

École polytechnique de Louvain

Design and Validation of a Radar–Camera Fusion Framework for Object Detection

Conception et validation d'une fusion bimodale
radar-caméra pour la détection d'objets

Author: **Léonel NGAGUE**

Supervisors: **Benoît MACQ, Christophe CRAEYE**

Readers: **Gauthier ROTSART DE HERTAING, Nicolas BIOUL,
Christophe DE VLEESCHOUWER**

Academic year 2025–2026

Master [120] in Electrical Engineering

Generative artificial intelligence tools, including ChatGPT and Gemini, were used during the preparation of this thesis to assist with language improvement, sentence reformulation, grammar correction, and the generation of schematic illustrations. The technical content, experimental analyses, and final interpretations remain the responsibility of the author and were carefully reviewed and validated.

Acknowledgements

This thesis marks the end of an important chapter in my academic journey. My three bachelor years in Charleroi and my two master years at the École polytechnique de Louvain have been both demanding and enriching. This path was not always easy, but it allowed me to grow, to discover different aspects of engineering, and to develop both technical and personal skills that will remain valuable beyond this work.

I would like to express my sincere gratitude to my supervisors, Professor Benoît Macq and Professor Christophe Craeye, for their guidance and support throughout this thesis. I would also like to give special thanks to Gauthier Rotsart de Hertaing for his availability, his attentive support, and his valuable help during the project, especially regarding the experimental setup and technical difficulties encountered along the way. I am also grateful to OpenHub for providing access to the experimental platform used in this work.

Finally, I would like to thank Samuel Kouff, with whom I shared part of this project through our parallel work. I am also deeply grateful to my family, friends, and classmates for their encouragement and support throughout these years. Their presence made this journey more meaningful and helped me reach the end of this thesis with gratitude and pride.

Abstract

This thesis studies radar-camera multimodal fusion for object detection in autonomous driving. Reliable perception remains a major challenge for autonomous systems, especially when they must operate in complex or degraded environments. While single-modality approaches can achieve strong performance, they are inherently limited by the weaknesses of the sensor on which they rely. Multimodal fusion offers a promising way to combine complementary information, but its benefits depend on the ability of the architecture to effectively exploit each modality. Through a theoretical and experimental study of recent perception methods, this work analyzes the potential and limitations of radar-camera fusion compared with camera-only perception. The study provides insight into the conditions under which multimodal fusion can be beneficial, while also highlighting the challenges that remain for designing robust perception systems.

Contents

1	Introduction	3
1.1	General Context	3
1.2	Problem Statement	4
1.3	Aims of this Thesis	4
2	Experimental Setup	6
2.1	Acquisition Platform	6
2.1.1	Camera	7
2.1.2	Radar	8
2.1.3	ASRock Acquisition Computer	9
2.2	Temporal Alignment	10
2.2.1	Timestamping Strategy	10
2.2.2	Frame Alignment	12
2.3	Limitations of the Synchronization Strategy	12
3	Multimodal Fusion Strategies	14
3.1	Introduction to Multimodal Fusion	14
3.2	Fusion Levels	16
3.2.1	Early Fusion	17
3.2.2	Late Fusion	19
3.2.3	Intermediate Fusion	21
3.3	Fusion Architectures	26
3.3.1	Attention-based Fusion	26
3.3.2	Transformer-based Fusion	28
4	Comparative Performance Analysis	31
4.1	Datasets for Multimodal Perception	31
4.1.1	Overview of public Datasets	31
4.1.2	nuScenes Dataset	32
4.1.3	Locally Acquired Louvain-la-Neuve Dataset	34
4.2	Evaluation Metrics	37

4.3	Model Selection	38
4.3.1	State-of-the-Art Architectures on nuScenes Benchmark . . .	38
4.3.2	Selected Architectures	40
5	Experimental Validation	46
5.1	Introduction	46
5.2	Global Benchmark Comparison	47
5.3	Backbone Training Experiments	48
5.3.1	Training Procedure	49
5.3.2	Stage 1 Training	50
5.3.3	Stage 2 Training	51
5.3.4	Comparison with the Original Backbone	52
5.4	Impact of Illumination Conditions	52
5.4.1	Day vs Night Global Comparison	53
5.4.2	Per-Class Performance Analysis	54
5.5	Multimodal Robustness Analysis	55
5.5.1	Weather Analysis	55
5.5.2	Sensor Failure Analysis	57
5.6	Discussion	60
6	Future Work	62
7	Conclusion	63

Chapter 1

Introduction

1.1 General Context

Autonomous driving has become a major research topic in intelligent transportation systems, where reliable perception is essential for safe navigation and decision-making. Modern perception systems must interpret complex and dynamic environments in real time, often by combining complementary sensing modalities such as cameras, radars, and LiDARs [4, 14].

In this context, multimodal perception has gained increasing attention because it can improve scene understanding, object detection, localization, and robustness under challenging driving conditions [25, 40]. However, effectively combining heterogeneous sensor data remains a complex problem, especially when the modalities differ in resolution, noise characteristics, and sensitivity to environmental conditions.

This thesis was carried out in the framework of the MUSE project, supported by the OpenHub initiative¹. OpenHub supports the development of innovative projects from the initial idea to concrete technical solutions, while the MUSE project focuses on the acquisition, synchronization, and exploitation of multimodal data for intelligent transportation and autonomous driving perception applications.

This work is also related to the parallel thesis conducted by Samuel Kouff on radar data processing for robust object detection and tracking². While his work focuses on radar representation and processing, the present thesis investigates the contribution of radar information within radar-camera fusion architectures.

¹<https://www.openhub.be/>

²Samuel Kouff, *Efficient processing of radar data through transformers for robust object detection and tracking*, Master's thesis, UCLouvain, 2025–2026.

1.2 Problem Statement

Reliable perception remains a central challenge in autonomous driving, as perception systems must operate in complex, dynamic, and variable environments. When relying on a single sensing modality, these systems may be limited by the specific weaknesses of that modality, especially under degraded or unusual operating conditions. This raises the question of whether combining multiple sources of information can lead to more robust and reliable perception systems [14, 40].

Multimodal fusion aims to address this problem by exploiting complementary information from different sensing modalities. However, adding an additional modality does not automatically guarantee better performance. Fusion strategies must be able to combine heterogeneous information effectively, while avoiding situations where the model remains dominated by one modality or fails to fully exploit the additional data [9].

The main problem addressed in this thesis is therefore to evaluate the actual contribution of multimodal fusion compared with monomodal perception. More specifically, this work investigates whether multimodal architectures provide measurable benefits in object detection, in which conditions these benefits appear, and under which circumstances fusion may fail to improve robustness.

1.3 Aims of this Thesis

This thesis aims to evaluate the contribution of multimodal perception compared with monomodal approaches in the context of autonomous driving. More specifically, the work focuses on radar-camera fusion for object detection and investigates whether the combination of multiple sensing modalities can improve perception robustness under challenging conditions.

The thesis combines a theoretical study of multimodal fusion strategies with an experimental evaluation of recent perception architectures. The first part introduces the main concepts required to understand multimodal perception, sensor fusion, autonomous driving datasets, and modern object detection architectures. The second part evaluates selected monomodal and multimodal models in order to analyze their behavior under different evaluation conditions, including illumination

changes, adverse weather, and sensor degradation.

The main contributions of this thesis can be summarized as follows:

- introduction of the main theoretical concepts related to multimodal perception and sensor fusion;
- comparative study of recent monomodal and multimodal architectures for object detection;
- experimental evaluation of the robustness of these architectures under challenging conditions;
- analysis of the effective contribution and limitations of radar-camera fusion compared with camera-only perception.

Through these different analyses, this work aims to provide a broader understanding of current multimodal perception systems and the challenges associated with robust sensor fusion for autonomous driving applications.

Chapter 2

Experimental Setup

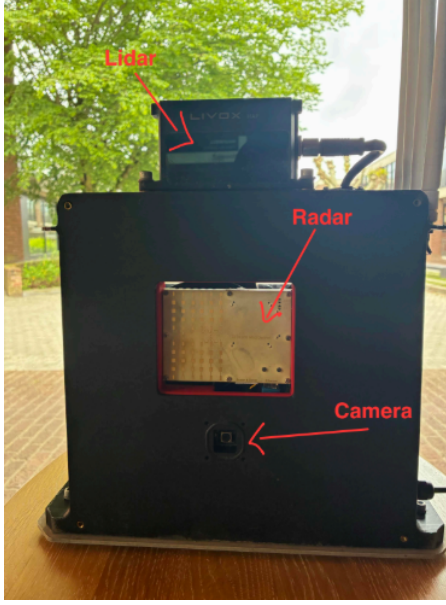
This chapter presents the acquisition platform and the temporal alignment strategy used throughout this work. The setup is composed of a monocular RGB camera, a K-MD2 FMCW radar, and an ASRock industrial computer used for data acquisition and storage. The objective of this platform is to acquire synchronized multimodal data for camera-radar fusion experiments in road environments. Compared to previous works conducted within the MUSE project [17, 10], the current setup focuses exclusively on radar-camera multimodality.

2.1 Acquisition Platform

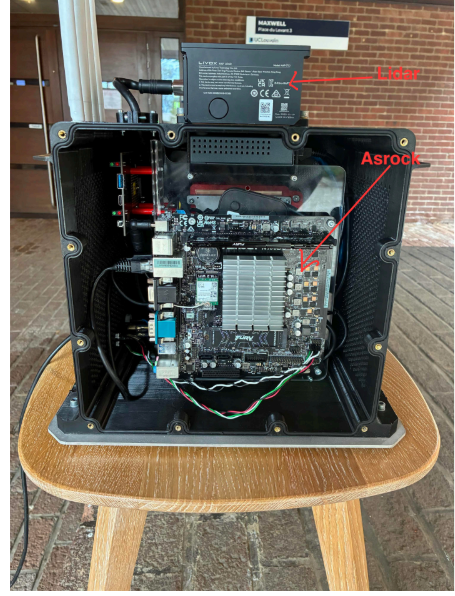
The acquisition setup contains three main components:

- a monocular RGB camera [1],
- a K-MD2 FMCW radar [27],
- an ASRock industrial computer.

An overview of the acquisition platform is shown in Fig. 2.1.



(a) Front view of the acquisition setup



(b) Rear view of the acquisition setup

Figure 2.1: Overview of the multimodal acquisition platform composed of the camera, the K-MD2 radar, and the ASRock acquisition computer. The left image shows the front view of the setup, while the right image shows the rear view containing the acquisition components. A LiDAR sensor can also be observed in the setup; however, it is not considered or analyzed within the scope of this work.

2.1.1 Camera

The visual sensor used in this work is the *ELP-USBFHD04H-L28* camera module equipped with an Aptina AR0330 CMOS sensor [1]. According to the manufacturer specifications, the camera supports Full HD image acquisition with a maximum resolution of 1920×1080 pixels at 30 fps. The camera communicates through a USB 2.0 interface and supports multiple image formats including H.264, MJPEG, and YUYV. It also supports the USB Video Class (UVC) protocol, allowing direct integration into Linux-based acquisition systems without requiring proprietary drivers.

The main characteristics of the camera are summarized below:

- Sensor: Aptina AR0330 CMOS,
- Resolution: up to 1920×1080 ,

- Frame rate: up to 30 fps,
- Lens: 2.8 mm wide-angle,
- Interface: USB 2.0,
- Output formats: H.264, MJPEG, YUYV.



Figure 2.2: ELP AR0330 camera module used for visual acquisition. Adapted from [1].

2.1.2 Radar

The radar used in this work is the *K-MD2* developed by RFbeam Microwave GmbH [27]. According to the manufacturer datasheet, the K-MD2 is a 24 GHz FMCW radar capable of estimating target distance and relative velocity. The K-MD2 was selected because it provides access to low-level FMCW radar representations that are easier to exploit for multimodal perception experiments. In contrast to more recent automotive radars providing only high-level detections, the K-MD2 allows direct access to radar processing outputs useful for research-oriented analysis.

The main radar characteristics provided in the datasheet are:

- Operating frequency: 24 GHz,
- Modulation type: FMCW,
- Maximum range: up to 250 m,
- Update rate: up to 20 Hz,
- Interfaces: Ethernet and serial communication.

The radar outputs processed radar information through Ethernet communication, including:

- range-Doppler maps,
- detected targets,
- tracked targets.

An overview of the K-MD2 radar is shown in Fig. 2.3.

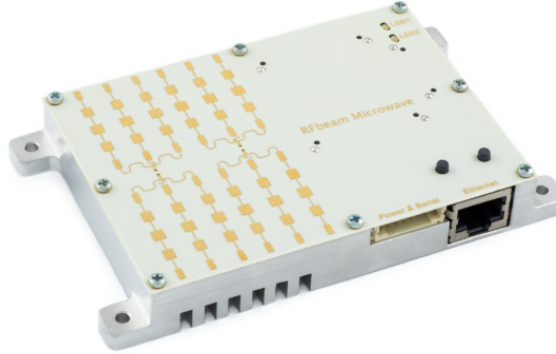


Figure 2.3: RFbeam K-MD2 FMCW radar used in the acquisition platform. Adapted from [27].

2.1.3 ASRock Acquisition Computer

The acquisition, synchronization, and storage processes are performed using an ASRock industrial computer running Linux. This computer acts as the central acquisition unit of the platform. Its main responsibilities are:

- camera acquisition,
- radar communication,
- timestamp generation,
- temporal alignment,
- data storage.

Unlike previous MUSE project setups [17], no Raspberry Pi is used in the current architecture. All acquisition and synchronization tasks are directly managed by the ASRock computer.

2.2 Temporal Alignment

Since the camera and the radar operate at different acquisition frequencies and rely on different time references, their data streams must be temporally aligned before multimodal fusion can be performed. Camera frames are timestamped on the ASRock acquisition computer, whereas the K-MD2 radar provides timestamps based on its own internal monotonic clock. These two time references are therefore not directly comparable. The objective of the synchronization strategy is to express both camera and radar measurements within a common temporal reference, allowing synchronized radar-camera pairs to be generated.

Figure 2.4 provides an overview of this strategy. It illustrates how camera frames are directly associated with the ASRock system clock, while radar timestamps are first expressed in the radar monotonic clock and then mapped to the ASRock temporal reference before nearest-neighbor frame alignment is performed.

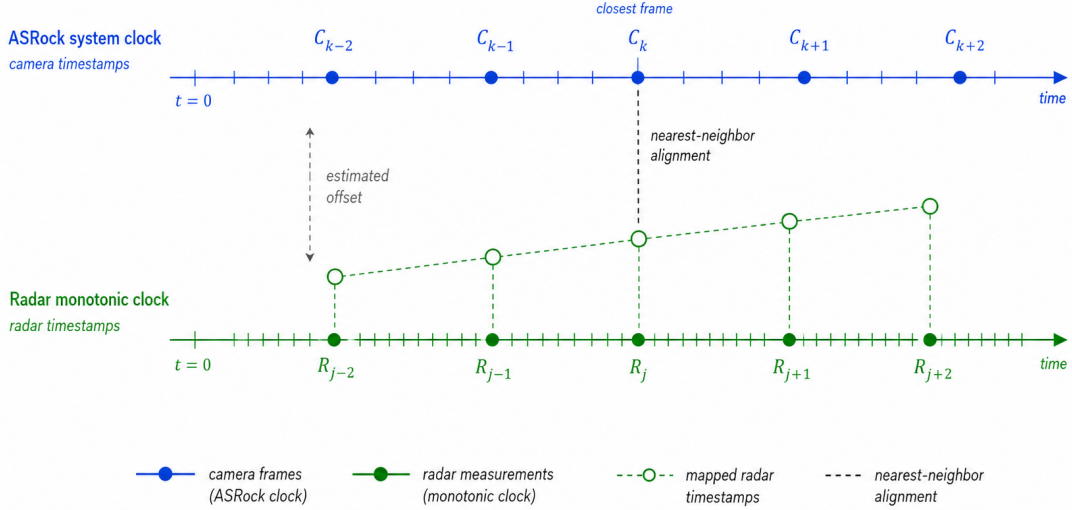


Figure 2.4: Overview of the timestamping and temporal alignment strategy used to synchronize radar and camera measurements. AI-assisted figure.

2.2.1 Timestamping Strategy

The camera stream is acquired directly on the ASRock computer through the USB interface using the `ffmpeg` acquisition tool. Video frames are recorded in MJPEG format at a fixed frame rate. Since the camera stream does not provide

hardware timestamps, each frame is timestamped on the acquisition computer during recording. The ASRock system clock therefore serves as the temporal reference for the camera stream, as illustrated in Fig. 2.4.

Radar timestamping requires an additional synchronization step. Each radar measurement is associated with the K-MD2 internal monotonic clock, which is independent from the ASRock system clock. The radar communicates with the acquisition computer through UDP packets over Ethernet. Although UDP transmission may introduce additional timing uncertainty due to variable latency and packet jitter, the main synchronization challenge comes from the fact that radar timestamps must first be related to the ASRock temporal reference.

To estimate this temporal relationship, timestamp exchanges are performed between the radar and the acquisition computer. The communication latency is approximated using the round-trip time:

$$L = \frac{t_{\text{return}} - t_{\text{send}}}{2} \quad (2.1)$$

where:

- t_{send} is the transmission time,
- t_{return} is the reception time,
- L is the estimated one-way latency.

The first synchronized radar frame is then used as a temporal anchor. Rather than correcting each packet independently, subsequent radar timestamps are expressed relative to this reference frame:

$$t_{\text{radar, corrected}} = t_{\text{anchor}} + (t_{\text{radar}} - t_{\text{radar},0}) \quad (2.2)$$

where:

- t_{anchor} is the ASRock timestamp associated with the first synchronized radar frame,
- t_{radar} is the current radar timestamp,
- $t_{\text{radar},0}$ is the first radar timestamp.

This correction maps radar measurements from the radar monotonic clock to the ASRock temporal reference, as represented by the mapped radar timestamps in Fig. 2.4.

2.2.2 Frame Alignment

Once radar timestamps have been expressed in the ASRock temporal reference, radar and camera frames can be aligned. Since the radar operates at a lower acquisition frequency than the camera, radar timestamps are used as the reference timeline. For each radar measurement, the closest camera frame in time is selected according to:

$$I^* = \arg \min_{I_i} |t_{\text{cam},i} - t_{\text{radar}}| \quad (2.3)$$

where:

- t_{radar} is the corrected radar timestamp,
- $t_{\text{cam},i}$ is the timestamp associated with camera frame I_i .

This nearest-neighbor alignment strategy, illustrated in Fig. 2.4, produces synchronized radar-camera pairs that can be used for multimodal perception experiments.

2.3 Limitations of the Synchronization Strategy

The proposed synchronization strategy provides a practical way to align radar and camera measurements, but it remains limited by the use of independent clocks. Even after the initial synchronization step, a clock drift naturally exists between the radar internal clock and the ASRock system clock due to oscillator inaccuracies. This drift may progressively introduce small temporal misalignments during long acquisition sessions.

Additional uncertainty can also be introduced by UDP transmission jitter, variable communication latency, and radar internal processing delays. Consequently, small synchronization errors may remain between modalities, especially for long recordings or highly dynamic scenes. Nevertheless, the implemented strategy is sufficient for generating approximately synchronized radar-camera pairs and

for supporting the exploratory multimodal acquisition work conducted in this thesis.

Chapter 3

Multimodal Fusion Strategies

3.1 Introduction to Multimodal Fusion

Motivation

Traditional perception systems relying on a single sensing modality remain inherently limited, since they only process one type of information and cannot exploit the complementary information provided by other sensors [13]. In autonomous driving applications, such limitations can significantly affect object detection and scene understanding, particularly in complex situations involving occlusions, adverse weather conditions, or degraded visibility [28, 40].

To address these limitations, multimodal fusion has emerged as a promising approach by combining information from multiple heterogeneous sources¹. By leveraging complementary sensor information, multimodal systems can provide richer scene representations and improve robustness compared with unimodal approaches [41, 14].

In this work, the focus is placed on camera-radar fusion, as these two modalities offer a relevant trade-off between perception performance, robustness, and cost. Although LiDAR sensors provide accurate 3D geometric information, they are not considered in this thesis due to practical and cost-related constraints. Camera-radar fusion therefore represents an attractive alternative for scalable autonomous driving perception systems [40, 25].

¹Heterogeneous modalities or sources refer to data sources with significantly different structures or representations, such as images, radar signals, or point clouds, which may differ in format, dimensionality, or statistical properties.

Cameras provide dense semantic information such as color, texture, and object appearance, but remain sensitive to illumination changes, glare, low-light conditions and occlusions [14, 34]. Radar, on the other hand, provides complementary physical information such as range, relative velocity, and object direction, while being more robust to poor illumination and degraded visibility [40, 29]. However, radar data are sparse, noisy, and have lower angular resolution, which limits their standalone semantic interpretation [40, 30]. The motivation for camera-radar fusion is therefore to combine the semantic richness of cameras with the geometric and motion-related cues of radar in order to improve perception robustness under challenging driving conditions.

Definition of Multimodal Fusion

With the rapid advancement of sensing technologies, modern systems are capable of collecting data from multiple heterogeneous sources, each providing a distinct perspective of the same observed object, activity, or phenomenon. These different sources of information, referred to as modalities, enable systems to capture richer and more diverse representations of the environment [41].

Multimodal fusion refers to the process of integrating information from multiple modalities into a unified representation. Given a set of modalities $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n$, the objective is to combine information from these different sources through a fusion function $F(\cdot)$ in order to obtain a more informative representation or prediction:

$$z = F(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n) \quad (3.1)$$

where z represents the resulting fused representation. The main motivation behind multimodal fusion is to exploit the complementary characteristics of different modalities in order to improve robustness, perception accuracy, and generalization capabilities compared with single-modality approaches [? 14].

In practice, multimodal fusion typically follows a processing pipeline in which each modality is first independently encoded into a feature representation, which are then combined through a fusion mechanism. According to [13], this integration can occur at different stages of the pipeline, leading to various fusion strategies that will be discussed in the following sections.

Recent advances in deep learning have significantly improved multimodal fusion techniques. While traditional methods relied on simple operations such as

concatenation or weighted averaging, modern approaches leverage attention mechanisms and transformer-based architectures to better capture interactions between modalities [13, 41].

3.2 Fusion Levels

After defining multimodal fusion, it is important to further distinguish the different levels at which fusion can be performed within a deep learning architecture. The position of the fusion stage in the processing pipeline strongly influences the model capacity to exploit cross-modal² relationships, the computational complexity, and the robustness of the final perception system.

Multimodal fusion strategies are commonly categorized into three main levels: *early fusion*, *intermediate fusion*, and *late fusion* [11, 13, 41]. In early fusion approaches, modalities are combined directly at the input or low-level feature stage, whereas late fusion approaches process each modality independently and combine only the final predictions. Although these approaches are generally simpler to implement, they often provide limited cross-modal interaction and may not fully exploit the complementary information available across sensors. In contrast, intermediate fusion performs feature fusion after partial unimodal feature extraction, allowing richer interactions between modalities while preserving modality-specific representations.

The choice of the fusion level is particularly important in autonomous driving and intelligent transportation systems, where camera and radar data exhibit strong complementarity but also significant differences in structure, resolution, and noise characteristics. A proper fusion strategy must therefore balance semantic richness, robustness to adverse conditions, computational efficiency, and the ability to model inter-modal dependencies. [2, 11, 12]. The following subsections present these three fusion strategies in detail.

²Cross-modal refers to interactions or information exchange occurring between different modalities, such as camera and radar data.

3.2.1 Early Fusion

Architectural Principle

Early fusion refers to the integration of multiple modalities at the initial stages of the processing pipeline, either directly at the raw data level or after a limited amount of low-level feature extraction. Instead of processing each modality independently until high-level representations are obtained, early fusion creates a shared representation from the beginning of the network [11].

In practice, raw data from different sensors are rarely perfectly aligned due to differences in spatial resolution, temporal sampling, and sensing principles. For this reason, early fusion is generally not performed directly on raw sensor outputs, but rather after an initial unimodal preprocessing stage where basic features are first extracted and projected into a common representation space. These low-level features are then concatenated or jointly processed to form a unified multimodal representation [2].

In [2] they describe early fusion as a strategy where “mixing the modalities starts in the very first layer” enabling the network to learn fused representations immediately instead of delaying cross-modal interactions. Similarly, the article [11] define early fusion as the combination of raw data or early features before deeper feature extraction, which allows the network to exploit low-level statistical correlations between modalities and improve robustness. As illustrated in Fig. 3.1, the fusion takes place at the beginning of the network, before the main convolutional processing stages.

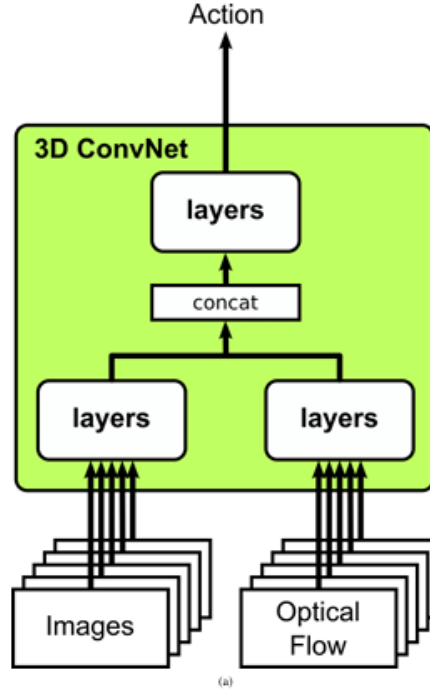


Figure 3.1: Example of an early fusion architecture where low-level features extracted from multiple modalities are concatenated before joint processing. Figure from [11]

Advantages

The main advantage of early fusion lies in its ability to exploit cross-modal correlations at an early stage of representation learning [11, 40, 30]. Since the modalities are combined before high-level abstraction occurs, the network can jointly learn complementary patterns that might otherwise be lost when each modality is processed independently. From an architectural perspective, early fusion also provides a relatively simple framework design, as information from different modalities is integrated into a single processing pipeline rather than requiring multiple independent branches and subsequent fusion mechanisms.

Limitations

One of the main limitations of early fusion is the requirement for accurate modality alignment. Since fusion is performed directly on low-level representations, the different modalities must be spatially and temporally synchronized before being jointly processed [40, 30]. This requirement becomes challenging in real-world sys-

tems where sensors often operate with different acquisition frequencies, resolutions, and coordinate systems. In camera-radar perception, accurate calibration and projection between sensor coordinate systems are necessary to establish meaningful correspondences between observations [40].

Another limitation concerns modality heterogeneity. Different sensor modalities often provide information with fundamentally different characteristics and structures, making direct integration more challenging [11, 3]. For example, images provide dense and semantically rich information, whereas radar measurements are generally sparse and noisy [40]. Consequently, early fusion usually requires modalities to be transformed into compatible representations before entering a shared network, introducing stronger architectural constraints and reducing flexibility compared with higher-level fusion approaches [30, 3].

3.2.2 Late Fusion

Architectural Principle

Late fusion is the process of combining modalities after each modality has been fully processed independently. Each sensor stream passes through its own dedicated network, producing either high-level features or final predictions, which are then merged to generate the final decision [11].

This is the simplest and historically most common multimodal fusion strategy because it allows each modality to be processed using highly specialized unimodal architectures without requiring strong synchronization constraints. As illustrated in Fig. 3.2, each modality follows its own independent processing branch before the outputs are combined at the final stage of the network.

In the simplest case, late fusion combines the outputs of different classifiers through averaging, weighted summation, voting mechanisms, or confidence-based decision rules. More advanced implementations may fuse logits or high-level embeddings before the final classification layer.

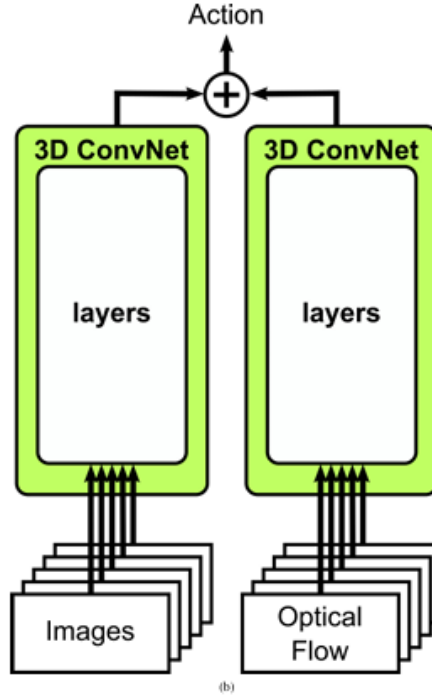


Figure 3.2: Example of a late fusion architecture where each modality is processed independently before the final predictions are combined to produce the final output. Figure from [11]

For camera-radar systems, this means that the camera detector and radar detector can operate almost independently, with the final object detection result obtained by combining both outputs. This strategy is easy to implement and allows each sensor to retain its own optimized processing pipeline.

Advantages

The main advantage of late fusion lies in its flexibility and modular design. Since each modality is processed independently, architectures specifically adapted to each sensor can be employed before combining their outputs [30, 40]. For example, image and radar data can be processed using dedicated networks designed according to the characteristics of each modality.

Late fusion also reduces the strict calibration and alignment requirements associated with lower-level fusion approaches. Since interaction between modalities occurs only at the final stage of the processing pipeline, the impact of spatial and

temporal inconsistencies between sensors is generally reduced [30].

In addition, the independent processing of modalities improves the interpretability and maintainability of the system, as each branch can be developed, analyzed, and optimized separately [11, 3]. Such modularity simplifies system design and facilitates the integration of new modalities without requiring major modifications to the complete architecture [40].

Limitations

The main limitation of late fusion is its poor ability to exploit cross-modal correlations. Since each modality is processed independently until the final stage, the model cannot learn deep interactions between complementary features. This significantly limits the capacity of the system to benefit from semantic and geometric relationships between modalities [11].

This limitation becomes particularly relevant in camera-radar fusion scenarios, where the two modalities provide complementary information. Since late fusion combines modalities only after independent processing, valuable interactions between intermediate features from different modalities may not be fully exploited [30]. As discussed in [30], decision-level fusion can therefore lose potentially useful information that could otherwise be captured through earlier feature interactions.

Finally, because decisions are made independently before fusion, errors produced by one branch may propagate directly to the final prediction without being corrected by complementary information from the other modality.

3.2.3 Intermediate Fusion

Architectural Principle

Intermediate fusion, also referred to as middle fusion or halfway fusion, consists of combining modalities after an initial stage of unimodal feature extraction but before the final decision layer. It represents a compromise between early fusion and late fusion by preserving modality-specific feature learning while still allowing cross-modal interactions during the representation process [11, 12].

In this strategy, each modality is first processed independently through dedicated subnetworks that extract relevant low-level and intermediate features [14, 30]. These

intermediate representations are then fused using operations such as concatenation, attention mechanisms, cross-attention, feature alignment, or transformer-based interaction modules [40, 41]. The fused representation is subsequently processed by a shared multimodal network to produce the final prediction. As illustrated in Fig. 3.3, the unimodal branches first generate modality-specific features before a dedicated fusion module combines them into a shared multimodal representation, which is then passed to the final prediction module.

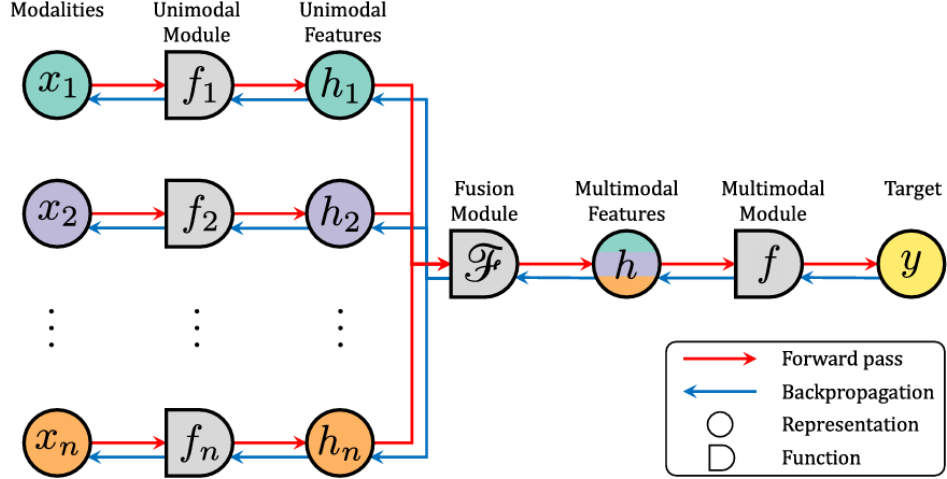


Figure 3.3: General architecture of intermediate fusion. Each modality is first processed independently by a dedicated unimodal module to extract modality-specific features. These intermediate representations are then combined through a fusion module to generate a shared multimodal representation, which is finally used by a multimodal prediction module to produce the target output. Adapted from [12].

This architecture is particularly suitable for camera-radar fusion because both modalities have very different statistical properties and information densities. The camera provides dense semantic and texture information, while radar provides sparse geometric and motion-related measurements. Allowing each modality to first develop its own representation improves robustness and avoids forcing premature alignment between incompatible data structures.

Cross-modal Interaction

The main strength of intermediate fusion lies in its ability to model cross-modal interactions while preserving modality-specific feature learning [30, 12].

Unlike early fusion, where low-level representations are directly merged, intermediate fusion allows each modality to first undergo an independent feature extraction stage in order to generate modality-adapted representations [30, 3]. Once these intermediate representations are obtained, the fusion module can learn relationships between modalities and exploit complementary information while preserving useful modality-specific characteristics [14, 9]. This enables richer interactions between heterogeneous modalities compared with direct low-level fusion.

Cross-modal interaction can be implemented through different fusion mechanisms that allow information from one modality to influence or refine the representation of another [9]. Such interactions are particularly important in intermediate fusion because they enable the model to exploit complementary information while preserving modality-specific representations.

Various mechanisms can be used to model cross-modal interactions, including attention-based approaches and transformer-based architectures. As illustrated in Fig. 3.4, cross-attention represents one possible implementation of this principle, where information is exchanged between modalities through attention mechanisms [9, 39]. The underlying principles of attention mechanisms and their role in transformer-based fusion architectures will be discussed in more detail in the following sections.

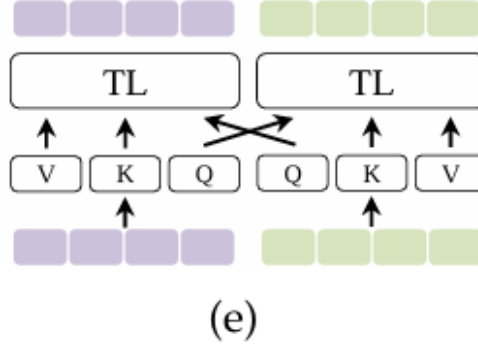


Figure 3.4: Cross-attention mechanism in transformer-based multimodal fusion. One modality provides the query embeddings (Q), while the second modality provides the key (K) and value (V) embeddings. This allows the model to selectively attend to the most relevant complementary information across modalities. TL denotes the transformer layer. Adapted from [39].

In the context of camera-radar fusion, cross-modal interactions are particularly relevant since the two modalities provide complementary information [40?]. Several studies indicate that intermediate fusion provides a favorable compromise between the representational capacity of early fusion and the flexibility of late fusion, particularly when modalities are heterogeneous and require dedicated feature extraction stages before fusion [30, 11, 12]. By enabling interactions between modality-specific features while preserving their individual characteristics, this strategy has become widely adopted in modern multimodal deep learning systems for perception tasks such as object detection and autonomous driving applications.

Advantages

Intermediate fusion offers several important advantages. First, it provides a strong balance between specialization and collaboration. Each modality can benefit from dedicated feature extraction adapted to its own characteristics before participating in multimodal learning. This is particularly important for camera-radar systems where the sensing principles are fundamentally different.

Second, it reduces the strict alignment requirements of early fusion. Since fusion occurs after feature extraction rather than directly on raw data, small spatial or temporal inconsistencies are easier to handle.

Third, intermediate fusion is highly compatible with attention mechanisms and transformer-based architectures, which currently represent the state-of-the-art for multimodal perception systems. These models allow dynamic feature selection instead of fixed concatenation, improving robustness and generalization.

Finally, intermediate fusion is generally more flexible than early fusion and more informative than late fusion, making it one of the most effective strategies for modern autonomous driving perception pipelines.

For these reasons, most recent state-of-the-art camera-radar and LiDAR-camera object detection systems adopt intermediate fusion rather than purely early or late fusion.

Limitations

Despite its strong performance, intermediate fusion also presents several challenges. The first limitation is architectural complexity. Compared to early or late fusion, intermediate fusion requires carefully designed multimodal interaction modules, often involving transformers, attention layers, or multiple alignment stages. This increases both implementation difficulty and computational cost.

Second, the design of the fusion point is not trivial. Choosing where and how to fuse modalities strongly affects performance, and there is no universal rule for selecting the optimal fusion depth.

The literature also highlights that many current fusion operations are performed without strong theoretical justification, often relying on empirical design choices rather than principled assumptions about modality relationships. This lack of standardized methodology remains an important research challenge [12].

Another limitation concerns missing modalities. Intermediate fusion architectures often assume that all modalities are available during both training and inference. In real-world applications, however, radar or camera data may be partially unavailable, noisy, or corrupted, which can reduce system robustness.

Finally, transformer-based intermediate fusion models often require significant computational resources, which may limit real-time deployment in embedded autonomous systems.

3.3 Fusion Architectures

After introducing the different levels of fusion, it becomes necessary to examine the architectural mechanisms that govern multimodal interactions. Recent studies categorize multimodal fusion methods according to the underlying deep learning architectures used to structure information exchange between modalities [41, 13]. The most prominent families include encoder-decoder architectures, attention-based models, graph neural networks (GNNs), and generative models [41].

In practice, these architectures rely on different neural backbones such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), and transformer-based models to process modality-specific information [12]. The choice of architecture generally depends on the characteristics of the input modalities and on the type of interactions that need to be modeled between them [13]. For perception tasks involving heterogeneous modalities such as camera-radar fusion, CNN-based feature extractors and attention-based interaction mechanisms are commonly employed due to their ability to capture spatial information and model relationships across modalities [40, 39].

Recent advances in multimodal learning have progressively shifted from simple fusion operations toward attention-based and transformer-based architectures [13, 39]. According to recent surveys, transformer-based models provide a flexible framework for modeling both intra-modal and inter-modal interactions through attention mechanisms, which explains their growing adoption in multimodal perception systems [39]. Therefore, the following sections mainly focus on attention-based and transformer-based architectures, as these approaches have become increasingly important in recent multimodal perception research.

3.3.1 Attention-based Fusion

Attention mechanisms provide a flexible and adaptive framework for modeling interactions between features, making them particularly suitable for multimodal fusion [33, 39]. Unlike fixed fusion operations such as concatenation or summation, attention mechanisms dynamically determine which information should receive more importance according to the current context. In multimodal systems, this allows the model to focus on the most relevant information coming from different modalities instead of treating all features equally [9, 39].

The attention mechanism relies on three main components: queries (Q), keys

(K), and values (V). Intuitively, the query can be interpreted as the information the model is currently searching for, the keys represent the available elements used for comparison, and the values contain the information that may contribute to the final representation. For example, in an image, different feature regions may correspond to roads, vehicles, pedestrians, or background elements. If the model aims to identify a vehicle, the query may represent the feature associated with the target object, while the keys correspond to all available image features and the values contain their associated information. The attention mechanism then determines which image regions are the most relevant and assigns them a higher contribution in the final representation.

The standard attention formulation, known as *scaled dot-product attention* [33], is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.2)$$

This formulation can be interpreted through several intuitive steps:

1. **Similarity computation:** The product QK^T evaluates how strongly each query matches the available keys. A high similarity indicates that the corresponding information may be relevant.
2. **Scaling:** The factor $\frac{1}{\sqrt{d_k}}$ prevents excessively large similarity values, which helps stabilize training and improve numerical behavior [33].
3. **Normalization:** The softmax operation converts these similarity scores into normalized attention weights. Larger weights indicate more relevant information.
4. **Weighted aggregation:** The final output is obtained by combining the values according to the computed attention weights. Features receiving higher weights contribute more strongly to the resulting representation.

Different variants of attention mechanisms have been proposed depending on how information is exchanged and processed within the network [39]. These mechanisms mainly differ according to how and when interactions occur within the architecture, while sharing the same objective of selectively emphasizing relevant information and preserving meaningful feature relationships. The most commonly used variants include:

- **Self-attention:** Queries, keys, and values are derived from the same representation, allowing the model to capture dependencies and interactions within a single modality or feature space [33].

- **Cross-attention:** Queries originate from one representation, while keys and values are obtained from another representation, enabling information exchange and interactions between different modalities [39].
- **Hierarchical attention:** Information is progressively combined across different processing streams or stages in order to model interactions at multiple representation levels [39].
- **Multi-head attention:** Instead of relying on a single attention operation, multiple attention heads are used in parallel, allowing the model to learn different relationships and focus on different aspects of the input representation simultaneously [33].

3.3.2 Transformer-based Fusion

Transformer architectures extend the attention-based paradigm by using attention mechanisms as the central component for representation learning [33, 39]. Unlike convolutional neural networks, which mainly process information through local operations, Transformers model relationships between all elements of an input representation. This ability to capture long-range dependencies has made them particularly attractive for perception tasks and multimodal learning, where interactions between heterogeneous sources of information must be modeled [33, 8, 39].

A key idea behind Transformers is to represent input data as sequences of tokens. Depending on the modality, these tokens may correspond to image patches, radar points, temporal features, or higher-level embeddings. Once converted into a common representation space, these tokens can be processed by attention layers, allowing the model to learn relationships within a modality or across multiple modalities [39, 36].

The original Transformer architecture proposed in [33] is illustrated in Fig. 3.5. It is composed of repeated encoder and decoder blocks based on attention layers, feed-forward networks, residual connections, and normalization layers. In the context of this work, the main interest of this architecture is not each individual component, but rather its ability to provide a flexible framework for modeling interactions between features and modalities.

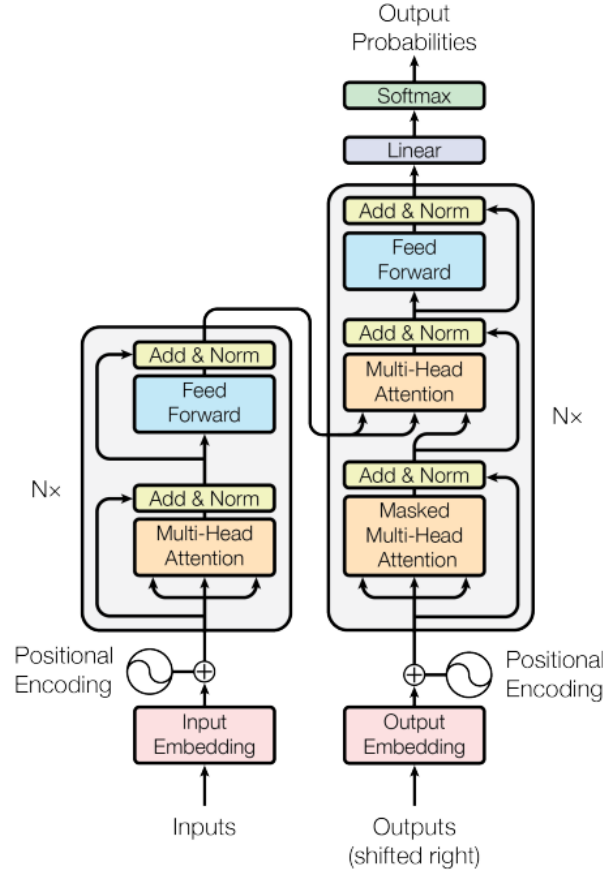


Figure 3.5: Original Transformer architecture from [33]. The model is composed of repeated encoder and decoder blocks combining attention mechanisms, feed-forward layers, residual connections, and normalization. In multimodal perception, this structure provides a flexible framework for modeling interactions between modality-specific representations.

Multimodal Fusion in Transformers Within Transformer-based architectures, multimodal fusion can be naturally integrated through attention mechanisms that allow information exchange between modality-specific representations [39, 9]. Depending on how modalities are represented and combined, several strategies can be distinguished:

- **Token-level fusion:** Tokens from different modalities are merged into a shared sequence and jointly processed by self-attention layers. This enables the model to learn cross-modal relationships from the beginning of the Transformer processing pipeline.

- **Cross-attention fusion:** Cross-attention fusion enables one modality to selectively attend to another by using queries from one representation and keys and values from another. This mechanism allows the model to exchange complementary information across heterogeneous modalities while preserving modality-specific representations [39, 9].
- **Hybrid fusion:** Modality-specific branches first extract separate representations before shared Transformer layers perform joint reasoning. This approach preserves modality-specific processing while still enabling multimodal interactions at a later stage.

These fusion strategies make Transformers especially relevant for modern multimodal perception systems. By relying on attention-based interactions, Transformer-based models can dynamically determine which features or modalities are most informative for a given scene, which is particularly useful in challenging autonomous driving conditions.

Chapter 4

Comparative Performance Analysis

This chapter presents the comparative framework used to evaluate multimodal perception architectures for autonomous driving. First, the main public datasets commonly used for radar-camera and multimodal perception are introduced. Particular attention is then given to the nuScenes dataset[4], which was selected as the main benchmark for this work.

Next, the evaluation metrics used for 3D object detection are presented, including the nuScenes Detection Score (NDS) and the different True Positive metrics. Finally, several state-of-the-art architectures are compared in order to motivate the selection of the models evaluated in the experimental chapter.

4.1 Datasets for Multimodal Perception

4.1.1 Overview of public Datasets

The development of deep learning methods for autonomous driving has been strongly supported by the availability of large-scale public datasets. Early datasets mainly focused on camera-based perception, while more recent datasets integrate multiple sensing modalities such as LiDAR, radar, GPS/IMU¹, and cameras [4, 40, 30].

The most widely used datasets include KITTI [7], Waymo Open Dataset [32], Argoverse [5], PandaSet [38], ONCE [24], and nuScenes [4]. These datasets differ

¹IMU (Inertial Measurement Unit) measures acceleration and rotational motion through inertial sensors.

in terms of sensing modalities, environmental diversity, annotation scale, and suitability for multimodal perception tasks.

Table 4.1 summarizes the main characteristics of representative datasets commonly used for 3D object detection and multimodal autonomous driving research.

Table 4.1: Comparison of representative autonomous driving datasets

Dataset	Modalities	Radar	Weather Diversity	Night Scenes	3D Detection
KITTI [7]	Camera + LiDAR	No	Limited	No	Yes
Waymo Open [32]	Camera + LiDAR + Radar	Limited	Moderate	Limited	Yes
Argoverse [5]	Camera + LiDAR	No	Limited	Limited	Yes
PandaSet [38]	Camera + LiDAR	No	Moderate	Limited	Yes
ONCE [24]	Camera + LiDAR	No	Moderate	Limited	Yes
nuScenes [4]	Camera + LiDAR + Radar	Yes	High	Yes	Yes

Compared to earlier datasets such as KITTI, more recent datasets place a stronger emphasis on multimodal fusion, large-scale 3D annotations, adverse weather robustness, nighttime perception, and bird’s-eye-view (BEV) perception tasks [40, 30].

4.1.2 nuScenes Dataset

The nuScenes dataset [4] was selected as the primary benchmark for this work due to its multimodal nature and its strong relevance for radar-camera fusion research. NuScenes is a large-scale autonomous driving dataset composed of 1000 driving scenes of 20 seconds each, recorded in urban environments in Boston and Singapore. The dataset contains approximately 1.4 million camera images together with synchronized LiDAR, radar, GPS, and IMU measurements [4].

The sensor setup used in nuScenes includes:

- six RGB cameras providing full 360° coverage,
- one LiDAR sensor,
- five FMCW radars,
- GPS and IMU sensors.

The dataset was specifically designed to support multimodal perception and autonomous driving research under diverse environmental conditions. In particular, it contains:

- daytime and nighttime scenes,
- rainy and clear-weather conditions,
- dense urban traffic,
- complex road interactions.

According to [4], approximately 11.6% of the scenes correspond to nighttime conditions, while 19.4% contain rainy weather conditions. These characteristics make nuScenes particularly relevant for robustness analysis under degraded visibility conditions. Another important advantage of nuScenes is the availability of synchronized radar data. Unlike many public datasets that only provide LiDAR and camera measurements, nuScenes integrates five automotive FMCW radars with a maximum range of 250 m [4].

The radar outputs are represented as sparse radar point clouds containing:

- target positions,
- radial velocities,
- radar cross-section information.

Camera data are provided as synchronized RGB images acquired at 12 Hz with a resolution of 1600×900 pixels [4].

In order to conduct a more rigorous robustness analysis, several custom subsets were extracted from the original nuScenes validation set. These subsets were generated using the metadata and scene annotations provided by nuScenes, allowing the categorization of samples according to specific environmental conditions such as daytime, nighttime, and rainy scenarios.

A first partition was created to separately evaluate:

- daytime scenes,
- nighttime scenes.

The resulting subsets contain:

- 5417 frames for the *day* subset,
- 602 frames for the *night* subset.

To further analyze the behavior of multimodal fusion under degraded sensing conditions, additional weather-based subsets were generated:

- day with rain,
- day without rain,
- night with rain,
- night without rain.

The final partition sizes are summarized below:

- *day rain*: 968 frames,
- *day no rain*: 4449 frames,
- *night rain*: 120 frames,
- *night no rain*: 482 frames.

Although these custom subsets introduce class distribution imbalances, particularly in nighttime scenarios where several categories such as *bus*, *trailer*, and *construction vehicle* are absent, they still provide a valuable framework for robustness analysis. These partitions enable a more detailed evaluation of the behavior and failure modes of both camera-only and radar-camera fusion architectures under specific environmental conditions. In particular, they make it possible to identify the situations in which each model performs best or suffers the strongest performance degradation, thereby offering a finer understanding of the impact of illumination and weather conditions on multimodal perception systems.

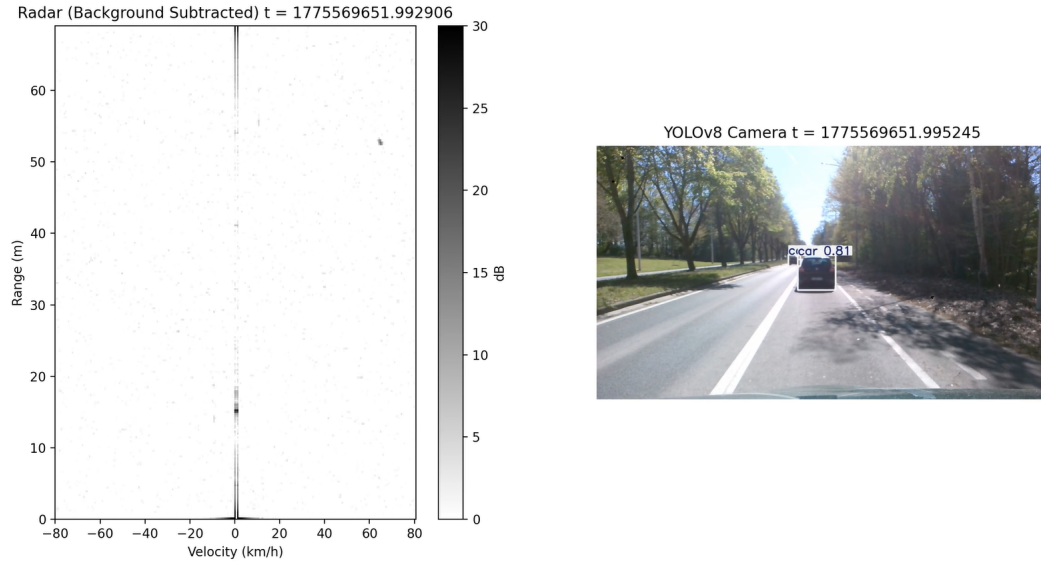
4.1.3 Locally Acquired Louvain-la-Neuve Dataset

Although the main experimental analyses in this thesis are conducted on the nuScenes dataset, additional camera-radar data were acquired locally in Louvain-la-Neuve using the experimental setup presented in Chapter 2. This acquisition campaign was carried out in collaboration with Samuel Kouff.

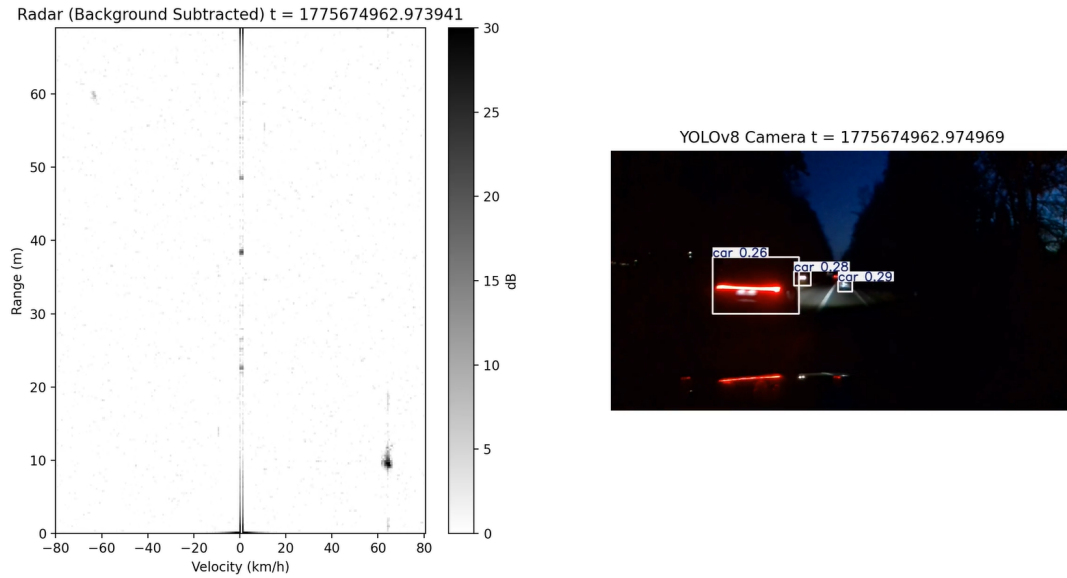
The objective of this local dataset was not to build a large-scale benchmark comparable to nuScenes, but rather to validate the acquisition platform, become familiar with the camera-radar setup, and obtain qualitative observations on the behavior of both modalities under real acquisition conditions. Several recordings were collected in road environments around Louvain-la-Neuve, including both daytime and nighttime sequences. A preprocessing pipeline was developed to exploit the acquired data. In particular, radar measurements were processed using a background removal step in order to reduce static clutter and highlight relevant

moving targets. The processed radar data were then analyzed together with the corresponding camera frames in order to qualitatively assess the behavior of both modalities.

To obtain a first qualitative indication of camera-only detection behavior on the local recordings, a YOLO detector was applied to the acquired RGB images. YOLO was selected because it is a widely used real-time object detection framework that predicts bounding boxes and class probabilities in a single forward pass [26]. In this subsection, YOLO is used only as an inspection tool to visualize the detections produced from the camera stream under different illumination conditions, rather than to establish a benchmark or compare it with the architectures evaluated later in this thesis.



(a) Daytime synchronized radar-camera acquisition



(b) Nighttime synchronized radar-camera acquisition

Figure 4.1: Examples of synchronized radar-camera acquisitions collected in Louvain-la-Neuve. In each image, YOLOv8 detections are shown on the camera stream, while the radar measurements displayed correspond to the processed radar output after background removal.

As illustrated in Fig. 4.1, daytime conditions provide a richer and more informative camera stream, allowing YOLOv8 to detect objects more clearly and

over longer distances. Under nighttime conditions, the camera signal becomes less reliable due to reduced illumination and lower contrast, which limits the quality and range of the detections. In contrast, radar measurements remain informative in both scenarios. The nighttime acquisition shows that radar is not degraded in the same way as the camera stream under low-light conditions.

These local acquisitions also support the robustness analyses conducted later on the nuScenes dataset. In particular, the visible difference between daytime and nighttime camera-only detections motivates the use of dedicated day and night partitions to assess the impact of illumination conditions on perception performance.

4.2 Evaluation Metrics

The performance of the evaluated architectures is assessed using the official nuScenes detection metrics [4]. These metrics are specifically designed for autonomous driving perception and provide a more comprehensive evaluation than conventional object detection metrics.

The main reported metrics are:

- **mAP (mean Average Precision):** evaluates the detection accuracy by measuring the precision-recall relationship across object classes;
- **NDS (nuScenes Detection Score):** a global evaluation metric combining detection accuracy and localization quality;
- **mATE (mean Average Translation Error):** measures the localization error between predicted and ground-truth object centers in meters;
- **mASE (mean Average Scale Error):** evaluates the scale mismatch between predicted and ground-truth bounding boxes;
- **MAOE (mean Average Orientation Error):** measures the orientation estimation error;
- **mAVE (mean Average Velocity Error):** evaluates the accuracy of the predicted object velocity;
- **mAAE (mean Average Attribute Error):** measures the error in attribute prediction such as moving or parked states.

Among these metrics, the two most important indicators are **mAP** and **NDS**. The mean Average Precision is defined as:

$$\text{mAP} = \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} AP_c \quad (4.1)$$

where $\mathbb{C}$ represents the set of evaluated object classes and AP_c denotes the Average Precision associated with class c .

The nuScenes Detection Score (NDS) is the primary benchmark metric used to rank models on the nuScenes leaderboard. It combines mAP with the different true positive quality metrics into a single score:

$$\text{NDS} = \frac{1}{10} \left[5 \cdot \text{mAP} + \sum_{mTP \in \mathbb{TP}} (1 - \min(1, mTP)) \right] \quad (4.2)$$

where $\mathbb{TP}$ includes:

$$\{\text{mATE}, \text{mASE}, \text{mAOE}, \text{mAVE}, \text{mAAE}\}.$$

This formulation ensures that the evaluation framework does not solely reward the ability of a model to detect objects, but also assesses the quality of the predicted spatial and dynamic properties, including object localization, scale estimation, orientation, velocity, and semantic attributes. Consequently, both mAP and NDS must be considered jointly during performance analysis, since models achieving similar detection accuracies may still exhibit significantly different behaviors in terms of localization precision and motion estimation robustness.

4.3 Model Selection

4.3.1 State-of-the-Art Architectures on nuScenes Benchmark

Since this work relies on the nuScenes dataset for evaluation, the selection of the studied architectures was guided by the official nuScenes object detection benchmark². In particular, this study focuses on recent high-performing architectures representative of the current nuScenes benchmark trends. Since top-performing

²<https://www.nuscenes.org/object-detection?externalData=all&mapData=all&modalities=Any>, accessed on April 2026.

radar-camera fusion methods such as RCBEVDet++ rely on attention-based or transformer-inspired fusion mechanisms, the selected models were chosen to follow similar design principles. Camera-only transformer-based architectures were therefore included to provide a more consistent comparison with radar-camera fusion approaches.

Table 5.1 summarizes several representative transformer-based or attention-driven architectures evaluated on the nuScenes benchmark.

Table 4.2: Transformer- or attention-based architectures evaluated on the nuScenes benchmark

Method	Year	Arch.	Mod.	Fusion / Mechanism	mAP	NDS
Camera-only Architectures						
StreamPETR [35]	2023	Transf.	C	Temporal query-based	0.620	0.676
SparseBEV [23]	2023	Transf.	C	Query-based sparse BEV detection	0.603	0.675
DualViewDistill [16]	2025	CNN+Transf.	C	Perspective-BEV dual-view distillation	0.621	0.695
Radar-Camera Fusion Architectures						
RacFormer [6]	2024	Transf.	R+C	Radar-aware attention fusion	cross- 0.649	0.702
RCBEVDet++ [20]	2024	CNN+Att.	R+C	Multi-layer attention BEV fusion	cross- 0.673	0.727
SpaRC [37]	2025	CNN+Att.	R+C	Sparse radar encoding + local self-attention	0.646	0.699
Legend: C = Camera, R = Radar; mAP = mean Average Precision; NDS = nuScenes Detection Score.						

Based on the architectures presented in Table 5.1, three models were selected for the experimental evaluation conducted in this work:

- **SparseBEV**,
- **StreamPETR**,
- **RCBEVDet++**.

The selected architectures include both camera-only and radar-camera fusion models in order to highlight the contribution of multimodal perception under challenging environmental conditions such as low illumination and adverse weather

[40, 30]. The selection was mainly guided by two criteria: the performance achieved on the nuScenes benchmark and the architectural relevance of the models. Particular attention was given to recent transformer-based and attention-driven approaches, since these architectures currently dominate state-of-the-art BEV perception and 3D object detection methods [35, 23, 21, 6].

Among multimodal approaches, RCBEVDet++ [20] was initially selected due to its strong performance on the nuScenes benchmark and its direct architectural relationship with SparseBEV [23], which is used as its camera branch. This makes the framework particularly relevant for comparing camera-only and radar-camera fusion approaches within a similar BEV-based architecture. However, since the official implementation of RCBEVDet++ is not publicly available, this work relies on the original RCBEVDet [21] architecture as the radar-camera fusion baseline. In addition to being publicly accessible, RCBEVDet remains compatible with BEV-based camera backbones, which motivated its selection alongside SparseBEV for the robustness analysis conducted in this work.

Finally, StreamPETR [35] was selected as an additional camera-only architecture because of its competitive performance and its fully transformer-based temporal modeling strategy. Unlike hybrid approaches such as DualViewDistill [16], StreamPETR relies entirely on transformer mechanisms for feature extraction and temporal reasoning, making it particularly interesting for the present analysis. In addition, its architecture offers promising perspectives for future radar-camera fusion extensions based on transformer-driven camera branches. The following subsections present the selected architectures in more detail and describe their main design principles, feature extraction strategies, and BEV perception mechanisms.

4.3.2 Selected Architectures

Before presenting the selected architectures, it is important to briefly introduce the Bird’s Eye View (BEV) representation, which has become a central paradigm in modern autonomous driving perception systems [22]. Recent surveys have highlighted the growing importance of BEV representations for multi-view perception and multimodal fusion in autonomous driving [18].

BEV representations project sensor observations into a unified top-down spatial representation of the environment, as illustrated in Figure 4.2. Compared with perspective-view image representations, BEV representations preserve geometric consistency and facilitate spatial reasoning for tasks such as 3D object detection, object tracking, occupancy prediction, and trajectory forecasting [18]. In particular,

BEV representations alleviate the perspective distortion inherent to image-based representations, where distant objects appear significantly smaller than nearby objects.

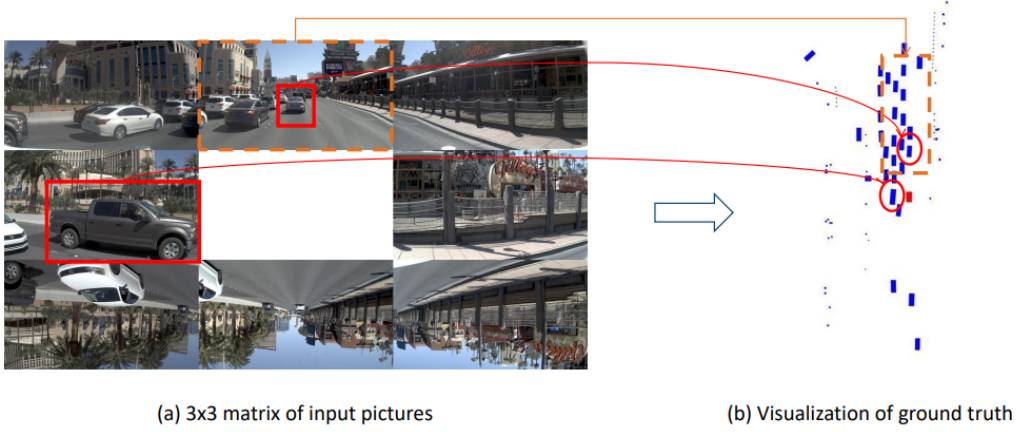


Figure 4.2: Illustration of Bird’s Eye View (BEV) object localization. Multi-view image observations are transformed into a unified top-down representation of the environment. Objects detected in image space are projected into the BEV domain, enabling geometrically consistent spatial reasoning. Figure adapted from [22].

Recent works such as BEVFusion [19] and RCBEVDet++ [20] demonstrated that BEV representations simplify feature alignment and facilitate fusion between heterogeneous sensors such as cameras, LiDAR, and radar sensors. By projecting all modalities into a unified spatial representation, BEV-based approaches enable complementary information from different sensors to be exploited more effectively. This unified representation is particularly relevant for robust 3D perception, as it facilitates multimodal fusion and improves spatial consistency across sensor modalities.

RCBEVDet

RCBEVDet [21] is a radar-camera fusion architecture based on the BEV representation for 3D object detection. The framework first extracts BEV features independently from multi-view camera images and radar point clouds before performing fusion directly in the BEV space. As illustrated in Figure 4.3, the image branch projects multi-view image features into a BEV representation, while the radar branch relies on the proposed *RadarBEVNet* backbone to generate radar

BEV features.

Unlike previous radar-camera fusion methods that reuse LiDAR-oriented radar encoders, RCBEVDet introduces a dedicated radar backbone combining a point-based stream and a transformer-based stream in order to capture both local radar features and global contextual information [21]. The architecture also introduces a Cross-Attention Multi-layer Fusion (CAMF) module, which aligns radar and camera BEV features through deformable cross-attention before applying channel and spatial fusion operations.

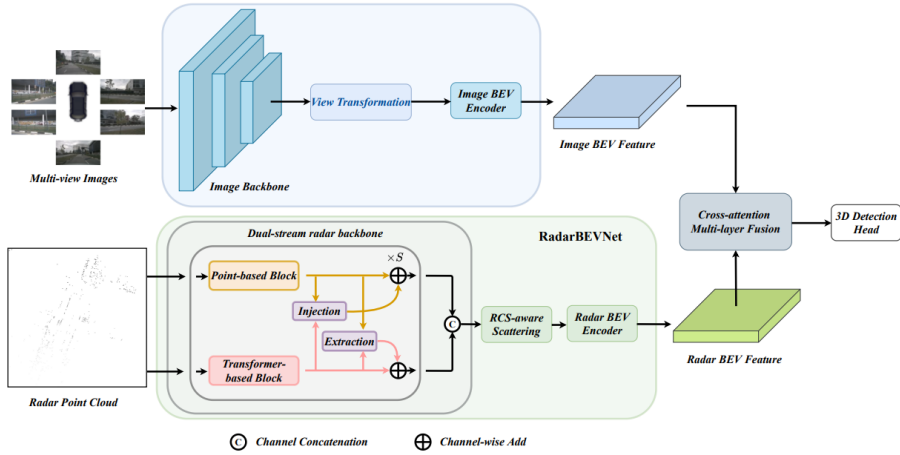


Figure 4.3: Overview of the RCBEVDet architecture adapted from [21]. Multi-view image features and radar features are independently projected into the BEV space before being fused through cross-attention-based fusion modules.

In addition as mentioned above, RCBEVDet++ [20] extends the same general architecture by replacing the camera branch with SparseBEV [23].

SparseBEV

SparseBEV [23] is a sparse query-based BEV detector designed for efficient multi-view 3D object detection. Unlike dense BEV-based approaches that explicitly construct dense BEV feature maps, SparseBEV directly models object instances through a set of sparse learnable BEV queries, as illustrated in Figure 4.4. Multi-view image features are first extracted using an image backbone and FPN before interacting with sparse BEV queries through adaptive spatio-temporal feature sampling.

A key contribution of SparseBEV is the proposed Scale-Adaptive Self-Attention (SASA) module, which allows queries to adaptively aggregate multi-scale contextual information directly in the BEV space [23]. Instead of relying on dense BEV encoders, SparseBEV uses query interactions to model spatial relationships while reducing computational complexity. The architecture also introduces adaptive spatio-temporal sampling mechanisms to exploit temporal information across video frames and compensate for ego-motion and object motion.

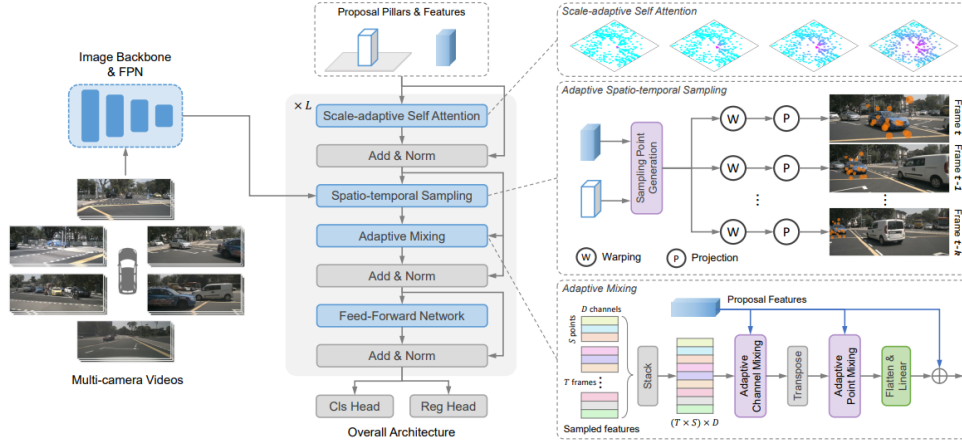


Figure 4.4: Overview of the SparseBEV architecture adapted from [23]. Sparse BEV queries interact with multi-view image features through adaptive spatio-temporal sampling and attention mechanisms.

StreamPETR

StreamPETR [35] is a sparse query-based multi-view 3D object detector designed for efficient temporal modeling in autonomous driving scenarios. Unlike dense BEV-based approaches that explicitly construct and temporally fuse dense BEV feature maps, StreamPETR adopts an object-centric temporal modeling strategy based on sparse object queries. As illustrated in Figure 4.5, multi-view image features are extracted using a standard image backbone before interacting with object queries through a propagation transformer and a memory queue mechanism.

A key contribution of StreamPETR is the proposed memory queue, which recursively stores object query information from previous frames in order to propagate temporal priors efficiently across time [35]. The architecture also introduces a propagation transformer combining hybrid attention and motion-aware layer normalization to model temporal interactions while compensating for ego-motion

and object motion. According to the authors, this object-centric temporal modeling strategy improves computational efficiency compared with dense temporal BEV fusion approaches while maintaining strong detection performance.

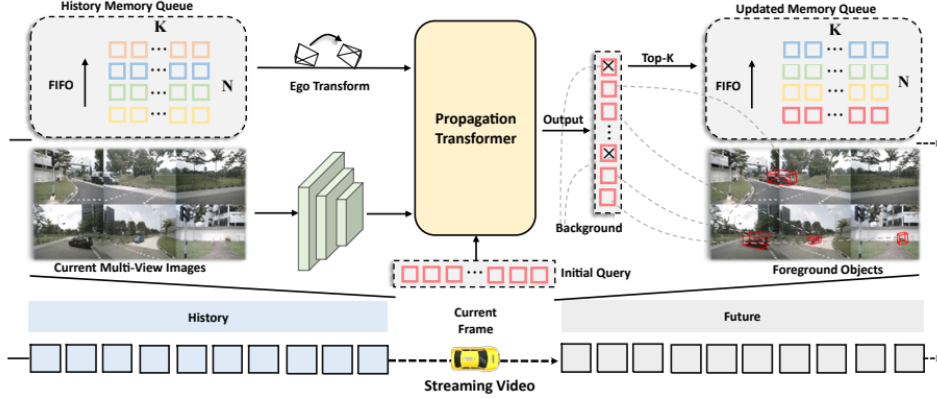


Figure 4.5: Overview of the StreamPETR architecture adapted from [35]. Temporal information is propagated through sparse object queries using a memory queue and a propagation transformer.

discussion

The architectures selected in this work represent different design strategies for multi-view 3D perception and multimodal fusion. SparseBEV [23] focuses on sparse query-based BEV perception with adaptive spatial and temporal feature aggregation, while StreamPETR [35] introduces an object-centric temporal modeling strategy based on sparse object queries and memory propagation. RCBEVDet [21] extends BEV-based perception to radar-camera fusion through dedicated radar feature extraction and cross-attention fusion mechanisms.

These architectures were selected because they provide complementary approaches for studying robustness in autonomous driving perception systems. In addition, all three frameworks are publicly available as open-source implementations, which facilitates reproducibility and experimental comparisons:

- StreamPETR: <https://github.com/exiawsh/StreamPETR/tree/main>
- SparseBEV: <https://github.com/MCG-NJU/SparseBEV/tree/main>
- RCBEVDet: <https://github.com/VDIGPKU/RCBEVDet/tree/main>

To ensure a fair comparison between architectures, experiments were conducted using the ResNet-50 backbone configuration for all models, since this configuration is available for each selected framework through the official implementations and pretrained weights. The corresponding performance metrics obtained from the official repositories are summarized in Table 4.3.

Table 4.3: Performance comparison of the selected architectures using the official ResNet-50 pretrained configurations available in the corresponding open-source repositories.

Architecture	Modality	NDS $\uparrow$	mAP $\uparrow$
StreamPETR [35]	Camera	54.6	44.9
SparseBEV [23]	Camera	55.6	—
RCBEVDet [21]	Radar + Camera	56.8	45.3

Chapter 5

Experimental Validation

5.1 Introduction

The objective of this chapter is to experimentally evaluate the contribution of multimodal fusion for 3D object detection in autonomous driving scenarios. In particular, the experiments aim to analyze the influence of multimodal perception on detection performance and to investigate the benefits of incorporating additional sensing modalities into the perception pipeline. Multimodal fusion has been shown to provide complementary information that can improve scene understanding and perception robustness compared to unimodal systems [25, 40]. However, the actual contribution of an additional modality strongly depends on the fusion strategy, the characteristics of the sensors, and the considered environmental conditions [9, 34].

As discussed in the previous chapter, the experimental evaluation focuses on three representative architectures with different fusion strategies and design principles:

- **StreamPETR**, a Transformer-based camera-only architecture,
- **SparseBEV**, a sparse Transformer-based camera-only BEV detector,
- **RCBEVDet**, a camera-radar fusion architecture based on a BEV representation that combines a dedicated radar Transformer backbone and cross-attention mechanisms to enable multimodal feature interactions.

In addition, the camera-only architecture **BEVDepth4D**, which serves as the camera branch baseline of the RCBEVDet framework, is evaluated in order to isolate the contribution of radar information and to better understand the impact

of introducing an additional sensing modality.

Several experiments were conducted on custom nuScenes subsets to investigate the behavior of these architectures under different sensing conditions. The experimental analysis focuses on:

- the global performance comparison of the considered architectures,
- the contribution of the multimodal backbone training strategy,
- the impact of illumination conditions (day versus night),
- the influence of adverse weather conditions,
- the robustness of multimodal systems under sensor degradation scenarios,
- the individual contribution of each sensing modality.

Beyond performance comparison alone, the objective is to better understand how multimodal fusion influences robustness and whether the integration of additional sensing modalities effectively improves perception capabilities in challenging driving scenarios.

5.2 Global Benchmark Comparison

Before analyzing the behavior of the architectures under specific conditions, an initial evaluation is first performed on the complete nuScenes validation set. The purpose of this experiment is not to conduct a detailed performance analysis, but rather to establish a common baseline and verify the correctness of the implemented experimental environments.

Unlike the performance values reported in the previous section 4.3, the results presented in Table 5.1 are obtained by directly running inference using the publicly available pre-trained weights provided by the corresponding implementations [35, 23, 21]. Consequently, these results serve primarily as a sanity check to ensure that the different frameworks, dependencies, and inference pipelines have been correctly configured before conducting further experiments.

Table 5.1: Global benchmark comparison on the complete nuScenes validation set

Method	Year	Architecture	Mod.	Fusion / Mechanism	mAP	NDS
BEVDepth4D	2022	CNN	C	Camera-only BEV	0.406	0.519
StreamPETR	2024	Transformer	C	Object-centric queries	0.449	0.546
SparseBEV	2023	Transformer	C	Query-based sparse BEV	0.454	0.556
RCBEVDet	2024	CNN + Attention	R+C	Multi-layer cross-attention fusion	0.452	0.567

Legend: C = Camera, R = Radar.

As expected, the obtained performances remain consistent with the trends reported in the original works [35, 23, 21]. Transformer-based camera-only architectures already achieve strong performance on the nuScenes benchmark, with SparseBEV obtaining the highest mAP among the camera-based approaches, while RCBEVDet achieves the highest NDS score overall.

At this stage, however, these observations should not be interpreted as definitive conclusions regarding the superiority of one architecture over another. The architectures rely on different sensing modalities, backbone designs, and fusion mechanisms, making direct comparisons difficult. Instead, the purpose of this preliminary benchmark is to establish a reliable experimental baseline from which the subsequent analyses can be conducted under controlled conditions.

5.3 Backbone Training Experiments

During the preliminary analysis of the evaluated architectures, it was observed that the publicly available implementation of RCBEVDet relies on camera backbone weights initialized from standard ImageNet pre-training. While such initialization is commonly adopted for visual feature extraction, the camera-only architectures considered in this work (SparseBEV and StreamPETR) rely on backbones pre-trained on the *nuImages* dataset, which is more closely related to the target autonomous driving domain. Using different initialization datasets may introduce additional performance variations unrelated to the multimodal fusion strategy itself. Consequently, directly comparing architectures trained under different initialization conditions may lead to biased conclusions regarding the contribution of multimodal

fusion.

In deep learning, training refers to the process by which a model progressively adjusts its learnable parameters using annotated data and an optimization algorithm. In the context of 3D object detection, the backbone corresponds to the feature extraction part of the network, which transforms raw sensor inputs into intermediate representations used by the detection head. The training process is generally organized into epochs, where one epoch corresponds to a complete pass through the training dataset. Since the full dataset cannot usually be processed at once, it is divided into smaller groups of samples called batches. The batch size therefore defines the number of samples processed before each parameter update and can influence both training stability and generalization [15].

To provide a more consistent comparison framework, the complete RCBEVDet training procedure was therefore reproduced using a backbone pre-trained on *nuImages*.

5.3.1 Training Procedure

As discussed previously, the RCBEVDet backbone was retrained in order to ensure a more consistent comparison with the other evaluated architectures. For the training procedure, the backbone weights were selected from the official MMDetection3D implementation¹. These weights are generated through a multi-stage process where a ResNet50 backbone is first initialized using ImageNet weights and subsequently used to train a *Cascade Mask R-CNN* detector on *nuImages*. Multiple training versions are available in the official implementation, and the model trained for 20 epochs was selected.

Following the original RCBEVDet implementation [21], the complete two-stage training strategy was reproduced using these initialization weights. The original implementation trains both stages for 12 epochs using the AdamW optimizer. Due to computational limitations, the exact hardware configuration reported in the original work could not be reproduced. The original implementation relies on 8 GPUs with a batch size of 8 per GPU, corresponding to an effective batch size of 64, whereas the experiments in this work were conducted using 2 GPUs with a batch size of 8 per GPU, resulting in an effective batch size of 16.

¹MMDetection3D nuImages configuration README

5.3.2 Stage 1 Training

During the first training stage, only the camera branch inherited from BEVDepth4D is optimized. Figure 5.1 presents the evolution of the NDS and mAP metrics across training epochs using both standard checkpoints and EMA² checkpoints.

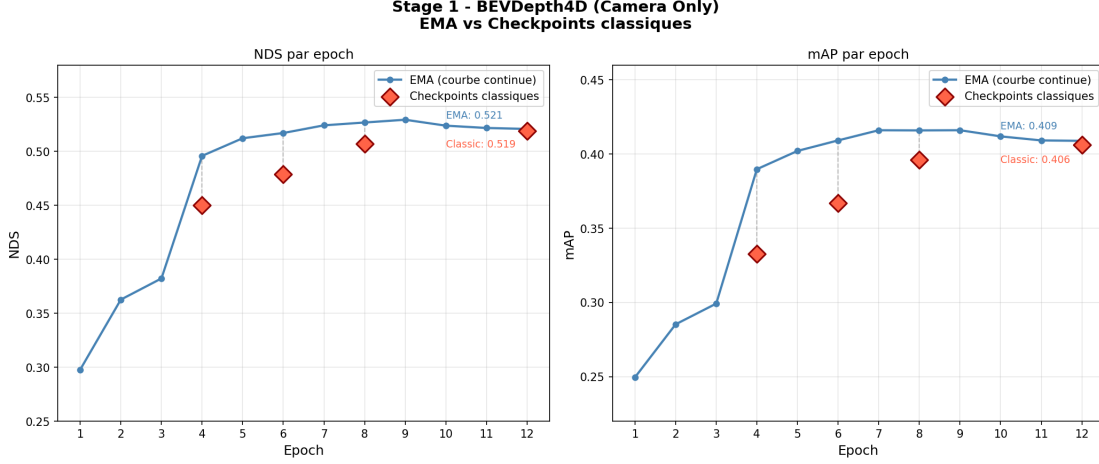


Figure 5.1: Evolution of NDS and mAP during Stage 1 training of the BEVDepth4D camera branch. The continuous curves correspond to EMA checkpoints, while the diamond markers represent standard checkpoints evaluated at selected epochs.

As observed in Figure 5.1, both metrics increase rapidly during the first epochs before progressively converging toward a stable region. A significant performance increase is observed between epochs 3 and 5, after which the improvement becomes more gradual. The final EMA model reaches an NDS of approximately 0.521 and a mAP of approximately 0.409, compared with 0.519 and 0.406 for the corresponding standard checkpoint.

The results also show that EMA consistently produces slightly higher performances than the standard checkpoints throughout training. The relatively large improvement observed during this stage indicates that the camera branch learns most of the scene representation from visual information alone. Since this stage starts from pre-trained initialization weights and optimizes the complete camera branch, substantial feature adaptation occurs during the training process.

²EMA (Exponential Moving Average) is a smoothing technique that computes a weighted average of model parameters across training iterations.

5.3.3 Stage 2 Training

During the second stage, the complete radar-camera fusion architecture is trained while keeping the camera branch frozen. Only the radar branch and multimodal fusion modules are optimized.

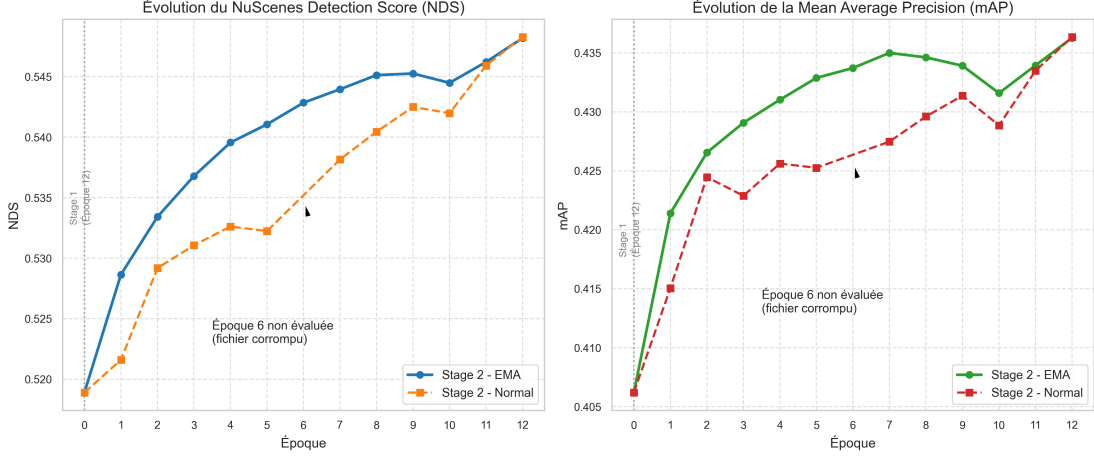


Figure 5.2: Evolution of NDS and mAP during Stage 2 training of the complete RCBEVDet architecture. Epoch 0 corresponds to the final performance obtained after Stage 1 training. The subsequent epochs represent the optimization of the radar branch and multimodal fusion modules while the camera branch remains frozen.

Figure 5.2 shows the evolution of the NDS and mAP metrics during this second stage. Compared with the first training stage, the observed performance increase is significantly smaller. Starting from the Stage 1 baseline (NDS = 0.519 and mAP = 0.406), the final model reaches approximately NDS = 0.548 and mAP = 0.436. This corresponds to an improvement of approximately:

$$[\Delta NDS = +0.029] \quad (5.1)$$

and

$$[\Delta mAP = +0.030] \quad (5.2)$$

Compared with the substantial learning observed during Stage 1, the smaller performance increase suggests that Stage 2 behaves similarly to a fine-tuning process rather than a complete feature learning stage. This interpretation is consistent with the original RCBEVDet training procedure [21], where the camera

branch parameters are frozen during the second stage. Consequently, the optimization process mainly focuses on learning complementary information from radar measurements and adapting the multimodal fusion mechanisms. The limited performance gain should not necessarily be interpreted as a weak contribution of radar information. Since the camera-only baseline already achieves relatively strong performances after Stage 1, the remaining margin for improvement is naturally reduced.

5.3.4 Comparison with the Original Backbone

Overall, the final performances obtained after reproducing the complete training procedure remain slightly below the values reported in the original implementation. A likely explanation is the difference in training conditions, in particular the reduced effective batch size used in this work (16 instead of 64). Since batch size and optimization settings can influence convergence and final model performance, the small performance gap should be interpreted with caution [15]. Despite this limitation, the reproduced model remains relatively close to the original reported performances, which indicates that the training procedure was successfully reproduced under more constrained computational conditions. However, the resulting absolute metrics should not be used to directly rank RCBEVDet against the camera-only architectures evaluated in this work, since these models were trained with different initialization strategies, batch sizes, and optimization configurations.

5.4 Impact of Illumination Conditions

The robustness of perception systems under varying illumination conditions remains a major challenge for autonomous driving applications. Camera-based perception models are particularly sensitive to low-light environments since the quality of visual information can significantly degrade during nighttime scenes [34, 29]. In contrast, radar sensors are generally less affected by illumination variations, which makes multimodal fusion particularly relevant for improving perception robustness in challenging conditions [25].

In this section, the different architectures are evaluated under day and night conditions in order to analyze the impact of illumination on both camera-only and multimodal perception systems. Particular attention is given to the ability of radar-camera fusion to maintain detection performances when visual conditions deteriorate.

5.4.1 Day vs Night Global Comparison

As described previously in the nuScenes dataset presentation, custom day and night subsets were extracted from the validation set in order to evaluate the robustness of the different architectures under varying illumination conditions. Separate inference experiments were conducted on these subsets to compare the evolution of detection performances between daytime and nighttime scenes.

Table 5.2: Global detection performances obtained under day and night conditions for all evaluated architectures. The best performances for each metric are highlighted in bold.

Model	mAP (%)			NDS (%)		
	Day	Night	Δ	Day	Night	Δ
BEVDepth4D (Camera)	41.01	23.68	-17.33	52.29	30.63	-21.66
StreamPETR (Camera)	45.29	25.53	-19.76	54.92	32.45	-22.47
SparseBEV (Camera)	45.86	25.41	-20.45	55.91	34.05	-21.86
RCBEVDet (Radar+Camera)	44.05	26.14	-17.91	55.28	32.78	-22.50

All architectures experience a significant performance degradation under nighttime conditions, confirming the strong dependency of camera-based perception systems on illumination quality [34, 29]. Reduced illumination degrades visual feature extraction and consequently affects object localization and classification performances.

BEVDepth4D exhibits the lowest overall performances, highlighting the limitations of conventional CNN-based camera-only approaches under low-light conditions. In contrast, Transformer-based architectures such as StreamPETR and SparseBEV achieve higher performances thanks to their ability to model global scene interactions through attention mechanisms [33, 39]. SparseBEV notably obtains the highest NDS scores in both daytime and nighttime scenarios, suggesting improved localization stability.

RCBEVDet achieves the highest nighttime mAP performance among all evaluated architectures (**26.14%**), while maintaining performances close to SparseBEV. However, despite the integration of radar information, the relative degradation between daytime and nighttime conditions remains comparable to the other architectures. Since radar measurements are significantly less sensitive to illumination variations [34, 29], one could expect the multimodal model to exhibit a smaller performance drop under low-light conditions. To better understand whether radar

information effectively compensates for the degradation of visual features, the following section presents a more detailed per-class average precision analysis.

5.4.2 Per-Class Performance Analysis

To better evaluate the contribution of multimodal fusion under low-light conditions, a more detailed per-class analysis is conducted in Table 5.3. Studying the evolution of Average Precision for each object category allows a more precise understanding of how the different architectures behave under nighttime conditions and whether radar information effectively improves robustness for specific classes.

Several categories are intentionally excluded from this analysis. Classes such as *Bus*, *Trailer*, and *Construction Vehicle* obtain zero nighttime performances because the corresponding night subset contains very few or no ground truth instances due to the strong imbalance of the nuScenes dataset [4]. In addition, classes such as *Traffic Cone* and *Barrier* are not considered further since they are less relevant for analyzing the contribution of multimodal fusion under degraded illumination conditions.

Table 5.3: Per-class AP comparison under day and night conditions

Class	BEVDepth			StreamPETR			SparseBEV			RCBEVDet		
	Day	Night	Δ	Day	Night	Δ	Day	Night	Δ	Day	Night	Δ
Car	0.627	0.589	-0.038	0.634	0.591	-0.043	0.634	0.594	-0.040	0.703	0.714	+0.011
Truck	0.344	0.279	-0.065	0.419	0.378	-0.041	0.388	0.395	+0.007	0.380	0.281	-0.099
Pedestrian	0.487	0.500	+0.013	0.507	0.562	+0.055	0.533	0.385	-0.148	0.544	0.590	+0.046
Motorcycle	0.406	0.204	-0.202	0.472	0.285	-0.187	0.519	0.291	-0.228	0.463	0.244	-0.219
Bicycle	0.359	0.325	-0.034	0.462	0.351	-0.111	0.473	0.301	-0.172	0.365	0.328	-0.037

The results presented in Table 5.3 show that Transformer-based architectures generally maintain stronger performances under nighttime conditions compared with the CNN-based BEVDepth4D baseline. In particular, StreamPETR and SparseBEV exhibit larger degradations for certain classes. Both architectures show a significantly stronger performance drop for the *Bicycle*, while SparseBEV also experiences a substantially larger degradation for *Pedestrian* detections.

The multimodal RCBEVDet architecture achieves the highest nighttime AP for the *Car* and *Pedestrian* classes, while also improving the average nighttime AP compared with its camera-only counterpart BEVDepth4D. More specifically, the nighttime AP increases from 0.589 to 0.714 for *Car* and from 0.500 to 0.590 for *Pedestrian*. Such improvements may indicate that radar information provides

complementary cues for large reflective objects and vulnerable road users under low-light conditions.

However, the improvements remain less visible for other categories such as *Truck*, *Motorcycle*, and *Bicycle*, where the nighttime performances remain relatively close to the camera-only baseline. This observation is particularly interesting since radar sensors are expected to provide robust detections for highly reflective objects through their Radar Cross Section (RCS).

5.5 Multimodal Robustness Analysis

The previous experiments highlighted the impact of illumination variations on both camera-based and multimodal perception systems. Beyond low-light conditions, multimodal fusion is also expected to improve robustness in more challenging environments where one or several sensing modalities may become degraded or unreliable [34, 25, 29].

This section further investigates the robustness of the evaluated architectures under adverse weather conditions and sensor ablation scenarios in order to better understand the contribution and limitations of multimodal fusion for autonomous driving perception.

5.5.1 Weather Analysis

The previous experiments mainly focused on illumination variations through day and night evaluations. In this section, the analysis is extended to more challenging environmental conditions by combining illumination changes with adverse weather scenarios. Such conditions are known to affect the reliability of autonomous driving perception systems [34, 29, 25].

To further investigate the robustness of multimodal fusion, additional experiments were conducted on several weather and illumination subsets of nuScenes using RCBEVDet. Four conditions were considered:

- day without rain,
- day with rain,
- night without rain,

- night with rain.

These experiments aim to analyze how radar-camera fusion behaves when both sensing modalities are simultaneously challenged. While camera sensors, as discussed previously, remain highly sensitive to illumination changes and visibility degradation, radar measurements can also be affected by adverse weather conditions such as rain, which may introduce noise and additional reflections in the received signals [34, 29]. Table 5.4 presents the global detection performances obtained under the different weather and illumination conditions.

Table 5.4: RCBEVDet performance under different weather and illumination conditions

Condition	mAP	NDS
Day – No Rain	43.32	54.61
Day – Rain	47.11	59.13
Night – No Rain	27.32	33.45
Night – Rain	14.97	17.06

The results show that the model achieves its best performances under the *Day – Rain* condition, which appears somewhat surprising since *Day – No Rain* would intuitively represent a more favorable sensing scenario for both modalities. In contrast, as expected, the performances drop significantly under nighttime conditions, with the *Night – Rain* subset producing the lowest mAP and NDS scores. This behavior is consistent with the fact that nighttime rain simultaneously degrades both sensing modalities, making object perception and multimodal fusion substantially more challenging. To provide a finer evaluation of how adverse weather and illumination conditions affect the different object categories, a more detailed per-class analysis is presented in Table 5.5.

Table 5.5: Per-class AP comparison for the main object categories under different weather and illumination conditions using RCBEVDet

Class	Day No Rain	Day Rain	Night No Rain	Night Rain
Car	0.694	0.742	0.732	0.593
Truck	0.374	0.403	0.293	0.000
Pedestrian	0.546	0.502	0.624	0.505
Motorcycle	0.467	0.395	0.234	0.399
Bicycle	0.365	0.386	0.328	0.000

The per-class results confirm the observations obtained from the global metrics. Surprisingly, the *Day – Rain* condition often achieves performances comparable to, or even better than, the *Day – No Rain* condition for several object categories such as *Car*, *Truck*, and *Bicycle*. This behavior may appear counterintuitive since rain is generally expected to degrade both visual and radar measurements [34, 29].

One possible explanation is related to the distribution of the nuScenes subsets. As discussed previously in the dataset description section, the *Day – Rain* subset contains significantly fewer scenes and therefore fewer object instances than the much larger *Day – No Rain* subset. Since the performance differences between both conditions remain relatively limited, this behavior may suggest that the multimodal architecture remains strongly dependent on visual information, while radar features contribute only partially to the final predictions. In such a scenario, the smaller size of the *Day – Rain* subset could artificially increase the average performances, especially if the considered scenes are globally less diverse or less challenging.

Nevertheless, these observations should be interpreted carefully since multiple factors may influence the obtained performances, including dataset imbalance, scene complexity, sensor synchronization, and the fusion strategy itself. To further investigate these hypotheses and better evaluate the effective contribution of each sensing modality, the following section analyzes the behavior of the multimodal architecture under sensor failure scenarios by independently removing the camera or radar inputs.

5.5.2 Sensor Failure Analysis

One of the main motivations behind multimodal fusion is the ability to improve robustness when one sensing modality becomes degraded or temporarily unavailable [25, 9]. Evaluating the behavior of multimodal architectures under sensor failure conditions therefore provides important insights into the effective contribution of each modality and the robustness of the fusion strategy itself [34, 29]. To investigate this aspect, several inference experiments were performed using RCBEVDet while selectively disabling one sensing modality:

- camera-only inference,
- radar-only inference,
- full multimodal inference.

In addition, the camera-only BEVDepth4D architecture is also evaluated in order to compare the behavior of a dedicated monocular BEV detector with the camera

branch inherited from the multimodal RCBEVDet framework. This comparison makes it possible to better understand the impact of multimodal training and the effective contribution of radar information within the fusion architecture.

Table 5.6: Sensor ablation analysis for RCBEVDet and BEVDepth4D

Configuration	mAP	NDS
RCBEVDet (Radar + Camera)	43.60	54.80
RCBEVDet (Camera Only)	40.49	51.72
RCBEVDet (Radar Only)	0.00	0.00
BEVDepth4D	40.60	51.90

As seen in 5.6 the radar-only inference produces null performances on all evaluated metrics, suggesting that the architecture is unable to perform meaningful detections when visual information is entirely removed.

In addition, the camera-only inference of RCBEVDet remains extremely close to the performances obtained with the standalone BEVDepth4D architecture. This observation suggests that the multimodal training procedure does not significantly improve the robustness or representation capability of the camera branch when radar information is unavailable. Combined with the radar-only results, these observations may indicate that the fusion process remains strongly camera-centered and that radar features contribute only partially to the final predictions. To better understand the contribution of each modality for different object categories, a more detailed per-class analysis is presented in Table 5.7.

Table 5.7: Per-class AP comparison for the main object categories during sensor ablation experiments

Class	RCBEVDet Radar + Camera	RCBEVDet Camera Only	RCBEVDet Radar Only	BEVDepth4D Camera Only
Car	0.703	0.626	0.000	0.627
Truck	0.380	0.342	0.000	0.344
Pedestrian	0.544	0.490	0.000	0.487
Motorcycle	0.463	0.378	0.000	0.406
Bicycle	0.365	0.350	0.000	0.359

The per-class results presented in Table 5.7 confirm the observations obtained from the global metrics. The camera-only inference of RCBEVDet produces performances that remain very close to those obtained with the standalone BEVDepth4D

architecture across all evaluated object categories. This suggests that multimodal training does not significantly improve the robustness or representation capability of the camera branch when radar information is removed.

By contrast, the radar-only configuration completely fails to detect objects for all evaluated classes. Such strictly null performances may at first appear surprising, since one could still expect a small number of low-confidence detections from radar measurements alone. To verify that this behavior is not caused by the complete absence of predictions, a qualitative inspection of the radar-only outputs was performed.



Figure 5.3: Qualitative visualization of the radar-only predictions produced by RCBEVDet. The six camera views are displayed together with the projected bounding boxes obtained when only radar information is used at inference time. Although detections are produced, the predicted bounding boxes are clearly mislocalized which is consistent with the null quantitative results reported in Table 5.7.

As illustrated in Fig. 5.3, the model does produce bounding boxes when only radar information is provided. In the displayed example, the six camera views are shown together with the projected radar-only detections, and the predicted bounding boxes appear poorly localized and semantically inconsistent. This qualitative result confirms that the null quantitative performance is not due to a missing output, but rather to the inability of the radar-only configuration to generate meaningful detections under the considered inference setting. This behavior suggests that the architecture remains strongly dependent on visual information and that the radar branch alone is not sufficiently exploited to support reliable object detection.

5.6 Discussion

The experimental results obtained throughout this chapter highlight several important observations regarding the robustness and limitations of camera-based and multimodal perception architectures for autonomous driving.

First, the experiments confirm the strong performances of Transformer-based camera-only architectures such as StreamPETR and SparseBEV. These architectures consistently achieve higher global performances than the CNN-based BEVDepth4D baseline, particularly during daytime conditions [33, 39]. Nevertheless, the per-class analyses also reveal that some categories remain strongly affected by nighttime conditions, confirming that these architectures still largely depend on the quality of visual information.

The experiments also demonstrate that introducing an additional sensing modality improves the overall detection performances. RCBEVDet consistently achieves higher global mAP and NDS scores than its camera-only counterpart BEVDepth4D, both during daytime and nighttime conditions. Similar observations are frequently reported in recent multimodal perception studies, where radar-camera fusion is identified as an effective strategy for improving perception robustness in challenging environments [25, 34].

However, the detailed robustness analyses reveal that the contribution of multimodal fusion remains more limited than initially expected. In particular, the weather experiments show that the multimodal architecture does not significantly reduce the performance degradation under adverse conditions where both sensing modalities are simultaneously challenged. The sensor failure experiments further reinforce this observation: the radar-only inference completely fails to produce meaningful detections, while the camera-only inference of RCBEVDet remains extremely close to the performances obtained with the standalone BEVDepth4D architecture. These results suggest that the multimodal model remains strongly camera-centered, which is consistent with recent studies highlighting the difficulty of effectively exploiting complementary information in multimodal fusion systems [9, 25]. The per-class analyses support this interpretation as well. Although object categories such as cars, trucks, and motorcycles are generally associated with relatively strong Radar Cross Section (RCS) signatures [31], the improvements brought by radar information remain relatively limited in several challenging scenarios, suggesting that radar features may not be fully exploited during the fusion process despite their potential robustness under degraded visibility conditions.

Several factors may explain these limitations. As discussed in the training section the second stage behaves more like a fine-tuning process, where radar information only slightly refines the visual representations learned during the first stage. In addition, the training conditions used in this work differ from those reported in the original implementation, which may also influence the final fusion quality. The original RCBEVDet work additionally proposes sensor dropout strategies during training in order to improve robustness against sensor degradation [21], suggesting that more robust training procedures may play an important role in balancing the contribution of the different sensing modalities.

Overall, the experiments conducted in this chapter highlight both the potential and the current limitations of radar-camera fusion for autonomous driving perception. While multimodal fusion improves global performances compared with camera-only approaches, the detailed robustness analyses show that effectively leveraging radar information under challenging conditions remains a difficult problem and strongly depends on the fusion strategy and training procedure.

Chapter 6

Future Work

The experiments conducted in this work highlight several perspectives for improving multimodal perception systems for autonomous driving. Although radar-camera fusion improves global detection performances, the robustness analyses reveal that effectively exploiting radar information under degraded environmental conditions remains a difficult challenge.

A first direction concerns the design of stronger multimodal architectures based on more robust camera backbones. The results obtained with Transformer-based architectures such as StreamPETR and SparseBEV suggest that integrating radar information on top of stronger camera-only models could potentially improve multimodal robustness. Recent architectures such as RCBEVDet++ follow this direction by combining radar fusion with more advanced BEV representations [20].

Another important perspective concerns the multimodal training strategy itself. The experiments suggest that the evaluated architecture remains strongly camera-centered, indicating that radar information may not be fully exploited during inference. More robust training strategies, such as sensor dropout approaches proposed in RCBEVDet [21], could help encourage better modality balancing under degraded sensing conditions.

Finally, future work could also investigate different fusion strategies and larger multimodal datasets in order to better understand how fusion design impacts robustness and generalization in autonomous driving perception systems.

Chapter 7

Conclusion

This thesis investigated the contribution of multimodal fusion compared with monomodal perception for autonomous driving object detection. The objective was not only to compare absolute detection performances, but also to analyze whether the addition of a sensing modality leads to more robust perception under challenging conditions.

The experimental results show that multimodal fusion can improve global detection performance, but also that this improvement is not automatic. Adding an additional modality does not necessarily guarantee better robustness if the fusion strategy does not effectively exploit the complementary information provided by each sensor. In particular, the evaluated radar-camera architecture remained strongly dependent on visual information in several experiments, and the radar-only analysis showed that the radar branch alone was not able to provide reliable detections in the considered setting.

These results must also be interpreted with caution, since the evaluated architectures were not trained under strictly identical conditions. Therefore, the goal of this work was not to establish a definitive ranking between camera-only and radar-camera models based solely on absolute mAP or NDS values. Instead, the analysis focused on the behavior of each model under specific conditions, such as illumination changes, adverse weather, and sensor degradation. Overall, this work shows that the benefit of multimodal fusion depends not only on the presence of additional sensor data, but also on the quality of the fusion mechanism, the training strategy, and the ability of the model to balance modality contributions. Future work should therefore focus on more robust training procedures, better exploitation of radar information, and fusion architectures designed to reduce dependence on a single dominant modality.

Bibliography

- [1] Ailipu Technology Co., Ltd., “Elp-usb fhd04h-l28 camera module specifications,” <https://www.webcamerausb.com/elp-1080p-h264-aptina-ar0330-color-cmos-pc-webcam-board-mini-uvic-otg-cctv-full-hd-28mm-wi.html>, 2023.
- [2] G. Barnum, S. Talukder, and Y. Yue, “On the benefits of early fusion in multimodal representation learning,” *arXiv preprint arXiv:2011.07191*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.07191>
- [3] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, “Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition,” *Machine Vision and Applications*, vol. 32, no. 121, 2021.
- [4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nusenes: A multimodal dataset for autonomous driving,” *arXiv preprint arXiv:1903.11027*, 2019.
- [5] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, “Argoverse: 3d tracking and forecasting with rich maps,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8748–8757. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Chang_Argoverse_3D_Tracking_and_Forecasting_With_Rich_Maps_CVPR_2019_paper.pdf
- [6] X. Chu, J. Deng, G. You, Y. Duan, H. Li, and Y. Zhang, “Racformer: Towards high-quality 3d object detection via query-based radar-camera fusion,” *arXiv preprint arXiv:2412.12725*, 2024, accepted to CVPR 2025. [Online]. Available: <https://arxiv.org/abs/2412.12725>
- [7] J.-E. Deschaud, “Kitti-carla: A kitti-like dataset generated by carla simulator,” *arXiv preprint arXiv:2109.00892*, 2021. [Online]. Available: <https://arxiv.org/abs/2109.00892>

- [8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [9] Y. Fan, W. Xu, F. Huo, X. Li, H. Wang, N. Cheng, H. Zhou, D. Niyato, and X. S. Shen, “Inter-modal interactions in multimodal learning: A comprehensive survey,” 2025. [Online]. Available: https://www.researchgate.net/profile/Huo-Fushuo/publication/396854780_Inter-modal_Interactions_in_Multimodal_Learning_A_Comprehensive_Survey/links/68fb36c37d9a4d4e870c2e4d/Inter-modal-Interactions-in-Multimodal-Learning-A-Comprehensive-Survey.pdf
- [10] S. François, “Multimodal-based vehicle tracking for traffic video-surveillance and machine learning analysis,” Master’s thesis, École polytechnique de Louvain, UCLouvain, 2025, unpublished master’s thesis.
- [11] K. Gadzicki, R. Khamsehashari, and C. Zetsche, “Early vs late fusion in multimodal convolutional neural networks,” in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–8. [Online]. Available: <https://ieeexplore.ieee.org/document/9190246>
- [12] V. Guarrasi, F. Aksu, C. M. Caruso, F. D. Feola, A. Rofena, F. Ruffini, and P. Soda, “A systematic review of intermediate fusion in multimodal deep learning for biomedical applications,” *Image and Vision Computing*, vol. 158, p. 105509, 2025.
- [13] S. Hangloo and B. Arora, “Multimodal fusion techniques: Review, data representation, information fusion, and application areas,” *Neurocomputing*, vol. 649, p. 130827, 2025.
- [14] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, “Multi-modal sensor fusion for auto driving perception: A survey,” *arXiv preprint arXiv:2202.02703*, 2022. [Online]. Available: <https://arxiv.org/pdf/2202.02703>
- [15] I. Kandel and M. Castelli, “The effect of batch size on the generalizability of convolutional neural networks on a histopathology dataset,” *ICT Express*, vol. 6, no. 4, pp. 312–315, 2020.
- [16] M. Käppeler, Ö. Çiçek, D. Cattaneo, C. Gläser, Y. Miron, and A. Valada, “Bridging perspectives: Foundation model guided bev maps for 3d object

- detection and tracking,” *arXiv preprint arXiv:2510.10287*, 2025. [Online]. Available: <https://arxiv.org/abs/2510.10287>
- [17] F. Ledent, “Synchronization of multimodal data flows for real-time ai analysis,” Master’s thesis, École polytechnique de Louvain, UCLouvain, 2022. [Online]. Available: <http://hdl.handle.net/2078.1/thesis:35599>
- [18] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, H. Tian, E. Xie, J. Xie, L. Chen, T. Li, Y. Li, Y. Gao, X. Jia, S. Liu, J. Shi, D. Lin, and Y. Qiao, “Delving into the devils of bird’s-eye-view perception: A review, evaluation and recipe,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2151–2172, 2024.
- [19] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, “Bevfusion: A simple and robust lidar-camera fusion framework,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [20] Z. Lin, Z. Liu, Y. Wang, L. Zhang, and C. Zhu, “Rcbevdet++: Toward high-accuracy radar-camera fusion 3d perception network,” *arXiv preprint arXiv:2409.04979*, 2024.
- [21] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. Zhu, “Rcbevdet: Radar-camera fusion in bird’s eye view for 3d object detection,” *arXiv preprint arXiv:2403.16440*, 2024. [Online]. Available: <https://arxiv.org/pdf/2403.16440>
- [22] C. Liu, L. Zhou, Y. Huang, and A. Knoll, “Yolo-bev: Generating bird’s-eye view in the same way as 2d object detection,” *arXiv preprint arXiv:2310.17379*, 2023.
- [23] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, “Sparsebev: High-performance sparse 3d object detection from multi-camera videos,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 18 580–18 590. [Online]. Available: <https://arxiv.org/abs/2308.09244>
- [24] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, H. Xu, and C. Xu, “One million scenes for autonomous driving: Once dataset,” *arXiv preprint arXiv:2106.11037*, 2021. [Online]. Available: <https://arxiv.org/abs/2106.11037>
- [25] H. Qian, M. Wang, M. Zhu, and H. Wang, “A review of multi-sensor fusion in autonomous driving,” *Sensors*, vol. 25, no. 19, p. 6033, 2025. [Online]. Available: <https://doi.org/10.3390/s25196033>

- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *arXiv preprint arXiv:1506.02640*, 2016.
- [27] RFbeam Microwave GmbH, “K-md2 engineering sample datasheet,” <https://rfbeam.ch/product/k-md2-engineering-sample/>, 2018, revision A.
- [28] D. Roy, Y. Li, T. Jian, P. Tian, K. Chowdhury, and S. Ioannidis, “Multi-modality sensing and data fusion for multi-vehicle detection,” *IEEE Transactions on Multimedia*, vol. 25, pp. 2169–2183, 2023.
- [29] F. Sezgin, D. Vriesman, D. Steinhauser, R. Lugner, and T. Brandmeier, “Safe autonomous driving in adverse weather: Sensor evaluation and performance monitoring,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/IV55152.2023.10186596>
- [30] K. Shi, S. He, Z. Shi, A. Chen, Z. Xiong, J. Chen, and J. Luo, “Radar and camera fusion for object detection and tracking: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 3488–3530, 2026.
- [31] W. Shi, X. Zhang, Y. Shen, H. Chen, and Z. Yu, “Rcs estimation and synthesis of typical traffic participants,” in *2021 Photonics & Electromagnetics Research Symposium (PIERS)*. IEEE, 2021, pp. 2474–2478. [Online]. Available: <https://doi.org/10.1109/PIERS53385.2021.9694942>
- [32] 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)*, 2020, pp. 2446–2454. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Sun_Scalability_in_Perception_for_Autonomous_Driving_Waymo_Open_Dataset_CVPR_2020_paper.pdf
- [33] 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 (NeurIPS)*, vol. 30, 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [34] J. Wang, Z. Wu, Y. Liang, J. Tang, and H. Chen, “Perception methods for adverse weather based on vehicle infrastructure cooperation system: A review,” *Sensors*, vol. 24, no. 2, p. 374, 2024. [Online]. Available: <https://doi.org/10.3390/s24020374>

- [35] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, “Exploring object-centric temporal modeling for efficient multi-view 3d object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3621–3631.
- [36] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, “Multimodal token fusion for vision transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 186–12 195. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Multimodal_Token_Fusion_for_Vision_Transformers_CVPR_2022_paper.html
- [37] P. Wolters, J. Gilg, T. Teepe, F. Herzog, F. Fent, and G. Rigoll, “Sparc: Sparse radar-camera fusion for 3d object detection,” *arXiv preprint arXiv:2411.19860*, 2024. [Online]. Available: <https://arxiv.org/abs/2411.19860>
- [38] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, “Pandaset: Advanced sensor suite dataset for autonomous driving,” in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 3095–3101.
- [39] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal learning with transformers: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, arXiv:2206.06488v2 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2206.06488>
- [40] S. Yao, R. Guan, X. Huang, Z. Li, X. Sha, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu, and Y. Yue, “Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review,” *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2094–2123, 2024.
- [41] F. Zhao, C. Zhang, and B. Geng, “Deep multimodal data fusion,” *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, 2024.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl