Benchmarking object detection robustness performance in adverse weather conditions using scenario attentional score aggregation

Hanyue Liu†, Linze Li†, Hewen Deng, Jing Tian, Cheung-Chi Leung

NUS-ISS, National University of Singapore, Singapore 119615
Email: hanyue.liu@u.nus.edu, e1221775@u.nus.edu, hewen.deng@u.nus.edu, tianjing@nus.edu.sg, cc.leung@nus.edu.sg
† *These authors contributed equally to this work.*

Abstract

Accurate object detection under adverse weather conditions is critical for autonomous vehicles to effectively perceive information from the surrounding environment. Despite the notable progress in object detection models, there remains an absence of a comprehensive framework for assessing the robustness of these models in adverse weather scenarios. This paper proposes a new object detection robustness performance evaluation protocol in adverse weather conditions including (i) a new benchmark dataset and (ii) a new score aggregation method. Firstly, we construct a new dataset, which contains 12,000 images and 24 weather scenarios including four severities of fog, rain, and snow. Secondly, in contrast to traditional score aggregation methods that treat various scenarios *equally*, we propose a new score aggregation method, called *scenario attentional score aggregation* (SASA), to assess models' overall robustness performance. It not only considers the amount of test images in experiments but also considers the severity of weather conditions. We evaluate three representative object detection models using the proposed performance evaluation protocol. We release our dataset and associated code at https://github.com/hanyuesgithub/ObjectDetection-FRAS for advancing the development of robust object detection models in challenging adverse weather conditions.

Key words: object detection; robustness; adverse weather; score aggregation

1. Introduction

Object detection is a critical and fundamental technology in computer vision-based autonomous driving solutions [1–4]. Its primary objective is to localize and recognize objects in the surrounding environment, enabling autonomous systems to make informed navigation decisions. In recent years, deep learning-based techniques have achieved significant advancements in object detection for autonomous driving, owing to two primary benefits, including improved accuracy and real-time processing capabilities [5].

However, current deep learning-based object detection techniques are usually developed and trained using images captured under normal weather conditions [6]. These typical scenarios do not account for the diverse and challenging conditions encountered in real-world driving, such as adverse weather conditions including fog, rain, and snow. Consequently, object detectors trained on these normal scenes suffer from significant performance degradation when exposed to adverse weather.

The need for robust object detection performance in adverse weather conditions is critical for two main reasons. First, the safety of autonomous vehicles and their occupants depends on the reliable detection of objects under all driving conditions. Second, the ability to maintain high detection performance in adverse weather is essential for the widespread adoption and public trust in autonomous driving technologies. Robust object detection ensures that autonomous vehicles can navigate safely and effectively, regardless of the environmental conditions they face [7, 8].

To address the aforementioned challenges, this paper proposes a comprehensive evaluation framework specifically tailored for object detection under adverse weather conditions. More specifically, the contributions of this paper are two folds.

- Considering that there is a lack of a benchmark dataset of object detection models under adverse weather conditions, we develop a new dataset, called *KITTI-FRAS*, which considers three representative types of adverse weather conditions (i.e., fog, rain, and snow). For each type of weather condition, two different image augmentation techniques are employed to generate images with four intensity levels.

- Leveraging on the new dataset *KITTI-FRAS*, we propose a unified evaluation protocol to evaluate the performance of object detection models by developing a new score aggregation method that combines performance scores of various scenarios to provide a unified performance for object detection models.

This paper is organized as follows. Section 2 provides an overview of the related works on datasets and evaluation methods under adverse weather condition. Section 3 presents the new dataset *KITTI-FRAS*, followed by the new evaluation protocol proposed in Section 4. Then, three representative object detection models are evaluated in Section 5 using the proposed evaluation protocol. Finally, Section 6 summarizes our work and findings.

2. Related works and motivation of our works

This section presents a brief overview of related work on object detection benchmark datasets and evaluation methodologies. In addition, the motivations for our work are emphasized following each review.

2.1. Related works on datasets

Table 1 summarizes a list of relevant object detection datasets. The detailed descriptions are provided as follows.

- *KITTI* [9] is one of the most popular datasets in mobile robotics and autonomous driving. The dataset encompasses diverse scenarios, capturing real-world traffic situations across freeways, rural areas, and inner-city scenes with numerous static and dynamic objects. The dataset was captured during clear daylight hours and therefore does not include bad weather conditions.

Table 1: A comparison between the existing benchmark datasets and our new dataset developed in this paper.

Dataet	# Images	Image scenes	Adverse weathers	# Weather intensities	Real or synthetic	# Simulation methods
KITTI [9]	7,481	city, rural area, highway	No	-	real	-
Cityscapes [10]	25,000	city	No	-	real	-
Weather KITTI [11]	7,481	city, rural area, highway	fog, rain	7	synthetic	1
Foggy Cityscapes [12]	25,000	city	fog	3	synthetic	1
Snow100K [13]	100,000	scene from Flickr	snow	3	synthetic	1
DAWN [14]	1,000	city, highway	fog, snow, rain, sandstorms	-	real	-
MSLS [15]	1,680,000	city, rural area, highway	fog, rain, snow	-	real	-
Oxford RobotCar [16]	20,000,000	city	rain, snow	-	real	-
BDD100K [17]	120,000,000	city, rural area, highway	fog, rain, snow	-	real	-
Our *KITTI-FRAS*	12,000	city, rural area, highway	fog, snow, rain	4	synthetic	2

- *Cityscapes* [10] is a large-scale database which focuses on semantic understanding of urban street scenes. The dataset consist of around 5000 fine annotated images and 20000 coarse annotated ones. Data was captured in 50 cities during several months, daytimes, and good weather condition, without any adverse weather conditions.

- *Weather KITTI* [11] contains the rainy and foggy augmentation of KITTI dataset. For each sequence, the dataset provides more than 7 rain levels (from dizzle to storm conditions) and 7 fog intensities (from light to dense fog).

- *Foggy Cityscapes* [12] is a synthetic foggy dataset that simulates fog on real scenes. Each foggy image is rendered with a clear image and depth map from Cityscapes, containing three fog severity levels. Each severity level is characterized by a constant attenuation coefficient which determines the fog density and the visibility range.

- *Snow100K* [13] consists 100k synthesized snowy images and corresponding snow-free ground truth images downloaded from Flickr. This dataset consists of three subsets as per the variations inside single image: Snow100K-S, Snow100K-M and Snow100K-L. Each subset contains around 33k images.

- *DAWN* [14] emphasizes a diverse traffic environment (urban, highway and freeway) as well as a rich variety of traffic flow. The DAWN dataset comprises a collection of 1000 images from real-traffic environments, which are divided into four sets of weather conditions: Fog, snow, rain and sandstorms. The dataset is annotated with object bounding boxes for autonomous driving and video surveillance scenarios.

- *MSLS* [15] is a large and diverse dataset for lifelong place recognition from image sequences in urban and suburban settings. It contains more than 1.6 million images from 30 major cities across six continents, with all images tagged with sequence information and geo-located with GPS and compass angles. The dataset spans all seasons over a nine-year period, capturing different weather conditions, cameras, daylight variations, and structural settings.

- *Oxford RobotCar* [16] contains over 100 repetitions of a consistent route through Oxford, UK, captured over a period of over a year. The dataset captures many different combinations of weather, traffic and pedestrians, along with longer term changes such as construction and roadworks.

- *BDD100K* [17] contains 100K videos and 10 tasks to evaluate the exciting progress of image recognition algorithms on autonomous driving. The dataset possesses geographic, environmental, and weather diversity.

Motivation for a new dataset: Our new dataset *KITTI-FRAS* is different from the aforementioned datasets in the following two aspects. Firstly, datasets *DAWN* [14], *MSLS* [15], *Oxford RobotCar* [16], and *BDD100K* [17] contain images taken from real-world adverse weather scenarios without detailed annotations of weather severity levels. Consequently, they provide a limited description of adverse weather conditions for model evaluation. This limitation restricts their utility as benchmark datasets for evaluating object detection models under adverse weather conditions. Secondly, datasets *Weather KITTI* [11], *Foggy Cityscapes* [12], and *Snow100K* [13] provide weather severity information; however, they only provide a limited range of adverse weather types. Furthermore, these datasets employ only a single method for simulating adverse weather, resulting in limited data diversity.

On the contrary, our *KITTI-FRAS* dataset address these limitations by employing two distinct simulation methods to generate fog, rain, and snow weathers. Each weather condition is further categorized into four severity levels (see Section 3 for details). This comprehensive simulation enhances the diversity and robustness of the dataset, making it a more effective benchmark for evaluating object detection models under various adverse weather scenarios.

2.2. Related works on evaluation protocol

The object detection performance evaluation could be conducted on individual experiment or multiple experiments. For the individual experiment, *Intersection over Union* (IoU) is a popular metric for quantifying the overlap between two sets. Then, *Average Precision* (AP) and *mean Average Precision* (mAP), incorporating trade-offs between precision and recall, serve as standard metrics for evaluating object detection algorithms, in benchmark datasets such as PASCAL VOC [18] and COCO [19]. For the multiple experiments, score aggregation plays a critical role in evaluating algorithm performance. Existing score aggregation methods, such as simple averaging, weighted averaging, and voting, have been widely applied in numerous studies. For example, in the object detection task of YOLO [20], average scores are used to assess the overall performance of the model. Additionally, a weighted average scores to evaluate the performance of detection models is presented in [19]. However, studies that use weighted average scores and voting methods to evaluate object detection models are relatively rare. The concept of weighted average scores is illustrated in [21], which uses the *Weighted Boxes Fusion* (WBF) method to fuse detection boxes with a weighted average score and evaluate model performance. Similarly, the *Non-Maximum Suppression* (NMS) technique [22] applies the voting concept by selecting the best bounding box among overlapping ones based on their scores.

Motivation for a new evaluation protocol: While these aforementioned metrics are effective for evaluating object detection under general conditions where all scenarios are treated *equally*, their application in assessing a model's robustness under adverse weather conditions involving varying scenarios is less emphasized. To address this limitation, we propose a new score aggregation method. This method not only considers the amount of images in the test dataset but also considers the severity level of weather conditions. This ensures that the aggregated results are fairer and more representative. Experimental results on various object detection models and simulated datasets demonstrate that our approach accurately reflects the actual performance of algorithms.

3. Proposed new benchmarking dataset

To construct the new dataset *KITTI-FRAS*, we leverage the KITTI dataset by applying two different image augmentation methods for fog, rain and snow, to simulate realistic weather effects. For each category, we create four discrete weather severity levels, including light, moderate, heavy and severe. The details are provided in the following sections.

3.1. KITTI-FRAS: Fog

To generate high-quality synthetic fog images, we apply two representative methods [23, 24], where the results are illustrated in Figure 1.

The first method is a physics-based optical model [23], which has been commonly applied in [25–27] and can be mathematically formulated as

$$I(x) = I_0(x)t(x) + L_\infty(1 - t(x)), \tag{1}$$

where $I(x)$ is the observed foggy image at the pixel x, $I_0(x)$ is the clear scene radiance and L_∞ is the horizon radiance or the atmospheric light. Furthermore, the transmission $t(x)$ determines the amount of scene radiance that reaches the camera as

$$t(x) = e^{-\alpha d(x)}, \tag{2}$$

where α is an extinction coefficient and $d(x)$ is the distance the light travels through the fog. To simulate fog utilizing this optical model, depth images are employed to represent the parameter $d(x)$. The controlled variable α is set to 2, 4, 6, and 8, corresponding to four distinct levels of severity: light, moderate, heavy, and severe.

The second method for simulating fog images involves the utilization of a generative model [24], which has the same conceptual framework of CycleGAN, to superimpose fog effects onto clear images. The controlled variable *intensity* is set to 0.1, 0.3, 0.5, and 0.7, corresponding to four distinct levels of fog severity.

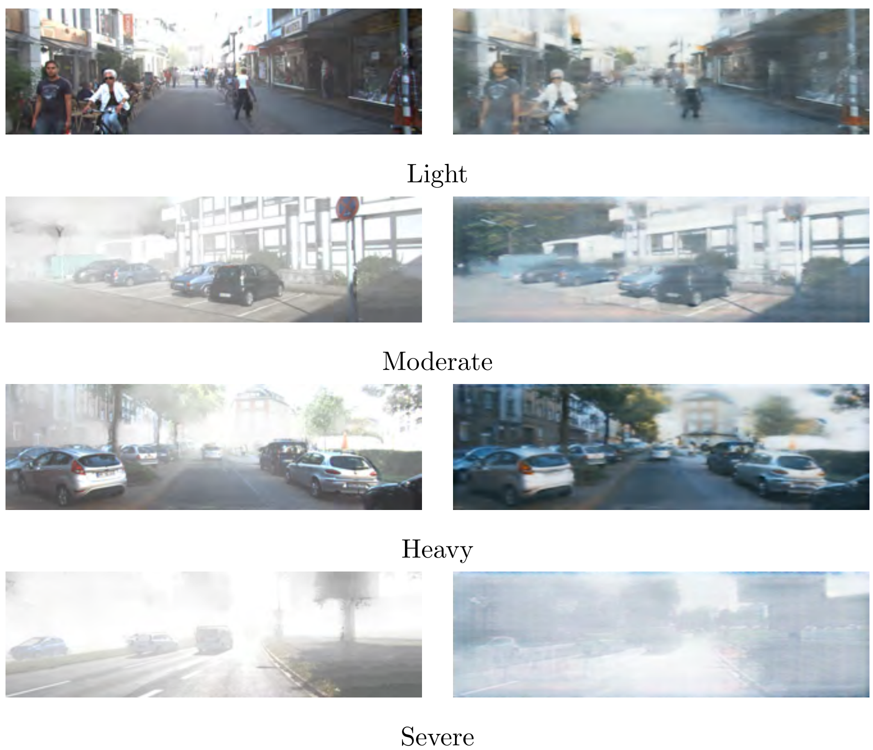

Figure 1: Examples of generated *fog* images with four intensity levels using two methods [23] (left column) and [24] (right column).

3.2. KITTI-FRAS: Rain

To generate high-quality synthetic rain images, we apply two representative methods [11, 28], where the results are illustrated in Figure 2.

The first method combines physical models and image-to-image translation to create visually convincing rain simulations. We explore two realistic rendering approaches including a physics-based rendering method and a hybrid model combining a GAN-based approach with physics-based technique. The physics-based rendering technique [11] simulates the appearance of rain in images by estimating scene depth and overlaying fog-like attenuation layers with individual rain streaks. For the fog-like attenuation layers, we render the volumetric attenuation using the model described in [29], where the per-pixel attenuation $I_{\text{att}}(x)$ is expressed as the sum of the extinction $L_{\text{ext}}(x)$ caused by the volume of rain, and the airlight scattering $A_{\text{in}}(x)$, which results from the environmental lighting as

$$I_{\text{att}}(x) = IL_{\text{ext}}(x) + A_{\text{in}}(x), \tag{3}$$

$$L_{\text{ext}}(x) = e^{-0.312R^{0.67}d(x)}, \tag{4}$$

$$A_{\text{in}}(x) = \beta_{HG}(\theta)\overline{E}_{\text{sun}}(1 - L_{\text{ext}}(x)), \tag{5}$$

where R denotes the rainfall rate R (in mm/hr), $d(x)$ the pixel depth, β_{HG} represents the standard Heynyey-Greenstein coefficient, and $\overline{E}_{\text{sun}}$ represents the average sun irradiance which we estimate from the image-radiance relation[30]. In the rain streak rendering process, utilizing the rain streak database [31], we achieved a more realistic raindrop effect by selecting the streak S from the streak database $\mathcal{S}$, and then wrap it as $S' = \mathcal{H}(S)$ to match the drop dynamics from the physical simulator, where $\mathcal{H}$ is the homography computed from the start and end points in image space given by the physical simulator and the corresponding points in the database streak image. Generative adversarial networks (GANs)[32] further enhance realism by learning visual characteristics such as wetness and reflections. Initially, images are translated into rainy versions using GANs, followed by [11] to overlay rain layers, resulting in comprehensive and nuanced rainfall simulations.

In the second method, we apply Adobe After Effects' Simulation-CC Rainfall tutorial [28] to successfully simulate raindrops with various properties, including size, speed, opacity, and scene depth.

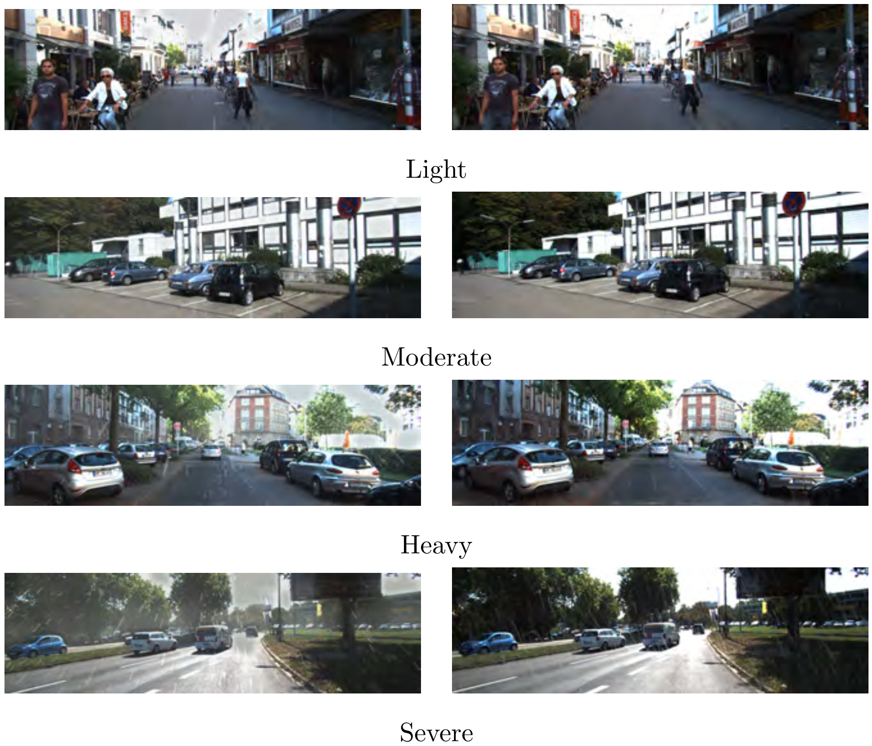

Figure 2: Examples of generated *rain* images with four intensity levels using two methods [11] (left column) and [28] (right column).

3.3. KITTI-FRAS: Snow

To generate high-quality synthetic snow images, we apply two representative methods [28, 33], where the results are illustrated in Figure 2.

In the first method, snowflakes with different attributes (namely snowflake size, density, scene depth, and falling speed) are simulated based on Adobe After Effects' Simulation-CC Snowfall synthesis tutorial [28]. Additionally, Automold library [34] is employed to better simulate real-world snow scenes. The Automold Road Augmentation library searches for pixels in images with brightness values less than or equal to a specified threshold (one value chosen per image) in the hue, saturation, and lightness (HLS) color space and multiplies their brightness by a sampling factor (once per image) to simulate the accumulation of snow in the environment through brightness augmentation as

$$outputpixel(i,j) = \begin{cases} inputpixel(i,j) \times samplingfactor, & intensity \leq threshold; \\ inputpixel(i,j), & otherwise; \end{cases} \tag{6}$$

where $inputpixel(i,j)$ represents the original pixel intensity value at position (i,j) in the image, $threshold$ is the specified threshold value that changes according to the strength of the snow cover, the $samplingfactor$ is the factor used for brightness augmentation.

In the second approach, we introduce an enhanced approach inspired by the methodology outlined in [33]. We utilize hexagonal crystals to simulate snowflake morphology, distributing them across the canvas. To simulate the aggregation effect of multiple snowflakes, we randomly select 10% of the snowflakes each time for mutual attachment. The attachment process iterates randomly between 1 to 5 times. Additionally, snowflake intensity is adjusted based on the size of the snow particles, and motion blur is applied to simulate descent. Finally, Gaussian blur was used to add fuzziness, simulating snowy conditions.

$$\begin{aligned} Snowflake_k &= \bigcup_{k=1}^{iterations} Hexagram_i(center_k, size_k), && (7) \\ center_k &= random(p \times Snowflake_{k-1}(random(vertex)), && (8) \\ size_k &= size \times random(0.5, 3.0), && (9) \\ iterations &= random(1,5), && (10) \end{aligned}$$

where $Snowflake$ represents the final generated snowflake, $Hexagram$ is a function used to draw a hexagram with the specified center $center_k$ and size $size_k$, and p represents the percentage of new snowflakes randomly selected from $Snowflake_{k-1}$.

4. Proposed new evaluation protocol

To provide a comprehensive performance evaluation of object detection model across various simulation, this section proposes a new score aggregation method,

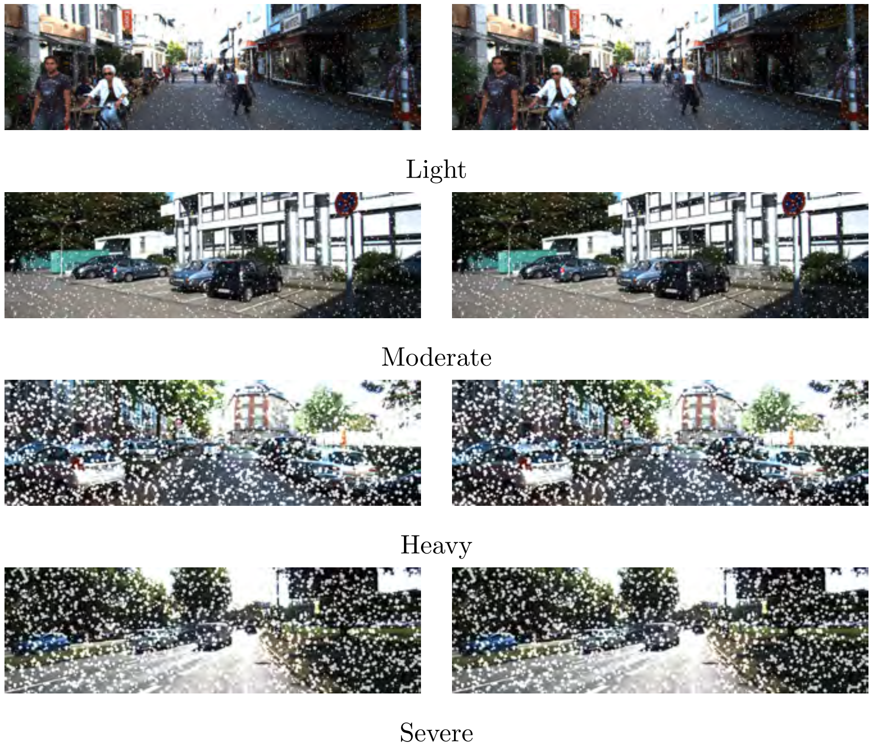

Figure 3: Examples of generated *snow* images with four intensity levels using two methods [28] (left column) and [33] (right column).

called *Scenario Attentional Score Aggregation* (SASA). This new method not only considers the amount of images in the test dataset but also considers the severity of weather conditions. By this way, it *adaptively* applies different weights for scores in various scenarios and combines them in a weighted manner. This is different from traditional methods that consider various simulations *equally* and simply averages their scores.

Considering a general performance evaluation, where M simulation methods are employed to simulate an adverse weather, such as fog; and each method produces N weather datasets of discrete severity levels, spanning from light to severe. The $M \times N$ dataset matrix, denoting a specific adverse weather, can be structured as in Table 2. For each experiment test T_{MN}, we can evaluate the mAP performance of the object detection model. Then the objective is to combine the mAP scores across all experiment tests in Table 2 to obtain a summarized score. For that we could aggregate scores based on specific simulation method (row) or specific severity level (column).

Table 2: An overview of experimental setup of a specific weather simulation. The objective is to combine the mAP scores across all experiment tests to obtain a summarized score.

Simulation	Severity level			
method	1	2	...	N
1	T_{11}	T_{12}	...	T_{1N}
2	T_{21}	T_{22}	...	T_{2N}
...	...	...	...	...
M	T_{M1}	T_{M2}	...	T_{MN}

4.1. Score aggregation based on specific simulation method

Table 3: An illustration of score aggregation based on specific simulation method m.

Index k	1	2	...	$N+1$	...	2^N-1
	$\{T_{m1}\}$	$\{T_{m2}\}$	...	$\{T_{m1}, T_{m2}\}$	...	$\{T_{m1}, T_{m2}, \ldots, T_{mN}\}$

For a given simulation method m, we establish the various experiment tests combinations, such as $\{T_{m1}\}$, $\{T_{m2}\}$, $\{T_{m1}, T_{m2}\}$, $\{T_{m1}, T_{m2}, \ldots, T_{mN}\}$, as illustrated in Table 3. For each test k, we can obtain the mAP score of the object detection model. Then, all these 2^N-1 mAP scores can be aggregated as

$$\text{SASA}(m) = \sum_{k=1}^{2^N-1} \left(\frac{w_i(k) + w_s(k)}{2} \times mAP_k \right), \tag{11}$$

where $w_i(k)$ and $w_s(k)$ are two weighting factors considering the number of images in the test dataset and the severity of weather conditions, respectively.

- Firstly, the weighting factor $w_i(k)$ is adaptive to the contribution of the number of test images because the score obtained from a larger dataset should contribute more towards the final summarized score.

$$w_i(k) = \frac{L_k}{\sum_{t=1}^{2^N-1} L_t} \tag{12}$$

where L_k represents the number of test images in the k-th experiment test in Table 3.

- Next, the weighting factor $w_s(k)$ is adaptive to the severity level of bad weather because the score obtained from a more severe weather should contribute more towards the final summarized score.

$$w_s(k) = \frac{c_k}{\sum_{t=1}^{2^N-1} c_t}, \tag{13}$$

where c_k represents the cumulative distribution function of a multivariate Gaussian distribution controlling weather severity, $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 denote the mean and variance, respectively. For the lowest severity level 1 and highest severity level N, they are assigned to the range of $\mu - 3\sigma, \mu + 3\sigma$. By this way, we compute the probability density function for each severity level k.

4.2. Score aggregation based on specific severity level

Table 4: An illustration of score aggregation based on specific severity level n.

Index k	1	2	...	$M+1$	...	$2^M - 1$
	$\{T_{1n}\}$	$\{T_{2n}\}$	...	$\{T_{1n}, T_{2n}\}$	...	$\{T_{1n}, T_{2n}, \ldots, T_{Mn}\}$

For a given severity level n, we establish the various experiment tests combinations, such as $\{T_{1n}\}$, $\{T_{2n}\}$, $\{T_{1n}, T_{2n}\}$, $\{T_{1n}, T_{2n}, \ldots, T_{Mn}\}$, as illustrated in Table 4. For each test k, we can obtain the mAP score of the object detection model. Then, all these $2^M - 1$ mAP scores can be aggregated as

$$\mathrm{SASA}(n) = \sum_{k=1}^{2^M-1} \left(\frac{w_i(k) + w_m(k)}{2} \times mAP_k \right), \tag{14}$$

where $w_i(k)$ and $w_m(k)$ are two weighting factors considering the number of images in the test dataset and the image simulation method, respectively. Firstly, the weighting factor $w_i(k)$ is defined as (12) by applying the same intuition that the score obtained from a larger dataset should contribute more towards the final summarized score. Next, the weighting factor $w_m(k)$ is adaptive to the image simulation methods, which are treated equally in this study.

5. Experimental results

5.1. Object detection models

With the augmented weather dataset, we evaluate and compare the performance of three trained object detection models to quantify how adverse weather conditions affect the overall robustness of these models in autonomous driving scenarios. We choose YOLOv7, SSD, and Faster R-CNN as examples. These three methods are widely recognized and extensively used in both academic research and industry. YOLOv7 [35] represents the latest version of one-stage detection methods, SSD [36] is a classic example of single-stage detection, and Faster R-CNN [37] is representative of two-stage detection methods. By selecting these three methods, we can comprehensively evaluate the performance differences between single-stage and two-stage detection methods under adverse weather conditions.

5.2. Implementation details

We use the dataset *object_image_2*, which is a subset of the KITTI dataset[9]. It contains $7,481$ images, covering driving scenarios including urban areas, rural areas, and highways. There are 8 classes in this dataset, where each image contains up to 15 cars and 30 perdestrians, along with various levels of occlusion and truncation. The dataset undergoes preprocessing to establish comprehensive labels. Specifically, the categorizations pertaining to "Car," "Van," and "Truck" remain unaltered, while "Pedestrian," "Person_sitting," and "Cyclist" are merged into a singular category denoted as "Person". Subsequently, categories denoted as "Tram," "Misc," and "DontCare" are omitted from the dataset.

The first 5,000 images of this dataset are used for model training, and the subsequent 2000 image are selected to simulate adverse weather conditions: Rain, fog and snow. We create four intensity levels (light, moderate, heavy, severe) for each weather condition. The original images are grouped into 4 groups using a cyclic screening approach with a three-image interval: for instance, the 5,000th image is grouped as light, the 5,001st as moderate, the 5,002nd as heavy, the 5,003rd as severe, and the 5,004th as light, continuing following this cycle till the last image is reached. Then the weather intensity for each group of 500 images is simulated accordingly. This approach alleviates scene redundancy, as adjacent images in the original dataset usually derive from consecutive frames of the same autonomous driving scenario, thus avoiding overly homogeneous scenes in weather images of the same intensity.

The PyTorch framework is employed to facilitate the training and evaluation phases. Specifically, the models YOLOv7, SSD, and Faster R-CNN are individually deployed on A100 40G, RTX 3090 24G, and RTX 3090 24GB GPU platforms, respectively. Training iterations are conducted over varying epochs, with YOLOv7 undergoing 100 epochs, SSD undergoing 200 epochs, and Faster RCNN also undergoing 200 epochs.

5.3. Experimental results

Following the training phase, the three models are subsequently applied for inference datasets on augmented data to assess their performance in fogging, raining and snowing weather conditions. Inference results are illustrated in Figure 4. Their objective performance is evaluated using the proposed SASA method, which is summarized for specific simulation method (two methods for each type of weather in our study) or a specific severity level (four levels for each type of weather in our study). Therefore, we can obtain six SASA scores shown in Table 5and nine SASA scores shown in Table 6. As seen from these results, YOLOv7 consistently demonstrates superior performance across all weather conditions. Conversely, SSD and Faster R-CNN exhibit mixed performance under different weather conditions, each with its own strengths and weaknesses.

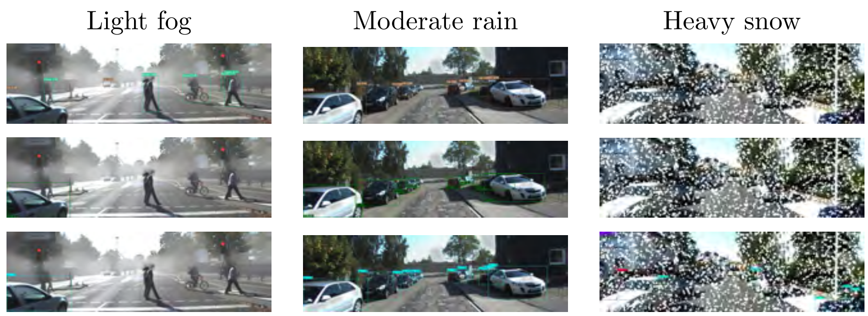

Figure 4: Examples of object detection results obtained by YOLOv7 (left column), SSD (middle column), and Faster R-CNN (right column) on our *KITTI-FRAS* dataset.

Table 5: The SASA performance evaluation of object detection models aggregated for different simulation methods.

Model	Fog (simulation)		Rain (simulation)		Snow (simulation)	
	1	2	1	2	1	2
YOLOv7	0.752	0.328	0.937	0.921	0.765	0.631
SSD	0.211	0.108	0.688	0.654	0.269	0.291
Faster RCNN	0.065	0.007	0.699	0.213	0.317	0.274

Moreover, as weather conditions worsens, the overall detection performance exhibits a declining trend. Notably, fog emerges as the most influential weather condition affecting object detection performance, posing a significant challenge due to its blurring effect on objects.

Table 6: The SASA performance evaluation of object detection models aggregated for different severity levels.

Model	Fog				Rain				Snow			
	Light	Moderate	Heavy	Severe	Light	Moderate	Heavy	Severe	Light	Moderate	Heavy	Severe
YOLOv7	0.694	0.600	0.527	0.433	0.942	0.931	0.931	0.920	0.937	0.857	0.714	0.419
SSD	0.304	0.163	0.125	0.090	0.713	0.703	0.675	0.634	0.635	0.479	0.211	0.147
Faster RCNN	0.104	0.049	0.020	0.010	0.491	0.475	0.466	0.419	0.635	0.464	0.234	0.052

6. Conclusions

In this study, we have investigated evaluating the robustness of object detection models under adverse weather conditions by proposing a comprehensive evaluation framework. This framework introduces the new *KITTI-FRAS* dataset, specifically designed for adverse weather scenarios, and a detailed model performance evaluation protocol. Using this framework, we have evaluated three classical object detection models, revealing varying degrees of robustness across the models. Our proposed evaluation framework effectively addresses the challenge of assessing model robustness in adverse weather conditions and holds potential as a standardized evaluation tool within the field of object detection. Moreover, this protocol is generalizable to any object detection model, providing researchers and developers with a consistent and quantitative approach for evaluating model performance in challenging environments.

References

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, Proceedings of the IEEE 111 (3) (2023) 257–276.

[2] R. Kaur, S. Singh, A comprehensive review of object detection with deep learning, Digital Signal Processing 132 (2023) 103812.

[3] Y. Li, N. Miao, L. Ma, F. Shuang, X. Huang, Transformer for object detection: Review and benchmark, Engineering Applications of Artificial Intelligence 126 (2023).

[4] Y. Sun, Z. Sun, W. Chen, The evolution of object detection methods, Engineering Applications of Artificial Intelligence 133 (2024).

[5] C. Zhao, R. W. Liu, J. Qu, R. Gao, Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons, Engineering Applications of Artificial Intelligence 128 (2024).

[6] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, J. Han, Towards large-scale small object detection: Survey and benchmarks, IEEE Trans. on Pattern Analysis and Machine Intelligence 45 (11) (2023) 13467–13488.

[7] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, L. Zhang, Image-adaptive YOLO for object detection in adverse weather conditions, in: Proc. AAAI Conf. on Artificial Intelligence, Vol. 36, 2022, pp. 1792–1800.

[8] M. J. Mirza, C. Buerkle, J. Jarquin, M. Opitz, F. Oboril, K.-U. Scholl, H. Bischof, Robustness of object detectors in degrading weather conditions, in: IEEE Int. Intelligent Transportation Systems Conference, IEEE, 2021, pp. 2719–2724.

[9] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[11] M. Tremblay, S. S. Halder, R. De Charette, J.-F. Lalonde, Rain rendering for evaluating and improving robustness to bad weather, International Journal of Computer Vision 129 (2021) 341–360.

[12] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, T. Yao, Exploring object relation in mean teacher for cross-domain detection, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019, pp. 11457–11466.

[13] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, J.-N. Hwang, DesnowNet: Context-aware deep network for snow removal, IEEE Trans. on Image Processing 27 (6) (2018) 3064–3073.

[14] M. A. Kenk, M. Hassaballah, DAWN: Vehicle detection in adverse weather nature dataset, arXiv preprint arXiv:2008.05402 (2020).

[15] M. Leyva-Vallina, N. Strisciuglio, N. Petkov, Data-efficient large scale place recognition with graded similarity supervision, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2023, pp. 23487–23496.

[16] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, P. Newman, Real-time kinematic ground truth for the Oxford robotcar dataset, arXiv preprint arXiv:2002.10152 (2020).

[17] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, BDD100K: A diverse driving dataset for heterogeneous multitask learning, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 2636–2645.

[18] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge, International Journal of Computer Vision 88 (2010) 303–338.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: European Conf. on Computer Vision, Zurich, Switzerland, 2014, pp. 740–755.

[20] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition), Las Vegas, NV, 2016.

[21] R. Solovyev, W. Wang, T. Gabruseva, Weighted boxes fusion: Ensembling boxes from different object detection models, Image and Vision Computing 107 (2021) 104117.

[22] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-NMS–improving object detection with one line of code, in: IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 5561–5569.

[23] H. Koschmieder, Theorie der horizontalen sichtweite, Beitrage zur Physik der freien Atmosphare (1924) 33–53.

[24] G. Zaher, Simulating weather conditions on digital images (2020).

[25] A. v. Bernuth, G. Volk, O. Bringmann, Simulating photo-realistic snow and fog on existing images for enhanced CNN training and evaluation, in: IEEE Intelligent Transportation Systems Conference, 2019, pp. 41–46.

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

[27] S. Kahraman, R. de Charette, Influence of Fog on Computer Vision Algorithms, Tech. rep., Inria Paris (Sep. 2017).

[28] M. Christiansen, Adobe after effects CC visual effects and compositing studio techniques, Adobe Press, 2013.

[29] Y. Weber, V. Jolivet, G. Gilet, D. Ghazanfarpour, A multiscale model for rain rendering in real-time, Computers & Graphics 50 (2015) 61–70.

[30] B. Horn, Robot vision, MIT press, 1986.

[31] K. Garg, S. K. Nayar, Vision and rain, International Journal of Computer Vision 75 (2007) 3–27.

[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE signal processing magazine 35 (1) (2018) 53–65.

[33] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I.-H. Chen, J.-J. Ding, S.-Y. Kuo, All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss, in: IEEE/CVF Int. Conf. on Computer Vision, 2021, pp. 4176–4185.

[34] U. Saxena, Automold-road-augmentation-library, https://github.com/UjjwalSaxena/Automold-Road-Augmentation-Library (2018).

[35] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single shot multibox detector, in: European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 21–37.

[37] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).