

# Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather

Mario Bijelic<sup>1,3</sup> Tobias Gruber<sup>1,3</sup> Fahim Mannan<sup>2</sup> Florian Kraus<sup>1,3</sup> Werner Ritter<sup>1</sup> Klaus Dietmayer<sup>3</sup> Felix Heide<sup>2,4</sup>

<sup>1</sup>Mercedes-Benz AG    <sup>2</sup>Algolux    <sup>3</sup>Ulm University    <sup>4</sup>Princeton University

## Abstract

The fusion of multimodal sensor streams, such as camera, lidar, and radar measurements, plays a critical role in object detection for autonomous vehicles, which base their decision making on these inputs. While existing methods exploit redundant information in good environmental conditions, they fail in adverse weather where the sensory streams can be asymmetrically distorted. These rare “edge-case” scenarios are not represented in available datasets, and existing fusion architectures are not designed to handle them. To address this challenge we present a novel multimodal dataset acquired in over 10,000 km of driving in northern Europe. Although this dataset is the first large multimodal dataset in adverse weather, with 100k labels for lidar, camera, radar, and gated NIR sensors, it does not facilitate training as extreme weather is rare. To this end, we present a deep fusion network for robust fusion without a large corpus of labeled training data covering all asymmetric distortions. Departing from proposal-level fusion, we propose a single-shot model that adaptively fuses features, driven by measurement entropy. We validate the proposed method, trained on clean data, on our extensive validation dataset. Code and data are available here <https://github.com/princeton-computational-imaging/SeeingThroughFog>.

## 1. Introduction

Object detection is a fundamental computer vision problem in autonomous robots, including self-driving vehicles and autonomous drones. Such applications require 2D or 3D bounding boxes of scene objects in challenging real-world scenarios, including complex cluttered scenes, highly varying illumination, and adverse weather conditions. The most promising autonomous vehicle systems rely on redundant inputs from multiple sensor modalities [59, 6, 74], including camera, lidar, radar, and emerging sensor such as FIR [30]. A growing body of work on object detection using convolutional neural networks has enabled accurate 2D and 3D box estimation from such multimodal data, typically relying on camera and lidar data [65, 11, 57, 72, 67, 43, 36].



Figure 1: Existing object detection methods, including efficient Single-Shot detectors (SSD) [41], are trained on automotive datasets that are biased towards good weather conditions. While these methods work well in good conditions [19, 59], they fail in rare weather events (top). Lidar-only detectors, such as the same SSD model trained on projected lidar depth, might be distorted due to severe backscatter in fog or snow (center). These asymmetric distortions are a challenge for fusion methods, that rely on redundant information. The proposed method (bottom) learns to tackle unseen (potentially asymmetric) distortions in multimodal data without seeing training data of these rare scenarios.

While these existing methods, and the autonomous system that performs decision making on their outputs, perform well under normal imaging conditions, they fail in adverse weather and imaging conditions. This is because existing training datasets are biased towards clear weather conditions, and detector architectures are designed to rely only on the redundant information in the undistorted sensory streams. However, they are not designed for harsh scenarios that distort the sensor streams asymmetrically, see Figure 1. Extreme weather conditions are statistically rare. For example, thick fog is observable only during 0.01 %

of typical driving in North America, and even in foggy regions, dense fog with visibility below 50 m occurs only up to 15 times a year [62]. Figure 2 shows the distribution of real driving data acquired over four weeks in Sweden covering 10,000 km driven in winter conditions. The naturally biased distribution validates that harsh weather scenarios are only rarely or even not at all represented in available datasets [66, 19, 59]. Unfortunately, domain adaptation methods [45, 29, 42] also do not offer an ad-hoc solution as they require target samples, and adverse weather-distorted data are underrepresented in general. Moreover, existing methods are limited to image data but not to multisensor data, e.g. including lidar point-cloud data.

Existing fusion methods have been proposed mostly for lidar-camera setups [65, 11, 43, 36, 12], as a result of the limited sensor inputs in existing training datasets [66, 19, 59]. These methods do not only struggle with sensor distortions in adverse weather due to the bias of the training data. Either they perform late fusion through filtering after independently processing the individual sensor streams [12], or they fuse proposals [36] or high-level feature vectors [65]. The network architecture of these approaches is designed with the assumption that the data streams are consistent and redundant, i.e. an object appearing in one sensory stream also appears in the other. However, in harsh weather conditions, such as fog, rain, snow, or extreme lighting condition, including low-light or low-reflectance objects, multimodal sensor configurations can fail asymmetrically. For example, conventional RGB cameras provide unreliable noisy measurements in low-light scene areas, while scanning lidar sensors provide reliable depth using active illumination. In rain and snow, small particles affect the color image and lidar depth estimates equally through backscatter. Adversely, in foggy or snowy conditions, state-of-the-art pulsed lidar systems are restricted to less than 20 m range due to backscatter, see Figure 3. While relying on lidar measurements might be a solution for night driving, it is not for adverse weather conditions.

In this work, we propose a multimodal fusion method for object detection in adverse weather, including fog, snow, and harsh rain, without having large annotated training datasets available for these scenarios. Specifically, we handle asymmetric measurement corruptions in camera, lidar, radar, and gated NIR sensor streams by departing from existing proposal-level fusion methods: we propose an adaptive single-shot deep fusion architecture which exchanges features in intertwined feature extractor blocks. This deep early fusion is steered by measured entropy. The proposed adaptive fusion allows us to learn models that generalize across scenarios. To validate our approach, we address the bias in existing datasets by introducing a novel multimodal dataset acquired on three months of acquisition in northern Europe. This dataset is the first large multimodal driving

dataset in adverse weather, with 100k labels for lidar, camera, radar, gated NIR sensor, and FIR sensor. Although the weather-bias still prohibits training, this data allows us to validate that the proposed method generalizes robustly to unseen weather conditions with asymmetric sensor corruptions, while being trained on clean data.

Specifically, we make the following contributions:

- We introduce a multimodal adverse weather dataset covering camera, lidar, radar, gated NIR, and FIR sensor data. The dataset contains rare scenarios, such as heavy fog, heavy snow, and severe rain, during more than 10,000 km of driving in northern Europe.
- We propose a deep multimodal fusion network which departs from proposal-level fusion, and instead adaptively fuses driven by measurement entropy.
- We assess the model on the proposed dataset, validating that it generalizes to unseen asymmetric distortions. The approach outperforms state-of-the-art fusion methods more than 8% AP in hard scenarios independent of weather, including light fog, dense fog, snow, and clear conditions, and it runs in real-time.

## 2. Related Work

**Detection in Adverse Weather Conditions** Over the last decade, seminal work on automotive datasets [5, 14, 19, 16, 66, 9] has provided a fertile ground for automotive object detection [11, 8, 65, 36, 41, 20], depth estimation [18, 40, 21], lane-detection [27], traffic-light detection [33], road scene segmentation [5, 2], and end-to-end driving models [4, 66]. Although existing datasets fuel this research area, they are biased towards good weather conditions due to geographic location [66] and captured season [19], and thus lack severe distortions introduced by rare fog, severe snow, and rain. A number of recent works explore camera-only approaches in such adverse conditions [52, 7, 1]. However, these datasets are very small with less than 100 captured images [52] and limited to camera-only vision tasks. In contrast, existing autonomous driving applications rely on multimodal sensor stacks, including camera, radar, lidar, and emerging sensor, such as gated NIR imaging [22, 23], and have to be evaluated on thousands of hours of driving. In this work, we fill this gap and introduce a large scale evaluation set in order to develop a fusion model for such multimodal inputs that is robust to unseen distortions.

**Data Preprocessing in Adverse Weather** A large body of work explores methods for the removal of sensor distortions before processing. Especially fog and haze removal from conventional intensity image data have been explored extensively [68, 71, 34, 54, 37, 7, 38, 47]. Fog results in a distance-dependent loss in contrast and color. Fog removal methods have not only been suggested for display

application [25], it has also been proposed as preprocessing to improve the performance of downstream semantic tasks [52]. Existing fog and haze removal methods rely on scene priors on the latent clear image and depth to solve the ill-posed recovery. These priors are either hand-crafted [25] and used for depth and transmission estimation separately, or they are learned jointly as part of trainable end-to-end models [38, 32, 73]. Existing methods for fog and visibility estimation [58, 60] have been proposed for camera driver-assistance systems. Image restoration approaches have also been applied to deraining [10] or deblurring [37].

**Domain Adaptation** Another line of research tackles the shift of unlabeled data distributions by domain adaptation [61, 29, 51, 28, 70, 63]. Such methods could be applied to adapt clear labeled scenes to demanding adverse weather scenes [29] or through the adaptation of feature representations [61]. Unfortunately, both of these approaches struggle to generalize, because, in contrast to existing domain transfer methods, weather-distorted data in general, not only labeled data, is underrepresented. Moreover, existing methods do not handle multimodal data.

**Multisensor Fusion** Multisensor feeds in autonomous vehicles are typically fused to exploit varying cues in the measurements [44], simplify path-planning [15], to allow for redundancy in the presence of distortions [48], or solve for joint vision tasks, such as 3D object detection [65]. Existing sensing systems for fully-autonomous driving include lidar, camera, and radar sensors. As large automotive datasets [66, 19, 59] cover limited sensory inputs, existing fusion methods have been proposed mostly for lidar-camera setups [65, 56, 11, 36, 43]. Methods such as AVOD [36] and MV3D [11] incorporate multiple views from camera and lidar to detect objects. They rely on the fusion of pooled regions of interest and hence perform late feature fusion following popular region proposal architectures [50]. In a different line of research, Qi et al. [49] and Xu et al. [65] propose a pipeline model that requires a valid detection output for the camera image and a 3D feature vector extracted from the lidar point-cloud. Kim et al. [35] propose a gating mechanism for camera-lidar fusion. In all existing methods, the sensor streams are processed separately in the feature extraction stage, and we show that this prohibits learning redundancies, and, in fact, performs worse than a single sensor stream in the presence of asymmetric measurement distortions.

### 3. Multimodal Adverse Weather Dataset

To assess object detection in adverse weather, we have acquired a large-scale automotive dataset providing 2D and 3D detection bounding boxes for multimodal data with a fine classification of weather, illumination, and scene type in rare adverse weather situations. Table 1 compares our

Dataset Sensor Setup	KITTI [19]	BDD [69]	Waymo [59]	NuScenes [6]	Ours
RGB CAMERAS	2	1	5	6	2
RGB RESOLUTION	1242×372	1280×720	1920×1080	1600×900	1920×1024
LIDAR SENSORS	1	x	5	1	2
LIDAR RESOLUTION	64	0	64	32	64
RADAR SENSOR	x	x	x	4	1
GATED CAMERA	x	x	x	x	1
FIR CAMERA	x	x	x	x	1
FRAME RATE	10Hz	30Hz	10Hz	1 Hz/10 Hz	10 Hz
DATASET STATISTICS					
LABELED FRAMES	15K	100k	198k	40K	13.5K
LABELS	80k	1.47M	7.87M	1.4M	100K
SCENE TAGS	x	✓	x	✓	✓
NIGHT TIME	x	✓	✓	✓	✓
LIGHT WEATHER	x	✓	x	✓	✓
HEAVY WEATHER	x	x	x	x	✓
FOG CHAMBER	x	x	x	x	✓

Table 1: Comparison of the proposed multimodal adverse weather dataset to existing automotive detection datasets.

dataset to recent large-scale automotive datasets, such as the Waymo [59], NuScenes [6], KITTI [19] and the BDD [69] dataset. In contrast to [6] and [69], our dataset contains experimental data not only in light weather conditions but also in heavy snow, rain, and fog. A detailed description of the annotation procedures and label specifications is given in the supplemental material. With this cross-weather annotation of multimodal sensor data and broad geographical sampling, it is the only existing dataset that allows for the assessment of our multimodal fusion approach. In the future, we envision researchers developing and evaluating multimodal fusion methods in weather conditions not covered in existing datasets.

In Figure 2, we plot the weather distribution of the proposed dataset. The statistics were obtained by manually annotating all synchronized frames at a frame rate of 0.1 Hz. We guided human annotators to distinguish light from dense fog when the visibility fell below 1 km [46] and 100 m, respectively. If fog occurred together with precipitation, the scenes were either labeled as snowy or rainy depending on the environment road conditions. For our experiments, we combined snow and rainy conditions. Note that the statistics validate the rarity of scenes in heavy adverse weather, which is in agreement to [62] and demonstrates the difficulty and critical nature of obtaining such data in the assessment of truly self-driving vehicles, i.e. without the interaction of remote operators outside of geo-fenced areas. We found that extreme adverse weather conditions occur only locally and change very quickly.

The individual weather conditions result in asymmetrical perturbations of various sensor technologies, leading to asymmetric degradation, i.e. instead of all sensor outputs being affected uniformly by a deteriorating environmental condition, some sensors degrade more than others, see Figure 3. For example, conventional passive cameras perform well in daytime conditions, but their performance degrades in night-time conditions or challenging illumination settings such as low sun illumination. Meanwhile, active scanning sensors as lidar and radar are less affected by ambient light changes due to active illumination and a narrow bandpass

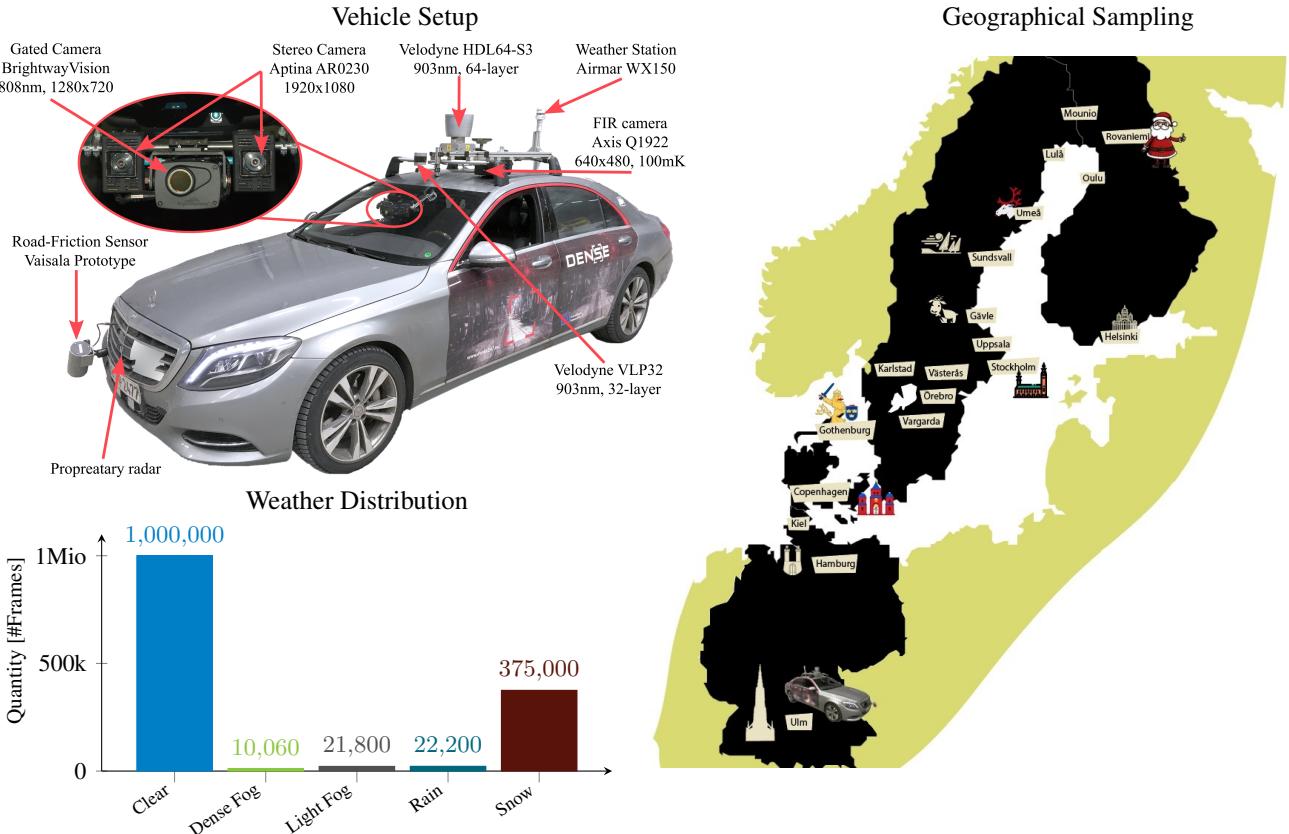


Figure 2: Right: Geographical coverage of the data collection campaign covering two months and 10,000 km in Germany, Sweden, Denmark, and Finland. Top Left: Test vehicle setup with top-mounted lidar, gated camera with flash illumination, RGB camera, proprietary radar, FIR camera, weather station, and road friction sensor. Bottom Left: Distribution of weather conditions throughout the data acquisition. The driving data is highly unbalanced with respect to weather conditions and only contains adverse conditions as rare samples.

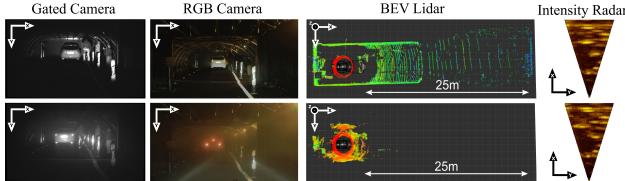


Figure 3: Multimodal sensor response of RGB camera, scanning lidar, gated camera, and radar in a fog chamber with dense fog. Reference recordings under clear conditions are shown in the first row, recordings in fog with visibility of 23 m are shown in the second row.

on the detector side. On the other hand, active lidar sensors are highly degraded by scattering media as fog, snow or rain, limiting the maximal perceivable distance at fog densities below 50 m to 25 m, see Figure 3. Millimeter-wave radar waves do not strongly scatter in fog [24], but currently provide only low azimuthal resolution. Recent gated images have shown robust perception in adverse weather [23], provide high spatial resolution, but are lacking color information compared to standard imagers. With these sensor-specific weaknesses and strengths of each sensor, multimodal data can be crucial in robust detection methods.

### 3.1. Multimodal Sensor Setup

For acquisition we have equipped a test vehicle with sensors covering the visible, mm-wave, NIR, and FIR band, see Figure 2. We measure intensity, depth, and weather condition.

**Stereo Camera** As visible-wavelength RGB cameras, we use a stereo pair of two front-facing high-dynamic-range automotive RCCB cameras, consisting of two On-Semi AR0230 imagers with a resolution of  $1920 \times 1024$ , a baseline of 20.3 cm and 12 bit quantization. The cameras run at 30 Hz and are synchronized for stereo imaging. Using Lensagon B5M8018C optics with a focal length of 8 mm, a field-of-view of  $31.1^\circ \times 21.7^\circ$  is obtained.

**Gated camera** We capture gated images in the NIR band at 808 nm using a BrightwayVision BrightEye camera operating at 120 Hz with a resolution of  $1280 \times 720$  and a bit depth of 10 bit. The camera provides a similar field-of-view as the stereo camera with  $31.1^\circ \times 17.8^\circ$ . Gated imagers rely on time-synchronized camera and flood-lit flash laser sources [31]. The laser pulse emits a variable narrow pulse, and the camera captures the laser echo after an adjustable

delay. This enables to significantly reduce backscatter from particles in adverse weather [3]. Furthermore, the high imager speed enables to capture multiple overlapping slices with different range-intensity profiles encoding extractable depth information in between multiple slices [23]. Following [23], we capture 3 broad slices for depth estimation and additionally 3-4 narrow slices together with their passive correspondence at a system sampling rate of 10 Hz.

**Radar** For radar sensing, we use a proprietary frequency-modulated continuous wave (FMCW) radar at 77 GHz with 1° angular resolution and distances up to 200 m. The radar provides position-velocity detections at 15 Hz.

**Lidar** On the roof of the car, we mount two laser scanners from Velodyne, namely HDL64 S3D and VLP32C. Both are operating at 903 nm and can provide dual returns (strongest and last) at 10 Hz. While the Velodyne HDL64 S3D provides equally distributed 64 scanning lines with an angular resolution of 0.4°, the Velodyne VLP32C offers 32 non-linear distributed scanning lines. HDL64 S3D and VLP32C scanners achieve a range of 100 m and 120 m, respectively.

**FIR camera** Thermal images are captured with an Axis Q1922 FIR camera at 30 Hz. The camera offers a resolution of  $640 \times 480$  with a pixel pitch of  $17\text{ }\mu\text{m}$  and a noise equivalent temperature difference (NETD)  $< 100\text{ mK}$ .

**Environmental Sensors** We measured environmental information with an Airmar WX150 weather station that provides temperature, wind speed and humidity, and a proprietary road friction sensor. All sensors are time-synchronized and ego-motion corrected using a proprietary inertial measurement unit (IMU). The system provides a sampling rate of 10 Hz.

## 3.2. Recordings

**Real-world Recordings** All experimental data has been captured during two test drives in February and December 2019 in Germany, Sweden, Denmark, and Finland for two weeks each, covering a distance of 10,000 km under different weather and illumination conditions. A total of 1.4 million frames at a frame rate of 10 Hz have been collected. Every 100th frame was manually labeled to balance scene type coverage. The resulting annotations contain 5.5k clear weather frames, 1k captures in dense fog, 1k captures in light fog, and 4k captures in snow/rain. Given the extensive capture effort, this demonstrates that training data in harsh conditions is rare. We tackle this approach by training only on clear data and testing on adverse data. The train and test regions do not have any geographic overlap. Instead of partitioning by frame, we partition our dataset based on independent recordings (5-60 min in length) from different locations. These recordings originate from 18 different major cities illustrated in Figure 2 and several smaller cities along the route.

**Controlled Condition Recordings** To collect image and range data under controlled conditions, we also provide measurements acquired in a fog chamber. Details on the fog chamber setup can be found in [17, 13]. We have captured 35k frames at a frame rate of 10 Hz and labeled a subset of 1,5k frames under two different illumination conditions (day/night) and three fog densities with meteorological visibilities  $V$  of 30 m, 40 m and 50 m. Details are given in the supplemental material, where we also do comparisons to a simulated dataset, using the forward model from [52].

## 4. Adaptive Deep Fusion

In this section, we describe the proposed adaptive deep fusion architecture that allows for multimodal fusion in the presence of unseen asymmetric sensor distortions. We devise our architecture under real-time processing constraints required for self-driving vehicles and autonomous drones. Specifically, we propose an efficient single-shot fusion architecture.

### 4.1. Adaptive Multimodal Single-Shot Fusion

The proposed network architecture is shown in Figure 4. It consists of multiple single-shot detection branches, each analyzing one sensor modality.

**Data Representation** The camera branch uses conventional three-plane RGB inputs, while for the lidar and radar branch, we depart from recent bird’s eye-view (BeV) projection [36] schemes or raw point-cloud representations [65]. BeV projection or point-cloud inputs do not allow for deep early fusion as the feature representations in the early layers are inherently different from the camera features. Hence, existing BeV fusion methods can only fuse features in a lifted space, after matching region proposals, but not earlier. Figure 4 visualizes the proposed input data encoding, which aids deep multimodal fusion. Instead of using a naive depth-only input encoding, we provide depth, height, and pulse intensity as input to the lidar network. For the radar network, we assume that the radar is scanning in a 2D-plane orthogonal to the image plane and parallel to the horizontal image dimension. Hence, we consider radar invariant along the vertical image axis and replicate the scan along vertical axis. Gated images are transformed into the image plane of the RGB camera using a homography mapping, see supplemental material. The proposed input encoding allows for a position and intensity-dependent fusion with pixel-wise correspondences between different streams. We encode missing measurement samples with zero value.

**Feature Extraction** As feature extraction stack in each stream, we use a modified VGG [55] backbone. Similar to [36, 11], we reduce the number of channels by half and cut the network at the conv4 layer. Inspired by [41, 39], we use six feature layers from conv4-10 as input to SSD detection

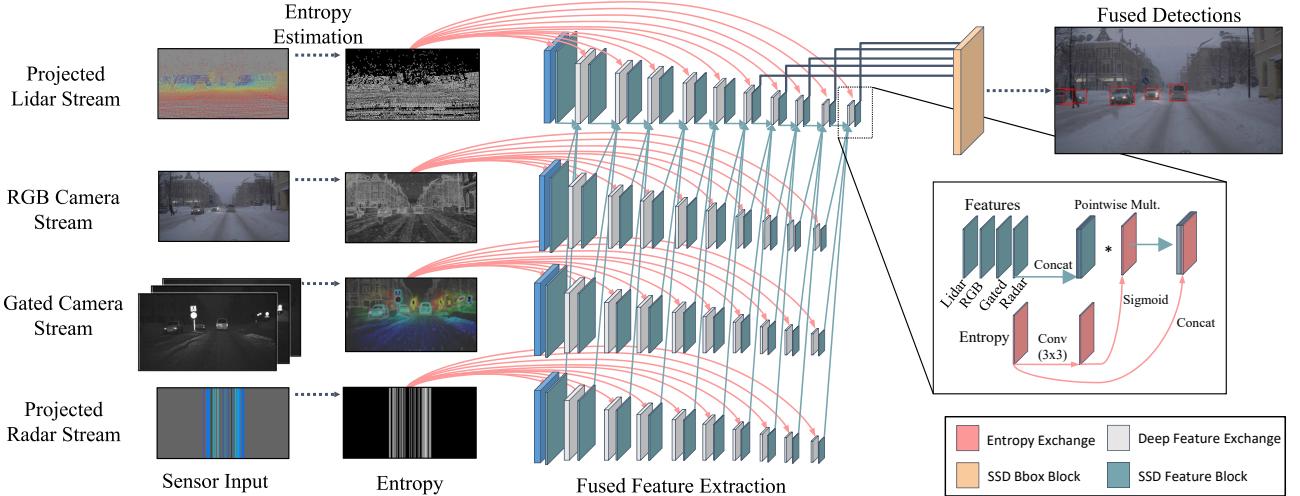


Figure 4: Overview of our architecture consisting of four single-shot detector branches with deep feature exchange and adaptive fusion of lidar, RGB camera, gated camera, and radar. All sensory data is projected into the camera coordinate system following Sec. 4.1. To steer fusion in-between sensors, the model relies on sensor entropy, which is provided to each feature exchange block (red). The deep feature exchange blocks (white) interchange information (blue) with parallel feature extraction blocks. The fused feature maps are analyzed by SSD blocks (orange).

layers. The feature maps decrease in size<sup>1</sup>, implementing a feature pyramid for detections at different scales. As shown in Figure 4, the activations of different feature extraction stacks are exchanged. To steer fusion towards the most reliable information, we provide the sensor entropy to each feature exchange block. We first convolve the entropy, apply a sigmoid, multiply with the concatenated input features from all sensors, and finally concatenate the input entropy. The folding of entropy and application of the sigmoid generates a multiplication matrix in the interval [0,1]. This scales the concatenated features for each sensor individually based on the available information. Regions with low entropy can be attenuated, while entropy rich regions can be amplified in the feature extraction. Doing so allows us to *adaptively fuse features* in the feature extraction stack itself, which we motivate in depth in the next section.

## 4.2. Entropy-steered Fusion

To steer the deep fusion towards redundant and reliable information, we introduce an entropy channel in each sensor stream, instead of directly inferring the adverse weather type and strength as in [58, 60]. We estimate local measurement entropy,

$$\rho = \sum_{m,n} \sum_{i=0}^{w,h} p_i^{mn} \log(p_i^{mn}) \quad , \text{with} \\ p_i^{mn} = \frac{1}{MN} \sum_{j,k}^{M,N} \delta(I(m+j, n+k) - i). \quad (1)$$

The entropy is calculated for each 8 bit binarized stream  $I$  with pixel values  $i \in [0, 255]$  in the proposed image-space

data representation. Each stream is split into patches of size  $M \times N = 16 \text{ px} \times 16 \text{ px}$  resulting in a  $w \times h = 1920 \text{ px} \times 1024 \text{ px}$  entropy map. The multimodal entropy maps for two different scenarios are shown in Figure 5: the left scenario shows a scene containing a vehicle, cyclist, and pedestrians in a controlled fog chamber. The passive RGB camera and lidar suffer from backscatter and attenuation with decreasing fog visibilities, while the gated camera suppresses backscatter through gating. Radar measurements are also not substantially degraded in fog. The right scenario in Figure 5 shows a static outdoor scene under varying ambient lighting. In this scenario, active lidar and radar are not affected by changes in ambient illumination. For the gated camera, the ambient illumination disappears, leaving only the actively illuminated areas, while the passive RGB camera degenerates with decreasing ambient light.

The steering process is learned purely on clean weather data, which contains different illumination settings present in day to night-time conditions. No real adverse weather patterns are presented during training. Further, we drop sensor streams randomly with probability 0.5 and set the entropy to a constant zero value.

## 4.3. Loss Functions and Training Details

The number of anchor boxes in different feature layers and their sizes play an important role during training and are given in the supplemental material. In total, each anchor box with class label  $y_i$  and probability  $p_i$  is trained using the cross entropy loss with softmax,

$$H(p) = \sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)). \quad (2)$$

<sup>1</sup>We use a feature map pyramid [(24, 78), (24, 78), (12, 39), (12, 39), (6, 20), (3, 10)]

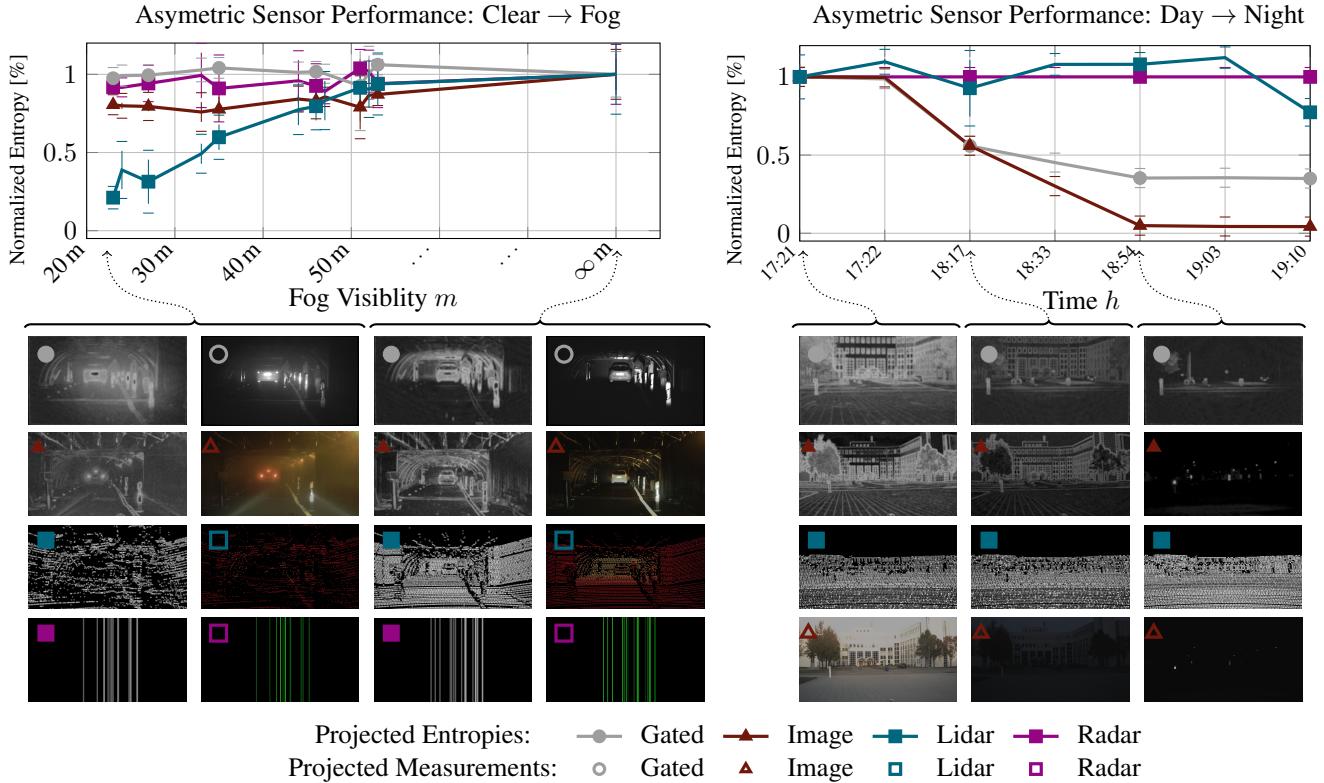


Figure 5: Normalized entropy with respect to the clear reference recording for a gated camera, RGB camera, radar, and lidar in varying fog visibilities (left) and changing illumination (right). The entropy has been calculated based on a dynamic scenario within a controlled fog chamber illustrated in Figure 3 (left) and a static scenario with changing natural illumination settings (right). Quantitative numbers have been calculated following Eq. (1). Note the asymmetric sensor failure for different sensor technologies. Qualitative results are given below and are connected via arrows to their corresponding fog density/daytime.

The loss is split up for positive and negative anchor boxes with a matching threshold of 0.5. For each positive anchor box, the bounding box coordinates  $x$  are regressed using a Huber loss  $H(x)$  given by,

$$H(x) = \begin{cases} x^2/2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{if } |x| > 1 \end{cases} \quad (3)$$

The total number of negative anchors is restricted to  $5 \times$  the number of positive examples using hard example mining [41, 53]. All networks are trained from scratch with a constant learning rate and L2 weight decay of 0.0005.

## 5. Assessment

In this section, we validate the proposed fusion model on unseen experimental test data. We compare the method against existing detectors for single sensory inputs and fusion methods, as well as domain adaptation methods. Due to the weather-bias of training data acquisition, we only use the clear weather portion of the proposed dataset for training. We assess the detection performance using our novel multimodal weather dataset as a test set, see supplemental data for test and training split details.

We validate the proposed approach in Table 2, which we

dub Deep Entropy Fusion, on real adverse weather data. We report Average Precision (AP) for three different difficulty levels (easy, moderate, hard) and evaluate on cars following the KITTI evaluation framework [19] at various fog densities, snow disturbances, and clear weather conditions. We compare the proposed model against recent state-of-the-art lidar-camera fusion models, including AVOD-FPN [36], Frustum PointNets [49], and variants of the proposed method with alternative fusion or sensory inputs. As baseline variants, we implement two fusion and four single sensor detectors. In particular, we compare against late fusion with image, lidar, gated, and radar features concatenated just before bounding-box regression (*Fusion SSD*), and early fusion by concatenating all sensory data at the early beginning of one feature extraction stack (*Concat SSD*). The *Fusion SSD* network shares the same structure as the proposed model, but without the feature exchange and the adaptive fusion layer. Moreover, we compare the proposed model against an identical SSD branch with single sensory input (*Image-only SSD*, *Gated-only SSD*, *Lidar-only SSD*, *Radar-only SSD*). All models were trained with identical hyper-parameters and anchors.

Evaluated on adverse weather scenarios, the detection

WEATHER DIFFICULTY	clear			light fog			dense fog			snow/rain		
	easy	mod.	hard									
DEEP ENTROPY FUSION (THIS WORK)	89.84	<b>85.57</b>	<b>79.46</b>	90.54	<b>87.99</b>	<b>84.90</b>	87.68	<b>81.49</b>	<b>76.69</b>	88.99	<b>83.71</b>	<b>77.85</b>
DEEP FUSION (THIS WORK)	<b>90.07</b>	80.31	77.82	<b>90.60</b>	81.08	79.63	86.77	77.28	73.93	<b>89.25</b>	79.09	70.51
FUSION SSD	87.73	78.02	69.49	88.33	78.65	76.54	74.07	68.46	63.23	85.49	75.28	67.48
CONCAT. SSD	86.12	76.62	68.61	87.98	78.24	70.17	77.99	69.16	67.07	83.63	73.65	66.26
ADDA [61] <sup>1</sup>	85.27	70.51	67.86	87.83	78.68	70.38	87.64	78.12	74.37	84.17	74.25	66.86
CYCADA [29] <sup>1</sup>	88.50	77.84	69.56	89.08	79.36	75.58	87.24	77.04	73.38	85.56	74.80	67.22
IMAGE-ONLY SSD	85.43	75.75	67.79	87.76	78.52	70.43	<b>87.89</b>	78.25	74.96	84.33	74.38	67.01
GATED-ONLY SSD	77.10	61.95	58.27	80.65	69.64	61.75	75.16	66.76	61.68	77.32	61.31	57.23
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
RADAR-ONLY SSD	10.26	8.54	8.23	16.92	13.24	12.66	16.33	13.57	13.00	12.94	10.95	10.40
AVOD-FPN [36]	66.47	58.71	51.63	60.40	52.51	51.92	33.95	26.29	26.17	59.55	51.91	50.54
FRUSTUM POINTNET [49]	80.06	75.89	67.70	84.06	76.88	75.44	76.69	73.62	68.49	78.34	74.34	66.52

Table 2: Quantitative detection AP on real unseen weather-affected data from dataset split across weather and difficulties easy/moderate/hard following [19]. All detection models except domain adaptation approaches are trained solely on clean data without weather distortions. The best model is highlighted in bold.

performance decrease for all methods. Note that assessment metrics can increase simultaneously as scene complexity changes between the weather splits. For example, when fewer vehicles participate in road traffic or the distance between vehicles increases in icy conditions, fewer vehicles are occluded. While the performance for image and gated data is almost steady, it decreases substantially for lidar data while it increases for radar data. The decrease in lidar performance can be described by the strong backscatter, see Supplemental Material. As a maximum of 100 measurement targets limits the performance of the radar input, the reported improvements are resulting from simpler scenes.

Overall, the large reduction in lidar performance for foggy conditions affects the lidar only detection rate by a drop in 45.38 % AP. Furthermore, it also has a strong impact on camera-lidar fusion models *AVOD*, *Concat SSD* and *Fusion SSD*. Learned redundancies no longer hold, and these methods even fall below image-only methods.

Two-stage methods, such as Frustum PointNet [49], drop quickly. However, they asymptotically achieve higher results compared to AVOD, because the statistical priors learned for the first stage are based on *Image-only SSD* that limits its performance to image-domain priors. AVOD is limited by several assumptions that hold for clear weather, such as the importance sampling of boxes filled with lidar data during training, achieving the lowest fusion performance overall. Moreover, as the fog density increases, the proposed adaptive fusion model outperforms all other methods. Especially under severe distortions, the proposed adaptive fusion layer results in significant margins over the model without it (*Deep Fusion*). Overall the proposed method outperforms all baseline approaches. In dense fog, it improves by a margin of 9.69 % compared to the next-best feature-fusion variant.

For completeness, we also compare the proposed model to recent domain adaptation methods. First, we adapt our *Image-Only SSD* features from clear weather to adverse weather following [61]. Second, we investigate the style

transfer from clear weather to adverse weather utilizing [29] and generate adverse weather training samples from clear weather input. Note that these methods have an unfair advantage over all other compared approaches as they have seen adverse weather scenarios sampled from our validation set. Note that domain adaptation methods cannot be directly applied as they need target images from a specific domain. Therefore, they do also not offer a solution for rare edge cases with limited data. Furthermore [29] *does not model distortions*, including fog or snow, see experiments in the Supplemental Material. We note that synthetic data augmentation following [52] or image-to-image reconstruction methods that remove adverse weather effects [64] do neither affect the reported margins of the proposed multimodal deep entropy fusion.

## 6. Conclusion and Future Work

In this paper we address a critical problem in autonomous driving: multi-sensor fusion in scenarios, where annotated data is sparse and difficult to obtain due to natural weather bias. To assess multimodal fusion in adverse weather, we introduce a novel adverse weather dataset covering camera, lidar, radar, gated NIR, and FIR sensor data. The dataset contains rare scenarios, such as heavy fog, heavy snow, and severe rain, during more than 10,000 km of driving in northern Europe. We propose a real-time deep multimodal fusion network which departs from proposal-level fusion, and instead adaptively fuses driven by measurement entropy. Exciting directions for future research include the development of end-to-end models enabling the failure detection and an adaptive sensor control as noise level or power level control in lidar sensors.

## Acknowledgment

The authors would like to acknowledge the funding from the European Union under the H2020 ECSEL Programme as part of the DENSE project, contract number 692449, and thank Jason Taylor for fruitful discussion.

<sup>1</sup>Requires large corpus of adverse weather data for training.

## References

- [1] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 867–8678, 2018. [2](#)
- [2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, Dec 2017. [2](#)
- [3] M. Bijelic, T. Gruber, and W. Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1773–1779, 2018. [5](#)
- [4] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. [2](#)
- [5] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 44–57. Springer, 2008. [2](#)
- [6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liog, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. [1, 3](#)
- [7] 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](#)
- [8] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 354–370. Springer, 2016. [2](#)
- [9] M. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays. Argoverse: 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8740–8749, 2019. [2](#)
- [10] J. Chen, C.-H. Tan, J. Hou, L.-P. Chau, and H. Li. Robust video content alignment and compensation for rain removal in a cnn framework. *CVPR*, 2018. [3](#)
- [11] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. [1, 2, 3, 5](#)
- [12] H. Cho, Y.-W. Seo, B. V. Kumar, and R. R. Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1836–1843. IEEE, 2014. [2](#)
- [13] 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. [5](#)
- [14] 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. [2](#)
- [15] D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel. Path planning for autonomous vehicles in unknown semi-structured environments. *The International Journal of Robotics Research*, 29(5):485–501, 2010. [3](#)
- [16] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. [2](#)
- [17] 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. [5](#)
- [18] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. [2](#)
- [19] 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. [1, 2, 3, 7, 8](#)
- [20] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1440–1448, 2015. [2](#)
- [21] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. [2](#)
- [22] Y. Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. [2](#)
- [23] 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*, 2019. [2, 4, 5](#)
- [24] S. Hasirlioglu, A. Kamann, I. Doric, and T. Brandmeier. Test methodology for rain influence on automotive surround sensors. In *IEEE International Conference on Intelligent Transportation Systems*, pages 2242–2247. IEEE, 2016. [4](#)
- [25] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. [3](#)
- [26] R. Heinzler, P. Schindler, J. Seekircher, W. Ritter, and W. Stork. Weather influence and classification with automotive lidar sensors. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1527–1534, 2019.
- [27] A. B. Hillel, R. Lerner, D. Levi, and G. Raz. Recent progress in road and lane detection: a survey. *Machine vision and applications*, 25(3):727–745, 2014. [2](#)
- [28] J. Hoffman, T. Darrell, and K. Saenko. Continuous manifold based adaptation for evolving visual domains. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–874, 2014. [3](#)
- [29] 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. [2, 3, 8](#)

- [30] P. Hurney, P. Waldron, F. Morgan, E. Jones, and M. Glavin. Review of pedestrian detection techniques in automotive far-infrared video. *IET intelligent transport systems*, 9(8):824–832, 2015. 1
- [31] S. Inbar and O. David. Laser gated camera imaging system and method, may 2008. 4
- [32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 3
- [33] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi. Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):1800–1815, 2016. 2
- [34] S. Ki, H. Sim, J.-S. Choi, S. Kim, and M. Kim. Fully end-to-end learning based conditional boundary equilibrium gan with receptive field sizes enlarged for single ultra-high resolution image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 817–824, 2018. 2
- [35] J. Kim, J. Choi, Y. Kim, J. Koh, C. C. Chung, and J. W. Choi. Robust camera lidar sensor fusion via deep gated information fusion network. In *IEEE Intelligent Vehicle Symposium*, pages 1620–1625. IEEE, 2018. 3
- [36] 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. 1, 2, 3, 5, 7, 8
- [37] 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*, pages 8183–8192, 2018. 2, 3
- [38] 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. 2, 3
- [39] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, July 2017. 5
- [40] F. Liu, C. Shen, G. Lin, and I. D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2016. 2
- [41] 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 IEEE European Conference on Computer Vision*, pages 21–37. Springer, 2016. 1, 2, 5, 7
- [42] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 2
- [43] W. Luo, B. Yang, and R. Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 1, 2, 3
- [44] O. Mees, A. Eitel, and W. Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *IEEE International Conference on Intelligent Robots and Systems*, pages 151–156. IEEE, 2016. 3
- [45] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2
- [46] U. D. of Commerce / National Oceanic and A. Administration. Fog definition. *Federal Meteorological Handbook No. 1: Surface Weather Observations and Reports. U.S. Department of Commerce / National Oceanic and Atmospheric Administration*, 2005. 3
- [47] D. Park, H. Park, D. K. Han, and H. Ko. Single image dehazing with image entropy and information fidelity. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4037–4041, 2014. 2
- [48] C. Premebida, O. Ludwig, and U. Nunes. Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, 26(9):696–711, 2009. 3
- [49] 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. 3, 7, 8
- [50] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 3
- [51] C. Sakaridis, D. Dai, and L. V. Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7374–7383, 2019. 3
- [52] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, pages 1–20, 2018. 2, 3, 5, 8
- [53] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2016. 7
- [54] H. Sim, S. Ki, J.-S. Choi, S. Seo, S. Kim, and M. Kim. High-resolution image dehazing with respect to training losses and receptive field sizes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 912–919, 2018. 2
- [55] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 201. 5
- [56] V. A. Sindagi, Y. Zhou, and O. Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282, 2019. 3
- [57] S. Song and J. Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2016. 1
- [58] R. Spinneker, C. Koch, S. Park, and J. J. Yoon. Fast fog detection for camera based advanced driver assistance systems.

- In *IEEE International Conference on Intelligent Transportation Systems*, pages 1369–1374, Oct 2014. 3, 6
- [59] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset, 2019. 1, 2, 3
- [60] J.-P. Tarel, N. Hautiere, A. Cord, D. Gruyer, and H. Halmouai. Improved visibility of road scene images under heterogeneous fog. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 478–485. Citeseer, 2010. 3, 6
- [61] 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. 3, 8
- [62] G. J. van 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(10):4597–4609, 2010. 2, 3
- [63] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 3
- [64] 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. 8
- [65] D. Xu, D. Anguelov, and A. Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. 1, 2, 3, 5
- [66] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2174–2182, 2017. 2, 3
- [67] B. Yang, W. Luo, and R. Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 1
- [68] M.-h. Yang, V. M. Patel, J.-s. Choi, S. Kim, B. Chanda, P. Wang, Y. Chen, A. Alvarez-gila, A. Galdran, J. Vazquez-corral, M. Bertalmo, H. S. Demir, and J. Chen. NTIRE 2018 Challenge on Image Dehazing : Methods and Results. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–12, 2018. 2
- [69] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 3
- [70] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2039–2049, 2017. 3
- [71] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2
- [72] Y. Zhou and O. Tuzel. Voxelenet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. 1
- [73] 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 Conference on Computer Vision and Pattern Recognition*, pages 2223–2232, 2017. 3
- [74] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller, E. Kaus, R. G. Herrtwich, C. Rabe, D. Pfeiffer, F. Lindner, F. Stein, F. Erbs, M. Enzweiler, C. Knöppel, J. Hipp, M. Haueis, M. Trepte, C. Brenk, A. Tamke, M. Ghanaat, M. Braun, A. Joos, H. Fritz, H. Mock, M. Hein, and E. Zeeb. Making bertha drive—an autonomous journey on a historic route. *IEEE Intelligent Transportation Systems Magazine*, 6(2):8–20, 2014. 1