Seeing Through Fog Without Seeing Fog: Deep Sensor Fusion in the Absence of Labeled Training Data

Mario Bijelic * Fahim Mannan Tobias Gruber Werner Ritter Klaus Dietmayer Felix Heide

Abstract

The fusion of color and lidar data plays a critical role in object detection for autonomous vehicles, which base their decision making on these inputs. While existing methods exploit redundant and complimentary information under good imaging conditions, they fail to do this in adverse weather and imaging conditions where the sensory streams can be asymmetrically distorted. These rare “edge-case” scenarios are not represented in available data sets, and existing fusion architectures are not designed to handle severe asymmetric distortions. We present a deep fusion architecture that allows for robust fusion in fog and snow without having large labeled training data available for these scenarios. Departing from proposal-level fusion, we propose a real-time single-shot model that adaptively fuses features driven by temporal coherence of the distortions. We validate the proposed method, trained on clean data, in simulation and on unseen conditions of in-the-wild driving scenarios.

1. Introduction

Object detection is a fundamental computer vision problem in autonomous robotics including self-driving vehicles and autonomous drones. In such autonomous systems, 2D or 3D bounding boxes of scene objects have to be recovered in challenging real-world scenarios, including complex cluttered scenes, highly varying illumination, and adverse weather conditions. Recently, a growing body of work on convolutional neural networks for object detection has enabled accurate 2D and 3D box estimation from RGB-D data [51, 16, 58], or even single monocular images [46, 15, 20]. While these existing methods, and the autonomous system that performs decision making based on these methods, perform well under normal imaging conditions, they fail in adverse weather and imaging conditions. This is because relevant data sets are biased towards normal conditions, and existing detector architectures are not designed for harsh scenarios. Challenging conditions are statistically rare, e.g. thick fog is observable only during 0.01% of typical driving in North America, and hence not represented in available data sets [60, 16]. Existing fusion methods for lidar-camera setups either perform late fusion through filtering after independently processing the individual sensor streams [8], or they fuse proposals [31] or high-level feature vectors [58]. Common to all of these approaches is the assumption that depth and image data are consistent, i.e. an object appearing in one sensory stream also appears in the other.

In harsh weather conditions, such as fog, rain, snow,
or difficult lighting condition, including low-light or low-reflectance objects, conventional RGB-D sensor stacks can fail asymmetrically. For example, conventional RGB cameras provide unreliable noisy measurements in low-light scene areas, i.e., increasingly on overcast or rainy days, while scanning lidar sensors provide reliable depth through focused, active NIR illumination. In rain and snow, particles affect the color image and lidar depth estimates equally through occlusion and backscatter. Adversely, in foggy conditions, state-of-the-art pulsed lidar systems are restricted to less than 20 m range at visibilities of 50 m due to severe back-scatter at the fog particles, resulting in severe pulse-broadening. So, while relying on lidar-only measurements may appear as a solution to camera-only distortions – and in fact portrays lidar systems as an essential sensory fallback system for vision-only stacks in recent real-world deployments – this is not the case for foggy scene conditions. While our work is applicable to general asymmetric failures in RGB-D object detection, we focus on foggy scene conditions as a critical application of RGB-D data beyond 3D bounding-box extraction.

We tackle robust detection in fog without having large labeled training data sets available for these scenarios. Specifically, we handle asymmetric measurement corruptions in the image or depth data by departing from existing proposal-level fusion methods: we propose an adaptive single-shot deep fusion architecture. Our fusion model exchanges feature tensors in intertwined Siamese feature extraction networks in a single-shot detection approach. This deep early fusion is gated by analysis of a temporal coherence of the measurement distortions. The proposed gated fusion allows us to learn highly adaptive models that generalize across imaging scenarios which we validate through simulation, measured data in a controlled fog chamber and real-world driving data. Moreover, we show that the use of synthetic data for adverse conditions provides only limited improvement on high-fidelity real-world scenes, even when introducing accurate measurement model for measurement distortions. The proposed single-shot fusion model runs at real-time frame-rates on consumer GPU hardware.

Specifically, we make the following contributions:

- We propose a real-time fusion network for lidar and camera measurement in fog without having labeled training data present. The proposed architecture departs from proposal-level fusion, and instead performs adaptive fusion in an end-to-end fashion while exploiting the temporal consistency of distortions.
- We derive a model that accurately characterizes image and lidar distortions in fog. We verify this model under controlled conditions in a fog chamber.
- We introduce a measured validation set of automotive road captures in fog, and a real-time implementation of the proposed model.
- We validate the proposed approach, both in simulation and on real-world captures, where it outperforms state-of-the-art fusion methods by 32% mAP in severe fog.

2. Related Work

Detection in Adverse Weather Conditions Over the last decade, seminal work on automotive data sets [3, 10, 17, 12, 60] has provided a fertile ground for pedestrian/vehicle object detection [7, 5, 58, 31, 35, 19], depth estimation [15, 34, 20], lane-detection [23], traffic-light detection [28] road scene segmentation [3, 1], and end-to-end driving models [2, 60]. While fueling this research area and promising broad adoption in self-driving vehicles, existing data sets are biased towards good weather conditions due to geographic location [60] and capture season [17], thus lacking severe distortions introduced by rare fog, severe snow and rain. A number of recent works provide camera-only data [44, 4]. However, captured data sets are very small with less than 100 captured images [44], greatly limiting their suitability for image-based fog removal, and evaluation on camera-only vision tasks. In contrast, existing autonomous driving applications that are certified for driver-less road testing rely on lidar depth-sensing fused with camera measurement and these have to be evaluated on thousands of hours of driving data. In this work, we fill this gap and provide a fusion model for camera and lidar data that is robust to unseen distortions and we provide a representative large scale evaluation data set.

Pre-processing for Fog and Haze Removal A large body of work explores methods for fog and haze removal from conventional intensity image data [62, 64, 29, 49, 32, 4, 33]. Fog results in a distance-dependent loss in contrast and color. Image processing methods aiming to recover a clear image from foggy images have not only been used for display application [22]. Preprocessing has also been proposed improve the performance of downstream semantic tasks in a pipeline fashion [44], where the input image is first pre-processed for signal enhancement before feeding the result into the downstream model. Existing fog and haze removal methods rely on scene priors on the latent clear image and depth to solve the ill-posed recovery problem. These priors are either hand-crafted [22] and used for depth and transmission estimation separately, or they are learned jointly as part of trainable end-to-end models [33, 27, 66]. Existing methods for fog and visibility estimation [52, 53] have been proposed for camera-based driver-assistance systems. In contrast to image data, existing scanning lidar systems only allow for enhancement during the acquisition, as part of the peak-finding in the returned photon-echoes [57].
making them particularly susceptible to distortions introduced through fog, rain and snow. In this work we demonstrate that current lidar-camera fusion methods are severely limited by the lidar performance in these scenarios, that is camera-only stacks indeed outperform existing fused stacks for detection tasks.

**Lidar-camera Fusion.** The information from autonomous multi-modal sensing systems is typically fused to exploit varying cues in the measurements [36], simplify path-planning [11], to allow for redundancy in the presence of distortions [38], or solve for joint vision tasks, such as 3D object detection [58]. Most existing fully autonomous sensing systems include lidar, camera and radar sensors. Although radar systems can penetrate fog, they offer only very sparse spatial resolution [17], hence allow stopping a vehicle in the presence of an obstacle but do not facilitate dense semantic scene understanding [18]. In this work, we focus on the more challenging lidar-camera fusion as the backbone of a recent line of successful research [58, 39, 7, 31]. Methods such as AVOD [31] and MV3D [7] incorporate multiple views from camera and lidar to detect objects. These methods rely on fusion of pooled regions of interest and hence perform late feature fusion following the popular region proposal architectures [42]. In a different line of research, Qi et al. [39] and Xu et al. [59] 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. These are then pooled to identify the corresponding 3D box coordinates. [30] proposes a gating mechanism for fusing different sensor streams. In all existing methods, the sensor streams are processed independently in the feature extraction stage. We show that this approach prohibits learning redundancies early in the feature extraction stage and leads to lower performance of existing fusion models compared to a single sensor input in the presence of asymmetric measurement distortions.

3. Lidar and Intensity Image Formation

Over the past few years, a large number of datasets have become available for autonomous driving applications. However, these datasets are mostly captured in clear visibility conditions and are not suitable for training deep learning models that need to perform robustly in adverse weather conditions. Capturing data in adverse weather condition is also difficult as extreme conditions are relatively rare. For example even in foggy regions, heavy fog with visibility below 50m occur up to 15 times a year [54]. Fig. 2 shows the distribution of real driving data acquired over a two-week period in Sweden covering 3500km and driven under different weather conditions such as, clear, rainy, light/dense foggy, and snow. While this data is sufficient to validate a trained model, it is not sufficient for training a deep neural network. We tackle this issue by training our models on large clean datasets with a parametrized deep-fusion approach. We validate the proposed model using real driving data and simulated environments in a fog chamber. Next, we describe the measurement distortions introduced by fog in the image and lidar measurements.

3.1. Intensity Imaging in Fog

In foggy conditions, light is scattered by the suspended water droplets before falling on the image sensor to form an image. This scattering phenomena has two primary effects. First, the chief ray is attenuated before falling on the sensor, and second, a signal floor of scattering light is present. Both effects reduce contrast, and the observed foggy image can be modeled as

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

where $I_{\text{clear}}$ is the latent clear image, depth-dependent transmissivity $t$, the global ambient component $L$, following [43, 45]. The transmission coefficient $t(x) = \exp(-\beta d(x))$, 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, see Supplemental Material.

3.2. Pulsed Lidar in Fog

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

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

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. Note 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 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 [65], yielding the maximum distance of

$$d_{\text{max}} = \ln \left( \frac{n}{L_{\text{foggy}} + g} \right) \beta. \quad (3)$$

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. (2) and gain. Hence, in fog, lidar measurements suffer not only from loss in peak intensity. In addition back-scattering may result in a peak-shift inside of the fog volume and thus all information on the target scene point is lost. Fig. 3 shows a camera-lidar measurement in dense fog. Although visibility in this scenario can be accurately estimated [52, 53], and hence we can quantify the lidar and camera distortions accurately according Eq. (3) and Eq. (1), we show that existing RGB-D fusion methods fail under these asymmetric distortions.

4. Adaptive Deep Fusion

In this section, we describe the proposed adaptive deep fusion architecture that allows for RGB-D fusion under unseen asymmetric sensor distortions. Our architecture is designed under the constraint of real-time processing required for self-driving vehicles and autonomous drones. Specifically, we propose an efficient single-shot fusion architecture, adopting the single-shot detection architecture [35] as the backbone architecture for fused camera and lidar depth data.

4.1. Deepified Single-Shot Fusion

The proposed network architecture is shown in Fig. 4. It consists of two single-shot detection branches, one with color intensity images input and the other with the lidar measurements as input. While the RGB branch uses conventional three-plane RGB inputs, for the lidar branch, we depart from recent bird-eye-view (BeV) projection [31] or raw point-cloud input [59], and use lidar depth projected into the image space. We do this because 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 RGB features. Hence, existing fusion methods can only fuse features in a lifted space, subsequent to matching region proposals, but not earlier. Instead of a naive depth-only input encoding, we provide depth, height and pulse intensity as three feature layers to the lidar network. This encoding allows for position and intensity-dependent fusion. Pulse-broadened low intensity measurements as well as missing measurements, which we encode with zero intensity, are hence represented early in the feature extraction stack. The feature extraction network in both color and lidar streams is a modified VGG16 [50] feature extraction stack. In particular, we rely on a reduced stride of 3 for block 9, which we found essential to increase the feature map sizes for large object sizes. While we adopt the SSD detection head with 7 detection layers from [35], we optimize the anchor positions as described in Sec. 4.3. The described network architecture and input representation allows us to perform deepified adaptive feature fusion in the feature extraction stack itself. As shown in Fig. 4, each feature extraction layer is concatenated with the corresponding layer for the other respective sensory input. This approach increases the input depth to each block by a factor of two, while the output depth remains constant. This deep concatenation allows for intensity, position and color cues to be fused early in the feature extraction process. We demonstrate that this deepified fusion outperforms proposal-level fusion even for larger feature extraction backbones, while being computationally cheap.

4.2. Adaptive Fusion via Drop-Out

To adapt the feature level fusion to unseen distortions of either the RGB or lidar stream, we introduce a scalar fusion parameter $\rho \in [0, 1]$. This parameter allows us to specialize the fusion process during inference for unseen distortions of one of the two sensory streams. Specifically, we concatenate an additional feature layer with scalar value $\rho$ for each pixel to every deep feature exchange block. During training, we sample random drop-out probabilities $\rho$, set the adaptive fusion weight to this value, and drop-out the camera branch with probability $\rho$ and the lidar branch with probability $1 - \rho$, modeling asymmetric distortions without training data. This adaptive drop-out drives the scalar fusion parameter to be correlated with increasingly less reliable sensor data. During inference, we do not use drop-out, but set the adaptive fusion weight to the output of the visibility estimation model. The parameter can either be set
Figure 4: Overview of our architecture consisting of two single-shot detector branches with deep feature exchange and adaptive fusion of lidar and camera. A lidar projection to a 3-channel (height, distance, intensity) image and an RGB color image are the input. The noise channel (red) is concatenated to the deepified layers (white) which interchange information (blue) with parallel computing blocks. Bounding boxes are detected at the given feature maps by SSD blocks (brown).

Table 1: Optimized Anchor boxes for the proposed model.

| Anchor Boxes | (x, y, w, h) |
|--------------|-------------|
| BLOCK4       | (28.33, 34.02, 34.76, 44.35), (30.94, 60.00), (50.0, 50.0) |
| BLOCK7       | (39.99, 94.49, 44.54, 55.99), (60.0, 40.0), (56.03, 72.95), (52.65, 128.12) |
| BLOCK8       | (80.0, 80.0), (70.70, 97.79), (90.0, 90.0), (115.0, 115.0) |
| BLOCK9       | (93.24, 133.29), (73.02, 174.52), (130.0, 130.0) |
| BLOCK10      | (39.99, 94.49), (44.54, 55.99), (60.0, 40.0), (56.03, 72.95), (52.65, 128.12) |
| BLOCK11      | (150.0, 150.0), (153.60, 212.16), (168.37, 130.0), (138.41, 168.37) |
| BLOCK12      | (180.0, 180.0), (180.34, 308.03), (250.0, 250.0) |

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. To optimally represent the training data distribution, we use the K-means clustering algorithm with the intersection over union distance metric adopted from [41]. The training bounding boxes are clustered based on their width and height, and the optimal number of anchor boxes per feature layer was found by a search over different configurations resulting in a total of 24 anchor boxes. 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. To increase the performance further, an additional single-stride feature map was added for the 9th SSD block. The final bounding box distribution per feature map is reported in Table 1.

In total, each anchor box is trained using the cross entropy loss with softmax,

\[
H(p) = \sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)).
\]  

The total number of negative anchors is restricted to 5 times the number of positive examples using hard example mining [35, 48]. The single network training procedure is initialized with pretrained ImageNet VGG weights. The network is then fine-tuned using the ADAM optimizer with one training phases, with a constant learning rate of 0.0005. To further regularize the network, L2 weight decay of 0.0005 is used. The fused networks differ in weight initialization and learning rate as their pretrained weights are loaded from the single networks stacks and the common feature maps are fine-tuned with a lower learning rate of 0.00025. The training split are adopted from [6], and these are consistent across all experiments.

5. Datasets

While we train the proposed model on image and lidar data with clear visibility, we evaluate the model performance on data captured in experimental and simulated fog scenarios. Assessing the proposed method on measured fog data covers realistic weather and lighting conditions, which are not repeatable, whereas simulated data with known ground truth allows for repeatable quantitative evaluation for a range of different fog densities on the same scenes. The evaluation and training datasets are described in detail below.

5.1. Measured Dataset

The experimental consists of, 1725 labeled images captured under controlled conditions in a fog chamber, as well
We have equipped a test vehicle with a monocular front-facing RGB camera, that is an Aptina AR0230 imager with a resolution of $1920 \times 1024$. For range data acquisition, we use a Velodyne HDL64 S3D laser scanner. Both lidar and camera data are time synchronized and ego motion corrected utilizing a proprietary IMU. The overall system provides a sampling rate of 10Hz implemented in ROS [40]. The laser scanner provides the strongest and the last echo from a reflected signal. As the sensor characteristic is different to the Velodyne HDL64 S2 used in [17], fine-tuning on clean measured data is required for domain adaptation.

**Driving Captures** The driving dataset was captured over a two-week-long test drive in northern Germany, Sweden and Denmark covering a distance of 3500 km under different weather and lighting conditions. A total of 1.1 million images were collected at a framerate of 10Hz. To show the weather distribution every 300th frame was labeled by a human annotator in Fig. 2. We use a random subset of 1035 completely clear images for domain-adaptation via fine-tuning, and a subset of 4054 in fog and snow for evaluation. Example data for various distortions are shown in the supplementary material.

**Fog Chamber Captures** To collect image and range data in controlled conditions, we also provide a fog chamber evaluation measurements. Details on the fogchamber setup can be found in [14, 9]. This provides us with crucial validation data as real world fog densities are hard to measure in real world environments. We captured in total approx. 35k frames at a framerate of 10Hz and labeled 1745 images, under 4 different illuminations and 6 fog densities resulting in visibilities $V$ of 10 m to 60 m ($V = -\log(0.05)/\beta$). Visualization and evaluation result for the fogchamber measurements can be found in the supplemental document.

### 5.2. Synthetic Dataset

We have presented forward models for lidar and intensity image data in Sec. 3. We use these models now to simulate measurement data under fog using clear KITTI [17] data as input. Simulating intensity images in fog (Eq. 1) requires a dense depth map per image. To this end, we used the Pyramid Scene Parsing network [65] to obtain high quality depth maps. As a preprocessing step, we apply [61] to image data on the LAB color space to remove residual haze and lower the general image brightness. For lidar, the precise depth is available in the KITTI lidar measurement and the proposed back-scatter model from Eq. 2. In addition, we also model typical pointcloud wobbling distortions analogous to lidar light rays being scattered by exhaust gases [21]. To account for this typical behavior, the visibility is periodically changed over time for the azimuth and height, which is described in detail in the supplementary material. All simulation parameters, including noise floor, gain, and the wobbling frequency, are chosen to qualitatively match the behavior visible in the prototype measurements from Sec. 5.1. While published results, e.g. [45], uses fog attenuation coefficients 0, 0.005, 0.01, and 0.02, while only providing captured 100 test images, we extend this range to cover 0.04, 0.06, 0.08, and 0.16, and provide an order of magnitude more evaluation data.

### 6. Assessment

In this section, we validate the proposed deep fusion model on unseen experimental and simulated test data. From a signal-processing perspective, an immediate approach to handling foggy image streams seem to preprocess images to reduce fog and restore contrast. Next, we will discuss this baseline approach, then present methods for fog-density estimation from image data. Finally, we will demonstrate qualitative and quantitative results of the proposed method for varying fog levels ranging from light mist to dense fog and heavy snow.

**Image Enhancement** We demonstrate that it is possible to take a simulated foggy image and obtain a good quality high-resolution de-hazed image with a few minor extensions of existing image-to-image translation techniques. To this end, we adopt the popular Pix2PixHD [55] method, which is a recently proposed 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 [33] (Pix2Pix2HD-AOD). We trained all the models using the simulated fog dataset and report results in Tab. 2 for various image quality metrics, that measure distance to the known ground truth, including LPIPS [63], MSE, PSNR, SSIM [56], and VIF [47]. In simulation, the proposed extensions improve on vanilla Pix2PixHD in the VIF metric, while being on par in other metrics. However, the proposed variant adds substantial robustness on experimental data. Fig. 6 shows qualitative image reconstruction results on measured data. The proposed variant suffers from less artefacts and achieves relatively stable fog removal and contrast enhancement. Please see the supplemental material for additional discussion and results.

Table 2: Image quality scores of image-only fog removal (image size 1248 × 384) on synthetic KITTI data.

| Network     | n   | 0.00 | 0.005 | 0.01 | 0.02 | 0.04 | 0.06 | 0.08 | 0.16 |
|-------------|-----|------|-------|------|------|------|------|------|------|
| VGG [50]    | 1   | 0.999| 0.999 | 0.999| 0.999| 0.998| 0.992| 0.996| 0.999|
| SqueezeNet [26] | 10  | 0.922| 0.592 | 0.511| 0.678| 0.692| 0.490| 0.630| 0.992|
| SqueezeNet [26] | 10  | 0.936| 0.700 | 0.791| 0.869| 0.788| 0.641| 0.794| 0.997|

Table 3: Fog density classification using n temporal images. Evaluated classwise based on mAP [37].

Fog Density Estimation The fog density parameter \( \beta \) can be very accurately estimated from image data only. Tab. 3 shows estimation results on synthetic KITTI data, recall Sec. 5.2. For these evaluations, we took models pretrained on ImageNet and fine-tuned for fog density estimation with an identical split as in [6] for all images where a history of at least 10 frames was available. Using temporal image sequences as input – in contrast to a single frame – is beneficial for small, efficient networks, such as SqueezeNet [26], with a margin of 30% over single image density estimation.

6.1. Robust Fusion in Fog

Next, we validate the proposed approach, which we dub Adaptive Deepified SSD, on both simulated and real fog data. We report mean Average Precision for three different difficulty levels easy/moderate/hard as specified in the KITTI evaluation framework [17]. We also consider only cars as object category since there are significantly more car instances present in the KITTI dataset and our Sweden dataset. We compare the proposed model against recent state-of-the-art automotive fusion models, including AVOD, AVOD-FPN, MV3D, Frustum Nets, and variations of the proposed method. As baseline variants, we implement two fusion and two single sensor detectors. In particular, we compare against late fusion with image and lidar features concatenated just before bounding-box regression (which we dub Fusion SSD), and early fusion by concatenating image and lidar point projected into image space (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 an image-only input (Image-only SSD), and an identical SSD architecture with the projected lidar-only input (Lidar-only SSD). Note that these models are trained with same hyper-parameters and anchors as the proposed method.

Synthetic Evaluation Tab. 4 shows detection results for synthetic data at various different fog density levels. On clear data, AVOD-FPN has the best performance as it utilizes multiple views from the top and front, a feature pyramid and a region-proposal approach (in contrast to the single shot architecture [25] of the proposed method). However, as soon as fog is introduced the performance de-
creases rapidly along with the Lidar-only SSD and Concat SSD. MV3D even drops below the single image SSD method. Two-stage methods, including [39], drop quickly but asymptotically achieve slightly higher results compared to AVOD and MV3D as the statistical priors learned for the first stage are not disturbed. Concatenating image and lidar features, as implemented in Fusion SSD, performs better than existing methods, but still ranks below the proposed fusion layer. Two-stage methods, including [39], drop quickly but asymptotically achieve slightly higher results compared to AVOD and MV3D. The proposed adaptive fusion layer results in significant margins over the model without adaptive fusion (Deepified SSD), while being on-par for very light fog.

**Experimental Evaluation**

We validate the simulation results from the previous section on the experimental test data described in Sec. 5.1. To account for the domain shift, we fine-tune all proposed models on clear weather acquired with the same camera system, and test it on unknown fog data. Tab. 5 summarized the evaluation results, split into two different fog levels. The fog densities have been labeled by human annotators on camera images estimating the approximate maximal viewing distance. The proposed method outperforms all baseline approaches. Compared to the best naive fusion approach, Concat-SSD, it improves by a margin of of 32.5%, and in dense fog it improves by a margin of 12.5% on the feature fusion variant. Consistent with the simulation results, especially under severe distortions, the proposed adaptive fusion layer results in significant margins larger than 6% over the model without adaptive fusion.

We also validate that image enhancement methods and domain adaptation methods do not perform well on experimental data. Even though models trained on simulated data can generalize to real-world scenes to a certain extent, residual artefacts are often present, see Fig. 6. As a result, image enhancement does only slightly improve detection scores in Tab. 5. In addition, we demonstrate the effect of using re-

---

**Table 4**: mAP scores for different models (row) at varying fog densities (columns) on synthetic KITTI data. Best SSD models are labeled magenta and second best blue. The image pre-processing comparisons use the Pix2PixHD-CJ (best from Tab. 2) for fog removal before object detection. Note that these results do not generalize to real-world data, see Tab. 5.

**Table 5**: Quantitative detection mAP on unseen fog data from the Sweden dataset split into different distortion levels and difficulties easy/moderate/hard [17]. The proposed model is trained solely on clean data without distortions. The best model is marked magenta and the second-best in blue. The image pre-processing comparisons use the Pix2PixHD-CJ (best from Tab. 2) for fog removal.
do not model distortions different, but semantically identical scenes. These methods directly apply as they model a mapping between stylistically domain adaption methods on the image data. Please note that existing domain adaption methods cannot be di-
verse transfer. The reverse transfer does not recover enough information to model a clear scene correctly.

cent domain adaption methods on the image data. Please note that existing domain adaption methods cannot be di-
rectly applied as they model a mapping between stylistically different, but semantically identical scenes. These methods do not model distortions, such as fog, which we validate in Fig. 8, using the recent CyCADA [24]. We have setup Cy-
cada here to learn a mapping from KITTI data to our experi-
mental scenes. Using simulated foggy KITTI data, one may then generate large amounts of foggy domain-transformed training data. However, due to the lack of modeling dis-
tortions, Tab. 5 shows that training on this domain-adapted data does not improve detection performance. Finally, note also that adapting simulated fog to measured fog data requires again a large fog data corpus – the problem at hand.

7. Conclusion and Future Work

In this paper, we address a critical problem in au-
tonomous driving: multi-sensor detection in scenarios where labeled data is sparse and difficult to obtain. We propose a deep model that learns to fuse camera and lidar data, and allows for robust, real-time object detection under fog. Based on deep feature exchange and adaptive fusion via drop-out, our model can be trained on easily available clear data, and can not only generalize to foggy conditions, but outperform existing fusion models under different lev-
els of fog. We validate the proposed approach in simulation and on an extensive real world driving dataset, which we will publish. Interesting directions for future research include the application of the proposed adaptive fusion model to other unseen distortions such as debris, sand, local occlu-
sions and hardware-induced partial and full-sensor failures.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561, 2015. 2
[2] 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
[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Seg-
mentation and recognition using structure from motion point clouds. In European conference on computer vision, pages 44–57. Springer, 2008. 2
[4] 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
[5] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In European Conference on Computer Vision, pages 354–370. Springer, 2016. 2
[6] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals using stereo imagery for accurate object class detection. In Arxiv, 5, 7
[7] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-View 3D Object Detection Network for Autonomous Driving. In CVPR, nov 2017. 2, 3, 8
[8] 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 Automa-
tion (ICRA), 2014 IEEE International Conference on, pages 1836–1843. IEEE, 2014. 1
[9] M. Colomb, K. Hirech, P. Morgane, B. J.J., P. Lacote, and D. J. Innovative Artificial Fog Production Device, a Technical Facility for Research Activities. 2004. 6
[10] 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 Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2
[11] 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
[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedes-
trian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4):743–761, 2012. 2
[13] A. Dudhane and S. Murala. C2MSNet: A Novel approach for single image haze removal, jan 2018. 7
[14] P. Dutton, F. Bernardino, F. Chausse, and M. Colomb. Methodology Used to Evaluate Computer Vision Algorithms in Adverse Weather Conditions. In Transportation Research Procedia, volume 14, 2016. 6
[15] 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. 1, 2

Figure 8: Qualitative Domain Adaptation Results using CyCADA[24]. Adapting KITTI (left) to our experimental data adapts to winter scenes with snow, but does not properly model fog distortions (middle). See Tab. 5 for detection scores in the last row. For completeness, (right) shows a reverse transfer. The reverse transfer does not recover enough information to model a clear scene correctly.
