Thesis for The Degree of Licentiate of Engineering

Learning-Based Sensor Fusion for 3D Scene Understanding in ADAS and AD

Amer Mustajbasic

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2026

Learning-Based Sensor Fusion for 3D Scene Understanding in ADAS and AD

Amer Mustajbasic

© Amer Mustajbasic, 2026
except where otherwise stated.
All rights reserved.

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

"The lamps are different, but the light is the same."
- Rumi

Learning-Based Sensor Fusion for 3D Scene Understanding in ADAS and AD

Amer Mustajbasic

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Robust perception is a central requirement for advanced driver assistance systems (ADAS) and autonomous driving (AD), which must operate reliably in complex, continuously changing environments. To achieve this, information from multiple sensors such as cameras, lidar, and radar must be integrated to exploit their complementary strengths while remaining robust to sensor degradation and partial observability. This thesis studies learning-based multimodal perception focusing on architectural design choices for sensor fusion, spatial scene representations, and geometry-centric representation learning.

In multimodal perception, a key design question is at what stage information from different sensors should be fused. We investigate how mid-level fusion enables effective learning of both modality-specific feature extraction and cross-modal interaction. In structured spatial representations such as Bird's-Eye-View (BEV) grids, we show that attention-based fusion allows models to dynamically weight sensor contributions depending on context, improving robustness and scene understanding compared to early feature collapse.

While BEV representations provide a convenient fusion space, they impose fixed spatial discretization and scale poorly to full three-dimensional reasoning. To address this, we explore probabilistic scene representations based on learnable 3D Gaussian particles, showing that sparse distance measurements from lidar and radar serve as inductive priors for stable multimodal learning in less structured scene representations.

Finally, we study geometry-centric pre-training using occupancy estimation as a supervision signal and show that while geometric structure yields strong spatial reasoning, it requires complementary feature separation mechanisms to achieve semantic discriminability in fine-grained classification tasks.

In general, the results suggest that the multimodal perception that emerges from the joint design of fusion strategies, scene representations, and learning objectives can form a robust and scalable foundation for scene understanding in safety-critical automotive applications.

Keywords

ADAS, AD, Multimodal Sensor Fusion, Multimodal Learning, Camera, Lidar, Radar, Bird's-Eye-View, 3D Gaussian particles, Scene Representation, Pre-training, Foundational model

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **A. Mustajbasic**, S. Chen, E. Stenborg, and Selpi, *SMAB:Simple Multimodal Attention for Effective BEV Fusion*
IEEE Intelligent Vehicles Symposium (IV), Proceedings, 2025, pp. 1766-1772.

[**Paper II**] **A. Mustajbasic**, H. Fu, J. Xu, S. Chen, E. Stenborg, and Selpi, *Guided Gaussians: Enhancing 3D Occupancy Estimation with Sparse Sensor Priors*
Frontiers in Artificial Intelligence and Applications, Vol. 413. 28th European Conference on Artificial Intelligence, ECAI 2025.

[**Paper III**] **A. Mustajbasic**, S. Chen, E. Stenborg, and Selpi, *GeoPriors: Learning Latent 3D Structure via Occupancy Pre-Training for Efficient Multi-Task Scene Understanding*
Paper under submission.

Acknowledgment

The work presented in this thesis was supported both financially and through the provision of computational resources by Zenseact AB. This research was also partially supported by the Swedish Agency for Innovation Systems, VINNOVA, funded by the Swedish Government as well as National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

I would like to express my deepest gratitude to my supervisors, Dr. Selpi, Dr. Shuangshuang Chen, and Dr. Erik Stenborg, whose mentorship and guidance were instrumental in the development of this work, constantly challenging me to refine my ideas and expand my technical horizons. I also thank Prof. Dag Wedelin and Prof. Peter Damaschke for their valuable feedback.

I am also immensely grateful to my company, Zenseact, for its heavy investment in research as being part of this organization provides the distinct feeling of being at the constant spearhead of state-of-the-art innovation. My sincere thanks go to my manager, Linh Trang, who was always there to support my work and highlight my contributions to the company.

I would also like to thank my SRP team at Zenseact for providing such a pleasant and inspiring environment that made every day a learning experience.

On a personal note, I am profoundly grateful to my family and friends, without whose support and patience this academic journey would not have been possible.

Beyond the kindness of those around us, every path of learning ultimately returns to its first source. At the edge of every understanding stands the One who first whispered knowledge into silence. Rabbi zidni 'ilma — increase me in knowledge, let it shape me, and let its benefit reach others.

Contents

Part I

Summary

Chapter 1

Introduction

Traffic safety has been a persistent societal challenge since the introduction of motorized vehicles on public roads. Over the past century, substantial efforts have been devoted to reducing traffic related risks through improvements in infrastructure, regulation, and vehicle design. These efforts include safer road layouts, the separation of vulnerable road users from high-speed traffic, and the widespread adoption of passive and active vehicle safety systems. Through extensive research [1], advances in automotive engineering have been closely guided by an increasing understanding of passenger safety, resulting in innovations such as crumple zones, airbags, child seats, whiplash protection, and electronic stability control. A notable example is the three-point seat belt, invented by Volvo [2], which demonstrates how a single safety innovation can have a profound and lasting impact on traffic injury and fatality reduction.

1.1 From Traffic Safety to Automated Driving

Despite these sustained investments, road traffic accidents remain a major global public health concern [3]. International statistics show that traffic accidents are among the leading causes of death in children and young adults aged 5 to 29, resulting in approximately 1.2 million fatalities annually worldwide [3]. More than half of these fatalities involve vulnerable road users, such as pedestrians, cyclists, and motorcyclists, who lack physical protection in traffic environments. Such incidents often occur unexpectedly, in situations where drivers do not anticipate danger, making safety interventions particularly impactful in rare but critical moments. Human factors such as fatigue, inattention, delayed reaction times, and misjudgment account for the majority of traffic accidents [3].

While human drivers are highly capable, their performance degrades under cognitive load or stress. Automated driving functions have therefore been developed to support or override human actions in safety-critical situations, compensating for these limitations rather than fully replacing the driver. These developments reflect a shift from purely passive protection toward intelligent systems that proactively perceive, reason about, and mitigate risk in complex

traffic environments. Advanced Driver Assistance Systems (ADAS) and ultimately Autonomous Drive (AD) technologies [4] represent a promising pathway toward further reducing traffic related injuries and fatalities. The European Commission estimates that mandatory ADAS features could help save over 25,000 lives by 2034 [5]. Complementing this policy perspective, and considering a broader scope in terms of time horizon and technology deployment, a simulation-based study by the AAA Foundation for Traffic Safety projects substantial long-term safety benefits from ADAS deployment, suggesting that under the most probable adoption scenario, these technologies could prevent over 249,000 traffic fatalities, 14 million nonfatal injuries, and 37 million crashes in the United States alone between 2021 and 2050 [6].

These projected safety benefits are fundamentally enabled by the ability of ADAS to assist human drivers in complex driving situations. ADAS support human drivers by continuously monitoring the vehicle's surroundings, identifying hazards, and issuing warnings or intervening when necessary. The effectiveness of such systems critically depends on robust perception, the ability to accurately and reliably interpret the surrounding environment under diverse operating conditions, including variations in weather, lighting, traffic density, and sensor availability [7],[8],[9].

1.2 Challenges in Vehicle Perception

Recent advances in machine learning, particularly deep learning [10], have enabled significant progress in vehicle perception. Data driven models can extract geometric structure and category level information [11] from high dimensional sensor data, supporting tasks such as object detection, scene understanding, and occupancy estimation [12],[13]. At the same time, modern vehicles are increasingly equipped with multimodal sensor suites [4], including cameras, lidar, and radar, as well as onboard computing platforms capable of processing large data volumes in real time [14],[15].

However, the availability of multiple sensors alone does not guarantee reliable perception. Each sensing modality exhibits distinct strengths and limitations with respect to range, resolution, robustness, and failure modes [16]. Effectively integrating these heterogeneous sensor streams into a coherent scene representation therefore remains a central challenge in automated driving research.

Addressing this challenge requires more than simply adding sensors or increasing model capacity. The reliability of multimodal perception depends on both how sensor signals are fused and how the fused information is spatially represented. In practice, many approaches rely on structured spatial representations such as grid-based Bird's-Eye-View (BEV) maps, which provide a 2D top-down representation of the scene, or voxel grids, which discretize the environment in 3D space. Both representations offer computational efficiency and architectural simplicity. However, these discretized representations introduce trade-offs in scalability and representational flexibility when extending to full three-dimensional scene understanding. Fixed spatial discretization constrains

the modeling of range variability, scene sparsity, and uncertainty, factors that become especially pronounced in pre-training settings (see Figure 3.2), where the objective is to learn general and transferable scene representations that are not tightly coupled to specific tasks or predefined spatial layouts [17],[18],[19].

These observations highlight that robust multimodal perception cannot be achieved by fusion mechanisms or scene representations in isolation. Instead, their joint design fundamentally shapes a system's capacity to cope with partial observability and to learn representations that generalize beyond task-specific supervision. An additional consideration in the design of scene representations concerns interpretability. While human-understandable representations can facilitate debugging, validation, and trust in safety-critical systems, they are not a strict requirement for effective perception. In learning-based systems, intermediate representations are typically optimized for task performance rather than human interpretability.

1.3 Research Questions

Motivated by the challenges described in the Section 1.2, how sensors are fused, how the resulting information is spatially structured, and how representations are learned and interpreted, this thesis investigates how learning-based multimodal perception systems can be systematically designed from the perspectives of sensor fusion architecture, robustness, spatial scene representations and representation learning, prioritizing representation effectiveness and transferability over explicit interpretability.

As a core design choice, mid-level multimodal fusion is adopted as the architectural foundation, reflecting a deliberate architectural decision rather than a comparative survey of fusion levels. In this paradigm, each sensing modality is first processed by modality-specific feature extractors, after which intermediate representations are projected into a shared spatial domain. Compared to early fusion, which aggregates raw inputs, or late fusion, which combines task-level predictions, mid-level fusion enables structured cross-modal interaction while preserving modality-specific characteristics, see Figure 3.1. This architectural choice provides a controlled setting for analyzing how cross-modal information exchange influences robustness and performance.

Building upon this foundation, the thesis studies how heterogeneous sensor information from cameras, lidar, and radar can be integrated in a robust manner under partial observability and sensor degradation. It further investigates how adaptive geometric scene representations with embedded inductive priors shape learning dynamics and spatial reasoning. Finally, it explores how geometry-centric pre-training can enable transferable scene representations that generalize across downstream perception tasks with limited supervision. Through this unified perspective on fusion architecture, spatial representation, and representation learning, the contributions of this thesis are developed around the following research questions:

RQ1. How can multimodal fusion architectures be designed to effectively integrate heterogeneous automotive sensor data?

RQ2. How can multimodal fusion architecture(s) in RQ1 be made robust to partial observability, sensor degradation, and variations in sensor reliability?

RQ3. How do different spatial scene representations, such as grid-based BEV versus adaptive 3D probabilistic models, affect multimodal perception performance?

RQ4. How can inductive priors embedded in spatial scene representations improve the accuracy, robustness and effectiveness of multimodal fusion?

RQ5. Can geometry-centric pre-training of multimodal scene representations reduce the dependence on labeled data during downstream semantic fine-tuning?

RQ6. To what extent do pre-trained multimodal representations generalize across diverse downstream automotive perception tasks?

To contextualize these research questions within the current state of the field, it is important to examine how multimodal fusion is commonly realized in existing systems. Simpler approaches operate directly on raw or lightly processed inputs [20], whereas mid-level fusion methods first extract modality-specific features before combining them in a shared spatial domain. However, even within mid-level fusion, many existing systems rely on simple feature aggregation [21] or tightly coupled cross-dependent fusion strategies [22]. Although effective, these designs limit adaptability under changing sensor conditions and may underutilize modality-specific redundancies, thereby motivating **RQ1** and **RQ2**.

At the scene representation level, grid-based formulations introduce inherent discretization constraints that hinder scalability to full three dimensional reasoning and continuous spatial modeling. These limitations shift attention toward the broader question of how spatial scene representation influences multimodal perception leading to **RQ3** and how inductive priors embedded within spatial representations can guide learning dynamics, directly addressing **RQ4**.

In addition, most perception systems are trained in a task-specific manner, tightly coupling representation learning to annotated downstream objectives. Despite the fact that recent advances in large-scale pre-training have demonstrated the value of transferable representations, geometric structure is often treated as a secondary signal rather than a primary learning objective [17],[18],[19]. These observations raise the question of whether learning scene geometry alone can serve as a unifying foundation for diverse perception tasks in autonomous driving, thereby motivating **RQ5** and **RQ6**.

The research questions are addressed in the included papers as follows, with the mapping summarized in Table 1.1. **Paper I** primarily investigates the design of multimodal fusion architectures at the mid-level feature representation, directly addressing **RQ1**. By studying attention-based fusion within structured

BEV representations, it demonstrates how learned cross-modal interactions enable effective integration of heterogeneous automotive sensors. In addition, the work systematically analyzes robustness under varying sensing conditions, contributing to **RQ2** by showing how adaptive feature weighting mitigates the effects of partial observability and sensor degradation. Because the study is grounded in a structured BEV formulation, it also provides empirical insight into **RQ3**, highlighting how grid-based spatial representations shape multimodal perception performance and fusion behavior.

Paper II shifts the emphasis from fusion mechanisms alone to the role of spatial scene representations, contributing primarily to **RQ3** and **RQ4**. The paper systematically studies multimodal perception within an adaptive probabilistic 3D Gaussian particle formulation and introduces inductive biases in the form of guided initialization from sparse distance measurements, demonstrating how representation-level structure can stabilize learning and improve robustness in multimodal fusion. By exploring fusion strategies within this adaptive representation context, the paper also provides complementary architectural insight related to **RQ1**.

Paper III builds upon the same probabilistic scene representation. It investigates geometry-centric pre-training of multimodal scene representations, using scene occupancy estimation as a supervision signal to learn transferable 3D features. Further, this paper investigates whether such pre-training reduces dependence on labeled semantic data, directly addressing **RQ5**. The pre-trained model was then evaluated across multiple downstream automotive perception tasks to address **RQ6**.

Table 1.1: Overview of how each included paper contributes to the research questions (**RQ1**–**RQ6**).

	RQ1	**RQ2**	**RQ3**	**RQ4**	**RQ5**	**RQ6**
Paper I	✓	✓	✓			
Paper II	✓		✓	✓		
Paper III					✓	✓

Chapter 2

Background

Modern vehicles increasingly integrate driver assistance systems, including adaptive cruise control, automatic emergency braking, lane keeping assistance, and collision avoidance [4]. These systems reduce driver workload, maintain situational awareness, and intervene when necessary, acting as an additional safety layer. More advanced automated applications, such as automated shuttles and robotaxis [23], [24], further demonstrate the safety potential of automation in constrained environments.

Regardless of automation level, vehicle automation relies on a common architecture consisting of sensing, perception, decision making, and actuation, see Figure 2.1.

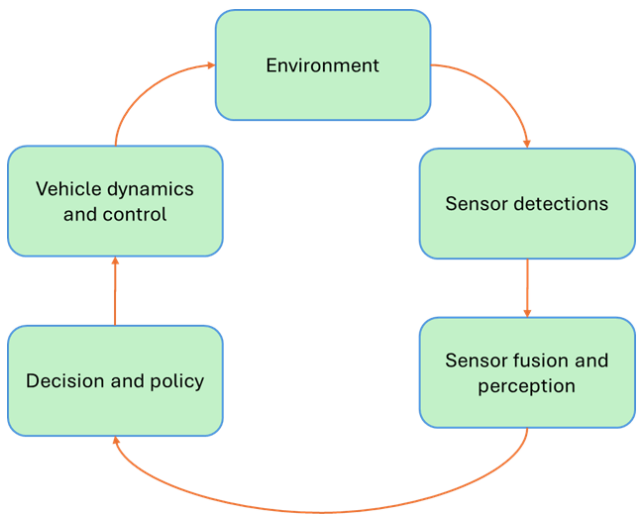


Figure 2.1: Circle of autonomy: sensing, understanding, decision making, and acting.

This functionality is realized through tightly integrated hardware and software stacks. Vehicles employ cameras, radar, and lidar alongside high-

performance computing platforms to process sensor data under strict latency and power constraints. On the software side, layered architectures and middleware support real-time processing and deterministic execution of safety-critical functions. Perception plays a central role by converting raw sensor data into structured environment representations consumed by planning and control modules. Consequently, advances in robust perception, particularly multimodal fusion and scene representation, are key enablers of vehicle automation.

The remainder of this chapter describes the automotive sensors considered in this thesis, and provides an overview of how data from such sensors are collected, curated, and annotated.

2.1 Automotive Sensors

Sensor configurations vary across vehicles depending on automation level, safety requirements, and cost. Cameras are the most widely deployed sensors, initially adopted for low-speed assistance and later extended to core ADAS functions. However, camera-only systems are vulnerable to low illumination, adverse weather, and occlusions [8], motivating the integration of complementary sensors such as radar and lidar [4], [25]. Increasing automation has further driven the adoption of multi-view and multimodal sensor setups [26].

Cameras

Cameras provide dense, high-resolution visual information and are cost-effective and intuitive for perception tasks. A camera system consists of a lens, image sensor, and control unit [27], [28]. Lens design determines field of view, while sensor characteristics such as shutter type, exposure, and gain affect image quality. Image Signal Processors perform demosaicing, noise reduction, and color correction [29], producing images suitable for perception algorithms.

Under optimal conditions, cameras provide rich visual information. However, outside their operational domain, for example in low-light environments, under direct sunlight, during adverse weather, or in the presence of fast motion, they become sensitive to environmental disturbances and may suffer significant degradation in signal quality. Also, depth information is permanently lost during image projections so an estimation of the scene depth from images is inherently ill-posed, motivating fusion with complementary sensing modalities.

Radar

Radar is an active sensing modality that provides robust performance in adverse weather and low-light conditions. By analyzing reflected radio waves, radar directly measures range, relative velocity, and angle, making it particularly effective for collision avoidance and adaptive cruise control. Modern automotive radars employ multiple input multiple output (MIMO) antenna configurations, which improve angular resolution and target separability and increasingly offer 4D sensing with elevation information [30]. However, radar data is sparse,

noisy, provide limited information about detected objects, and require extensive signal processing to suppress multipath reflections or ghost targets.

Lidar

Lidar is an active sensor that uses time-of-flight measurements of laser pulses to generate accurate 3D point clouds [31]. It provides precise distance information and operates independently of ambient lighting. Automotive lidars employ rotating or solid-state scanning mechanisms and are well suited for obstacle detection and spatial reasoning. Limitations include sparsity at long range, sensitivity to adverse weather and higher cost compared to cameras and radar.

Multi-view and multimodal sensor setup

Robust perception in complex traffic environments requires multi-view and multimodal sensor configurations. By combining multiple cameras, radars, and lidars, vehicles achieve extended spatial coverage and leverage complementary sensor characteristics [4], [25], [26], [32]. Overlapping views provide redundancy, while heterogeneous modalities reduce uncertainty and improve reliability.

Effective sensor fusion depends on accurate intrinsic and extrinsic calibration as well as precise temporal synchronization. Misalignment in space or time can lead to inconsistent environment representations, particularly in dynamic scenes. Differences in sensor viewpoints introduce parallax effects as shown in Figure 2.2, complicating data association but also providing valuable geometric cues.

Figure 2.2: Parallax effects in multiview sensor setups (© Zenseact AB). Due to sensor placement, the lidar beam overshoots the pedestrian, making the joint interpretation of objects between camera and lidar ambiguous.

While multimodal systems increase complexity, data volume, and maintenance requirements, they enable more reliable perception under occlusions and

adverse weather conditions. This has driven the adoption of powerful onboard computing platforms capable of real-time multimodal processing.

2.2 Data Collection, Curation, and Annotation

Learning-based perception systems are fundamentally shaped by the quality and diversity of training data. Large-scale datasets are collected using onboard and/or dedicated sensor rigs [26], [32] like the one shown in Figure 2.3, capturing diverse environments, weather conditions, and traffic scenarios.

Data Collection

Synchronized multimodal data streams are recorded during driving sessions to capture representative and rare safety-critical scenarios. Large-scale efforts spanning diverse regions are required to ensure coverage [32]. Sensor heterogeneity, environmental effects, legal constraints, and storage limitations significantly influence data collection strategies.

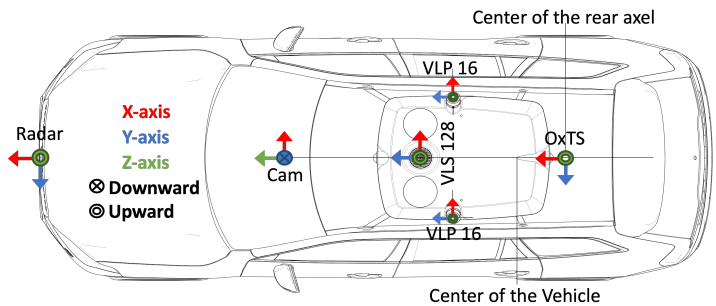


Figure 2.3: Zenseact Open Dataset (ZOD) data collection vehicle and sensor setup [32].

Data Curation

Raw sensor data must be filtered, synchronized, calibrated, and structured before use. Dataset imbalance, where common scenarios dominate rare but important events, poses a major challenge. Targeted sampling and scenario-based selection are commonly employed to improve coverage and generalization. Metadata describing sensor availability and conditions further supports reliable analysis.

Data Annotation

Supervised learning relies on large-scale annotations such as bounding boxes, segmentation masks, and occupancy labels [33]. Annotation is labor-intensive

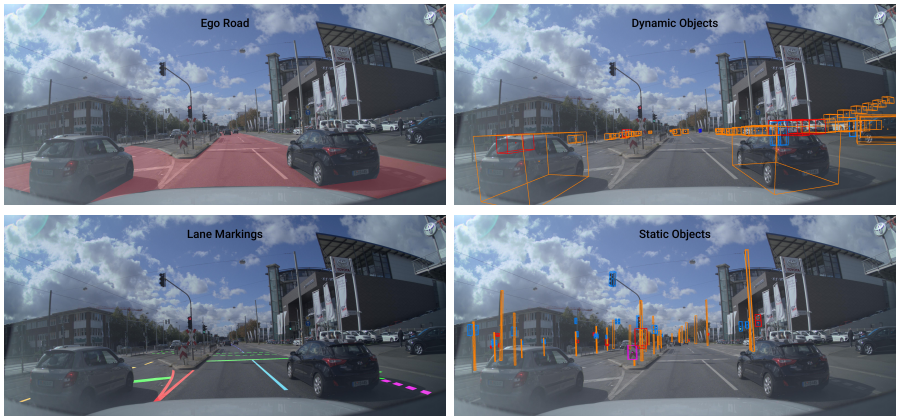


Figure 2.4: Object annotations in Zenseact Open Dataset [32] where same data point is annotated for ego road, dynamic objects, lane markings and static objects.

and often combines automated labeling with human verification, and validation by projecting labels across sensor modalities using calibrated geometric transformations, see Figure 2.4. Noise, ambiguity, and inconsistencies are unavoidable, motivating learning approaches that explicitly account for label uncertainty. Maintaining consistent annotation standards across large datasets is essential for robust and generalizable perception models.

Chapter 3

Learning-Based Multimodal Perception

This chapter reviews learning-based approaches to multimodal perception in autonomous driving. Starting from classical hand-crafted methods, it traces the development toward modern deep learning-based systems and covers the core concepts of scene representation, multimodal fusion strategies, attention-based fusion, and representation learning through pre-training.

Before the widespread adoption of deep learning [10], multimodal perception in autonomous driving was largely hand-crafted using classical approaches based on probabilistic filtering and feature extraction to process and fuse sensor data [34]. For example, multimodal sensor fusion was commonly formulated within Bayesian filtering frameworks such as Kalman filters [35],[36] and particle filters [37], where measurements from various sensors were combined through explicitly designed state-space models. Feature extraction from images relied on manually engineered descriptors such as SIFT [38] or HOG [39], whereas radar systems historically used Constant False Alarm Rate (CFAR) algorithms to distinguish potential targets from background noise and clutter, a process that effectively acts as the "object proposal" stage [40]. Fusion was typically performed through explicit probabilistic data association or late-stage decision fusion.

While these approaches provided interpretability and strong performance in constrained settings, they required significant manual engineering and were often brittle under varying environmental conditions. Hand-crafted features and predefined motion or object models struggled to scale to the diversity and long-tail phenomena present in traffic scenarios. Moreover, the separation between feature extraction, fusion, and decision-making limited the ability to fully optimize perception systems.

Learning-based perception, realized through deep learning models [10], addresses many of these limitations by leveraging large scale data to jointly model semantics, geometry, and uncertainty in the observed environment. Deep neural networks, such as convolutional [41],[10] and transformer-based architectures [42], enable optimization of perception pipelines by learning

hierarchical feature representations directly from data. This data driven formulation is particularly well suited to traffic scenarios, where road structures, road users, and environmental conditions vary widely. By learning directly from diverse observations, such systems achieve strong generalization across scenarios, making them a foundational component of modern driver assistance and automated driving technologies.

Multimodal perception builds on this paradigm by integrating information from multiple sensor modalities, including cameras, lidar, and radar. Each modality contributes distinct and complementary cues. For example, cameras encode rich appearance-based features such as color, texture, and shape, enabling detailed semantic interpretation of objects and scene context, while lidar and radar provide accurate geometric structure and motion estimates with robustness across a wide range of operating conditions. Their combination enables a more holistic and reliable understanding of the driving environment than any single modality can provide.

To effectively exploit this complementarity in learning-based systems, a unified spatial representation is required, providing a common reference frame that enables coherent reasoning across modalities and supports consistent downstream perception tasks. Within such unified spaces, attention mechanisms [42],[43] have emerged as an effective strategy for multimodal fusion by allowing networks to dynamically weight sensor features according to their contextual relevance. Combined with appropriate modeling of spatial and temporal relationships [44], attention-based fusion enables multimodal systems to integrate complementary sensor cues while preserving modality-specific strengths.

A key advantage of this learning-based formulation is its ability to adaptively prioritize the most informative sensor signals under varying environmental and sensing conditions. By learning cross-modal correlations, these models can exploit complementary modality characteristics and infer scene properties that may only be partially observable from individual sensors [21], [45], [46], [44].

Heterogeneous sensor data also provides redundancy, extended spatial coverage, and reduced uncertainty. By capturing cross modal correlations, learning-based models can infer latent scene attributes and improve overall perception reliability.

Formally, let $\mathcal{S} = \{s_1, \ldots, s_M\}$ denote a set of sensor modalities, where each sensor s_m produces observations $\mathbf{x}_m \in \mathcal{X}_m$. Learning-based multimodal perception can be formulated as learning a representation function

$$f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_M \to \mathcal{Y} \tag{3.1}$$

followed by a task-specific mapping

$$g : \mathcal{Y} \to \mathcal{T}. \tag{3.2}$$

where $\mathcal{Y}$ denotes a unified scene representation, such as a BEV feature map or an occupancy based spatial representation and $\mathcal{T}$ is a downstream task space, such as object detection or semantic segmentation. The overall perception system is thus given by the composition $g \circ f$, mapping multimodal inputs to a downstream task space $\mathcal{T}$

3.1 Scene Representations

One of the fundamental components of any perception system is the choice of scene representation, as it determines how information extracted from sensor data is organized, structured, and exposed to downstream reasoning tasks. In learning-based multimodal perception, principled scene representations that unify heterogeneous sensor data into a common spatial domain play a central role in enabling effective fusion and interpretation of information from heterogeneous sensors and multiple viewpoints, providing an explicit geometric reference frame that enables coherent reasoning across modalities and supports consistent downstream perception tasks.

Traditional object centric perception pipelines, which represent the environment as a set of detected objects with associated attributes, have proven effective in many scenarios. However, they often struggle in the settings characterized by partial observability or the presence of previously unseen, long tail objects, such as articulated trailers or objects of unknown category (e.g., roadside debris) [47]. These challenges have motivated a shift toward more holistic, spatially grounded scene representations that explicitly encode the geometry and occupancy of the environment.

Among these representations, BEV representations [48], [49], 3D occupancy grid [13] as well as probabilistic and explicit 3D Gaussian representations [50],[51] have emerged as effective abstractions for multimodal perception in autonomous driving scenarios.

Bird's-Eye-View Representation

BEV representations encode sensor information in a common, top-down coordinate frame aligned with the road surface. By transforming measurements from different sensor modalities and viewpoints into this shared spatial frame, BEV representations enable natural and consistent fusion while preserving the metric structure of the environment.

This representation is particularly well suited for multimodal learning, as it unifies heterogeneous sensor viewpoints and decouples perception from sensor specific viewpoint characteristics. As a result, downstream reasoning can be performed in a spatially consistent manner that is independent of individual sensor configurations.

As camera images do not directly provide depth information, the features extracted from images must be lifted from the image plane into the BEV space. Several approaches have been proposed to address this, including estimating depth distributions along each image ray as in LSS [52], attending to BEV query locations using image features like BEVFormer [53], or projecting image features into a voxel grid followed by compression along the height dimension in SimpleBEV [20].

In lidar-centric pipelines, a common use of BEV is to project point clouds into a dense top-down raster and then apply a CNN to perform detection directly in BEV space. BirdNet [48] follows this paradigm by constructing a BEV representation from lidar information and learning a 3D object detection

model operating in the BEV domain. SimpleBEV [20], on the other hand, rasterizes lidar or radar signatures directly into a voxel grid, which is then combined with image features and compressed along the height dimension to produce a BEV representation.

Through this unified spatial abstraction, BEV representations facilitate early or mid fusion (see Section 3.2) of multimodal features, allowing learning-based models to jointly exploit semantic context and geometric accuracy.

From a system level perspective, BEV representations provide a scalable and convenient abstraction that decouples perception from downstream tasks such as semantic segmentation and 3D object detection. By operating in a unified spatial domain, downstream modules can be designed independently of the original sensor configurations.

However, generating accurate BEV features from diverse sensor viewpoints is non trivial, as it requires addressing occlusions [54], depth estimation ambiguities [52],[53],[20], and sensor sparsity [55], [56]. Learning-based approaches tackle these challenges by optimizing depth inference, feature lifting, and multimodal fusion jointly for overall performance.

3D Occupancy Representation

A 3D occupancy grid represents the environment as a set of spatial cells $\{c_i\}$, where each cell is associated with an occupancy probability

$$P\left(c_i = \text{occupied} | \mathbf{x}_m\right), \tag{3.3}$$

where $\mathbf{x}_m$ denotes the observations from modality m.

Each cell belongs to a discretized spatial grid and encodes the probability of being occupied or free. 3D occupancy grids model space directly without making any assumptions about the existence of any objects, how many objects exist in the scene, or their shapes. This property allows occupancy based representations to naturally handle partial observability, occlusions, and previously unseen or ambiguous elements in the environment, as each cell independently represents uncertainty about its state.

Using occupancy and spatial scene representation where sensor measurements is expressed in a common three dimensional spatial domain naturally captures uncertainty, facilitates consistent fusion of heterogeneous sensor measurements and provide unified abstraction for multimodal perception. Expressing sensor measurements in a common domain makes fusion more consistent regardless of sensors original modality or data format.

Information from multiple sensor modalities can be integrated through spatio-temporal aggregation of multimodal features across the grid, enabling learning-based occupancy prediction models to infer coherent three-dimensional structures.

Certain sensor signals, such as lidar, have a natural advantage for 3D grid like representations, as the point measurements can be directly rasterized or processed within a structured scene representation. VoxelNet [57] leverages this by dividing a lidar point cloud into voxels and encoding the points within each voxel into a unified feature representation using voxel feature encoding.

A similar voxel-based approach has been applied to image data in Monoscene [58], where features from a single image are projected along the optical ray to all possible voxel locations, producing an initial 3D representation that is subsequently refined with a computationally intensive 3D UNet.

Despite these advantages, grid based occupancy representations introduce significant computational challenges. Increasing the spatial resolution of a 3D grid leads to cubic growth in memory and computational complexity, which directly impacts downstream processing. As a result, grid resolution must be carefully chosen to balance representational accuracy particularly with respect to high accuracy and resolution of active sensors such as lidar, against the affordable computational capacities of the system.

3D Gaussian particle Representation

An alternative approach to uniform discretization of the space as in grid based representations, is to represent the environment using a set of probabilistic elements, referred to as 3D Gaussian particles introduced in 3D Gaussian Splatting [59] and extended to automotive scene understanding in GaussianFormer [50]. This representation models the scene as a collection of localized Gaussian distributions enabling flexible, uncertainty-aware and geometry-preserving scene representations.

A 3D Gaussian particle representation describes the environment by a finite set

$$\mathcal{G} = \{(\boldsymbol{\mu}_i, \mathbf{R}_i, \mathbf{S}_i, \mathbf{f}_i, \alpha_i)\}_{i=1}^{N} \tag{3.4}$$

where each particle is parametrized by a mean position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, rotation of the Gaussian $\mathbf{R}_i \in SO(3)$, scale along principal axes $\mathbf{S}_i \in \mathbb{R}^{3\times3}$, feature vector $\mathbf{f}_i \in \mathbb{R}^C$ and opacity $\alpha_i \in [0, 1]$.

For any point $\mathbf{p} = (x, y, z)$ in space, the contribution of a single 3D Gaussian particle is defined as

$$d(\mathbf{p}; \boldsymbol{\mu}, \mathbf{S}, \mathbf{R}, \alpha) = \exp\left(-\frac{1}{2}(\mathbf{p} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{p} - \boldsymbol{\mu}) \right) \alpha, \tag{3.5}$$

where the covariance matrix is factorized as

$$\boldsymbol{\Sigma} = \mathbf{R}\,\mathbf{S}\mathbf{S}^\top\,\mathbf{R}^\top. \tag{3.6}$$

Occupancy prediction at point $\mathbf{p}$ is obtained then by summing the contributions of all 3D Gaussian particles in the set $\mathcal{G}$:

$$\hat{O}(\mathbf{p}; \mathcal{G}) = \sum_{i=1}^{N} d_i(\mathbf{p}; \boldsymbol{\mu}_i, \mathbf{S}_i, \mathbf{R}_i, \alpha_i) = \sum_{i=1}^{N} \exp\left(-\frac{1}{2}(\mathbf{p} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{p} - \boldsymbol{\mu}_i) \right) \alpha_i. \tag{3.7}$$

3D Gaussian particles are inherently anisotropic and can adapt to scene variability that arises from the presence of both known and unknown objects,

varying object scales, and diverse environmental structures. In learning-based perception frameworks, the parameters of each 3D Gaussian particle are learnable and jointly optimized, enabling the representation to conform to the observed scene rather than relying on a fixed spatial discretization. This adaptivity stands in contrast to 3D grid based representations, where the majority of grid cells may be unoccupied, leading to inefficient use of representational capacity.

Features associated with 3D Gaussian particles can be interpreted as spatially distributed feature densities, anchored to the corresponding Gaussian components. Beyond modeling occupancy alone, Gaussian particle representations can also encode semantic attributes of scene elements, allowing geometry and semantics to be represented jointly within unified probabilistic framework.

From a computational perspective, 3D Gaussian particle representations offer advantages in both memory footprint and processing efficiency compared to dense 3D grids. The ability to concentrate representational resources on occupied or informative regions of the scene, together with the direct association of particles with scene elements, enables more efficient scene modeling and inference.

In contrast to grid based representations, where multimodal features are explicitly assigned to fixed spatial cells, 3D Gaussian particle representations integrate multimodal sensor information indirectly through learning. Multimodal observations influence the parameters of the Gaussian particles, such as their position, shape, and associated features, via the learning process, resulting in a more flexible and adaptive fusion mechanism that is not constrained by a predefined spatial lattice.

3.2 Multimodal Fusion Strategies

Given modality specific feature representations, $\mathbf{F}_m$, derived either directly from raw sensor observations, $\mathbf{x}_m$ or from intermediate or high level processed features, multimodal fusion can be formulated as the integration of heterogeneous modality specific representations into a unified decision or feature space. Fusion is applied over modality-specific representations indexed by a set of entities

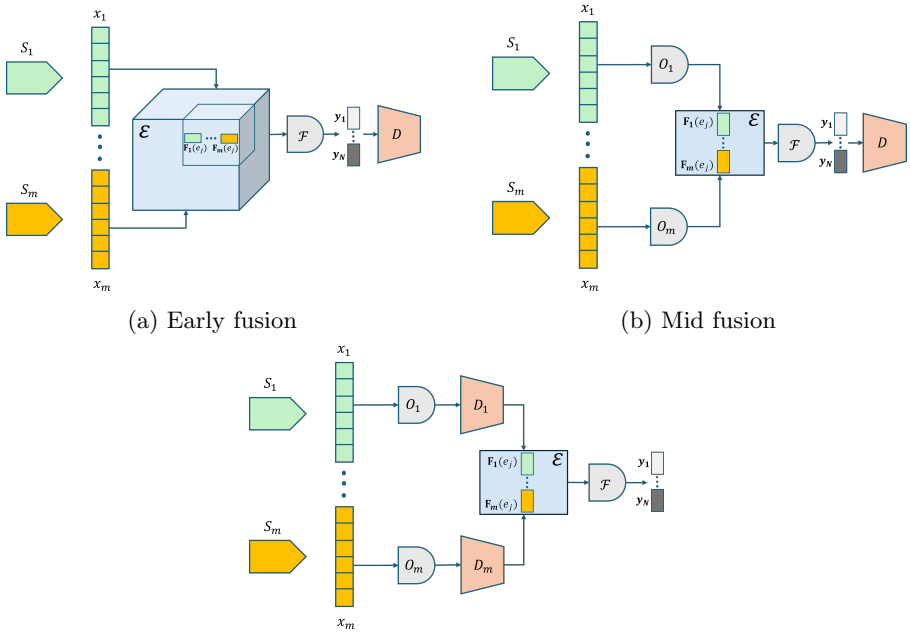
$$\mathcal{E} = \{e_j\}_{j=1}^{N}, \tag{3.8}$$

where each e_j denotes an abstract element in a common reference domain. Entities may correspond to spatial primitives, semantic regions or scene level instances. For each entity $e_j \in \mathcal{E}$, fusion aggregates the representations

$$\mathbf{y}_j = \mathcal{F}\Big(\{\mathbf{F}_m(e_j)\}_{m=1}^{M}\Big), \tag{3.9}$$

where $\mathbf{y}_j$ denotes the fused representation associated with the entity e_j, and $\mathcal{F}(\cdot)$ is a fusion operator.

Fusing and integrating information from multiple sensors, both within the same modality and across different modalities, can be performed at various levels of abstraction within the perception pipeline.



(a) Early fusion

(b) Mid fusion

(c) Late fusion

Figure 3.1: Comparison of early, mid, and late fusion strategies. (a) **Early fusion:** Raw sensor observations $\mathbf{x}_m$ are aligned to entities e_j to form low-level representations $\mathbf{F}_m(e_j)$, which are directly fused by $\mathcal{F}$ into $\mathbf{y}_j$ and optionally decoded by D. (b) **Mid fusion:** Sensor observations are first processed by modality-specific operators to obtain intermediate features $\mathbf{F}_m(e_j)$. Fusion $\mathcal{F}$ combines these embeddings into $\mathbf{y}_j$, followed by decoding through D. (c) **Late fusion:** Each modality independently produces predictions $\mathbf{F}_m(e_j)$ via its own processing and decoder. Fusion $\mathcal{F}$ then combines these high-level outputs into the final result $\mathbf{y}_j$.

Different fusion strategies are typically categorized into early fusion, mid level fusion and late fusion, see Figure 3.1. Early fusion enables rich interactions between modalities at low abstraction levels, while later fusion operates on higher level features or predictions. In this way, the choice of fusion strategy directly influences both perception performance and the system's resilience under various environmental conditions.

Early fusion techniques are more suited for dense spatial representations, as BEV feature maps, while late fusion prioritize redundancy and modularity by combining higher level outputs. Mid level fusion operates on the higher level feature abstractions exploiting cross-modal correlations without assumption of strict sensor signal alignment, which makes it more robust to sensor and calibration noise, modality-specific properties and formats.

The stage at which fusion is performed has a significant impact on the system's ability to exploit cross-modal correlations and to maintain robustness in the presence of sensor noise, partial observability, and modality specific failure. A straightforward approach is to combine sensor information in a

shared representation space, such as a BEV map or a 3D occupancy grid, where learning-based perception models can jointly process the data and reason consistently across sensor modalities.

Early Fusion

In a perception pipeline, sensor signals can be integrated directly at the feature level. This approach assumes that sensor data from different modalities are aligned to a common spatial and/or temporal reference frame. By performing fusion at this early stage, the learning model can exploit low level cross-modal correlations and jointly optimize semantic and geometric reasoning in a multimodal setting.

An example of early fusion is the aggregation of camera image features with raw radar or lidar measurements in a BEV representation, as explored in SimpleBEV [20]. In this approach, radar or lidar signals are rasterized into a 3D voxel grid and concatenated with lifted image features, followed by convolutional compression along the height dimension. While this design is simple, a fusion of sparse and dense modality features, prior to height aggregation, assumes spatial alignment and semantic compatibility. In addition, convolution-based BEV processing of fused signals imposes a limited receptive field, which can restrict broader contextual reasoning across modalities, particularly when cross-modal cues are spatially separated or require flexible range aggregation.

Some early fusion approaches attempt to use one sensor modality as a guide to transform another modality into a representation space that is more compatible for training the downstream task. Frustum PointNets [45] segments and classifies point cloud regions defined by 2D detection proposal in image space, effectively using camera detection to select 3D lidar points for further processing. This tightly couples the modalities at an early stage and can be effective when 2D proposals are reliable, however, it also exposes a key limitation of early coupling, signals depend on each other from the start, making the system vulnerable to sensor failures or lack of 2D detections.

Early fusion also presents several challenges. Specific sensor signal characteristics such as differing resolutions, noise profiles, and data densities can be difficult to reconcile at the feature level. Moreover, unless carefully designed, early fusion schemes may be less robust to sensor failures, calibration errors, or synchronization issues, particularly in systems that implicitly assume the continuous availability and reliability of all sensor modalities.

Mid Fusion

Mid-level fusion is particularly attractive in learning-based perception because it enables both modality-specific feature extraction and cross-modal interaction to be learned jointly from data. In this setting, raw sensor signals are first processed by dedicated encoders tailored to each modality, producing intermediate feature representations that capture higher-level semantic and geometric information. A common instantiation of mid-level fusion in autonomous driving lifts modality-specific features into a shared latent space, allowing the model to

learn how information from different sensors should be combined and weighted depending on the context and sensor reliability.

Approaches such as BEVFusion [21] align camera and lidar features in BEV space, concatenate them, and process the fused representation with a BEV encoder, enabling richer cross-modal interaction than late fusion while retaining geometric consistency. This formulation has become widespread due to its simplicity and compatibility with convolutional architectures. However, BEV-based mid fusion often relies on local convolutional operators, which limits interaction to nearby spatial regions and restricts the ability to model non-local cross-modal dependencies.

To address this limitation, recent works introduce attention-based interaction mechanisms. DeepInteraction [22], for example, employs attention to facilitate adaptive and non-local feature exchange between modalities, allowing the model to emphasize complementary cues across larger spatial extents. While such designs improve flexibility and robustness, they expose an important scalability challenge. Tightly coupled attention-based fusion is non-redundant and can become computationally expensive as spatial resolution or the number of modalities increases, making naive extensions to richer sensor setups impractical.

A central trade-off in mid-level fusion is to operate on meaningful intermediate representations that capture non-local context and modality complementarities, while remaining computationally tractable as scene complexity and sensor diversity grow.

Late Fusion

Late fusion is commonly adopted when modularity, redundancy, and fault tolerance are primary design objectives. In this paradigm, each sensor modality is processed independently by a dedicated perception pipeline, and fusion is performed only on high-level outputs such as object detections, occupancy estimates, or semantic maps. This design allows individual modality pipelines to be developed, trained, and validated separately, and mitigate the system degradation under temporary sensor signal loss or complete failure.

In learning-based perception, this modularity simplifies system maintenance and facilitates sensor-specific architectural choices. However, late fusion largely forgoes the benefits of joint representation learning. Because fusion occurs only after modality-specific predictions are generated, cross-modal geometric and semantic cues cannot be exploited during earlier stages of feature extraction. As a result, ambiguities that could be resolved through joint reasoning must instead be handled through post-hoc association and conflict resolution.

This trade-off is illustrated by RadarNet [60], which performs late fusion of lidar and radar detections. While the modular design improves robustness, the limited spatial resolution of radar and its susceptibility to spurious returns require extensive post-processing to suppress false positives. This highlights a recurring limitation of late fusion namely that complementary cues, such as radar velocity supporting camera or lidar semantics, are only weakly integrated once predictions are already formed.

While late fusion offers robustness and engineering simplicity, it restricts the effective use of multimodal complementarities.

3.3 Attention Based Multimodal Mid Fusion and Structured Initialization

Since their introduction, attention mechanisms [61],[42],[43] have been widely adopted as a central architectural component in learning-based perception systems. At a high level, attention is a feature selection and weighting mechanism that allows a model to dynamically focus on the most relevant parts of its input when computing a representation. Attention computes scores through learned similarity measures between query, key and value embeddings and then aggregates information according to the scores making an adaptive information flow within the neural network. By enabling data driven feature weighting, attention allows models to dynamically adapt to changing sensing conditions and has therefore become a dominant strategy for mid level information fusion, which is particularly interesting for multimodal perception where relevance and reliability of different sensor modalities vary across scenes.

Beyond the fusion operator itself, the effectiveness of attention is closely tied to the spatial representation on which it operates. In structured domains such as BEV grids, explicit spatial alignment across modalities provides a regular layout over which attention can aggregate cross-modal signals efficiently. The dense and well-defined grid structure simplifies correspondence between sensors, allowing attention mechanisms to exploit spatial redundancy and complementary cues in a geometrically consistent manner.

However, when moving toward more flexible or continuous scene representations[50], the interaction between fusion and representation becomes more intricate. Without fixed discretization, spatial relationships and feature correspondences must be encoded implicitly or guided through additional structural assumptions. In such settings, attention alone is often insufficient and geometric priors, initialization strategies, and learned latent structures jointly influence how multimodal information is integrated and stabilized during training.

3.4 Pre-training for Multimodal Scene Representation Learning

While sensor fusion and spatial representations determine how multimodal information is integrated, the learning paradigm governs what is ultimately captured within those representations. Beyond architectural design, an equally important question concerns how scene representations are acquired in the first place and to what extent their learning is tied to specific downstream objectives.

Learning-based perception systems for autonomous driving have traditionally been trained end-to-end i.e. using fully supervised task-specific objectives,

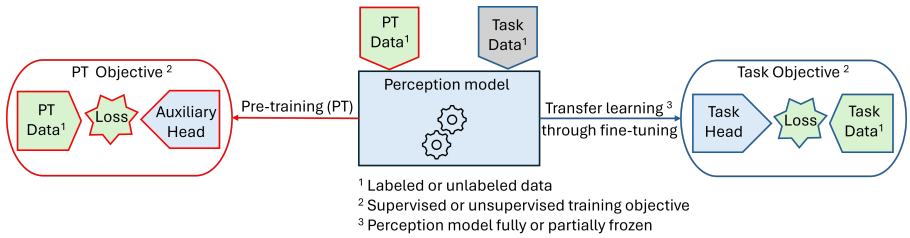


Figure 3.2: Pre-training paradigm: The perception model is first pre-trained using either supervised or unsupervised learning on labeled and/or unlabeled data. In the subsequent stage, the pre-trained perception backbone is (fully or partially) frozen, and a task-specific head is attached and fine-tuned for the downstream task.

such as object detection or semantic segmentation. In this paradigm, models learn representations directly for annotated labels, often requiring large amounts of carefully curated data. While this type of training can achieve strong performance on the target task, it often biases the learned features toward narrow label spaces and limits their transferability across tasks and datasets, generalizing poorly when applied beyond the original training objective and scope.

Pre-training has emerged as a strategy to mitigate these limitations by decoupling representation learning from narrow downstream task objectives as shown in the Figure 3.2. Instead of optimizing features solely for a specific annotated output, pre-training aims to learn more general representations before adapting them to a particular perception task through fine-tuning. This two-stage paradigm shifts part of the learning process from task-specific supervision to broader representation learning.

The effectiveness of large-scale pre-training has been clearly demonstrated by vision foundation models trained on vast and diverse datasets. Foundational models such as CLIP [62] learn joint vision–language representations through large-scale contrastive pre-training, enabling broad cross-modal generalization across downstream tasks. Similarly, DINO [17] and DINOv2 [18] leverage self-supervised objectives to acquire rich and transferable visual features without relying on explicit human annotations. Models such as SAM [19] further show that large-scale pre-training with generic objectives can produce versatile representations adaptable to diverse segmentation tasks. These line of works demonstrate that sufficiently expressive models can capture structural regularities that generalize beyond any single downstream objective.

Inspired by these advances, recent work in autonomous driving has explored generalizable scene representations through occupancy modeling and spatial consistency. Approaches such as 4DOcc [63] and UnO [64] learn temporally consistent occupancy representations, but are largely restricted to a single sensor modality (lidar) and do not explicitly demonstrate transferability across diverse downstream tasks. Other methods, such as GASP [65], leverage large foundation models during pre-training to reduce manual annotation effort.

However, they remain indirectly tied to semantic supervision through pseudo-labels generated by external models.

Applying foundation model paradigms to autonomous driving remains non-trivial. From an industrial perspective, strong reliance on external foundation models as supervisory signals introduces practical challenges. Foundation models improve over time and adopting a newer model version would require reprocessing large-scale datasets to regenerate the corresponding outputs. In real-world deployments where sensor data continuously grows and operational domains shift, regenerating pseudo-labels at scale can become computationally expensive or infeasible.

Moreover, semantic taxonomies, annotation protocols, and deployment requirements often evolve over time. Representations learned through explicit semantic supervision are therefore inherently dependent on the stability of label definitions and the continued availability of compatible supervisory signals.

Pre-training strategies grounded primarily in intrinsic geometric consistency, rather than external semantic supervision, offer a potentially more scalable and sustainable alternative. By anchoring representation learning in spatial structure derived directly from multimodal sensor measurements, such approaches reduce dependence on evolving label definitions and externally generated pseudo-labels, while being adaptable to diverse tasks and deployment settings.

Chapter 4

Summary of Included Papers

This chapter summarises the content and scientific contributions of the three papers included in this thesis, along with the contributions of the respective authors.

4.1 SMAB:Simple Multimodal Attention for Effective BEV Fusion

In **Paper I** (SMAB), we explore learning based multimodal fusion for perception using structured BEV scene representations, as detailed in Section 3.1. The work centers on architectural design choices for fusing signals from heterogeneous sensor modalities camera, radar, and lidar at the mid level of the perception pipeline, with a particular focus on multimodal attention based fusion mechanisms.

The core contribution of the paper is a lightweight, modular attention based fusion module that operates directly on modality specific BEV feature maps. Camera images are processed by a dedicated encoder and projected into a shared voxel based coordinate system using known calibration parameters. Lidar point clouds and radar signals are rasterized directly into the same voxel grid structure. Following height compression to obtain 2D BEV representations, these aligned feature maps serve as input to the BEV Feature Aggregator (BFA) module. There, multimodal deformable attention selectively aggregates cross-modal information at each spatial location and generates a unified BEV feature map used then for the downstream task.

Operating on structured spatial representations such as BEV enables attention to be applied directly over aligned grids, allowing modality-specific features to interact in a geometrically consistent manner. A key design principle of **Paper I** is the preservation of modality-specific redundancy rather than enforcing early feature collapse across modalities. This is enabled by a redundancy preserving training protocol that allows the attention mechanism

to dynamically reweight sensor contributions according to learned contextual cues and controlled signal degradation during training, while still maintaining modality specific features. As a result, the model can adaptively emphasize reliable sensors under favorable conditions and downweight degraded ones in challenging scenarios, such as poor visibility or sparse measurements creating robust predictions.

The proposed approach is trained and evaluated on the NuScenes [26] dataset, using the BEV vehicle segmentation task as the primary benchmark. Results demonstrate that SMAB achieves competitive or in several cases superior performance compared to more complex fusion architectures, while offering a simpler and more scalable design. The results further indicate that image based signals contribute strongly to overall perception performance, and that incorporating depth or distance related information leads to consistent improvements, even when such signals are sparse.

Contribution

The main idea came from A. Mustajbasic who also developed the method, implemented it, conducted all experiments and prepared first draft of the paper. S. Chen, E. Stenborg, and Selpi contributed with feedback on the idea, following up experiments, providing valuable interpretations and editing the paper.

4.2 Guided Gaussians: Enhancing 3D Occupancy Estimation with Sparse Sensor Priors

In **Paper II**, we extend the investigation of multimodal fusion to more flexible and adaptive scene representations, focusing on probabilistic 3D occupancy estimation using Gaussian particle representations, as introduced in Section 3.1. The paper addresses the underutilization of multimodal context in automotive sensor setups by explicitly leveraging distance measurements from lidar and radar as inductive priors for initializing and guiding probabilistic 3D Gaussian particle scene representations.

Unlike fixed resolution 3D grids, 3D Gaussian particles adapt to the spatial and semantic structure of the scene, enabling more efficient and expressive modeling of occupied and free space. The environment is modeled as a set of learnable anisotropic Gaussian primitives, each parametrized by spatial position, covariance and latent feature embeddings.

A central contribution of the work is the introduction of guided initialization strategies for probabilistic 3D Gaussian particles, leveraging sparse sensor priors from lidar/radar measurements. Farthest Point Sampling (FPS) [66] is used to select initialization points that provide improved spatial coverage of the scene. These priors provide inductive bias that anchors the initial particle distribution to physically meaningful regions of space, improving training stability and convergence. Multimodal sensor observations are then integrated through attention based mechanisms that update particle features and parameters, allowing the representation to refine itself over time based on learned multimodal evidence.

It is demonstrated in the paper that less structured representation space benefits from appropriate geometric guidance and that attention based fusion alone is not sufficient in this setup. By combining sparse sensor priors with learnable attention driven updates, the proposed method achieves improved 3D occupancy estimation performance while significantly reducing memory and computational requirements.

Experimental results in the paper show that the proposed guided Gaussian particle representation achieves performance comparable to state-of-the-art multimodal fusion methods on large scale benchmarks such as SurroundOcc-NuScenes [67] and SemanticKITTI [68] occupancy. Moreover, the experiments show that the guided Gaussian representation produces more precise and stable occupancy predictions, particularly in scenarios where ground truth supervision is noisy or imperfect. By explicitly leveraging multimodal sensor information during both initialization and refinement, the approach improves even perception robustness under partial observability due to varying weather conditions.

Contribution

Original idea came from A. Mustajbasic who also contributed to the method, implementation, running and designing experiments and making first draft of the paper. H. Fu and J. Xu worked with implementation, running experiments,

generating visualization and writing parts of the paper. S. Chen, E. Stenborg, and Selpi contributed with feedback on the idea, following up experiments, providing valuable interpretations and editing the paper.

4.3 GeoPriors: Learning Latent 3D Structure via Occupancy Pre-Training for Efficient Multi-Task Scene Understanding

Paper III investigates geometry-centric pre-training of multimodal 3D scene representations. The paper examines whether scene structure alone, learned through occupancy estimation, can serve as an effective foundation for downstream perception tasks in autonomous driving. Building on the continuous 3D Gaussian scene representation studied in **Paper II**, pre-training is performed by supervising the model to predict binary occupancy from multi-view camera observations without relying on semantic labels or features from external vision foundation models.

The proposed method emphasizes geometric consistency as a stable and label-independent signal. By grounding supervision in occupancy estimation, the model learns to encode spatial layout and free space in a modality-consistent manner.

The results reveal that purely geometric pre-training captures strong spatial reasoning, achieving competitive overall (Intersection over Union) IoU score on 3D occupancy prediction and enabling label-efficient transfer to BEV vehicle segmentation with only 40-60% labeled data. However, the approach exhibits significant limitations in semantic separability for fine-grained classification tasks. Class-wise analysis shows that while the model effectively captures object locations, it struggles to distinguish between semantically similar classes that occupy similar geometric structures. For instance, different vehicle types (cars, buses, trucks, construction vehicles) occupy rigid, vehicle-shaped volumes and overlapping spatial regions, and flat surfaces can correspond to drivable areas or sidewalks. These observations indicate that geometric features alone provide limited discriminative power for separating semantic categories in the absence of explicit supervision during pre-training.

These findings establish that effective pre-training for autonomous driving requires hybrid objectives combining geometric consistency with explicit feature separation mechanisms to ensure semantically discriminative features. The work demonstrates both the promise and current limitations of geometry-first approaches, identifying concrete directions for developing methods that bridge geometric reasoning and semantic understanding.

Contribution

The project proposal was initiated by A. Mustajbasic who also developed the method, worked on implementation, running and designing experiments and making first draft of the paper. S. Chen, E. Stenborg, and Selpi contributed with feedback on the idea, following up experiments, providing valuable interpretations and editing the paper.

Chapter 5

Discussion and Future Work

5.1 Discussion

This thesis investigates learning-based multimodal perception for autonomous driving, with particular emphasis on how architectural design choices influence sensor fusion effectiveness and scene representation. It demonstrates that achieving robust perception requires careful co-design of fusion mechanisms and scene representations, rather than treating these components independently.

The thesis further shows that multimodal perception benefits substantially from mid-level fusion strategies that preserve modality-specific information while enabling learned cross-modal interaction. In structured BEV representations, attention-based fusion exploits explicit spatial alignment to integrate complementary sensor cues, as demonstrated in **Paper I**. The results indicate that even relatively simple attention mechanisms can enhance BEV feature representations when properly designed. At the same time, the findings highlight that robustness to partial sensor degradation remains dependent on how complementary modalities are balanced, pointing to the importance of adaptive fusion strategies in real-world deployment, where sensor failures are inevitable.

Moving beyond structured BEV grids to more adaptive representations based on 3D Gaussian particles introduces new challenges and opportunities. The study in **Paper II** shows that coupling fusion mechanisms with appropriate inductive biases can improve stability in learning and enhance spatial reasoning. In particular, sensor-guided initialization based on distance measurements improves both efficiency and performance by focusing representation capacity on informative regions. This demonstrates that leveraging raw sensor signals as spatial priors is beneficial, but also reveals a dependency on the quality and availability of such signals, which may limit robustness under degraded sensing conditions.

Building on the same scene representation, **Paper III** extends the investigation to geometry-centric pre-training, revealing important limitations of

purely geometric supervision. While occupancy-based pre-training produces strong global spatial reasoning, evidenced by competitive overall IoU scores, it struggles with fine-grained semantic discrimination, particularly for classes with similar geometric properties. The analysis exposes a fundamental limitation in that geometric pre-training effectively captures "where objects are" but provides insufficient cues for distinguishing "what those objects are" when multiple semantic categories occupy similar spatial configurations.

Taken together, these findings illustrate a key trade-off across the works. While structured fusion and geometry-centric representations improve robustness and spatial reasoning, they do not inherently guarantee semantic discriminability. The decoupling of geometry and semantics during pre-training offers practical benefits, such as stability across changing semantic taxonomies and reduced reliance on labeled data, but also introduces limitations that must be addressed through hybrid learning objectives. The results therefore suggest that effective multimodal perception systems require not only strong geometric priors and fusion strategies, but also mechanisms that explicitly support semantic differentiation, highlighting key directions for future research explored in the following section.

5.2 Future Work

The insights gained in this work point toward several directions for future research. Related to **RQ5**, a natural next step is to extend geometry-centric pretraining with explicit temporal learning objectives. The presented method in **Paper III** rely on single timeframe signals to obtain geometric representations while driving scenes are inherently dynamic and provide strong temporal supervision signals. Future research should therefore investigate pre-training strategies that involve temporal data sequences and enforce temporal consistency in the learned representations. Possible approaches include predicting future occupancy states or enforcing motion-consistent latent features across frames. Such objectives could be implemented using lidar sweeps or image frames from adjacent time stamps while evaluation could be done in terms of label efficiency improvements during semantic fine-tuning.

In addition to improving label efficiency, **RQ6** motivates expanding the evaluation of geometry-aware representations to a broader set of automotive perception tasks. Future studies should benchmark the pre-trained representations developed in this thesis on tasks beyond semantic segmentation or 3D semantic occupancy by including 3D object detection, and motion forecasting.

Another promising direction is the development of fully self-supervised multimodal pre-training frameworks that jointly learn from raw lidar, camera, and potentially radar data. Future models could employ cross-modal prediction objectives, such as reconstructing lidar geometry from camera features or aligning latent representations across sensor modalities. Training such models on large-scale unlabeled driving datasets would allow the learned representations to capture both geometric structure and appearance information, moving toward scalable foundation models for automotive perception.

Finally, future work should place greater emphasis on long-tail scenarios and safety-critical corner cases. Rare traffic situations, unusual object configurations, and extreme environmental conditions remain underrepresented in public datasets, limiting the robustness of current perception systems. One possible direction is to design training strategies that explicitly test robustness under sensor degradation, occlusions, or adverse weather conditions. The geometry-aware and uncertainty-aware representations explored in this thesis provide promising starting points for this line of research. In particular, the robustness of **Paper I** to signal loss and the computational efficiency of the sparse representation introduced in **Paper II** could be leveraged to develop perception models that remain reliable even when sensor inputs are incomplete or degraded.

Bibliography

[1] M. Kolbenstvedt, R. Elvik, B. Elvebakk, A. Hervik and L. Braein, "Effects of swedish traffic safety research 1971–2004," VINNOVA – Swedish Governmental Agency for Innovation Systems, Tech. Rep. VA 2007:10, 2007 (cit. on p. 3).

[2] Volvo Group, *The Three-Point Safety Belt – The Three Points that Saved One Million Lives*, 2026 (cit. on p. 3).

[3] United Nations Office for Disaster Risk Reduction (UNDRR), *Road traffic accident (tl0405)*, Hazard Information Profile (HIP), United Nations Office for Disaster Risk Reduction (UNDRR), 2025 (cit. on p. 3).

[4] Volvo Cars, *Safe Space Technology – Safety Technology for Safer Driving*, 2026 (cit. on pp. 4, 9–11).

[5] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, *Mandatory Driver Assistance Systems Expected to Help Save Over 25,000 Lives by 2038*, Jul. 2024 (cit. on p. 4).

[6] AAA Foundation for Traffic Safety, *Safety Benefits of Advanced Driver Assistance Systems (ADAS)*, Jan. 2026 (cit. on p. 4).

[7] Y. Zhang, A. Carballo, H. Yang and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023, ISSN: 0924-2716 (cit. on p. 4).

[8] C.-G. Roh, J. Kim and I.-J. Im, "Analysis of impact of rain conditions on adas," *Sensors*, vol. 20, no. 23, 2020, ISSN: 1424-8220 (cit. on pp. 4, 10).

[9] J. Kim, B. J. Park, C. G. Roh and Y. Kim, "Performance of mobile lidar in real road driving conditions," *Sensors (Basel)*, vol. 21, no. 22, p. 7461, Nov. 2021 (cit. on p. 4).

[10] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015 (cit. on pp. 4, 15).

[11] C. Heipke and F. Rottensteiner, "Deep learning for geometric and semantic tasks in photogrammetry and remote sensing," *Geo-Spatial Information Science*, vol. 23, no. 1, pp. 10–19, 2020 (cit. on p. 4).

[12] T. Yin, X. Zhou and P. Krähenbühl, "Center-based 3d object detection and tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 779–11 788 (cit. on p. 4).

[13] X. Tian et al., "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23, New Orleans, LA, USA: Curran Associates Inc., 2023 (cit. on pp. 4, 17).

[14] NVIDIA Corporation, *Self-Driving Car Hardware – NVIDIA DRIVE Platform*, 2026 (cit. on p. 4).

[15] Qualcomm Technologies, Inc., *Automated Driving – Qualcomm Automotive Expertise*, 2026 (cit. on p. 4).

[16] Aptiv PLC, *What Is Sensor Fusion?* Mar. 2020 (cit. on p. 4).

[17] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640 (cit. on pp. 5, 6, 25).

[18] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, ISSN: 2835-8856 (cit. on pp. 5, 6, 25).

[19] A. Kirillov et al., "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003 (cit. on pp. 5, 6, 25).

[20] A. W. Harley, Z. Fang, J. Li, R. Ambrus and K. Fragkiadaki, "Simple-bev: What really matters for multi-sensor bev perception?" In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2759–2765 (cit. on pp. 6, 17, 18, 22).

[21] Z. Liu et al., "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2774–2781 (cit. on pp. 6, 16, 23).

[22] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, Curran Associates Inc., 2022, ISBN: 9781713871088 (cit. on pp. 6, 23).

[23] Zoox, Inc., *Zoox – It's Not a Car; It's a Robotaxi Designed Around You*, 2026 (cit. on p. 9).

[24] Waymo LLC, *Waymo – Self-Driving Cars and Autonomous Vehicle Technology*, 2026 (cit. on p. 9).

[25] BMW AG, *Automotive Sensors – Assistance Systems' Sense Organs*, Apr. 2021 (cit. on pp. 10, 11).

[26] H. Caesar et al., "Nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 618–11 628 (cit. on pp. 10–12, 28).

[27] ROHM Semiconductor, *ADAS Camera System – Automotive Solution for Advanced Driver Assistance Systems*, 2026 (cit. on p. 10).

[28] Sony Semiconductor Solutions Corporation, *Image Sensor for Automotive Use*, 2026 (cit. on p. 10).

[29] TechNexion, *What Is a Camera ISP? What Are Its Functions?* 2025 (cit. on p. 10).

[30] Aptiv PLC, *What Is 4D Imaging Radar?* Sep. 2021 (cit. on p. 10).

[31] A. Vaughan, "An Introduction to Automotive Lidar," Texas Instruments Incorporated, White Paper SLYY150D, 2025 (cit. on p. 11).

[32] M. Alibeigi et al., "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 20 121–20 131 (cit. on pp. 11–13).

[33] S. Thomas, *Data Annotation Types Used for Autonomous Vehicles*, Sep. 2025 (cit. on p. 12).

[34] S. Thrun, "Probabilistic robotics," *Commun. ACM*, vol. 45, pp. 52–57, 2002 (cit. on p. 15).

[35] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960 (cit. on p. 15).

[36] S. F. Schmidt, "Application of state-space methods to navigation problems," in *Advances in Control Systems*, vol. 3, Elsevier, 1966, pp. 165–222 (cit. on p. 15).

[37] N. Gordon, D. Salmond and A. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, pp. 107–113, 2 1993 (cit. on p. 15).

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, 91–110, Nov. 2004, ISSN: 0920-5691 (cit. on p. 15).

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, 886–893 vol. 1 (cit. on p. 15).

[40] M. A. Richards, *Fundamentals of Radar Signal Processing, Third Edition.* McGraw Hill, 2022, ISBN: 9781260468717 (cit. on p. 15).

[41] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989 (cit. on p. 15).

[42] A. Vaswani et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Curran Associates Inc., 2017, 6000–6010, ISBN: 9781510860964 (cit. on pp. 15, 16, 24).

[43] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021 (cit. on pp. 16, 24).

[44] H. Cai, Z. Zhang, Z. Zhou, Z. Li, W. Ding and J. Zhao, *Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation*, 2023. arXiv: 2303.17099 [cs.CV] (cit. on p. 16).

[45] C. R. Qi, W. Liu, C. Wu, H. Su and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927 (cit. on pp. 16, 22).

[46] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 569–17 580 (cit. on p. 16).

[47] Y. Shi et al., "Grid-centric traffic scenario perception for autonomous driving: A comprehensive review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 11 814–11 834, 2025 (cit. on p. 17).

[48] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. García and A. De La Escalera, "Birdnet: A 3d object detection framework from lidar information," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3517–3523 (cit. on p. 17).

[49] S.-L. Yu, T. Westfechtel, R. Hamada, K. Ohno and S. Tadokoro, "Vehicle detection and localization on bird's eye view elevation images using convolutional neural network," in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, 2017, pp. 102–109 (cit. on p. 17).

[50] Y. Huang, W. Zheng, Y. Zhang, J. Zhou and J. Lu, "Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction," in *Computer Vision – ECCV 2024*, Springer Nature Switzerland, 2025, pp. 376–393, ISBN: 978-3-031-73383-3 (cit. on pp. 17, 19, 24).

[51] Y. Huang, A. Thammatadatrakoon, W. Zheng, Y. Zhang, D. Du and J. Lu, "Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 27 477–27 486 (cit. on p. 17).

[52] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 194–210, ISBN: 978-3-030-58568-6 (cit. on pp. 17, 18).

[53] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision – ECCV 2022*, Springer Nature Switzerland, 2022, pp. 1–18, ISBN: 978-3-031-20077-9 (cit. on pp. 17, 18).

[54] C. Lu and G. Dubbelman, "Learning to complete partial observations from unpaired prior knowledge," *Pattern Recognition*, vol. 107, p. 107 426, 2020, ISSN: 0031-3203 (cit. on p. 18).

[55] A. Popov, P. Gebhardt, K. Chen and R. Oldja, "Nvradarnet: Real-time radar obstacle and free space detection for autonomous driving," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6958–6964 (cit. on p. 18).

[56] P. Wu, S. Chen and D. N. Metaxas, "Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 382–11 392 (cit. on p. 18).

[57] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499 (cit. on p. 18).

[58] A.-Q. Cao and R. de Charette, " MonoScene: Monocular 3D Semantic Scene Completion," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Jun. 2022, pp. 3981–3991 (cit. on p. 19).

[59] B. Kerbl, G. Kopanas, T. Leimkuehler and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, Jul. 2023, ISSN: 0730-0301 (cit. on p. 19).

[60] B. Yang, R. Guo, M. Liang, S. Casas and R. Urtasun, "Radarnet: Exploiting radar for robust perception of dynamic objects," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 496–512, ISBN: 978-3-030-58523-5 (cit. on p. 23).

[61] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015 (cit. on p. 24).

[62] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, PMLR, Jul. 2021, pp. 8748–8763 (cit. on p. 25).

[63] T. Khurana, P. Hu, D. Held and D. Ramanan, "Point cloud forecasting as a proxy for 4d occupancy forecasting," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1116–1124 (cit. on p. 25).

[64] B. Agro, Q. Sykora, S. Casas, T. Gilles and R. Urtasun, "Uno: Unsupervised occupancy fields for perception and forecasting," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14 487–14 496 (cit. on p. 25).

[65] W. Ljungbergh et al., *Gasp: Unifying geometric and semantic self-supervised pre-training for autonomous driving*, 2025. arXiv: 2503.15672 (cit. on p. 25).

[66] Y. Eldar, M. Lindenbaum, M. Porat and Y. Zeevi, "The farthest point strategy for progressive image sampling," *IEEE Transactions on Image Processing*, vol. 6, no. 9, pp. 1305–1315, 1997 (cit. on p. 29).

[67] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou and J. Lu, " SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, Oct. 2023, pp. 21 672–21 683 (cit. on p. 29).

[68] J. Behley et al., "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9296–9306 (cit. on p. 29).