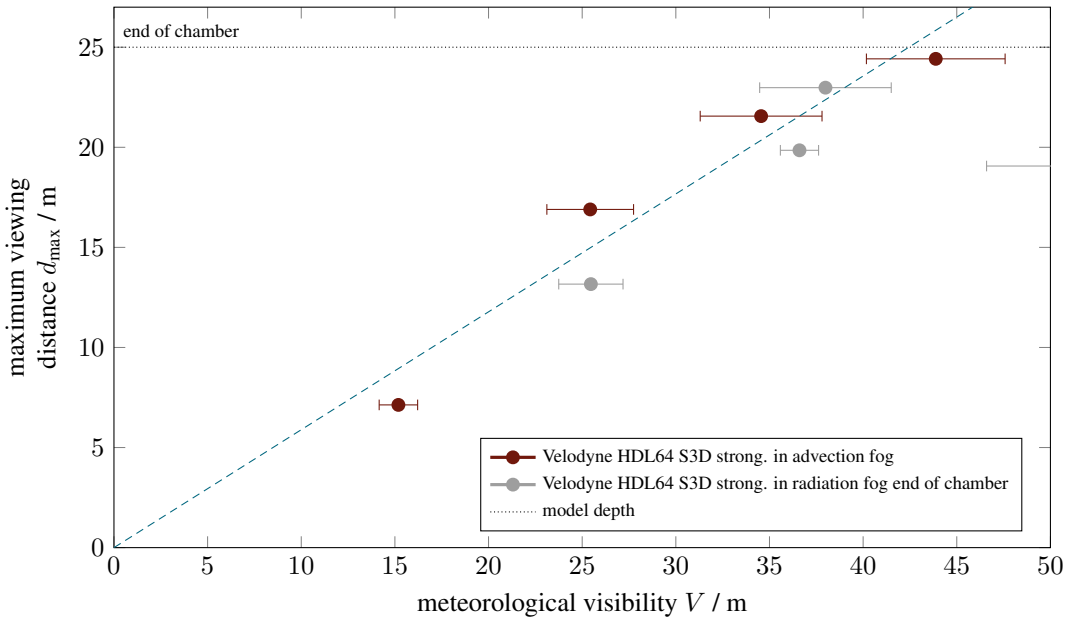


Figure 24: Here, the maximum viewing distance for the Velodyne HDL64 S3D lidar sensors in advection fog in a controlled fog chamber at different fog densities, fog types, and reflective targets with 90% reflectivity compared to our model prediction in the dashed line is presented.

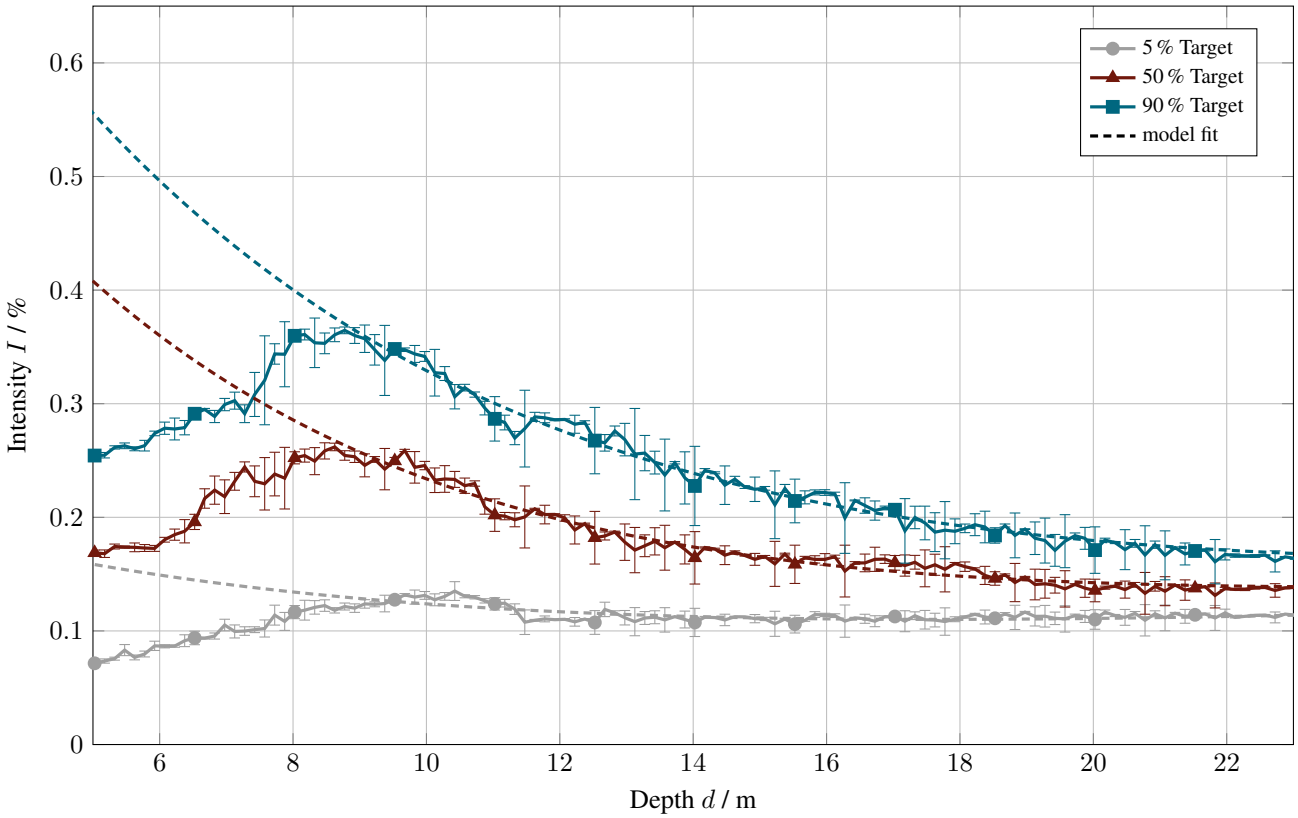


Figure 25: Fitted model to measured intensities $I$ in a controlled environment. The intensities have been measured with three calibrated diffusive reflective targets with values 5%, 50% and 90%. The scene is recorded with visibility $V \approx 50 - 60\,\mathrm{m}$, $\beta \approx 0.05 - 0.06\,\mathrm{m}^{-1}$. The intensity model from Eq. 2 is fitted into the measurement. The effect for distances closer than $8\,\mathrm{m}$ is due to the finite field of view of the light source.

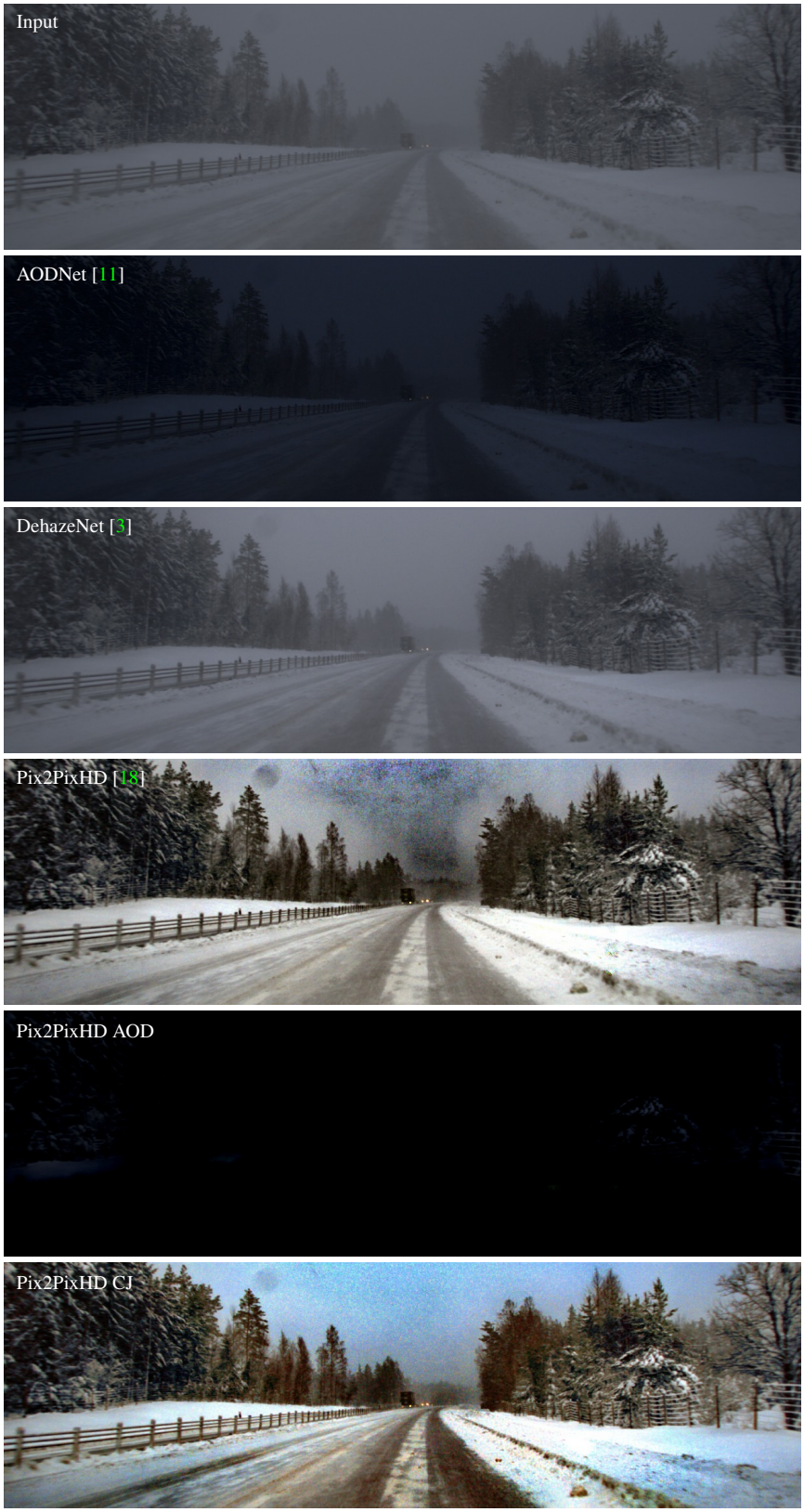


Figure 26: Additional image to image reconstruction results (top to bottom): Measured input image, AODNet [11], DehazeNet [3], Pix2PixHD [18], Pix2PixHD AOD and Pix2PixHD CJ in real adverse weather.


Figure 27: Failure cases for image reconstruction methods on unseen distortions. The learned distortion removal models were only trained on synthetic foggy scenes and therefore have limited generalization capabilities as different disturbance types can occur in combination as snow and fog.