Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

R&D Project

Object detection in adverse weather conditions using tightly-coupled data-driven multimodal sensor fusion

Kevin Patel

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr.-Ing. Sebastian Houben
M.Sc. Santosh Thoduka

November 2023

© 2023 Kevin Patel. All Rights Reserved

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

Date	Kevin Patel

Abstract

TODO: add abstract

Acknowledgements

TODO: add acknowledgements

Contents

1

Introduction

In the realm of autonomous driving, the ability to detect objects in challenging weather conditions remains a critical area of research. Consider a scenario where a self-driving vehicle navigates a winding mountain road at night amidst fog and rain. The limitations of the vehicle's visual cameras become evident as they struggle to detect objects due to reduced visibility, highlighting the crucial need for advanced object detection methods in adverse weather conditions. This is particularly significant in emergency situations, like when a deer suddenly appears on the road, necessitating quick and accurate object detection to prevent accidents [1] [2] [3].

As depicted in Figure 1.1, the performance of various sensors in automated systems under different conditions has been extensively analyzed. Cameras, for instance, excel in recognizing colors and signs but falter in dark or distant object measurement. Conversely, thermal sensors maintain efficiency in poor weather but lack color detection and texture information. Radar sensors are adept at speed measurement and are less hindered by visual obstructions, though they produce sparse and noisy data. LiDAR sensors, meanwhile, provide excellent object shape and size mapping but underperform in poor weather conditions [4].

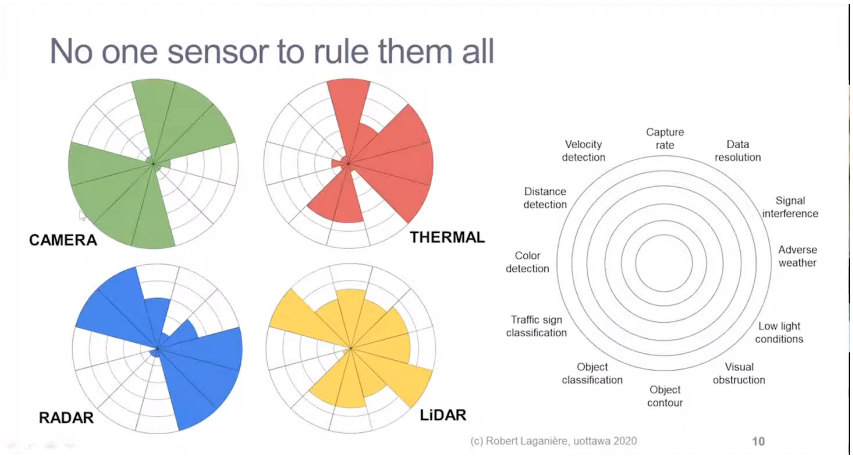


Figure 1.1: Sensors modality characteristics [5]

To overcome these limitations, this project proposes the development of a tightly-coupled multimodal

sensor fusion system, as exemplified in Figure 1.2. By integrating cameras, radar, and LiDAR sensors, this approach harnesses the unique strengths of each to create a comprehensive and reliable object detection framework. The fusion of these complementary sensors, coupled with advanced machine learning algorithms, aims to significantly enhance the range, accuracy, and reliability of detection in adverse weather conditions. Effectively synthesizing diverse data sources enables the creation of a robust and responsive system that overcomes the weaknesses of any individual sensor type. A sample from DENSE dataset [6] is shown in Figure 1.3 to illustrate the importance of fusion in adverse weather conditions.

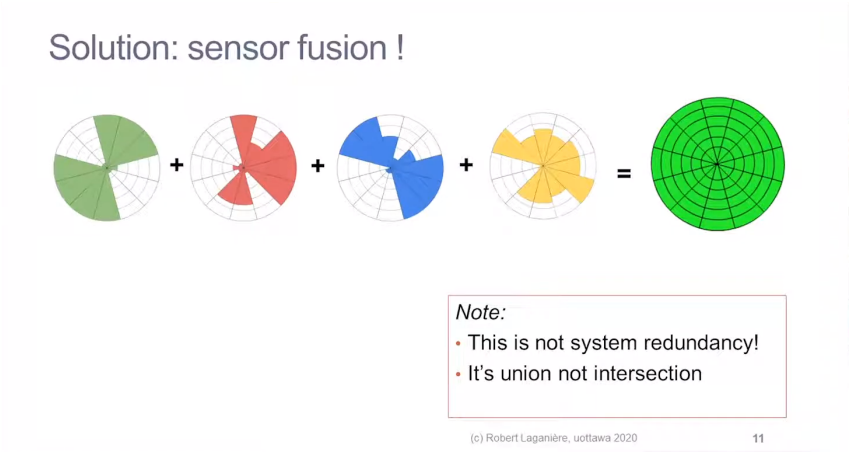


Figure 1.2: Sensors modality characteristics [5]

However, the integration of different sensor types presents its own set of challenges, such as varying resolutions, sampling rates, and the need for sophisticated calibration and alignment techniques. These hurdles necessitate the development of efficient and scalable algorithms and hardware architectures capable of processing large volumes of sensor data with minimal latency.

Despite these challenges, the anticipated outcomes of this project are transformative, with potential applications extending beyond autonomous vehicles to drones, surveillance, and security systems. By enhancing situational awareness through optimized sensor fusion, the project aims to foster safer and more efficient operations across various sectors. The primary objective is to ensure safe and efficient autonomous driving in adverse weather conditions, prioritizing the safety of passengers, other drivers, and pedestrians. This will be achieved through a sensor fusion system designed for minimal latency, enabling the processing of data from multiple sensors in near real-time.

In this research, the focus is specifically on 2D object detection to manage the intricacies of network and computational complexity. This scope simplifies the task of handling high-dimensional data, allowing for a more streamlined implementation while addressing the challenge of object detection in adverse weather conditions. Such conditions, including fog, snow, and rain, significantly hinder the visibility and recognition of various entities like cars, trucks, pedestrians, and cyclists. The 'tightly-coupled' approach involves the intricate integration of various data modalities, combining features at multiple levels for optimal results. The 'data-driven' aspect emphasizes leveraging existing datasets to enhance performance,

Figure 1.3: Importance of sensor fusion in adverse weather conditions, *Fusion of Camera, Lidar, Radar (Image adapted from [6])

while 'multimodal' pertains to the utilization of diverse sensor types. Finally, 'sensor fusion' is central to the project, entailing the amalgamation of data from different sensors to improve environmental perception and object detection capabilities.

1.1 Motivation

The advent of autonomous vehicles (AVs) marks a significant leap in transportation technology, but their operation in adverse weather conditions presents a formidable challenge. This research is motivated by the need to enhance the object detection capabilities of AVs under such conditions, a critical factor in ensuring their reliability and safety.

Adverse weather conditions, including rain, snow, sleet, and fog, significantly impact traffic safety. While these conditions have long been a concern for human-driven vehicles, leading to substantial accidents and fatalities globally, they pose a unique set of challenges for AVs. For instance, in the United States, over 30,000 vehicle crashes annually are attributed to snowy or icy roads, resulting in over 5,000 fatalities [7], [8]. Similarly, in Europe, 25% of all road accidents are weather-related, leading to numerous fatalities annually [9]. These statistics, while highlighting human driving risks, underscore the urgency for advanced object detection systems in AVs to address similar challenges.

Current AV technologies demonstrate limitations in inclement weather, which hampers their wider adoption and trust. For example, the Mcity shuttle ceases operation in conditions necessitating continuous use of windshield wipers [10], and Tesla's autopilot system struggles in heavy storms or when lane lines are obscured [11]. General Motors' Super Cruise also restricts usage in adverse weather [12]. These examples illustrate the need for improved perception systems in AVs, especially under challenging weather conditions.

This research, focusing on tightly-coupled data-driven multimodal sensor fusion for object detection, aims to address these limitations. By improving the ability of AVs to detect obstacles in diverse and adverse weather conditions, the research contributes significantly not only to road safety but also to the broader goal of achieving fully autonomous vehicular operation.

Moreover, the relevance of this research extends beyond road safety. It impacts various sectors including healthcare, precision agriculture, environmental monitoring, aerospace, defense, and industrial automation. For instance, advancements in sensor fusion technology can enhance efficiency and reduce production costs in these areas, further underlining the significance of this research.

In summary, this research is driven by the need to overcome environmental constraints that currently limit AVs, aiming to revolutionize not only the field of autonomous driving but also various technology-driven sectors. By focusing on developing robust object detection systems that can operate effectively in adverse weather conditions, the research holds the potential to significantly advance the state of AV technology and its applications.

1.2 Challenges and Difficulties

The field of multimodal sensor fusion, particularly in the context of adverse weather conditions, presents a series of distinct challenges that are critical to the advancement of reliable object detection systems. These challenges range from the scarcity of specialized datasets to the intricacies of fusion architecture, computational demands, and the search for effective network design and generalization. This section outlines the key challenges in this field.

Scarcity of Adverse Weather-Related Datasets: A fundamental challenge in this field is the lack of datasets that capture adverse weather conditions through multimodal sensors. Most existing datasets are oriented towards clear weather scenarios. For example, popular datasets for multimodal sensor fusion include camera and LiDAR but often lack radar sensors. These include KITTI [13], ApolloScape [14], and Waymo [15]. There are datasets that incorporate radar sensors, such as VoD [16], TJ4DRadSet [17], and PixSet [18], but they do not cover an extensive range of adverse weather conditions. This limitation restricts the development and validation of object detection models in more challenging environmental conditions and hinders the advancement of robust object detection systems capable of performing reliably under diverse weather conditions.

Limitations in Fusion Architecture: The predominant focus in the current landscape of fusion architectures is on middle or feature fusion, with only limited exploration in the area of tightly coupled fusion networks. This trend poses a particular challenge for sensor fusion that involves radar data, which tends to be noisier and sparser compared to the data produced by cameras and LiDAR. The scarcity of research on tightly coupled fusion methods for integrating such diverse data types is a significant obstacle in developing efficient and effective multimodal fusion systems.

Computational Constraints: Another significant barrier is the computational demand required to train large, complex networks. This requirement often exceeds the resources available to many researchers, restricting exploration in this field and slowing the advancement of more sophisticated fusion methods.

General Guidelines for Network Architecture: Currently, there is no standard or widely

accepted framework for the design of network architectures in multimodal sensor fusion, leading to several unanswered questions. Key considerations, as highlighted by Feng et al. [19], include:

- "What to fuse," such as LiDAR, radar, various camera types (color, thermal, event), or ultrasonic sensors;
- "How to fuse," with possibilities including addition or averaging, concatenation;
- "When to fuse," which can range from early, mid, late, or a combination of these fusion stages.

This absence of a clear guideline results in uncertainties regarding the optimal choices for integrating different modalities in sensor fusion.

Limited Generalization from Previous Studies: Many existing studies in multimodal sensor fusion have focused on results from their baseline models and custom datasets. This narrow scope limits the generalizability of their findings, as these models may not perform as well under different conditions or with alternative datasets.

Advanced Fusion Methods and Temporal Information: While recent datasets have begun to include baseline models featuring basic fusion methods, there is notable potential for significant performance enhancement. This can be achieved by incorporating advanced transformer-based architectures, known for their superior handling of complex data patterns and scalability. Additionally, employing sophisticated fusion techniques, such as tightly-coupled fusion, which integrates data more closely and efficiently, could further optimize the sensor fusion process. Additionally, the integration of temporal information in sensor fusion is an area that is yet to be extensively explored. Although not a focus of this project, it represents a promising direction for future research.

Addressing these challenges is critical for advancing the field of object detection in adverse weather conditions using multimodal sensor fusion. This research will contribute to this effort by exploring these underdeveloped areas, aiming to find more robust and effective object detection systems.

1.3 Problem Statement

The comprehensive analysis and practical application of state-of-the-art multimodal object detection methods under adverse weather conditions remain largely unexplored areas. The crux of this research project is to bridge this gap by providing an in-depth analysis and practical implementation of cutting-edge techniques in multimodal object detection. The project will explore the integration of multiple sensor modalities, including but not limited to cameras, LiDAR, and radar, to enhance detection capabilities in challenging weather scenarios.

One key aspect of this project is its focus on 2D object detection, a deliberate choice to reduce the complexities associated with network and computer processing. While simplifying the technical challenges, this focus does not diminish the project's primary objective: to explore object detection in adverse weather conditions through a multimodal approach and tightly coupled fusion architecture.

A key challenge is to explore fusion strategies which capitalize on the strengths of different sensors while mitigating their limitations. For example, the integration of visual camera data with radar information

presents a complex yet crucial research question. This research aims to explore an optimal fusion strategy, a technique that effectively combines the unique strengths of each sensor type, thereby creating a more robust and accurate detection system. The strategic combination of these modalities holds the potential to significantly enhance object detection performance, especially in conditions where traditional single-sensor systems fall short.

The performance of the selected multimodal object detection methods will be evaluated under a variety of adverse weather conditions on a common protocol on widely recognized datasets, specifically DENSE [6] and nuScenes [20]. These datasets are chosen for their relevance in testing robustness and reliability in challenging environmental conditions, providing a comprehensive benchmark for assessing the effectiveness of the methods.

Additionally, this study will compare the proposed methods to the best existing methods by comparing their outcomes. This comparison will not only highlight the efficacy of the new methods in adverse conditions but will also shed light on the strengths and weaknesses of each approach. Through this evaluation, the project aims to contribute significantly to the body of knowledge in multimodal object detection, providing valuable insights for future research and practical applications in this rapidly evolving field.

2

Background

2.1 Types of Fusion Architectures

TODO: write about different types of fusion architectures and their pros and cons.

3

Related Works

3.1 Adverse Weather Conditions Influence on Sensors

In the evolving landscape of autonomous robotics, particularly in the domains of self-driving vehicles and autonomous drones, object detection stands as a paramount challenge in computer vision. These cutting-edge applications necessitate precise 2D or 3D bounding boxes for objects within complex and often unpredictable real-world environments. These scenarios commonly involve cluttered scenes, varying lighting conditions, and notably, adverse weather conditions. To tackle these multifaceted challenges, state-of-the-art autonomous vehicle systems are increasingly reliant on a suite of redundant sensor modalities. Recent studies, such as those by Caesar et al. [20], Sun et al. [21], and Ziegler et al. [22], highlight this trend. These sensor modalities extend beyond traditional cameras and LiDAR, encompassing radar and emerging technologies like far-infrared (FIR) and near-infrared (NIR) sensors, which are proving instrumental in enabling reliable object detection in adverse conditions [6].

For standard perception systems in autonomous vehicles, the camera remains an indispensable, yet highly vulnerable sensor to adverse weather conditions. Despite its high resolution, a camera's functionality can be severely compromised by a single water drop on the lens or emitter during rain [23], as illustrated in Figure 3.1. In conditions like heavy snow or hail, image intensity can fluctuate, and object edges may become obscured, leading to detection failures [24]. Additionally, cameras are susceptible to strong light interference, either from direct sunlight or artificial sources like urban light pollution, causing significant operational challenges [25].

Table 3.1 systematically presents the quantitative impact of different weather conditions on a range of sensors. Notably, the influence classified as level 3, termed as moderate, is associated with perception errors occurring up to 30% of the time. This level of impact may translate to as much as 30% of the LiDAR point cloud being compromised by noise, or a similar proportion of camera image pixels experiencing distortion or obscurity. Similarly, the implications categorized under level 4, identified as serious, follow this trend as well [26].

LiDAR, the second-most common sensor in autonomous driving systems, exhibits a different response to adverse weather. As Fersch et al. [28] suggest, LiDAR sensors with small apertures are relatively unaffected by moderate rainfall. However, intense and uneven precipitation can generate fog clusters, potentially resulting in false obstacle detection by LiDAR systems. Hasirlioglu et al. [29] demonstrated

Table 3.1: The influence level of various weather conditions on sensors [26]

Modality	**Light rain** <4mm/hr	**Heavy rain** <25mm/hr	**Dense smoke** vis<0.1km	**Fog** vis<0.5km	**Haze** vis<2km	**Snow**	**Strong light** (over emitter)	**Cost**
LiDAR	2	3	5	4	0	5	2	high
Radar	0	1	2	0	0	2	0	medium
Camera	3	4	5	4	3	3	5	lowest
NIR	2	3	2	1	0	2	4	low
FIR	2	3	1	0	2	4	3	low

The effect level each phenomenon causes to sensors:

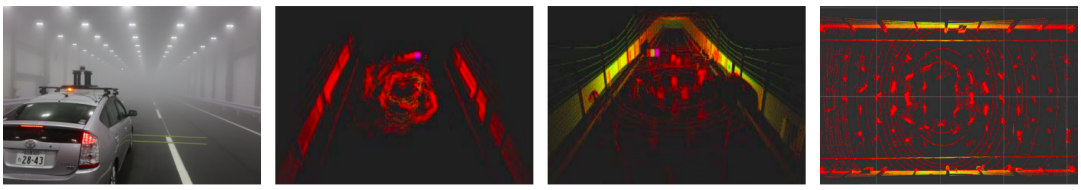
- 0 - negligible: influences that can almost be ignored;
- 1 - minor: influences that barely cause detection error;
- 2 - slight: influences that cause small errors on special occasions;
- 3 - moderate: influences that cause perception error up to 30% of the time;
- 4 - serious: influences that cause perception error more than 30% but lower than 50% of the time;
- 5 - severe: noise or blockage that cause false detection or detection failure;

Notes:

- NIR: Near-Infrared camera
- FIR: Far-Infrared camera
- Sensor's specifications:
  - LiDAR (850-950nm and 1550nm)
  - Radar (24, 77, and 122 GHz)
  - NIR Camera (λ 800-950 nm)
  - FIR Camera (λ 2-10 µm)

Figure 3.1: Van occluded by a water droplet on the lens (Image source [27])

that rainfall rates exceeding 40 mm/hr significantly reduce signal reflection intensity. Dense fog and smoke, along with strong light, can adversely affect LiDAR sensors in challenging conditions [25, 26]. This is exemplified in Figure 3.2, which showcases an instance of LiDAR's performance in fog, where it erroneously creates small false obstacle clouds. Similarly, Figure 3.4c illustrates how LiDAR's ability to measure distances is compromised in foggy environments. This contrasts with radar technology, whose outputs remain largely unaffected under similar conditions. Such discrepancies highlight the limitations of LiDAR in adverse weather and the need for integrating complementary sensor modalities for enhanced reliability in autonomous driving systems.



(a) Fog chamber (b) Point clouds in fog (c) Point clouds in rain (d) Rain pillars as detected by LiDAR

Figure 3.2: LiDAR performance test (a sample from LIBRE [2] dataset)

Radar, the third critical sensor in autonomous driving systems, is frequently used in mass-produced cars for active safety functions like Automatic Emergency Braking (AEB) and Forward Collision Warning (FCW). Its role in perception tasks for autonomous driving, however, is often undervalued. Unlike cameras operating in the visible light bands (384–769 THz) and LiDARs in the infrared bands (361–331 THz), radar utilizes longer wavelength radio bands (77–81 GHz). This attribute ensures its robust performance in adverse weather conditions [30]. Studies by Ijaz et al. [31] and Ismail [32] indicate radar's lower attenuation in rainy conditions compared to LiDAR. At 77 GHz, radar exhibits approximately 3.5 times less attenuation (10 dB/km) than LiDAR at 905 nm (35 dB/km), showcasing its superior robustness.

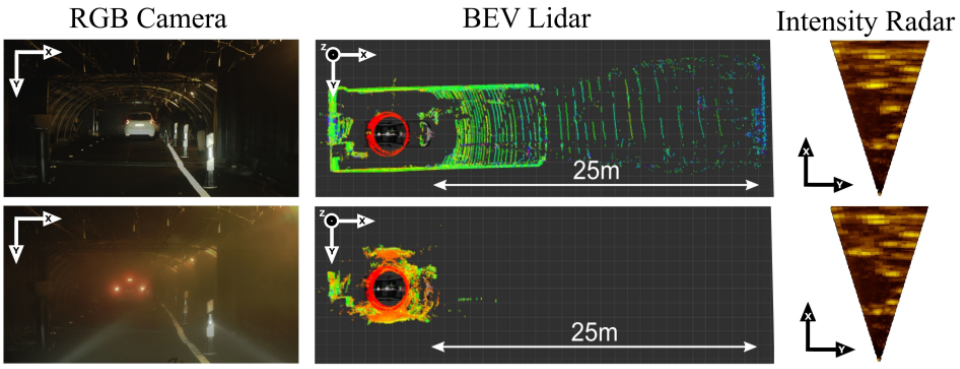


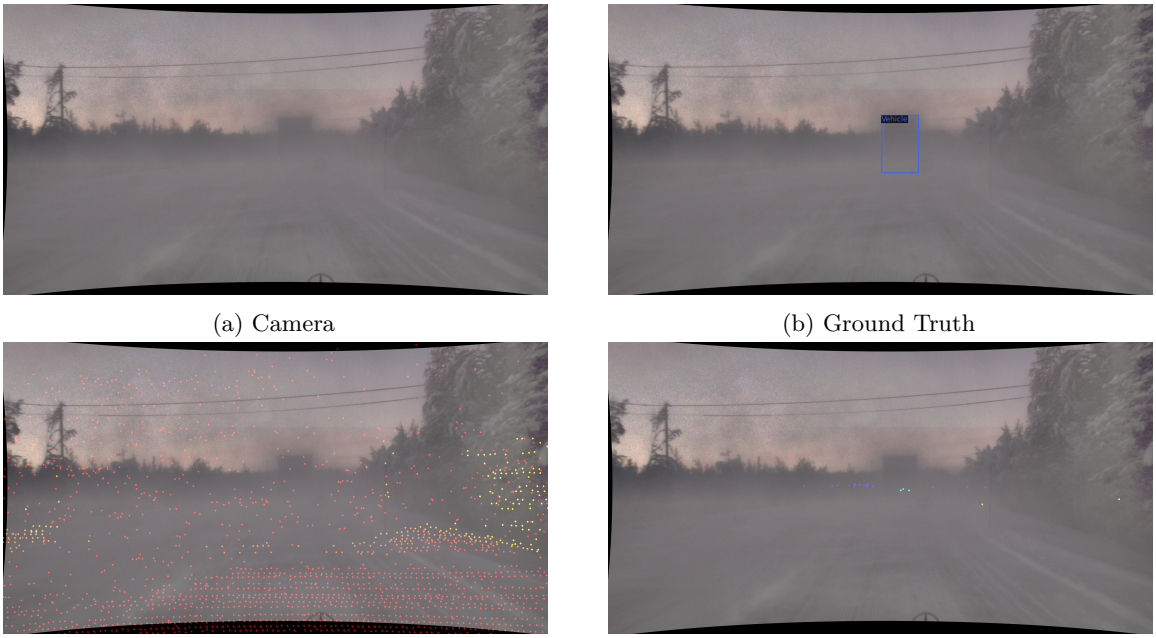
Figure 3.3: 1st row: clear weather condition, 2nd row: with fog. Shows that lidar affects by the fog but radar intensity remains the same (Image source [6])

Various experiments [24, 33–36] have demonstrated that radar's performance is minimally affected by dust, fog, snow, and light rain, although it degrades in heavy rainfall conditions. Figure 3.4 from the DENSE [6] dataset exemplifies this, showing radar's ability to detect vehicles under dense fog conditions, where cameras and LiDAR fail. Nevertheless, radar's low resolution and sparser point clouds compared to LiDAR limit its utility in perception tasks. The emerging 4D radar technology, while promising denser point clouds, lacks public datasets in adverse weather conditions for validation.

The combination of LiDAR and camera technologies alone has proven insufficient for navigating through adverse weather conditions with adequate safety assurance. However, the integration of these with radar, infrared cameras, gated cameras, stereo cameras, weather stations, and other weather-related sensors presents a new paradigm in autonomous vehicle perception. This multimodal sensor fusion, as evidenced in Figures 1.1 and 1.2, offers a composite strength that individual systems lack. Consequently, research groups worldwide are exploring various permutations and combinations of these sensors to enhance the reliability and safety of autonomous driving systems in challenging weather conditions.

3.2 Multimodal Sensor Fusion

The concept of multimodal sensor fusion has become increasingly pivotal in the field of object detection, particularly under adverse weather conditions. This approach integrates various sensor inputs to enhance perception and object detection capabilities, addressing the limitations inherent to individual sensors. A notable contribution in this field is the work of Radecki et al. [37], who conducted a thorough review of sensor efficacy across diverse weather conditions, including wet environments, varying light conditions, and dusty atmospheres. They developed a sophisticated system capable of object tracking and classification through a real-time, joint probabilistic perception algorithm. This algorithm dynamically selects the most appropriate sensor subsets based on prevailing weather conditions. By intelligently weighting sensors and accurately quantifying parameters specific to each weather scenario, the system not only improves the baseline of perception ability but also enhances its robustness and reliability. This work shows that multimodal sensor fusion is effective compared to individual sensors especially in the context of adverse



(a) Camera (b) Ground Truth

(c) Lidar (d) Radar

Figure 3.4: Dense fog influence on sensors (a sample from DENSE dataset [6]). Best viewed in zoomed-in view.

weather conditions.

However, the research of Radecki et al. [37] presents certain limitations. Firstly, it overlooks scenarios involving heavy traffic and urban environments, contexts that are crucial for real-world applications [38]. Furthermore, the study does not delve into deep learning-based fusion techniques, which could potentially offer more advanced data integration and interpretation capabilities [39]. Another constraint lies in the limited evaluation of weather samples on custom datasets, which may not accurately represent the more varied and unpredictable real-world weather conditions [40]. Lastly, the research primarily focuses on an early fusion approach with hand-tuned pre-processing of sensor data, which may impede the generalization of its findings to broader applications [41]. To address these challenges, this project will focus on developing advanced data-driven multimodal sensor fusion techniques for object detection in a variety of adverse weather conditions, employing publicly available datasets for comprehensive validation.

In 2019, FLIR System Inc. [42] and VSI Labs [43] tested the first-ever fused Automated Emergency Braking (AEB) sensor suite, comprising a thermal long-wave infrared (LWIR) camera, a radar, and a visible camera. The LWIR camera, operating in the 8 µm to 14 µm range at ambient temperature, was part of this groundbreaking sensor suite. The suite's performance was evaluated against standard AEB systems using radar and visible cameras under various conditions, including daytime, nighttime, and transitions from tunnel exits into sun glare. The results revealed that while most AEB systems function effectively during the day, they almost invariably failed under adverse conditions, often colliding with mannequins. In stark contrast, the LWIR sensor suite successfully avoided collisions in these challenging

scenarios, highlighting the efficacy of fusing camera and radar data in adverse weather conditions.

However, the use of a thermal camera as part of the sensor fusion raises concerns about the durability of such temperature-sensitive devices in real-world settings, necessitating further validation to ascertain their effectiveness in adverse weather [24].

Another significant development in multimodal sensor fusion is the CameraRadarFusionNet (CRF-Net) proposed by Nobis et al. [44]. Inspired by previous works on camera-LiDAR fusion [45, 46], the CRF-Net was designed to determine the most beneficial stage for sensor data fusion within a neural network architecture for detection tasks. Utilizing the nuScenes [20] dataset and their own TUM dataset, they introduced a novel training strategy, BlackIn, focusing on a specific sensor type. The fusion method employed, element-wise addition, demonstrated superior performance over an image-only network on both datasets, underscoring the significance of incorporating radar data into detection tasks.

Nevertheless, Nobis et al.'s approach [44] exhibited only marginal improvements in detection performance over the baseline image network. The absence of an RGB sensor ablation study in their work left questions about the system's robustness in case of camera failure. According to Safa et al. [47], pre-processing the radar data before fusion could further enhance performance, suggesting an area for potential improvement in future research endeavors.

Building on the concept of multimodal sensor fusion, Yang et al. [48] introduced RadarNet, a novel framework for object detection and velocity estimation. RadarNet is distinctive for its dual strategy of leveraging both radar and LiDAR sensors for perception. It employs an early fusion technique to learn joint representations from these sensors, while its late fusion phase incorporates radar's radial velocity evidence to enhance object velocity estimation. This approach was rigorously evaluated using the nuScenes dataset [20]. However, a limitation of RadarNet, as noted by Yang et al. [48], lies in the radar sensor data from the nuScenes dataset, which is characterized by low resolution, thus restricting its effectiveness in object detection. This issue of low resolution and erroneous elevation estimates, as highlighted in studies like Ulrich et al. [49] and Drews et al. [50], suggests a potential improvement for RadarNet: the integration of a higher-resolution radar, such as the one from the K-Radar dataset. This 4D radar, with a significantly wider elevation angle, could enhance the performance of RadarNet, overcoming the limitations of conventional radars.

In a similar vein, Bijelic et al. [6] from Mercedes-Benz AG conducted an extensive study focusing on enhancing detection performance in adverse weather through deep multimodal sensor fusion. Their experimental setup included a diverse array of sensors on their test vehicle, including stereo RGB cameras, a NIR camera, a 77 GHz radar, dual LiDARs, an FIR camera, a weather station, and a road-friction sensor. They introduced an innovative entropy-steered fusion approach, attenuating regions of low entropy while amplifying those with high entropy during feature extraction. This method, trained using clear weather data, showed impressive adaptability to adverse weather conditions. The fusion network was designed to maintain consistency across different scenarios, with all sensor data projected into the camera coordinate system. Their findings demonstrated that this fused approach significantly outperformed LiDAR or image-only methods, particularly under foggy conditions.

Additionally, Bijelic et al. [6] contributed to the research community by providing the SeeingThroughFog

or DENSE dataset. This comprehensive dataset, comprising 10,000 km of driving data recorded in Northern Europe, spans a range of weather and illumination conditions. The dataset includes detailed annotations for various weather frames, including clear weather, dense fog, light fog, and snow/rain, making it a valuable resource for future studies on multimodal sensor fusion in challenging weather scenarios.

However, the study by Bijelic et al. [6] also presents certain challenges. The projection transformation technique used in their approach may lead to the loss of crucial radar spatial information. Moreover, the extensive array of sensors required exceeds typical expectations for autonomous driving systems, posing implementation challenges in real-world scenarios. The large volume of data from multiple sensors could potentially impact the algorithm's response and reaction time. Additionally, the radar's performance was constrained by its limited azimuth and elevation resolution [26]. Future research could focus on enhancing the network architecture, possibly through the integration of higher resolution radars and transformer-based approaches, to further refine the performance of sensor fusion in adverse weather conditions.

Liu et al. [51] presented another innovative approach to target recognition and tracking by fusing radar and camera data. In their methodology, radar is the primary sensor, complemented by camera data as secondary information. This fusion was evaluated under challenging weather conditions, including rain, fog, and low visibility nighttime scenarios. The results demonstrated that radar-based detection was highly accurate in detecting moving targets in wet weather, while the camera excelled in target classification. The combined radar and camera data exhibited a superior performance, surpassing LiDAR-based methods by over 33%, highlighting the effectiveness of this fusion approach in challenging weather scenarios.

The exploration of multimodal sensor fusion for enhanced vehicle detection under adverse weather conditions continues with the work of Qian et al. [52], who developed the Multimodal Vehicle Detection Network (MVDNet). MVDNet uniquely incorporates LiDAR and radar data, utilizing a two-stage attention block within its fusion module. The network first applies self-attention to each modality to extract features, followed by the blending of these features with region-wise features through cross-attention mechanisms. This fusion technique has shown to be particularly effective in foggy conditions. The performance of MVDNet was rigorously tested and validated on the DENSE [6] and Oxford Radar Robotcar [53] datasets, where it demonstrated a notably improved performance compared to LiDAR-only systems in foggy environments.

Despite its robust performance, MVDNet's design by Qian et al. [52] has certain limitations. One significant issue is the misalignment between LiDAR and radar data within the dataset, which could potentially compromise the network's effectiveness. Furthermore, the simple label assignment strategy employed in the loss computation and the region-of-interest (ROI) assisted fusion design might limit the model's overall performance. These aspects suggest potential areas for improvement, possibly through the adoption of more advanced fusion techniques and refined label assignment strategies, as indicated in studies like Yang et al. [54].

Continuing in this vein, Rawashdeh et al. [55] proposed a CNN-based sensor fusion approach aimed at detecting drivable paths. This approach integrates data from cameras, LiDAR, and radar and was evaluated using the DENSE dataset [6]. Their multi-stream encoder-decoder network is designed to

counter the asymmetric degradation of the input sensors effectively. The depth and number of blocks allocated to each sensor in the architecture were determined by their respective input data densities, with the camera being the most dense, followed by LiDAR, and radar being the least dense. The fully connected network's outputs are transformed into a 2-D array for input into the decoder. The researchers demonstrated that their model could effectively ignore misleading road lines and edges, thereby accurately delineating the general drivable area.

However, Rawashdeh et al.'s approach [55] also presents certain challenges. The method lacks a comparative analysis with other state-of-the-art methods in the field, leaving its relative effectiveness somewhat uncertain. Additionally, the study does not delve into the real-time processing requirements or the computational costs of the proposed algorithm. These factors are crucial for practical applicability, especially in scenarios requiring rapid decision-making and processing, like autonomous driving in adverse weather conditions. Addressing these gaps could significantly enhance the feasibility and implementation potential of their sensor fusion approach in real-world applications.

3.3 Synthetic Data for Adverse Weather Conditions

The use of synthetic data and advanced image processing techniques has been a subject of considerable research focus. One prominent area is the application of de-hazing techniques to mitigate the impacts of adverse weather on visual data. Historically, physical priors have been employed for this purpose [56, 57]. However, with the advent of data-driven approaches, particularly deep learning, new methodologies have emerged. These deep de-hazing models, while innovative, often suffer from high computational complexity, making them less suitable for ultra-high-definition images. Chen et al. [58] highlighted a significant limitation of these models: their training on synthetic images does not effectively generalize to real-world hazy conditions. Conversely, Zhang et al. [59] leveraged temporal redundancy in video de-hazing, assembling a dataset comprising real-world hazy and haze-free videos. The challenge in this domain lies in obtaining paired hazy and haze-free ground-truth images, which is difficult in natural settings. However, this obstacle can be somewhat mitigated through the use of professional haze/fog generators that simulate real-world conditions [60, 61].

Another emerging trend in this field is the exploration of synthetic data generation for adverse weather conditions using Generative Adversarial Networks (GANs). Researchers such as Sun et al. [62], Zheng et al. [63], and Lee et al. [64] have investigated this approach, utilizing clean weather datasets like KITTI [13] and Cityscapes [65] as a basis. These methods predominantly involve the creation of artificial fog or rain images, supplemented by a limited selection of actual images captured under specific fog or rain conditions. However, the efficacy of these algorithms in diverse adverse weather scenarios remains somewhat ambiguous. A critical concern is whether these synthetic data-driven approaches can perform effectively under various real-world adverse weather conditions. Additionally, the methods for evaluating their real-world applicability and effectiveness in such scenarios are not fully established [66]. This gap in the research indicates a need for further investigation and development to enhance the reliability and applicability of synthetic data generation techniques in the context of adverse weather conditions for object detection.

3.4 The Role of Simulation in Autonomous Driving Research

The advent of autonomous driving technology, especially in challenging weather conditions, has significantly benefited from the utilization of simulation platforms and specialized experimental setups. One such noteworthy tool is the CARLA simulator [67], a widely recognized virtual platform. CARLA is particularly advantageous for researchers as it enables the creation of intricate road environments and the simulation of numerous non-ego entities in scenarios that would otherwise be impractical or prohibitively expensive to replicate in real-life experiments. This capability is crucial, considering that specific weather conditions, particularly those related to extreme climates or certain seasons, are not always readily available for testing. For example, tropical regions cannot easily conduct tests in snow conditions, and the unpredictability and brevity of natural rain showers may impede the collection of comprehensive experimental data. Most importantly, conducting tests in actual adverse weather conditions not only presents logistical challenges but also introduces significant safety risks. In contrast, simulators like CARLA offer a completely safe environment, eliminating the dangers associated with real-world testing [26].

However, the effectiveness of virtual datasets and simulation platforms in accurately representing real-world phenomena remains a topic of debate. The extent to which a simulator can truly mirror real-world conditions is an open question. Developing more realistic simulators is a key challenge in this domain. Additionally, determining the most effective methods for integrating real and virtual data is another critical area of ongoing research [19]. These aspects underline the need for continuous improvement in simulation technologies to ensure that they can effectively support the development and testing of autonomous driving systems, especially in the face of adverse weather conditions. The quest for more realistic simulators and the optimal blend of real and virtual data stand as important open questions, driving the future direction of research in autonomous driving simulations. Due to these reasons, this project will focus on the real-world datasets for adverse weather conditions.

4

Methodology

4.1 Available Datasets

The majority of deep multimodal perception approaches rely on supervised learning, which necessitates the use of high-quality, large-scale multimodal datasets with labeled ground truth for training deep neural networks. Several multimodal datasets, such as KITTI [13], ApolloScape [14], and Waymo [15], are prevalent in the domain of LiDAR-camera fusion. However, a significant number of these datasets are collected under clear weather conditions or lack a comprehensive array of sensors, including cameras, LiDAR, and Radar. A notable limitation is the scarcity of multimodal datasets that are collected under adverse weather conditions and incorporate at least all three of these essential sensors. Table 4.1 summarizes some of the available multimodal datasets[1] for evaluating the performance of deep multimodal perception techniques in adverse weather conditions. The dataset are sorted in ascending order with respect to year.

Table 4.1: Multimodal adverse weather conditions datasets. Sensors†: C-R-L-N-F denote Camera, Radar, LiDAR, Near-infrared, and Far-infrared sensors, respectively. Weather conditions‡: F-SN-R-O-SL-N denote Fog, Snow, Rain, Overcast, Sleet, and Night conditions, respectively. Note that highlighted datasets are used for the project.

Name	Sensors†	Weather Cond.‡	Size (GB)	Year	Citation Cnt.	Ref.
DENSE	**CRLNF**	**F, SN, R, O, N**	**582**	**2020**	**269**	**[6]**
nuScenes	**CRL**	**R, N**	**400**	**2020**	**3459**	**[20]**
Oxford RobotCar	CRL	R, SN, F	4700	2020	317	[53]
EU Long-term	CRL	SN, R, O, N	NA	2020	72	[68]
RADIATE	CRL	F, SN, R, O, SL, N	NA	2021	132	[69]
K-Radar	CRL	F, R, SN	13000	2022	15	[30]
Boreas	CRL	SN, R, O, N	4400	2022	38	[70]
aiMotive	CRL	R, O, N	85	2023	3	[71]

The selection of appropriate datasets is crucial. The criteria for selecting datasets cover several key areas, focusing on the availability of diverse sensors, specifically cameras, radar, and lidar, which are crucial for robust object detection in challenging environments. Additionally, the datasets must represent adverse

[1]For all the datasets, formal registration form is required to fill to access the dataset

weather conditions effectively, as this is a critical aspect of the research. Furthermore, the accessibility and thorough documentation of the datasets are considered, ensuring that the data can be easily understood and utilized in the research process. Another vital criterion is the dataset's popularity in existing research, as this allows for comparative analysis with publicly available methods, thereby validating the research findings. Moreover, the specific perception task, in this case, object detection, and the type of radar data, particularly in point cloud format, are essential considerations. The requirement for time-synchronized and calibrated data is also emphasized to ensure accuracy and reliability in sensor fusion and object detection algorithms.

After a comprehensive evaluation of these criteria, two datasets have been chosen for this research: DENSE and nuScenes. The DENSE dataset is particularly suited for this study as it includes data from various sensors under adverse weather conditions, which is crucial for testing the efficacy of multimodal sensor fusion in challenging environments. The nuScenes dataset, on the other hand, is widely used in the field, providing a rich source of data with camera, radar, and lidar sensors. Its extensive use in the community allows for a meaningful comparison with existing methods. Both datasets provide time-synchronized and calibrated data, which is essential for the accuracy of object detection algorithms in adverse weather conditions. The selection of these datasets aligns perfectly with the research objectives, offering a comprehensive platform for exploring and advancing the capabilities of data-driven multimodal sensor fusion in object detection under challenging weather scenarios.

4.1.1 DENSE dataset

The DENSE dataset, as detailed in Bijelic et al. (2020) [6], is a critical asset for evaluating multi-modal fusion algorithms in adverse weather conditions. Its standout feature is the extensive sensor array, including LiDAR, a stereo camera, a frontal long-range radar, a gated camera operating in the NIR band, a FIR camera, and a weather station sensor, as illustrated in Figure 4.1. These sensors allow for detailed data capture under various adverse weather conditions, such as rain, snow, light fog, and dense fog. Notably, the DENSE dataset uniquely offers a split for light and dense fog conditions, essential for assessing the detection performance of Lidar and Radar in varying visibility scenarios. The range of these conditions and their distribution are visually depicted in Figure 4.2. Additionally, the dataset includes data from a controlled lab environment within a fog chamber, offering a distinct view of sensor performance under simulated conditions. However, it's important to note that for the purposes of this project, only real-world data from the DENSE dataset is utilized. A few random samples from the dataset are shown in Figure 4.3. Note that LiDAR and Radar points are projected onto the camera image for visualization purposes.

The dataset covers a broad spectrum of environments, encompassing urban cities, suburban areas, highways, and tunnels. Its geographical scope is extensive, with data collection spanning over two months and covering 10,000 km across Germany, Sweden, Denmark, and Finland. This diverse environmental range enhances the dataset's applicability in various real-world scenarios.

Technically, the DENSE dataset offers radar targets in a point cloud format, aligning well with the Lidar data. Given the inherent noise in radar data, preprocessing has been performed to eliminate false points, thereby bolstering the dataset's accuracy and reliability. The radar data includes 3D information

- range, azimuth, and velocity. Moreover, it's noteworthy that the latest generation of radar sensors in the dataset provides 4D data, adding elevation to the existing dimensions. There are total 3 classes available in the dataset, including car, pedestrian, and cyclist. Object annotations in the DENSE dataset are provided in 2D as well in 3D format, and follow the COCO style format [72], with bounding box (bbox) parameters specified as x, y, width, and height. The annotation processed is well described in the supplementary material from the paper [73]. This meticulous approach to data collection, processing, and annotation positions the DENSE dataset as a powerful and adaptable tool for research in adverse weather conditions, especially in the domain of data-driven multimodal sensor fusion.

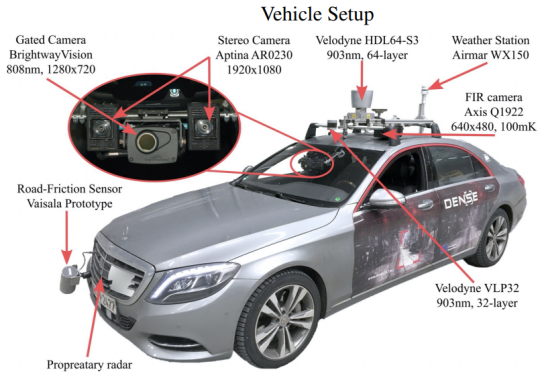


Figure 4.1: Test vehicle setup (Image source [6])

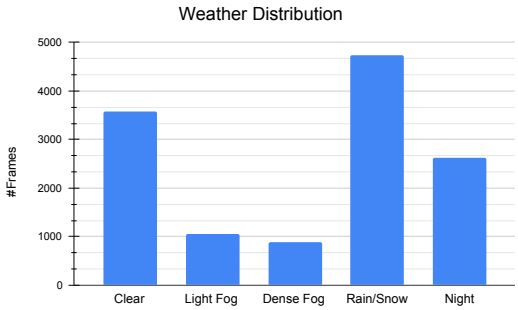


Figure 4.2: Distribution of weather conditions [6]

4.1.2 nuScenes dataset

In addition to the DENSE dataset, this project also uses the nuScenes dataset [20] as a benchmark. The nuScenes dataset stands out for its large-scale and diverse scenarios. The data collection vehicle for nuScenes, as depicted in Figure 4.4, is equipped with a comprehensive set of sensors, including a 32-beam LiDAR, six cameras, five long-range multi-mode radars, and a GPS/IMU system. This dataset provides 3D annotations for 23 classes of road users across 1,000 scenes, accumulating to a total of 1.3 million frames. Although the radar data in nuScenes is sparse, its extensive documentation makes it a good starting point for research in object detection. A few random samples from the dataset are shown in Figure 4.6.

The nuScenes dataset focuses on urban, suburban, and highway areas, but it covers fewer adverse weather conditions compared to the DENSE dataset, primarily rain and night scenarios, as shown in Figure 4.5. Like DENSE, nuScenes also provides radar data in point cloud format. There are total 23 classes but it can be categorize into 10 super classes, including car, truck, trailer, bus, construction vehicle, bicycle, motorcycle, pedestrian, traffic cone, and barrier. However, a notable distinction is that nuScenes does not provide 2D annotations. Researchers using this dataset typically generate their own 2D annotations based on the 3D annotations provided. Once these 2D annotations are created, the data is converted into the COCO style format [72], similar to DENSE, where the bbox format includes x, y,

Figure 4.3: Random samples from the DENSE dataset, where 1st column is Camera with ground truth, 2nd is LiDAR, 3rd is Radar (1st Row: Day, 2nd Row: Light Fog, 3rd Row: Dense Fog, 4th Row: Snow, 5th Row: Night). Note: Radar sparse points are highlighted with a red ellipse. Best viewed in zoomed-in view.

width, and height dimensions.

Table 4.2 highlights the overall comparison of the sensor setup and dataset statistics for datasets used in this project.

4.2 Evaluation Metrics

Metrics provide a standardized scale for comparing various methods in object detection tasks. Among these, Average Precision (AP) and Average Recall (AR) stand out as the most prominent. The understanding and computation of AP and AR are deeply rooted in more fundamental concepts such as Precision and Recall, which form the basis for these advanced metrics.

Average Precision (AP) is a pivotal metric in object detection tasks, offering a comprehensive

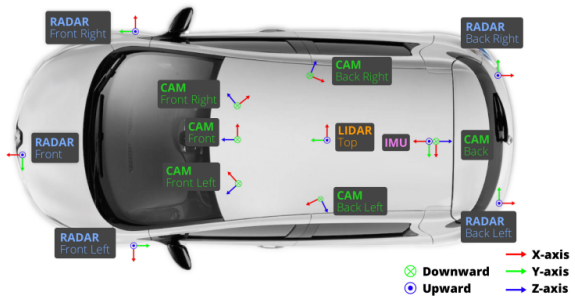


Figure 4.4: Test vehicle setup (Image source [20])

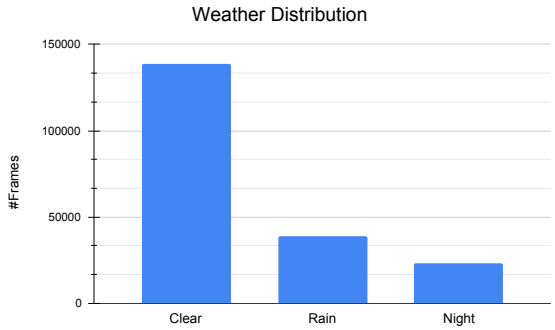
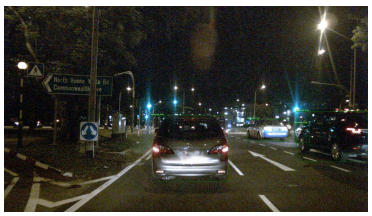


Figure 4.5: Distribution of weather conditions [20]

(a) Day time



(b) Night time

(c) Rain

Figure 4.6: Random samples from the nuScenes dataset

Table 4.2: Comparison of datasets features (Table adapted from [6])

Dataset	**NuScenes [20]**	**DENSE [6]**
RGB Cameras	6	2
RGB Resolution	1600x900	1920x1024
Lidar Sensors	1	2
Lidar Resolution	32	64
Radar Sensor	4	1
Gated Camera	x	1
FIR Camera	x	1
Frame Rate	1 Hz/10 Hz	10 Hz
Dataset Statistics		
Labeled Frames	40K	13.5K
Labels	1.4M	100K
Scene Tags	✓	✓
Night Time	✓	✓
Light Weather	✓	✓
Heavy Weather	x	✓
Fog Chamber	x	✓

measure of a model's precision and recall at various thresholds. A common misconception is that AP represents the average of Precision values, which is not accurate. A more precise interpretation of AP is that it reflects the area under the Precision-Recall curve. Precision and recall are fundamental concepts

in this context, defined as follows:

Precision (P) is the ratio of correctly predicted positive observations to the total predicted positive observations, formulated as

$$P = \frac{TP}{TP + FP}, \tag{4.1}$$

where TP is the number of true positives and FP is the number of false positives.

Recall (R) is the ratio of correctly predicted positive observations to all observations in the actual class, given by

$$R = \frac{TP}{TP + FN}, \tag{4.2}$$

with FN representing the number of false negatives.

In addition to these metrics, the COCO benchmark categorizes objects based on their size: small, medium, and large. Specifically, an object is considered *small* if its bounding box area is less than 32^2 pixels, allowing for more detailed assessment of model performance across different object scales.

AP can be calculated in numerous ways, but in the realm of object detection, it is typically computed class-wise and then averaged over all classes, yielding the mean Average Precision (mAP). In this project, we adopt the COCO style benchmarking, where mAP is referred to as AP. The COCO AP is calculated over a range of Intersection over Union (IoU) thresholds and is often denoted as $AP@[0.5 : 0.05 : 0.95]$, indicating multiple IoU thresholds from 0.5 to 0.95 with a step size of 0.05. The formula for calculating COCO AP is defined as:

$$AP@[0.5 : 0.05 : 0.95] = \frac{AP_{0.5} + AP_{0.55} + \ldots + AP_{0.95}}{10}, \tag{4.3}$$

where $AP_{0.5}$ is the area under the Precision-Recall curve for IoU ≥ 0.5.

The COCO detection benchmark encompasses 12 distinct metrics, as illustrated in Table 4.3.

Average Recall (AR) offers a vital alternative perspective to AP in the assessment of object detection models. AR evaluates the ability of a model to accurately recognize *all* relevant examples of the specified categories, disregarding the incidence of false positives. In the COCO benchmark, AR is calculated with varying numbers of detections per image, specifically 1, 10, or 100. Notably, AR at 1 considers only the detection with the highest confidence score for each image, focusing on the model's precision in identifying the most probable object. This contrasts with AR at 10 or 100, where multiple detections per image are considered, reflecting the model's capability to identify numerous objects with varying confidence levels. The importance of achieving a high recall in all categories, as indicated by AR, is particularly critical in scenarios where overlooking an object could lead to severe consequences. For instance, in the context of self-driving vehicles operating under adverse weather conditions such as fog or heavy rain, the failure to detect a pedestrian or an approaching vehicle could pose a threat to human life. AR, by focusing exclusively on the rate of detection and excluding considerations of precision, emphasizes these types of missed detections that might be neglected when only considering AP. Consequently, AR provides crucial insights about the reliability and effectiveness of object detection systems, complementing the focus on precision embodied by AP.

Table 4.3: COCO metrics (Table adapted from [72])

Metric	**Description**
Average Precision (AP):	
$AP^{IoU=.50:.05:.95}$	AP at IoU=.50:.05:.95 (primary COCO metric)
$AP^{IoU=.50}$	AP at IoU=.50 (PASCAL VOC metric)
$AP^{IoU=.75}$	AP at IoU=.75 (strict metric)
AP Across Scales:	
AP^{small}	AP for small objects: area $< 32^2$
AP^{medium}	AP for medium objects: $32^2 <$ area $< 96^2$
AP^{large}	AP for large objects: area $> 96^2$
Average Recall (AR):	
$AR^{max=1}$	AR given 1 detection per image
$AR^{max=10}$	AR given 10 detections per image
$AR^{max=100}$	AR given 100 detections per image
AR Across Scales:	
AR^{small}	AR for small objects: area $< 32^2$
AR^{medium}	AR for medium objects: $32^2 <$ area $< 96^2$
AR^{large}	AR for large objects: area $> 96^2$

Given the project's focus on object detection in adverse weather conditions, we will extend the evaluation of AP and AR to include performance under specific weather scenarios like fog, rain, and snow. This will provide a more comprehensive understanding of the model's robustness and effectiveness in varying environmental conditions.

In addition to these metrics, **Inference Time** and **FLOPs** (Floating Point Operations Per Second) or **GFLOPs** (GigaFLOPs) are crucial for assessing the computational efficiency and performance of the models. Inference time, significantly influenced by the hardware used, serves as a reliable indicator of a model's practical applicability in various scenarios. In this study, the inference time for all models is tested on an NVIDIA V100 GPU, ensuring a consistent and robust basis for comparison. Furthermore, the **Model Parameters** metric is instrumental in understanding the models' complexity, shedding light on their computational requirements and potential scalability.

4.3 Selected Methods

In the field of deep multimodal sensor fusion, most architectures focus on combining two key modalities: camera and lidar. This trend is largely because camera and lidar data are readily available. However, it's less common to find methods that use all three modalities: camera, lidar, and radar. This project, which aims at 2D object detection, has carefully chosen methods that are best suited for this area.

This research also looks into tightly-coupled fusion architectures, a different approach compared to the more typical early and middle/feature fusion architectures. The methods selected for this project had to meet two main criteria: they must be able to handle at least two modalities and must use either feature

fusion or tightly-coupled fusion techniques. We have identified three methods that meet these criteria for in-depth analysis in this study, as shown in Table 4.4. Note that some literature also refers to middle-level fusion as feature fusion. All methods follow PyTorch framework. These methods are at the forefront of multimodal sensor fusion, designed to work in adverse weather conditions. The following sections provide a detailed overview of each method, highlighting its key features and contributions to the field.

Table 4.4: Selected methods. Sensors†: C-R-L denote Camera, Radar, and LiDAR sensors, respectively

Name	Sensors†	Dataset Used	Fusion Arch.	Anchors Free?	Year	Ref.
SAF-FCOS	CR	nuScenes	Middle-level	Yes	2020	[74]
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled	No	2023	[75]
MT-DETR	CRL	DENSE	Tightly-coupled	Yes	2023	[76]

4.3.1 Method 1: SAF-FCOS

Overview The paper by Chang et al. [74] introduces a novel method for enhancing obstacle detection in autonomous driving systems. This method, called spatial attention fusion (SAF), effectively integrates data from two modalities, namely millimeter-wave (mmWave) radar and camera sensors. SAF addresses the sparsity of radar points by generating an attention weight matrix that distinctively fuses vision features, diverging from traditional concatenation or element-wise addition fusion methods. This method can be integrated into the feature-extraction stage of existing deep learning object detection frameworks, facilitating end-to-end training. The method follows the middle or feature fusion approach in the paper, as it extracts features from both modalities and fuses them in the middle layer.

One of the critical challenges identified in middle fusion schemes, such as the element-wise add operation between radar and vision feature maps, was their inadequacy in handling heterogeneous data [77]. Experiments revealed that such operations were unsuitable, primarily due to the distinct characteristics of radar and vision feature maps. The SAF-FCOS framework, built upon the Fully Convolutional One-Stage (FCOS) object detection system [78], provides a solution to this challenge.

2D Annotations The nuScenes dataset provides annotations exclusively in a 3D bounding box format, necessitating a conversion to 2D for compatibility with FCOS-based detection framework. This research utilizes an enhanced version of the Fully Convolutional One-Stage (FCOS) detector [78], augmented with ResNet-101 [79], to facilitate the generation of 2D annotations. These annotations undergo a meticulous manual examination and adjustment process to ensure accuracy and reliability. The effectiveness of this custom approach is illustrated in Figure 4.7, which depicts the comparative quality of annotations derived through this refined methodology [74]. The results are evaluated using the COCO benchmark, as detailed in Section 4.2.

Dataset Split The method limits its focus on detecting vehicular obstacles such as — bicycles, cars, motorcycles, buses, trailer, and trucks—collectively categorized as 'vehicles' for simplified classification.

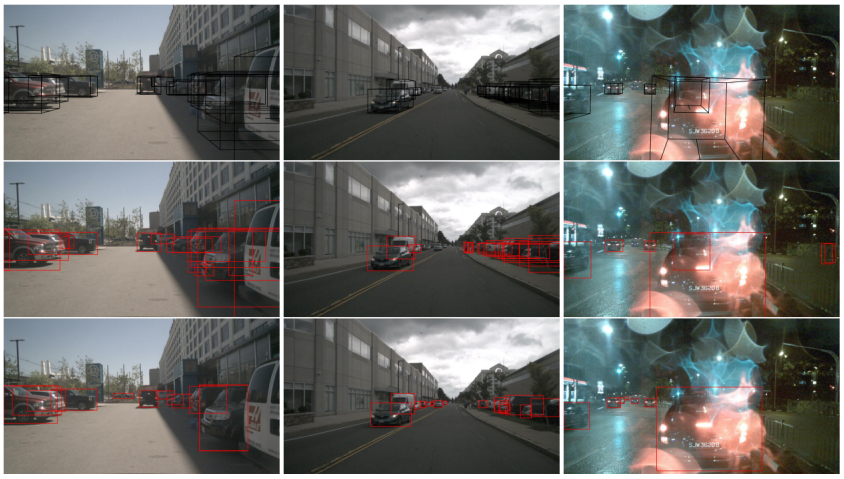


Figure 4.7: 2D annotations comparison of front camera in nuScenes dataset. Top row: the original annotations provided by the nuScenes, which are 3D bounding boxes colored by black. Middle row: the generated 2D annotations by converting the 3D bounding boxes. Bottom row: the 2D annotations generated by SAF-FCOS method. (Image source [74])

Pedestrians are excluded due to radar detection limitations in the nuScenes dataset. Emphasis is placed on front camera data from nuScenes, with the dataset split detailed in Figure 4.8.

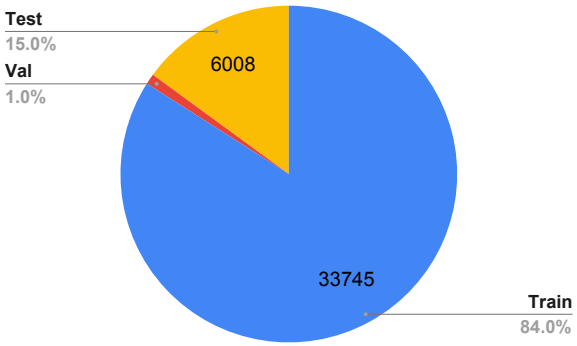


Figure 4.8: Data split of nuScenes dataset. Total samples: 40157. Note: this is a custom dataset split chosen by the authors.

Radar Imagery The process of translating radar data into visual representations is an intricate one, as depicted in Figure 4.9. It involves transforming 3D radar coordinates into camera coordinates and then rendering the radar data points in an image format suitable for neural network training.

Initially, the radar points in 3D radar coordinates (denoted as X_r) are transformed into the front camera's coordinate system using a transformation function, $X_i = X_r R + T$. Here, R represents the rotation matrix and T the translation matrix. This transformation allows us to locate the radar point

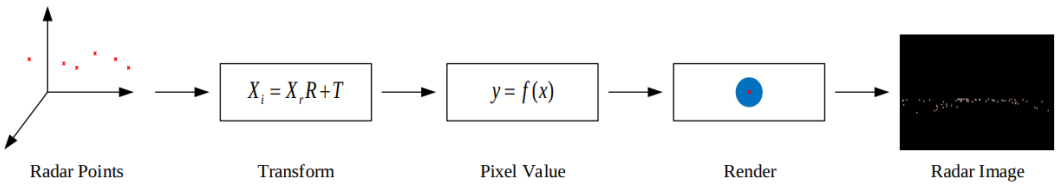


Figure 4.9: Radar image generation model (Image source [74]).

(X_i) in the camera coordinate system [74].

Following this, depth (d), longitudinal velocity (v_x), and lateral velocity (v_y) are converted into real pixel values in different color channels (Red, Green, Blue) using the equation:

$$R = \frac{128d}{250} + 127, \quad G = \frac{128(v_x + 20)}{40} + 127, \quad B = \frac{128(v_y + 20)}{40} + 127 \tag{4.4}$$

The term "rendering" in the below context refers to how individual radar data points are visually represented in the image. There are two primary cases to consider:

Rendering Case A: When two radar points, M and N, are significantly separated on the image plane (specifically, if the distance l between them exceeds twice the rendering radius r), they are rendered independently without overlap. Each radar point is visualized as a circle with radius r, ensuring no intersection between the circles due to sufficient distance [74].

Rendering Case B: Conversely, when the distance l between two radar points, M and N, is less than twice the rendering radius r, their rendered shapes overlap. In such instances, the "nearer the larger" principle is applied. If radar point M is nearer to the sensor (indicated by a smaller depth value d_M than d_N of point N), it is rendered larger or more prominently in the overlap area. This results in point M covering a more significant area in the overlap region compared to point N [74].

Both cases are illustrated in Figure 4.10.

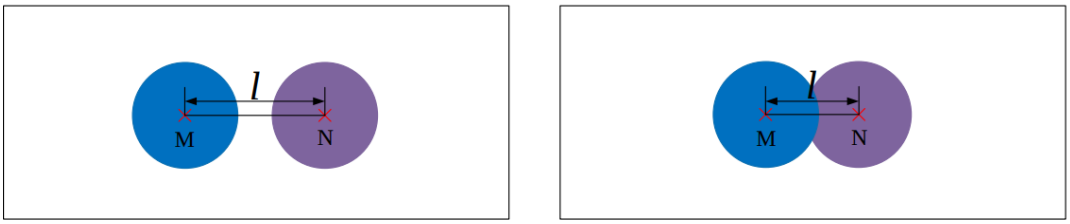


Figure 4.10: Radar image rendering condition (Image source [74]).

Model Architecture A significant aspect of this architecture is the introduction of the Spatial Attention Fusion (SAF) block, designed to comprehend the relationship between radar and vision data. This element is prominently featured in the network architecture, as illustrated in Figure 4.11. The middle fusion detection model is structured around the FCOS framework [78], comprising five primary components: the Radar Branch, Vision Branch, SAF block, Fusion Branch, and RetinaNet [78]. The FCOS prediction head is utilized here to make the architecture anchor-free.

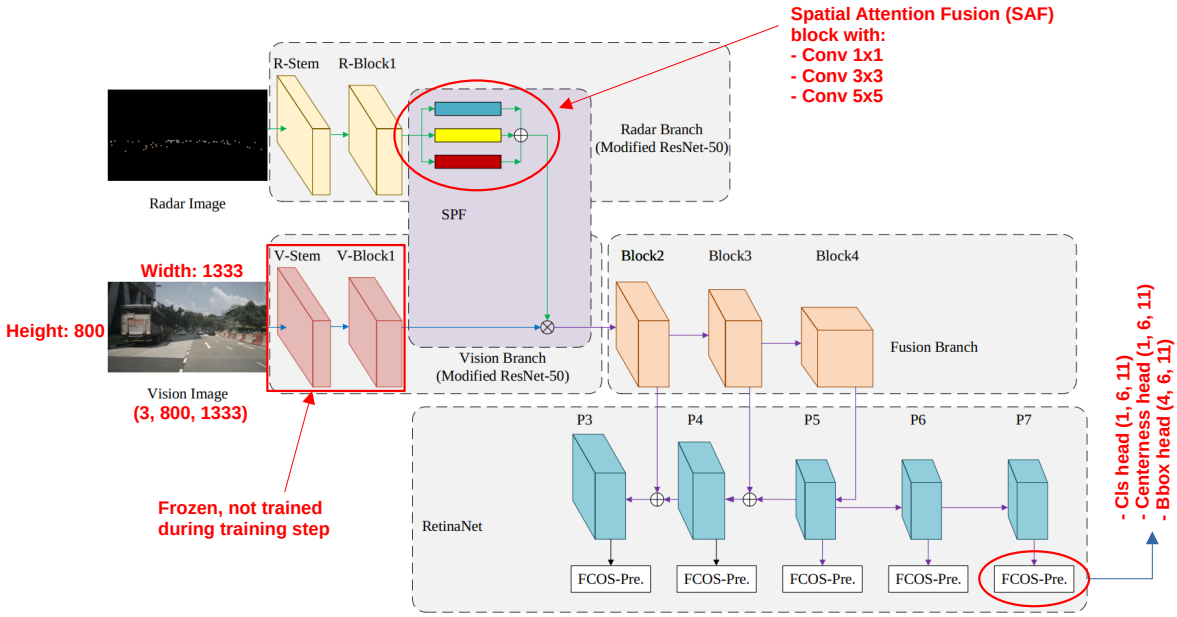


Figure 4.11: SAF-FCOS model architecture (Image adapted [74]).

Both the Radar and Vision Branches employ a modified version of the ResNet-50 [79] model, consisting of two convolution blocks, namely Stem and Block1. The Stem is the original stem module of ResNet-50, responsible for processing the input data. In contrast, Block1, mirroring the first stage in ResNet-50, differs by having only a single residual block instead of the three found in the standard ResNet-50 structure.

The Vision Branch and Fusion Branch used in SAF-FCOS are initially pre-trained on the ImageNet [80] dataset. Subsequently, all components are fine-tuned on the nuScenes [20] dataset. During the training phase, the Vision Branch is frozen and not updated. Additionally, Xavier initialization, which is the default in the PyTorch framework, is used for initializing the other blocks, except for the SAF block, which is initialized using the MSRA method [81].

The SAF block functions by encoding the radar image's feature maps into a spatial attention weight matrix. This matrix is then applied to re-weight the feature maps extracted by the vision sensor across all channels. Subsequently, the fused feature maps from both radar and vision sensors are processed by Block2, Block3, and Block4 in the Fusion Branch. These blocks are identical to the corresponding stages

in the ResNet-50 backbone used in the FCOS [78] framework. The SAF block is composed of three groups of convolution layers designed to extract the spatial attention matrix. The layer configurations include 'Conv 1×1' with kernel size $1 \times 1 \times 256 \times 1$, stride (1, 1), and padding [0, 0]. The 'Conv 3×3' and 'Conv 5×5' layers have configurations of $\{3 \times 3 \times 256 \times 1, (1, 1), [1, 1]\}$ and $\{5 \times 5 \times 256 \times 1, (1, 1), [2, 2]\}$, respectively. The design of these layers aims to generate an attention matrix with multi-scale receptive fields, thus facilitating the learning of radar points' representation and their environmental relationships. This matrix serves as a control mechanism to enhance the information flow within the vision sensor.

The rationale behind the SAF block stems from the inadequacy of other fusion blocks, such as add fusion and concatenation fusion in Figure 4.12, in effectively integrating radar and vision features. Previous studies, including the work of RVNet [82] and Nobis et al. [44], have experimented with these methods. However, these fusion techniques do not adequately address the non-homogeneous nature of radar and vision features, often neglecting the unique characteristics of radar signals. The SAF block, in contrast, utilizes radar points as gate cells to control the information flow from the vision sensor, thereby enhancing the detection of small or blurred objects and improving the recall rate. This approach differs from early or data fusion methods by considering areas devoid of radar points, thus offering a more comprehensive and effective fusion strategy.

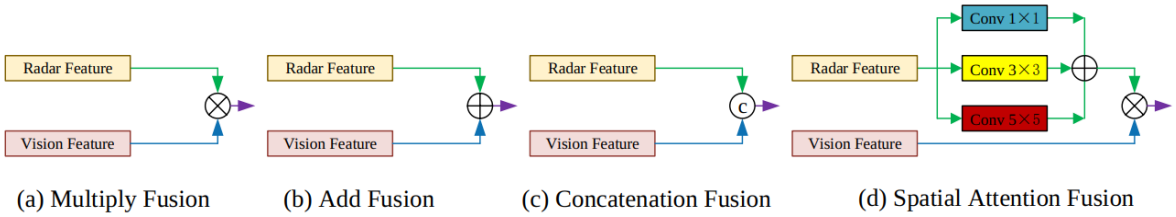


Figure 4.12: Different fusion blocks for feature fusion. From the left to right: Multiply Fusion (MUL) Block, Element-Wise Add Fusion (ADD) Block, Concatenation Fusion (CAT) Block and Spatial Attention Fusion (SAF) Block (Image source [74]).

Loss Function The loss function of the model is formulated with two principal components: the classification loss and the bounding box regression loss. For each spatial location i on the feature map, the classification loss compares the predicted class label c_i with the ground-truth label c_i^*. Simultaneously, the bounding box regression loss measures the discrepancy between the predicted bounding box coordinates t_i and the ground-truth bounding box t_i^*. The classification loss, L_{cls}, is computed using the focal loss method as detailed in [83], while the bounding box regression loss, L_{reg}, employs the IOU loss as per the UnitBox approach [84]. Here, N_{pos} signifies the number of positive samples. In this work, the balancing weight λ for L_{reg} is set to 1. The indicator function $\mathbb{1}_{c_i^*>0}$ is 1 if $c_i^* > 0$, and 0 otherwise [74].

The computation of the loss function is formalized as follows:

$$L(c_i, t_i) = \frac{1}{N_{\text{pos}}} \left(\sum_i L_{\text{cls}}(c_i, c_i^*) + \lambda \sum_i \mathbb{1}_{c_i^*>0} L_{\text{reg}}(t_i, t_i^*) \right) \tag{4.5}$$

4.3.2 Method 2: HRFuser

Overview Another work by Broedermann et al. [75] presents an extended work on HRNet [85] and HRFormer [86] to integrate multimodal sensors into a single network. It introduces HRFuser, a versatile, multi-resolution, multi-sensor fusion architecture that can efficiently integrate an arbitrary number of sensors like lidar, radar, and gated cameras, alongside standard cameras. HRFuser is built on the HRNet and HRFormer paradigms, preserving high-resolution representations and incorporating a novel multi-window cross-attention (MWCA) block for effective fusion across multiple resolutions. The system's generic design allows for easy scalability with various sensors without the need for specialized components for each sensor. Extensive testing on major autonomous driving datasets, including nuScenes [20], and DENSE [6], demonstrates HRFuser's superior performance over existing camera-only networks and sensor fusion methods, proving its efficacy in both standard and adverse weather conditions.

2D Annotations In addition to cameras, lidar sensors are also greatly impacted by harsh weather, as explored in Section 3.1. Such weather conditions lead to incomplete lidar-generated 3D object annotations. Figure 4.13 showcases both 2D and 3D tags from the DENSE dataset [6], illustrating that a substantial portion of objects (41.91% in the "dense fog" division of DENSE [6]) are only marked with 2D annotations, owing to the absence of accurate lidar readings caused by environmental elements like fog or rain. These factors result in inadequate and untrustworthy signals for generating 3D annotations. Nevertheless, detecting all critical safety objects is essential in challenging driving scenarios, even when their exact 3D placement is unattainable [75]. Therefore, the approach utilizes 2D annotations from the DENSE dataset.

For the nuScenes dataset [20], the creation of 2D annotations involves projecting 3D bounding boxes onto each image plane. This is done by calculating a convex hull rectangle from the corners of the projected boxes, a technique similar to that described in [87]. Initially, both the 3D annotations and point clouds are aligned with the vehicle's coordinate system. Subsequently, these 3D annotations are transformed into 2D bounding boxes through projection onto an image plane. During this process, any annotations categorized in the lowest visibility bin are discarded, effectively filtering out occluded boxes. For evaluation purposes, the paper adopts the KITTI style benchmarking [13] when analyzing the DENSE dataset. In contrast, for nuScenes, it employs the COCO style benchmarking method [72].

Table 4.5: Overview of sensor projection parameters. RCS stands for Radar Cross Section.

	DENSE	nuScenes
Sensor Imagery	**Sensor Parameters**	
Radar	Distance, Velocity	Distance, Velocity, RCS
Lidar	Distance, Intensity, Height	Distance, Intensity, Height

Radar and Lidar Imagery Before inputs are fed into HRFuser, the method projects all secondary modalities onto the camera's image plane. This projection ensures precise spatial correspondence between the input feature maps of various modalities. To achieve this, the approach combines depth, height, and

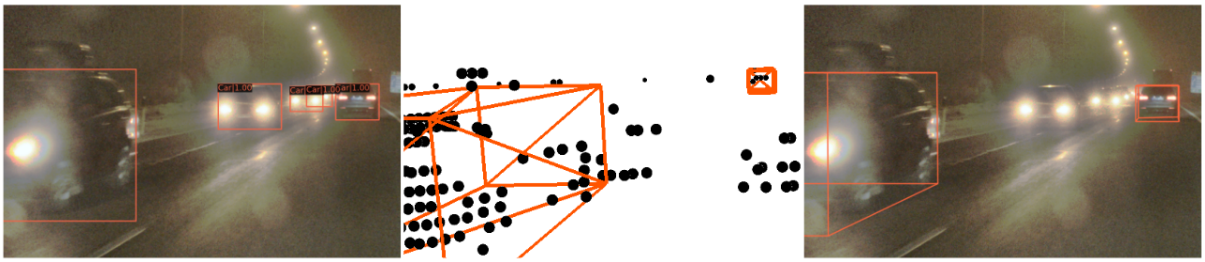


Figure 4.13: A scene from DENSE [6] illustrates (on the left) 2D and (on the right) 3D object labels. Due to adverse weather conditions affecting the (middle) point cloud, numerous critical objects are unaccounted for, resulting in them only being marked with 2D annotations and not in 3D. (Image source [75]).

pulse intensity for lidar image input, rather than relying solely on depth. In the case of radar imagery, the method posits that the radar scans in a two-dimensional plane orthogonal to the image plane and aligned with the horizontal axis of the image. This alignment implies that radar consistency is preserved along the image's vertical axis, necessitating the duplication of scans vertically. The proposed method for input encoding depends on both depth and velocity for the DENSE dataset, and includes Radar Cross Section (RCS) for the nuScenes dataset, enabling precise pixel matching across different data streams. Table 4.5 outlines the sensor projection parameters employed in both datasets. Notably, areas where measurements are missing are encoded with zero values. A representative image from the nuScenes dataset is illustrated in Figure 4.14.

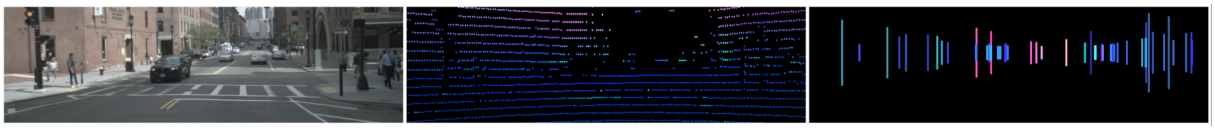


Figure 4.14: An example projection image from nuScenes [20]. Viewing sequence: RGB image, lidar point projection, radar point projection. Best viewed in zoomed-in view. (Image adapted from [75]).

Model Architecture The paper presents HRFuser, a design combining different resolutions and sensors for 2D object detection. The key idea of HRFuser is to keep high-quality, detailed images at every level of the processing steps, following the approach used in earlier research [85] [86]. This design is expanded to work with different types of data inputs such as lidar, radar, and it introduces an effective way to fuse data from various sensors. HRFuser stands out by having separate, simplified processing paths for each non-primary data input. The primary camera input, on the other hand, produces extra detailed, but less sharp features. The structure and details of HRFuser are illustrated in Figure 4.15 [75].

The study introduces the multi-window cross-attention (MWCA) block, a key component for integrating secondary sensors with the primary camera branch, enabling effective feature fusion at various resolutions. Illustrated in Figure 4.16, MWCA diverges from conventional fusion techniques like concatenation, addition or element-wise multiplication, employing transformer-based approach for more efficient integration. This approach efficiently combines features at various resolutions while addressing the quadratic complexity

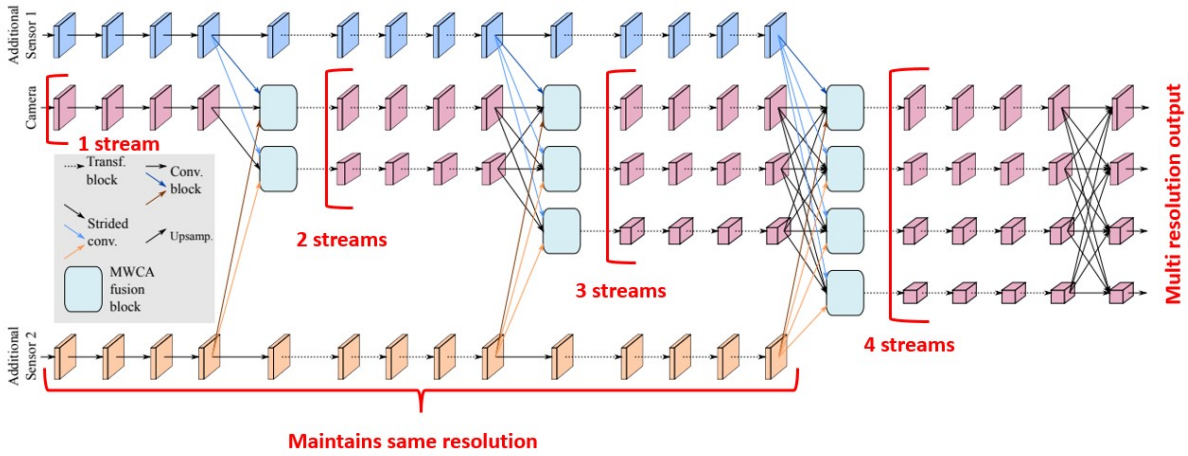


Figure 4.15: HRFuser architecture (Image adapted from [75]).

typically associated with attention mechanisms. By employing multi-head cross-attention on small, non-overlapping local windows, MWCA significantly reduces computational costs, making it feasible for high-resolution feature maps. Each window processes K^2 tokens of dimensionality D, based on the channel number of the fused stream, allowing for focused attention on pertinent sensor features and enhanced noise filtering. This method improves the performance of all sensors, especially those susceptible to high noise like radar. The computational overhead of this integration is relatively low, with only a +9.7% increase in floating-point operations (flops) and +1.9% in parameters [75].

To address the challenge of fusing high-resolution secondary branches with a lower resolution primary branch, as depicted in the annotated architecture in Fig. 4.15, it downsample the secondary branch with 7x7 convolutions. This approach ensures that the high-resolution branch aligns with the primary branch, preserving detail across all modalities. By leveraging local windows for fusion, the MWCA block efficiently combines local and more global relationships among different sensor data [75].

The HRFuser structure, depicted in Figure 4.15, incorporates a 'neck' that creates a feature pyramid. This is achieved by merging the enlarged outputs from all channels [85]. Following the neck is a Cascade R-CNN head [88], adhering to the common two-stage detection framework. Cascade R-CNN enhances detection by using a series of detectors, each trained at progressively higher Intersection over Union (IoU) thresholds, thus establishing a robust benchmark for the backbone [75]. It's important to note that the anchor-based Cascade R-CNN head can be substituted with anchor-free detection systems like FCOS [78].

To elaborate in more detail, the primary modality's input feature map, denoted as X for modality α, is segmented into P distinct spatial windows without any overlap: $X^\alpha \xrightarrow{\text{Split}} \{X_1^\alpha, X_2^\alpha, \ldots, X_P^\alpha\}$. Each window is vectorized, transforming the $K \times K$ spatial dimensions into a sequence of vectors, resulting in a total of K^2 spatial locations per window. The identical division and vectorization are also implemented on the feature maps Y^β of every secondary modality β, which encompasses $\beta \in \{1, \ldots, M\}$: $Y^\beta \xrightarrow{\text{Split}} \{Y_1^\beta, Y_2^\beta, \ldots, Y_P^\beta\}$, where M denotes the total number of secondary modalities.

The vectorization of all these input feature maps occurs across spatial dimensions, ensuring they

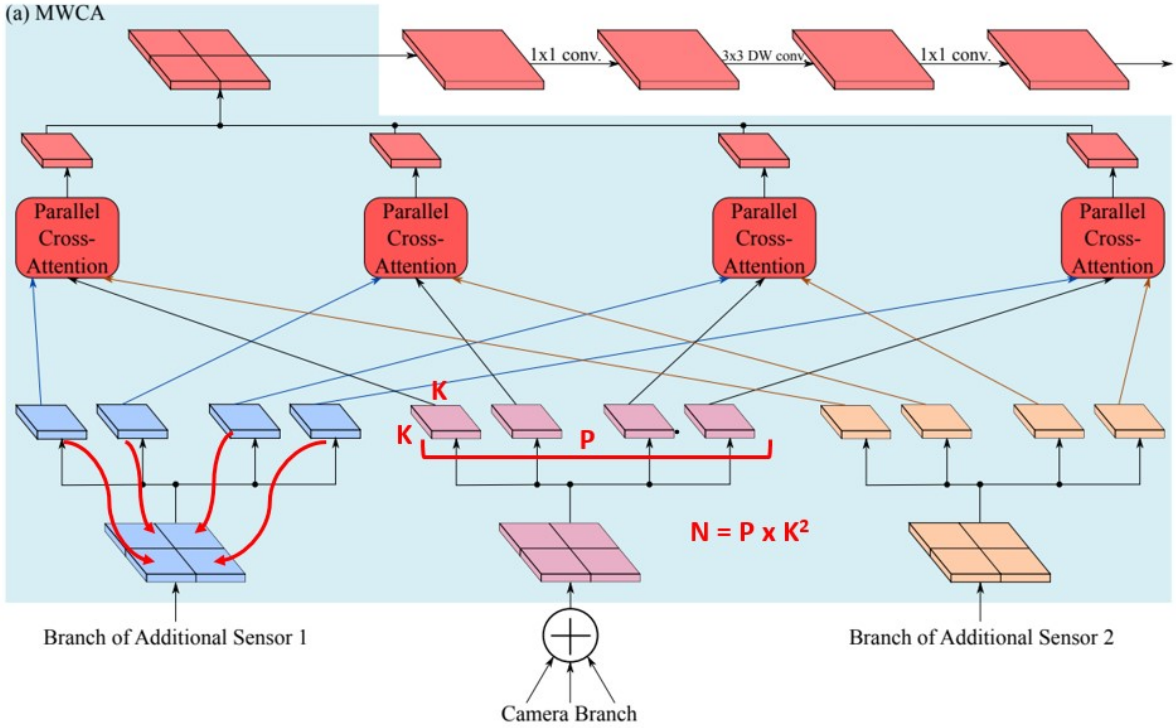


Figure 4.16: Proposed multi-window cross-attention block followed by a feed-forward network. DW conv. denotes depth-wise convolution. (Image adapted from [75]).

possess the uniform shape $X, Y \in \mathbb{R}^{N \times D}$. Here, N signifies the aggregate number of spatial locations after vectorization across all windows, while D represents the channel count. Specifically, for P windows each with $K \times K$ dimensions, $N = P \times K^2$, representing the total number of vectorized spatial locations available for attention mechanisms [75].

A localized transformer concurrently conducts cross-attention processes on individual, corresponding window groups. This approach of parallel cross-attention is demonstrated in detail in Fig. 4.17. The method for applying this parallel cross-attention to the p-th set of windows is outlined as follows [75]:

$$\text{MultiHead}(X_p^\alpha, Y_p^\beta) = \text{Concat}[head(X_p^\alpha, Y_p^\beta)_1, \ldots, head(X_p^\alpha, Y_p^\beta)_H] \in \mathbb{R}^{K^2 \times D}, \tag{4.6}$$

$$head(X_p^\alpha, Y_p^\beta)_h = \text{Softmax}\left(\frac{(X_p^\alpha W_q^{h,\beta})(Y_p^\beta W_k^{h,\beta})^T}{\sqrt{D/H}}\right) Y_p^\beta W_v^{h,\beta} \in \mathbb{R}^{K^2 \times D/H}, \tag{4.7}$$

$$\hat{X}_p = X_p^\alpha + \sum_{\beta=1}^{M} \left[Y_p^\beta + \text{MultiHead}(X_p^\alpha, Y_p^\beta) W_o^\beta\right] \in \mathbb{R}^{K^2 \times D} \tag{4.8}$$

where W_o^β is defined as $W_o^\beta \in \mathbb{R}^{D \times D}$, and the matrices $W_q^{h,\beta}, W_k^{h,\beta}, W_v^{h,\beta}$ for each head h, where $h \in$

$\{1, \ldots, H\}$, are represented as $W_q^{h,\beta}, W_k^{h,\beta}, W_v^{h,\beta} \in \mathbb{R}^{D \times \frac{D}{H}}$. These are the weight matrices, implemented by trainable linear projections. Here, H represents the total count of heads, and $\hat{X}_p$ is the resultant output from the concurrent cross-attention operations applied to the p-th window set.

The process culminates by reassembling the output from every set of P windows into a unified feature map, yielding the final MWCA output, denoted as X_{MWCA} [75].

$$\{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_P\} \xrightarrow{\text{Merge}} X_{\text{MWCA}}. \tag{4.9}$$

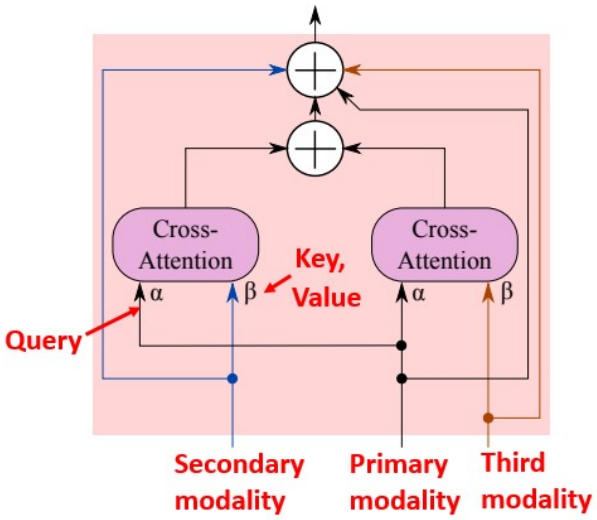


Figure 4.17: Parallel cross-attention block (Image adapted from [75]).

Loss Function The loss function aligns closely with the one delineated in SAF-FCOS 4.3.1. In this framework, CrossEntropy is employed for the calculation of classification loss, whereas SmoothL1 is utilized for regression loss. The formulation of the objective function is as follows [88]:

$$L(p_i, t_i) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{CE}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{SmoothL1}}(t_i, t_i^*). \tag{4.10}$$

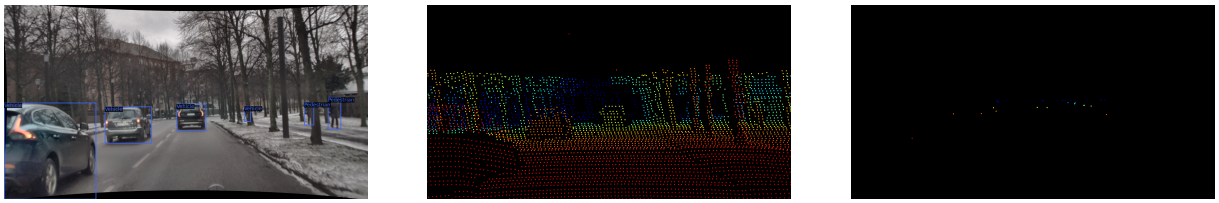
In this formula, p_i represents the probability estimate for the i^{th} anchor's objectness. Conversely, p_i^* is the actual ground truth score, assigned 1 for anchors with objects and 0 for those without. The parameter vector t_i captures the predicted dimensions and location of the bounding box, while t_i^* holds this information for the actual bounding box. The term L_{CE} denotes the cross-entropy loss for the classification task, and L_{SmoothL1} refers to the Smooth L1 loss for regression. Normalization is achieved via N_{cls} for classification and N_{reg} for regression. The scalar λ is instrumental in equilibrating the impact of the classification and regression losses within the total loss computation.

4.3.3 Method 3: MT-DETR

Overview Recent work by Chu et al. [76] proposes a novel end-to-end multimodal multistage object detection network called MT-DETR (MulTi-sensor MulTimodal DTtection TRansformer) that leverages data from multiple sensors - camera, lidar, radar and time - to achieve robust detection, especially in adverse weather conditions. Here time modality is an additional binary image input to inform model about day or night. It employs specialized fusion modules - Residual Fusion Module (RFM) and Confidence Fusion Module (CFM) for hierarchical cross-modal feature fusion. The network also uses a Residual Enhancement Module (REM) to strengthen individual sensor branches. A multi-stage loss function further regularizes feature learning across modalities. Extensive experiments on the publicly available DENSE [6] dataset demonstrate that MT-DETR significantly outperforms existing unimodal and multimodal detection methods.

2D Annotations This method employs the same 2D annotations as described in Section 4.3.2 for the DENSE dataset [6]. However, for evaluation purposes, it diverges from the HRFuser approach for the DENSE dataset [6] by utilizing the COCO-style benchmarking [72].

Radar and Lidar Imagery The approach described herein closely mirrors the HRFuser method 4.3.2 for the DENSE dataset [6], with a notable distinction: it exclusively utilizes a single sensor parameter, namely the depth (or range) values, for both radar and lidar imagery. Unlike the original method, where the authors provided a script for sensor imagery generation, this necessitated the development of our custom script. The code for generating projection images has been adapted from [6]. Representative examples of the projection images for lidar and radar generated using this adapted code are illustrated in Figure 4.18.



(a) Camera image with bbox (b) Lidar point projection (c) Radar point projection

Figure 4.18: Sample projection images for lidar and radar from DENSE dataset [6]. Each point is colored according to its depth value. Best viewed in zoomed-in view.

Model Architecture Prior to delving into the core structure of MT-DETR, this study also mentions various other object detection frameworks, including unimodal and middle fusion approaches, as depicted in Figures 4.19a and 4.19b. Unimodal techniques primarily process RGB camera imagery. These methods involve extracting features through a backbone network, followed by the utilization of multi-scale features in the detection head for object prediction. Conversely, middle fusion architectures independently extract

multimodal features through distinct branches and subsequently integrate these features for object detection. Building upon these concepts, the proposed MT-DETR innovates as a tightly-coupled network, uniquely designed for comprehensive end-to-end object detection. It distinctively processes inputs from multiple sensors by fusing their features concurrently.

(a) Unimodal (b) Middle fusion

Figure 4.19: Comparison of two types of object detection model architecture. (a) Unimodal architecture with camera image. (b) Middle fusion based multimodal architecture with three modalities (Image adapted from [76]).

The core of MT-DETR architecture builds upon the Transformer-based [89] approach, as utilized in the innovative DETR and its extension, Deformable DETR [90]. DETR conceptualizes object detection as a set prediction problem and simplifies the detection process by eliminating numerous conventional, manual designs. Similarly, Deformable DETR enhances this model by accelerating convergence and improving overall performance. Extending these advancements, MT-DETR employs a multimodal backbone, integrating data from various sensors and utilizing the Transformer [89] architecture in its detection head, akin to Deformable DETR. This approach allows for simultaneous processing of sensor data with fused feature sets, marking a significant step forward in object detection technology [76].

MT-DETR processes image data sourced from various sensors through its backbone, which yields multi-scale features as a result of multimodal fusion. The structural design of MT-DETR, illustrated in Fig. 4.20, encompasses four primary elements: a feature extraction unit, a fusion module, an enhancement module, and a detection head. Within this framework, ConvNeXt [91] is employed for feature extraction in distinct unimodal branches. These branches' outputs are then integrated in the fusion module. The enhancement module works by amalgamating these fused features with those from the unimodal branches, thereby enriching the unimodal features before the next scale of feature extraction. In the final stage, the detection head receives the integrated features from each scale, processed by the fusion module, to carry out the object prediction task.

In the model architecture, as depicted in one of its variations, an additional 'time image' is used as the fourth input, same dimension as the image. This time image is a straightforward binary representation designed to guide the model in differentiating between day and night. Examples of this can be seen in Figures 4.21 and 4.22.

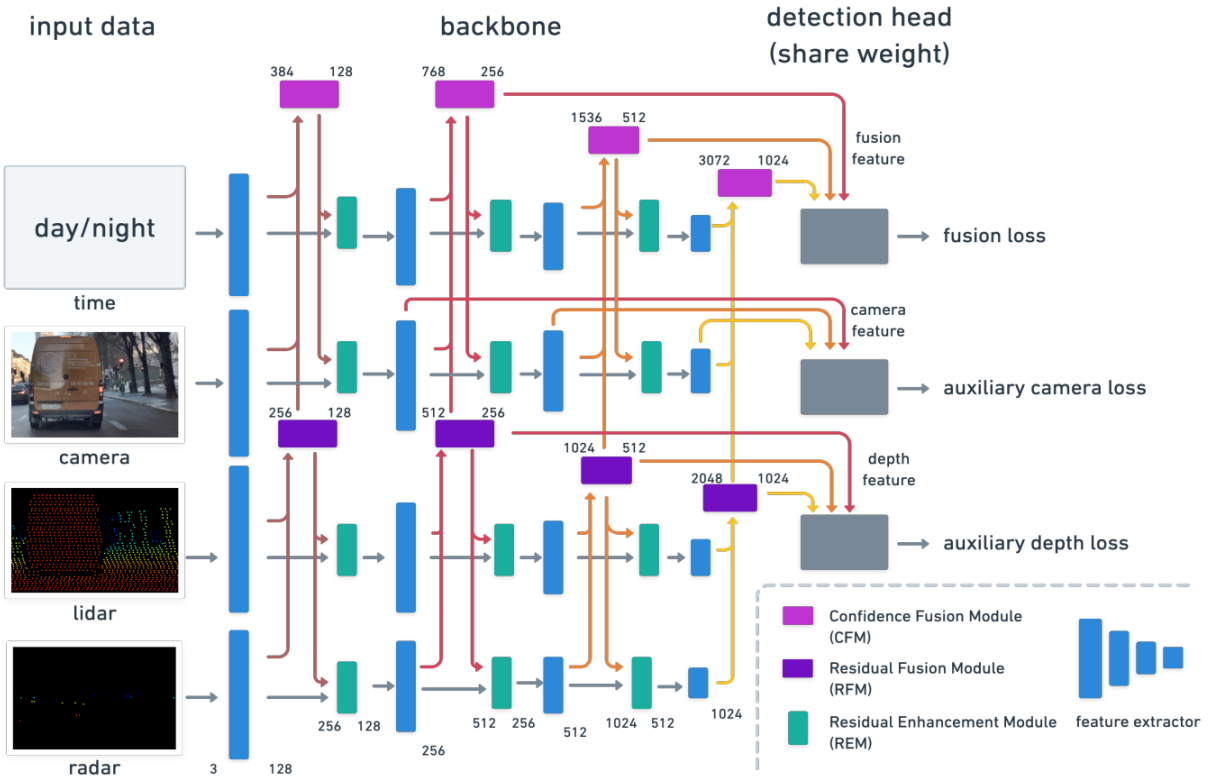


Figure 4.20: MT-DETR model architecture (Image source [76]).

The MT-DETR's 'Hierarchical Fusion Mechanism' addresses the potential loss of sensor relationships and priorities that may occur when features from all branches are merged simultaneously. This mechanism is particularly effective given the similarity in data types between lidar and radar. By initially fusing these two, the mechanism facilitates a more comprehensive and precise capture of depth information. Subsequently, this depth feature, derived from the lidar and radar fusion, is combined with data from the camera and time branches. This hierarchical approach in fusing modalities allows the model to gain a more nuanced understanding of depth, enhancing the overall efficacy of the fusion process.

The following text describes the two key fusion modules utilized in the architecture: the Residual Fusion Module (RFM) and the Confidence Fusion Module (CFM). Figure 4.23 illustrates the design of these modules.

In the context of fusion modules, the Residual Fusion Module (RFM) and Confidence Fusion Module (CFM) are introduced for fusion. RFM merges lidar and radar features to obtain depth features, emphasizing information from both by concatenating and reducing dimensions through convolution. Meanwhile, CFM combines depth features with camera and time features, creating a confidence map through concatenation and convolution. This map is then used for element-wise operations, resulting in the final fusion feature, as shown in Figure 4.23 (a)(b).

The depth feature F_i^{depth} and fusion feature F_i^{fusion} for the i-th stage can be calculated as follows:

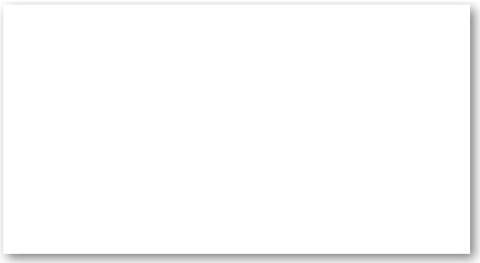


Figure 4.21: Day time image

Figure 4.22: Night time image

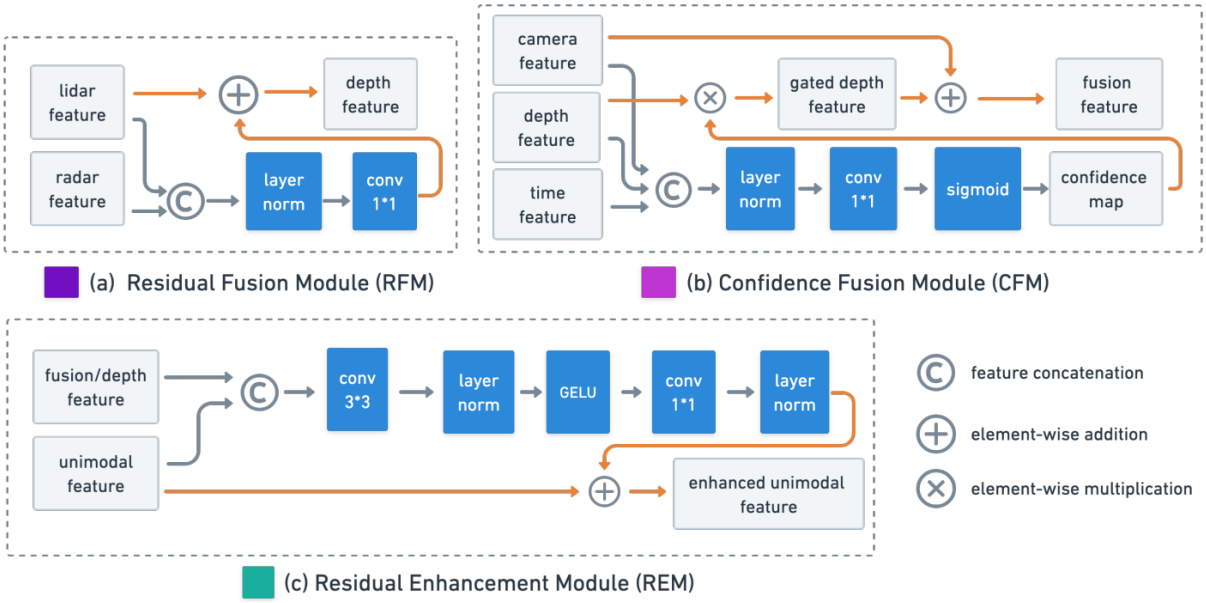


Figure 4.23: Fusion modules and enhancement module design (Image source [76]).

$$F_i^{\text{depth}} = RFM(F_i^{\text{lidar}}, F_i^{\text{radar}}) = F_i^{\text{lidar}} + \text{Conv}_{1\times1}(F_i^{\text{lidar}} \oplus F_i^{\text{radar}}), \tag{4.11}$$

$$\begin{aligned} F_i^{\text{fusion}} &= \text{CFM}\left(F_i^{\text{camera}}, F_i^{\text{depth}}, F_i^{\text{time}}\right) \\ &= F_i^{\text{camera}} + \left(F_i^{\text{depth}} * \sigma\left(\text{Conv}_{1\times1}\left(F_i^{\text{camera}} \oplus F_i^{\text{depth}} \oplus F_i^{\text{time}}\right)\right)\right), \end{aligned} \tag{4.12}$$

Where F_i^{camera}, F_i^{lidar}, F_i^{radar}, and F_i^{time} represent the feature outputs from the i-th stage ($i = 1, 2, 3, 4$) of each unimodal feature extractor. $\oplus$ denotes a feature concatenation operation, $*$ and $+$ indicate element-wise multiplication and addition, respectively. $\sigma(\cdot)$ and $Conv_{1\times1}$ represent the sigmoid function and a 1×1 convolution block, respectively.

The Residual Enhancement Module (REM) is designed to strengthen the unimodal branch. Illustrated

in Figure 4.23(c), REM shares structural similarities with RFM but incorporates a deeper convolution layer. Furthermore, REM places heightened emphasis on the characteristics of the unimodal branch, merging the output of the convolution block with the unimodal feature to create an enhanced feature. It is worth highlighting that the camera and time branches benefit from fusion features, whereas the lidar and radar branches are augmented through depth features.

The feature $\tilde{F}_i^m$ for each individual unimodal branch can be derived through the following process:

$$\begin{aligned} \tilde{F}_i^m &= \mathrm{REM}_m(F_i^m, F_i^{\mathrm{fusion}}), \text{ for } m \in \{\mathrm{camera}, \mathrm{time}\} \\ &= F_i^m + \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(\tilde{F}_i^m \oplus F_i^{\mathrm{fusion}})), \end{aligned} \tag{4.13}$$

$$\begin{aligned} \tilde{F}_i^m &= \mathrm{REM}_m(F_i^m, F_i^{\mathrm{depth}}), \text{ for } m \in \{\mathrm{lidar}, \mathrm{radar}\} \\ &= F_i^m + \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(F_i^m \oplus F_i^{\mathrm{depth}})), \end{aligned} \tag{4.14}$$

then, the feature at the subsequent scale, denoted as F_{i+1}^m, can be obtained as follows:

$$F_{i+1}^m = \mathrm{FE}_{i+1}^m(\tilde{F}_i^m), \text{ for } i \in \{1, 2, 3\}, \tag{4.15}$$

Where $REM_m(\cdot,\cdot)$ and $FE_i^m(\cdot)$ represent the REM module and feature extractor for each unimodal branch at the i-th stage.

For $m \in \{fusion, camera, depth\}$, the method gathers $F_m = \{F_i^m | i = 2, 3, 4\}$ and input them into the subsequent detection head for further prediction.

Loss Function In scenarios where sensor data is significantly unbalanced, a tendency may arise in the fusion mechanism to overly rely on camera-derived features, neglecting other unimodal inputs. Drawing inspiration from [88] [92], this approach involves integrating both the final fused features and the intermediate stage features into the detection head. This integration is crucial for the computation of the auxiliary loss function, which is instrumental in ensuring comprehensive information extraction across all branches. For the unimodal object detection, the model aligns with the Deformable DETR [90] in adopting a Focal loss [83] for classification, coupled with l1 loss and Generalized IoU loss [93] for bounding box regression, following the weightage pattern of Deformable DETR [90].

Subsequently, predictions from F_{fusion} lead to the computation of the fusion-specific loss $\mathcal{L}_{\mathrm{fusion}}$; those from F_{camera} result in the camera-specific loss $\mathcal{L}_{\mathrm{camera}}$; and predictions from F_{depth} contribute to the depth-specific loss $\mathcal{L}_{\mathrm{depth}}$. Here, $\mathcal{L}_{\mathrm{fusion}}$ serves as the primary loss component, whereas $\mathcal{L}_{\mathrm{camera}}$ and $\mathcal{L}_{\mathrm{depth}}$ function as supplementary loss components.

Let P_m represent the MT-DETR's modality-specific predictions and $\hat{P}$ denote the corresponding ground truth. For each modality m in the set {fusion, camera, depth}, the comprehensive loss function is defined as:

$$\mathcal{L}_m = 2\mathcal{L}_{\text{focal}}(P_m, \hat{P}) + 5\mathcal{L}_1(P_m, \hat{P}) + 2\mathcal{L}_{\text{GIoU}}(P_m, \hat{P}), \tag{4.16}$$

$$\mathcal{L}_{total} = \lambda_{fusion}\mathcal{L}_{fusion} + \lambda_{camera}\mathcal{L}_{camera} + \lambda_{depth}\mathcal{L}_{depth} \tag{4.17}$$

where $\mathcal{L}_{\text{focal}}(\cdot,\cdot)$ signifies the focal loss [83], $\mathcal{L}_1(\cdot,\cdot)$ represents the l_1 loss, and $\mathcal{L}_{\text{GIoU}}(\cdot,\cdot)$ denotes the GIoU loss [93]. The weights for these loss functions are in accordance with the Deformable DETR [90]. The coefficients $(\lambda_{\text{fusion}}, \lambda_{\text{camera}}, \lambda_{\text{depth}})$ are assigned values of (1, 1, 0.5) based on experimental findings.

5

Evaluation and Results

5.1 Experiment Description

Describe the experiments/evaluation you are performing to analyse your method.

5.2 Experimental Setup

Describe your experimental setup in detail.

5.3 Results

Describe the results of your experiments in detail.

6

Conclusions

6.1 Contributions

6.2 Lessons learned

6.3 Future work

Utilization of 4D Imaging Radar in Adverse Weather: There is a notable lack of research utilizing 4D imaging radar sensors, especially in adverse weather conditions [94]. Given the potential of these sensors in challenging environments, further exploration in this area is essential. The K-Radar dataset [30] is a step in the right direction, but yet to be explored.

A

Design Details

Your first appendix

B

Parameters

Your second chapter appendix

References

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[2] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda, "Libre: The multiple 3d lidar dataset," *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1094–1101, 2020.

[3] U. of Michigan, "Getting traction: Tips for traveling in winter weather," 2020. [Online]. Available: https://mcity.umich.edu/wp-content/uploads/2020/10/mcity-driverless-shuttle-whitepaper.pdf

[4] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[5] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

[6] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," pp. 11 682–11 692, 2020.

[7] Federal-Highway-Administration, "How Do Weather Events Impact Roads? - FHWA Road Weather Management." [Online]. Available: https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

[8] N. US Department of Commerce, "Getting traction: Tips for traveling in winter weather," Nov 2016. [Online]. Available: https://www.weather.gov/wrn/getting_traction

[9] G. Cookson, "Weather-Related Road Deaths in Europe £15bn+ Per Year — INRIX," 2 2022. [Online]. Available: https://inrix.com/blog/blog-cost-of-weather/

[10] U. Briefs, "Mcity grand opening," *Research Review*, vol. 46, no. 3, 2015.

[11] F. Lambert, "Watch tesla autopilot go through a snowstorm," https://electrek.co/2019/01/28/tesla-autopilot-snow-storm/, 2019, [Last accessed 10 May 2021].

[12] Cadillac General Motors, "Designed to take your hands and breath away," https://www.cadillac.com/ownership/vehicle-technology/super-cruise, 2021, last accessed 10 May 2021.

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

[14] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.

[15] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[16] A. Palffy, E. Pool, S. Baratam, J. F. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[17] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan *et al.*, "Tj4dradset: A 4d radar dataset for autonomous driving," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 493–498.

[18] J.-L. Déziel, P. Merriaux, F. Tremblay, D. Lessard, D. Plourde, J. Stanguennec, P. Goulet, and P. Olivier, "Pixset: An opportunity for 3d computer vision to go beyond point clouds with a full-waveform lidar dataset," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2987–2993.

[19] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," pp. 1341–1360, 2020.

[20] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.

[21] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller *et al.*, "Making bertha drive—an autonomous journey on a historic route," *IEEE Intelligent transportation systems magazine*, vol. 6, no. 2, pp. 8–20, 2014.

[23] R. Mardirosian, "Lidar vs. camera: driving in the rain," [Last accessed 10 April 2023], 2023. [Online]. Available: https://ouster.com/blog/lidar-vs-camera-comparison-in-the-rain/

[24] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, and M. A. Kaafar, "The impact of adverse weather conditions on autonomous vehicles: how rain, snow, fog, and hail affect the performance of a self-driving car," *IEEE vehicular technology magazine*, vol. 14, no. 2, pp. 103–111, 2019.

[25] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda, "LIBRE: The multiple 3d lidar dataset," *arXiv preprint arXiv:2003.06129*, 2020, (accepted for presentation at IV2020).

[26] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023.

[27] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection," *arXiv*, May 2020. [Online]. Available: https://arxiv.org/abs/2005.07431v1

[28] T. Fersch, A. Buhmann, A. Koelpin, and R. Weigel, "The influence of rain on small aperture lidar sensors," in *German Microwave Conference (GeMiC)*. IEEE, 2016, pp. 84–87.

[29] S. Hasirlioglu, I. Doric, C. Lauerer, and T. Brandmeier, "Modeling and simulation of rain for the test of automotive sensor systems," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 286–291.

[30] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, "K-Radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions," Jun. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.08171

[31] M. Ijaz, Z. Ghassemlooy, H. Le Minh, S. Rajbhandari, and J. Perez, "Analysis of fog and smoke attenuation in a free space optical communication link under controlled laboratory conditions," in *International Workshop on Optical Wireless Communications (IWOW)*. IEEE, 2012, pp. 1–3.

[32] I. Gultepe, "Measurements of light rain, drizzle and heavy fog," in *Precipitation: advances in measurement, estimation and prediction*. Springer, 2008, pp. 59–82.

[33] M. Adams, M. D. Adams, and E. Jose, *Robotic navigation and mapping with radar*. Artech House, 2012.

[34] G. Brooker, R. Hennessey, C. Lobsey, M. Bishop, and E. Widzyk-Capehart, "Seeing through dust and water vapor: Millimeter wave radar sensors for mining applications," *Journal of Field Robotics*, vol. 24, no. 7, pp. 527–557, 2007.

[35] R. Xu, W. Dong, A. Sharma, and M. Kaess, "Learned depth estimation of 3d imaging radar for indoor mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 260–13 267.

[36] R. Gourova, O. Krasnov, and A. Yarovoy, "Analysis of rain clutter detections in commercial 77 ghz automotive radar," in *European Radar Conference (EURAD)*. IEEE, 2017, pp. 25–28.

[37] P. Radecki, M. Campbell, and K. Matzen, "All weather perception: Joint data association, tracking, and classification for autonomous ground vehicles," *arXiv preprint arXiv:1605.02196*, 2016.

[38] P. Cai, S. Wang, Y. Sun, and M. Liu, "Probabilistic end-to-end vehicle navigation in complex dynamic environments with multimodal sensor fusion," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4632–4639, 2020. [Online]. Available: https://dx.doi.org/10.1109/LRA.2020.2994027

[39] I. Diaby, M. Germain, and K. Goïta, "Evidential data fusion for characterization of pavement surface conditions during winter using a multi-sensor approach," *Sensors*, vol. 21, no. 24, p. 8218, 2021. [Online]. Available: https://dx.doi.org/10.3390/s21248218

[40] L. G. Galvao, M. Abbod, T. Kalganova, V. Palade, and M. N. Huda, "Pedestrian and vehicle detection in autonomous vehicle perception systems—a review," *Sensors*, vol. 21, no. 21, p. 7267, 2021. [Online]. Available: https://dx.doi.org/10.3390/s21217267

[41] G. Rizzoli, F. Barbato, and P. Zanuttigh, "Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives," *Technologies*, vol. 10, no. 4, p. 90, 2022. [Online]. Available: https://dx.doi.org/10.3390/technologies10040090

[42] "Flir. fused aeb with thermal can save lives." [Online]. Available: https://www.flir.com/globalassets/industrial/oem/adas/flir-thermal-aeb-white-paper---final-v1.pdf

[43] "Research testing on adas autonomous vehicle technologies." [Online]. Available: https://www.vsi-labs.com/

[44] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[45] D. Yu, H. Xiong, Q. Xu, J. Wang, and K. Li, "Multi-stage residual fusion network for lidar-camera road detection," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2323–2328.

[46] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar-camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[47] A. Safa, T. Verbelen, I. Ocket, A. Bourdoux, F. Catthoor, and G. G. Gielen, "Fail-safe human detection for drones using a multi-modal curriculum learning approach," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 303–310, 2021.

[48] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "Radarnet: Exploiting radar for robust perception of dynamic objects," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 496–512.

[49] M. Ulrich, C. Gläser, and F. Timm, "Deepreflecs: Deep learning for automotive object classification with radar reflections," in *IEEE Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–6.

[50] F. Drews, D. Feng, F. Faion, L. Rosenbaum, M. Ulrich, and C. Gläser, "Deepfusion: A robust and modular 3d object detector for lidars, cameras and radars," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 560–567.

[51] Z. Liu, Y. Cai, H. Wang, L. Chen, H. Gao, Y. Jia, and Y. Li, "Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6640–6653, 2021.

[52] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 444–453.

[53] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6433–6438.

[54] Y. Yang, J. Liu, T. Huang, Q.-L. Han, G. Ma, and B. Zhu, "Ralibev: Radar and lidar bev fusion learning for anchor box free object detection system," *arXiv preprint arXiv:2211.06108*, 2022.

[55] N. A. Rawashdeh, J. P. Bos, and N. J. Abu-Alrub, "Drivable path detection using cnn sensor fusion for autonomous driving in the snow," in *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2021*, vol. 11748. SPIE, 2021, pp. 36–45.

[56] R. T. Tan, "Visibility in bad weather from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[57] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," in *IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2201–2208.

[58] Z. Chen, Y. Wang, Y. Yang, and D. Liu, "Psd: Principled synthetic-to-real dehazing guided by physical priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7180–7189.

[59] X. Zhang, H. Dong, J. Pan, C. Zhu, Y. Tai, C. Wang, J. Li, F. Huang, and F. Wang, "Learning to restore hazy video: A new real-world dataset and a new method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9239–9248.

[60] V. Mușat, I. Fursa, P. Newman, F. Cuzzolin, and A. Bradley, "Multi-weather city: Adverse weather stacking for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2906–2915.

[61] R. Timofte, S. Gu, J. Wu, and L. Van Gool, "Ntire 2018 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 852–863.

[62] T. Sun, J. Chen, and F. Ng, "Multi-target domain adaptation via unsupervised domain classification for weather invariant object detection," *arXiv preprint arXiv:2103.13970*, 2021.

[63] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 155–170.

[64] Y. Lee, Y. Ko, Y. Kim, and M. Jeon, "Perception-friendly video enhancement for autonomous driving under adverse weather conditions," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7760–7767.

[65] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[66] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE transactions on intelligent transportation systems*, vol. 22, no. 7, pp. 4230–4242, 2020.

[67] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[68] Z. Yan, L. Sun, T. Krajník, and Y. Ruichek, "Eu long-term dataset with multiple sensors for autonomous driving," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 697–10 704.

[69] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "Radiate: A radar dataset for automotive perception in bad weather," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1–7.

[70] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung *et al.*, "Boreas: A multi-season autonomous driving dataset," *The International Journal of Robotics Research*, p. 02783649231160195, 2022.

[71] T. Matuszka, I. Barton, Á. Butykai, P. Hajas, D. Kiss, D. Kovács, S. Kunsági-Máté, P. Lengyel, G. Németh, L. Pető *et al.*, "aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception," *arXiv preprint arXiv:2211.09445*, 2022.

[72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[73] F. Heide, "Adverse weather fusion supplement," https://www.cs.princeton.edu/~fheide/AdverseWeatherFusion/figures/AdverseWeatherFusion_Supplement.pdf, 2023, accessed: 2023-11-24.

[74] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," *Sensors*, vol. 20, no. 4, p. 956, 2020.

[75] T. Broedermann, C. Sakaridis, D. Dai, and L. Van Gool, "Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection," *arXiv preprint arXiv:2206.15157*, 2022.

[76] S.-Y. Chu and M.-S. Lee, "Mt-detr: Robust end-to-end multimodal detection with confidence fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5252–5261.

[77] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8311–8317.

[78] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[80] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[82] V. John and S. Mita, "Rvnet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments," in *Image and Video Technology: 9th Pacific-Rim Symposium, PSIVT 2019, Sydney, NSW, Australia, November 18–22, 2019, Proceedings 9*. Springer, 2019, pp. 351–364.

[83] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[84] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 516–520.

[85] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[86] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7281–7293, 2021.

[87] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv preprint arXiv:2009.08428*, 2020.

[88] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[89] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[90] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke

[91] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[92] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[93] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[94] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges," p. 4208, May 2022. [Online]. Available: https://doi.org/10.3390/s22114208