

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Multi-sensor fusion with attention mechanisms for visual perception in autonomous vehicles

Eduardo Sperle Honorato

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Eduardo Sperle Honorato

Multi-sensor fusion with attention mechanisms for visual perception in autonomous vehicles

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Denis Fernando Wolf

USP – São Carlos
June 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

H774m Honorato, Eduardo Sperle
Multi-sensor fusion with attention mechanisms
for visual perception in autonomous vehicles /
Eduardo Sperle Honorato; orientador Denis Fernando
Wolf. -- São Carlos, 2025.
105 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2025.

1. Attention Mechanism. 2. Multi-sensor Fusion.
3. Segmentation. 4. 3D Object Detection. 5.
Autonomous Vehicles. I. Wolf, Denis Fernando,
orient. II. Título.

Eduardo Sperle Honorato

**Fusão multi-sensorial com mecanismos de atenção para
percepção visual em veículos autônomos**

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestre em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Denis Fernando Wolf

USP – São Carlos
Junho de 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my parents for their unwavering support and for providing me with the opportunity to pursue this master's degree. Without their encouragement and sacrifices, none of this would have been possible.

I am also sincerely grateful to my advisor, whose guidance, patience, and expertise were instrumental in shaping the course of this research. His mentorship has been invaluable throughout this journey.

Additionally, I acknowledge the financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). This study was financed in part by CAPES – Finance Code 001 88887.841878/2023-00. The resources and opportunities made available through this funding were crucial to the successful completion of this research.

“Anything not saved will be lost.”

– Nintendo Wii

RESUMO

HONORATO, E. S. **Fusão multi-sensorial com mecanismos de atenção para percepção visual em veículos autônomos.** 2025. 105 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Veículos autônomos estão cada vez mais próximos de se tornarem parte do cotidiano urbano. No entanto, desafios significativos ainda precisam ser superados para garantir que esses veículos sejam seguros e eficientes. Um dos principais desafios está na percepção, especialmente na segmentação e detecção de objetos 3D, que utiliza múltiplos sensores para melhorar a precisão e operar em condições adversas. A fusão eficiente desses sensores é uma questão central, pois determina a qualidade da detecção e o custo computacional do sistema. Métodos modernos de fusão multissensorial fazem uso de técnicas de Aprendizado Profundo, e uma abordagem emergente nessa área é a utilização de mecanismos de atenção. Esses mecanismos permitem obter representações mais informativas dos mapas de características extraídos pelos sensores, destacando as informações mais relevantes e suprimindo aquelas menos significativas. Neste contexto, este trabalho investiga o uso de mecanismos de atenção para otimizar o modelo BEVFusion, que alcançou o estado da arte ao empregar uma fusão unificada Câmera-LiDAR na representação *Bird's Eye View* (BEV). O principal diferencial do BEVFusion é sua eficiente transformação da visão de perspectiva das câmeras para a representação BEV. No entanto, sua abordagem de fusão se limita à simples concatenação das características extraídas dos sensores, o que pode não ser a solução mais eficiente. Outro aspecto crítico do modelo BEVFusion é seu alto custo computacional, pois depende de redes neurais profundas que exigem hardware robusto, tornando sua aplicação em veículos autônomos mais desafiadora. Isso se deve ao fato de que o hardware embarcado desses veículos precisa ter baixo custo e alta eficiência energética. Diante desse cenário, este trabalho propõe o estudo e a implementação de mecanismos de atenção para aprimorar a fusão de sensores do BEVFusion nas tarefas de detecção de objetos 3D e segmentação, ao mesmo tempo em que busca tornar o modelo mais eficiente computacionalmente. Foram realizadas modificações para reduzir o consumo de VRAM e o tempo de processamento, garantindo um desempenho semelhante ao do modelo original, mas com menor demanda por recursos computacionais. Os resultados obtidos são promissores, demonstrando um aumento de 14.12% no IoU para a tarefa de segmentação e de 0.732% no mAP para a detecção de objetos 3D. Além disso, houve uma redução de 3,3 vezes no tempo de treinamento e uma diminuição de quase 50% no consumo de memória VRAM.

Palavras-chave: Mecanismo de Atenção, Fusão Multi-Sensorial, Segmentação, Detecção de Objetos 3D, Veículos Autônomos.

ABSTRACT

HONORATO, E. S. **Multi-sensor fusion with attention mechanisms for visual perception in autonomous vehicles.** 2025. 105 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Autonomous vehicles are increasingly becoming a part of urban life. However, significant challenges still need to be overcome to ensure these vehicles are both safe and efficient. One of the main challenges lies in perception, particularly in 3D object segmentation and detection, which relies on multiple sensors to enhance accuracy and operate under adverse conditions. The efficient fusion of these sensors is a crucial factor, as it directly impacts detection quality and computational cost. Modern multi-sensor fusion methods leverage Deep Learning techniques, and an emerging approach in this field is the use of attention mechanisms. These mechanisms enable more informative representations of feature maps extracted from sensors, highlighting the most relevant information while suppressing less significant data. In this context, this study explores the use of attention mechanisms to optimise the BEVFusion model, which has achieved state-of-the-art performance by employing a unified Camera-LiDAR fusion in a *Bird's Eye View* (BEV) representation. The key advantage of BEVFusion is its highly efficient transformation of camera-perspective views into the BEV representation. However, its sensor fusion approach is limited to a simple concatenation of extracted features, which may not be the most efficient solution. Another critical aspect of BEVFusion is its high computational cost, as it relies on deep learning models that demand powerful hardware, posing a challenge for deployment in autonomous vehicles. This is particularly relevant given the need for embedded hardware in such vehicles to be both cost-effective and energy-efficient. To address these issues, this study investigates and implements attention mechanisms to enhance the sensor fusion process in BEVFusion for 3D object detection and segmentation, while also improving computational efficiency. Modifications were made to reduce VRAM consumption and processing time, ensuring performance comparable to the original model but with lower computational demands. The results are promising, showing a 14.12% increase in IoU for the segmentation task and a 0.732% improvement in mAP for 3D object detection. Additionally, training time was reduced by a factor of 3.3, and VRAM consumption was nearly halved.

Keywords: Attention Mechanism, Multi-sensor Fusion, Segmentation, 3D Object Detection, Autonomous Vehicles.

LIST OF FIGURES

Figure 1 – Example of 3D object detection in an urban scenario, highlighting how objects are marked. Adapted from (SINDAGI; ZHOU; TUZEL, 2019).	24
Figure 2 – Example of semantic segmentation, adapted from (GAUTAM; MATHURIA; MEENA, 2022a).	24
Figure 3 – Example of lane detection, adapted from (HONDA; UCHIDA, 2023).	25
Figure 4 – Comparison of images obtained from the RGB camera (left) and thermal camera (right) taken from the ForesightAuto page. (DANZIGER, 2020).	31
Figure 5 – Comparison of 2D (left) and 3D (right) LiDAR Point Cloud Images (RO-BOSENSE..., 2022).	33
Figure 6 – Main sensors used in autonomous vehicles and their performance under different conditions. Adapted from (UDACITY, 2025).	34
Figure 7 – Illustration of fusion methods categorized by stage: a) Data Fusion, b) Decision Fusion, and c) Feature Fusion (ALABA; GURBUZ; BALL, 2022).	38
Figure 8 – Examples of the three types of segmentation. Image adapted from (COSTEA; PETROVAI; NEDEVSCHE, 2018).	41
Figure 9 – Example of an image recalibrated using attention mechanisms (right) compared to the original (left). Adapted from (WU <i>et al.</i> , 2023).	51
Figure 10 – Figure 10a illustrates the original BEVFusion fusion implementation, while Figure 10b shows the modified BEVFusion with the addition of attention and dropout mechanisms.	69
Figure 11 – Modifications to the BEVFusion pipeline. The image backbone (blue) was replaced with ResNet50, and dropout and an attention mechanism (red) were added to the sensor fusion process. Other components remain unchanged.	70
Figure 12 – Diagram of the CBAM module, showing sequential application of channel and spatial attention. Adapted from (WOO <i>et al.</i> , 2018).	75
Figure 13 – Sensor configuration used to generate the nuScenes Dataset (CAESAR <i>et al.</i> , 2020).	76
Figure 14 – Example of adverse conditions in the nuScenes dataset (CAESAR <i>et al.</i> , 2020).	77
Figure 15 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference.	83
Figure 16 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference.	84

Figure 17 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference..	85
Figure 18 – Visual comparison of segmentation results between cSE and BEVFusion, using the Ground Truth as reference.	88

LIST OF TABLES

Table 1 – Performance metrics for different models. Metrics with \uparrow indicate higher is better, and those with \downarrow indicate lower is better.	82
Table 2 – Configuration and performance of models during training.	85
Table 3 – Average mIoU Metrics by Category.	86
Table 4 – Configuration and performance of models for segmentation.	87

LIST OF ABBREVIATIONS AND ACRONYMS

AFNet	Attention Fusion Network
AFTR	Adaptive Fusion Transformer
AOE	Average Orientation Error
AP	Average Precision
ASCA	Adaptive Spatial Cross Attention
ASE	Average Scale Error
AVE	Average Velocity Error
BEV	Bird's Eye View
CA	Channel Attention
CAFA	Cross-Attention Feature Alignment
CBAM	Convolutional Attention Block Module
CMT	Cross Modal Transformer
CNN	Convolutional Neural Networks
cSE	Spatial Squeeze and Channel Excitation
csSE	Concurrent Spatial and Channel Squeeze and Excitation
EPMF	Efficient Perception-Aware Multi-Sensor Fusion
FFT	Fast Fourier Transform
FPN	Feature Pyramid Networks
GANs	Generative Adversarial Networks
GPU	Graphics Processing Unit
LFA	Local Feature Aggregation
LiDAR	Light Detection And Ranging
LIFT	LiDAR Image Fusion Transformer
LSTM	Long Short-Term Memory
MAFS	Modality-Agnostic Feature Sampler
mAP	mean Average Precision
mIoU	mean Intersection over Union
MMAF-Net	Multi-Modal Attention Fusion Network
mTP	mean True Positive
NDS	nuScenes Dataset Score
PMF	Perception-aware Multi-Sensor Fusion

PSPNet	Pyramid Scene Parsing Network
RADAR	Radio Detection And Ranging
RGB	Red Green Blue
RNNs	Recurrent Neural Networks
SA	Spatial Attention
SAF	Spatial Attention Frustum
SCFI	Self-supervised Cross-modal Feature Interaction
SE	Squeeze and Excitation
seq2seq	sequence-to-sequence
sSE	Channel Squeeze and Spatial Excitation
STSA	Spatial Temporal Self-Attention
TP	True Positive

LIST OF SYMBOLS

- \mathbb{R} — Real numbers set
- U — Input feature map
- \hat{U}_c — Recalibrated input feature map
- z_c — Aggregated information for the c -th channel
- z_{avg} — Average pooling aggregated information
- z_{max} — Max pooling aggregated information
- W_1 — Fully connected layers weights
- W_2 — Fully connected layers weights
- C — Number of channels
- H — Dimensional height
- W — Dimensional width
- r — Reduction rate
- W_{sq} — Weights of the convolution with kernel size 1
- $f^{7 \times 7}$ — Weights of the Convolution with kernel size 7
- σ — Sigmoid Activation
- Σ — Summation
- \cdot — Element-wise multiplication
- $ReLU$ — ReLU Activation

CONTENTS

1	INTRODUCTION	23
2	THEORETICAL BACKGROUND	29
2.1	Sensors in Autonomous Vehicles	29
2.1.1	<i>Camera</i>	30
2.1.2	<i>LiDAR</i>	32
2.1.3	<i>Radar</i>	33
2.2	Bird's-Eye View Representation	34
2.3	Multi-Sensor Fusion	35
2.3.1	<i>Probability Based Methods</i>	36
2.3.2	<i>Classification Based Methods</i>	36
2.3.3	<i>Inference Based Methods</i>	37
2.3.4	<i>Data Fusion Methods</i>	37
2.3.5	<i>Feature Fuse Methods</i>	38
2.3.6	<i>Decision Fusion Methods</i>	39
2.4	Segmentation	39
2.4.1	<i>Fully Convolutional Networks</i>	43
2.4.2	<i>Convolutional Models With Graphical Models</i>	43
2.4.3	<i>Encoder-Decoder Based Models</i>	43
2.4.4	<i>Multi-Scale and Pyramid Network Based Models</i>	43
2.4.5	<i>R-CNN Based Models</i>	44
2.4.6	<i>Dilated Convolutional Models and DeepLab Family</i>	44
2.4.7	<i>Recurrent Neural Network Based Models</i>	44
2.4.8	<i>Generative Models and Adversarial Training</i>	44
2.4.9	<i>RGB-D Based Segmentation</i>	45
2.4.10	<i>Projected Images Based Segmentation</i>	45
2.4.11	<i>Voxel-Based Segmentation</i>	45
2.4.12	<i>Point-Based Segmentation</i>	46
2.5	3D Object Detection	46
2.5.1	<i>Image based methods</i>	48
2.5.2	<i>Cloud Point Based Methods</i>	49
2.5.3	<i>Fusion-Based Method</i>	50
2.6	Attention Mechanisms	50

3	RELATED WORK	53
3.1	BEVFusion	53
3.2	Attention Mechanisms in Multi-Sensor Fusion for 3D Object Detection	55
3.3	Attention Mechanisms in Multi-Sensor Fusion for Segmentation	62
3.4	Final considerations	65
4	CONDUCTED RESEARCH	67
4.1	Modifications to BEVFusion	67
4.1.1	<i>Spatial Squeeze and Channel Excitation</i>	70
4.1.2	<i>Channel Squeeze and Spatial Excitation</i>	71
4.1.3	<i>Concurrent Spatial and Channel Squeeze and Excitation</i>	71
4.1.4	<i>Channel Attention</i>	72
4.1.5	<i>Spatial Attention</i>	73
4.1.6	<i>CBAM</i>	74
4.2	<i>nuScenes Dataset</i>	76
4.3	Materials and Resources	78
4.4	Evaluation	79
5	RESULTS	81
6	CONCLUSION	89
BIBLIOGRAPHY		93
APPENDIX A	PUBLISHED WORKS	103



INTRODUCTION

Advancements in autonomous vehicles have demonstrated significant potential to transform urban mobility, bringing us closer to a future where these vehicles will become a common presence in our cities. Over the years, there has been a growing and sustained interest in autonomous vehicle research. In recent years, this field has experienced remarkable progress, driven by continuous improvements in artificial intelligence algorithms, advancements in sensor technology, and increased computational capacity for processing and decision-making. These developments are paving the way for the practical feasibility and seamless integration of autonomous vehicles into daily life, fostering a revolution in how we move and interact within urban environments.

One of the most critical tasks for autonomous vehicles is perception, as it serves as the foundation for others such as motion and trajectory planning, obstacle avoidance, and more. Perception in autonomous vehicles involves the ability to interpret and understand the surrounding environment through sensors and data processing algorithms. This task is essential for ensuring safe and efficient driving, as it provides crucial information for motion and trajectory planning. By perceiving the environment, the vehicle can identify and recognize objects such as other vehicles, pedestrians, cyclists, traffic signs, and obstacles, enabling real-time decision-making to avoid collisions and comply with traffic rules.

Moreover, perception is also essential for autonomous navigation in complex and dynamic environments, such as urban areas and highways, where a variety of scenarios and conditions require the vehicle to respond adaptively and swiftly. Therefore, developing robust and accurate perception systems is a critical aspect of advancing autonomous vehicle technology and ensuring the safety and reliability of these systems. Within the field of perception for autonomous vehicles, several fundamental areas stand out, playing crucial roles in the vehicle's ability to understand and interact with its environment: 3D object detection, road and lane detection, and segmentation.

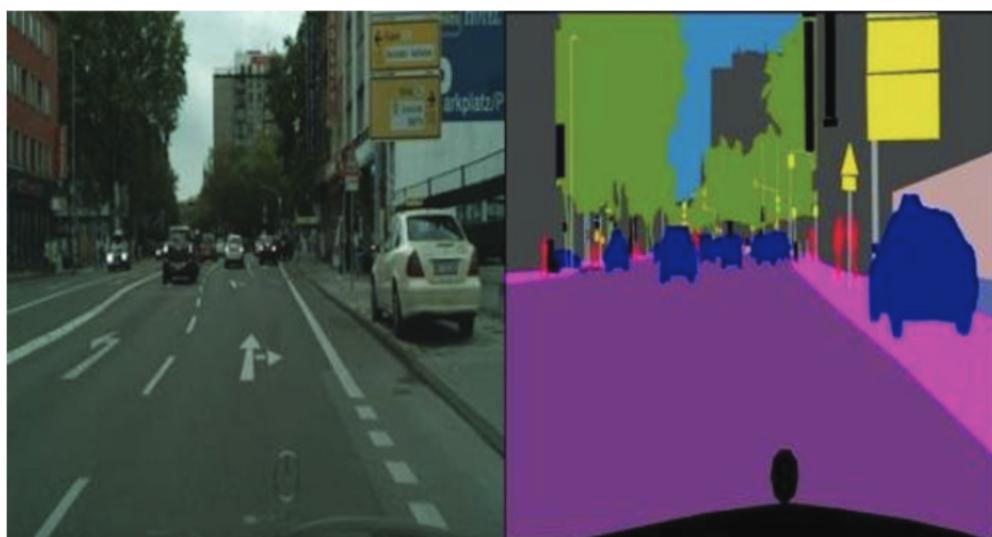
3D object detection is a key area within autonomous driving, responsible for identifying and localizing three-dimensional objects in the vehicle's surroundings, such as cars, pedestrians, cyclists, and obstacles. This task is fundamental to ensuring the safety and efficiency of autonomous vehicles by enabling them to recognize and appropriately respond to environmental elements that may pose risks or obstruct their path. Figure 1 illustrates an example of 3D object detection in an urban scenario, where the green bounding boxes represent the detected objects.

Figure 1 – Example of 3D object detection in an urban scenario, highlighting how objects are marked.
Adapted from ([SINDAGI; ZHOU; TUZEL, 2019](#)).



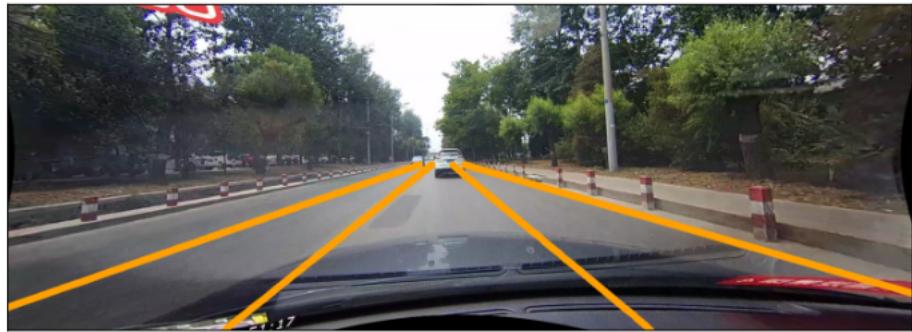
Segmentation is another critical area, aimed at dividing a scene into distinct regions or objects based on their visual or semantic attributes. This process enables a more detailed and nuanced understanding of the environment, supporting the precise identification and classification of individual elements. Additionally, it facilitates the delineation of areas of interest, such as roads, sidewalks, and pedestrian crossings, which are essential for autonomous navigation. Figure 2 illustrates an example of semantic segmentation in an urban environment, where objects belonging to the same category are represented by the same color.

Figure 2 – Example of semantic segmentation, adapted from ([GAUTAM; MATHURIA; MEENA, 2022a](#)).



Finally, road and lane detection is pivotal for autonomous navigation, as it involves identifying and mapping the paths available to the vehicle while determining their geometry, direction, and traffic conditions. This capability enables the vehicle to plan and execute safe and efficient trajectories, remain within the appropriate lanes, and adhere to traffic regulations. Together, these tasks form the cornerstone of perception in autonomous vehicles, delivering the critical information required for reliable and intelligent autonomous driving. Figure 3 illustrates an example of lane detection, showing the ground truth used to train a lane detection model.

Figure 3 – Example of lane detection, adapted from ([HONDA; UCHIDA, 2023](#)).



For all these tasks, the use of cameras is fundamental and highly intuitive, as vision is the primary sense humans rely on, and many elements in urban environments are designed to be visually distinguishable. However, there are limitations to relying exclusively on images, such as difficulties in accurately estimating distances or obtaining a comprehensive representation of the vehicle's surroundings solely through visual information. To overcome these limitations, other sensors, such as Light Detection And Ranging (LiDAR) and Radio Detection And Ranging (RADAR), can be employed alongside camera images. This combination enables the capture of additional nuances, such as distance, speed, and three-dimensional representations of objects around the vehicle. This multimodal approach provides a more complete and accurate view of the environment, thereby enhancing the perception and decision-making capabilities of autonomous vehicles.

The use of multiple sensors raises fundamental questions about how the data obtained from these devices can be effectively integrated to perform perception tasks. This is a broad and constantly evolving field of study, not only in the context of autonomous vehicles but also in many other areas. There are several approaches to sensor fusion; however, in recent years, the use of Machine Learning techniques, especially Deep Learning, has emerged as a promising way to achieve this integration intelligently and efficiently.

The field of Deep Learning is constantly evolving, with applications across a wide variety of domains. Often, a method is proposed to solve a specific problem, but due to the modular nature of algorithms, they can be adapted and modified to be applied in other areas. This flexibility and versatility of Deep Learning models have driven significant advancements

in various disciplines, constantly expanding the boundaries of what can be achieved with this technology.

In the current context, the attention mechanism, initially developed to address issues related to sequences and linguistic problems, has been successfully adapted to handle classification and detection problems in images. One of the most prominent attention methods is the *Transformer*([VASWANI et al., 2017](#)), widely known for its applications in natural language processing and pattern recognition in images.

The attention mechanism is based on the ability to process a signal and highlight the most relevant components for a given task, while disregarding the less relevant ones. In this sense, exploring these features can be crucial for identifying which sensors are most relevant in a specific situation, or even which specific parts of each sensor are crucial for the accurate detection of three-dimensional objects under certain circumstances. Thus, this thesis emerges as a study on how attention mechanisms can be employed to extract more meaningful information from sensors, aiming to enhance the fusion process between them.

This study is based on the BEVFusion model ([LIU et al., 2022](#)), which stands out for setting a new standard by converting multiple camera images and LiDAR point clouds into a unified representation known as Bird's Eye View (BEV). This representation provides an aerial view of the environment around the vehicle, facilitating the detection and segmentation of 3D objects, while optimizing processing time. However, the sensor fusion approach used by BEVFusion is relatively simple, consisting of concatenating the feature maps from the sensors and processing them through a convolutional network. Therefore, this thesis aims to explore more advanced fusion methods, utilizing attention mechanisms, to further improve the performance of the original model.

One of the major drawbacks of BEVFusion is its high computational cost, as it relies on deep neural networks to process multi-sensor data, demanding powerful hardware with high energy consumption. This poses a challenge for real-world deployment in autonomous vehicles, where onboard computational resources must be efficient, cost-effective, and energy-conscious. The reliance on hardware with excessive power consumption not only limits the scalability of such models but also contradicts the fundamental need for optimised embedded systems in autonomous driving. Therefore, improving computational efficiency is crucial to reducing inference time and energy consumption while maintaining high detection and segmentation performance. This study addresses these limitations by integrating attention mechanisms that enhance sensor fusion while simultaneously reducing the computational burden of the model.

The contributions of this work are twofold. First, it involves a comprehensive study of various attention mechanisms to perform multi-sensor fusion, aiming to enhance performance in 3D object detection and segmentation tasks. Second, it seeks to reduce the computational cost of the baseline state-of-the-art model without significant losses in these two tasks. Both objectives were successfully achieved, demonstrating the viability and impact of the proposed methods in

advancing the field of multi-sensor fusion and neural network optimization.

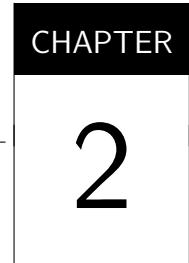
Experimental results demonstrated a significant performance improvement, showcasing the effectiveness and success of the proposed approach. The attention mechanisms explored in this work proved to be instrumental in enhancing multi-sensor fusion, highlighting their potential for improving tasks such as 3D object detection and semantic segmentation.

The structure of this thesis is organized as follows: Chapter 2 presents the theoretical background, where the main sensors used in autonomous vehicles are discussed, highlighting their utilities and limitations in different contexts. Additionally, fusion methods are presented, ranging from classical approaches to more modern ones, with taxonomic distinctions. The chapter also explores segmentation and 3D object detection, providing an introductory explanation and examining various approaches used in the field. At the end of Chapter 2, attention mechanisms are discussed.

Chapter 3 presents the related work, beginning with an explanation of BEVFusion, its functionality, and contributions. Then, other studies that employed attention mechanisms for multi-sensory fusion in autonomous vehicles, particularly for segmentation and 3D object detection tasks, are discussed.

Chapter 4 details the work carried out in this research, including the materials and resources used, as well as the evaluation methods employed.

Finally, Chapter 5 presents the final and conclusive results, providing an overview of the findings achieved during the development of this thesis.



THEORETICAL BACKGROUND

This chapter establishes the essential theoretical foundation for conducting the present research. We present fundamental concepts related to sensing in autonomous vehicles, elucidating the basic functioning of these sensors, the various types available, the data formats they generate, and their representations. We explore the individual advantages and disadvantages of each sensor and discuss how data fusion can enhance their complementarities. Special emphasis is placed on segmentation and 3D object detection, the central focus of this research, highlighting its applications in autonomous vehicles, the prominent challenges in this domain, and the rationale for directing resources and efforts toward this specific area. We conclude by discussing the attention mechanism, providing an overview and outlining the specific methods that will be explored in this thesis to perform multi-sensory fusion.

2.1 Sensors in Autonomous Vehicles

The growing demand for innovation in the automotive industry has driven the accelerated development of technologies aimed at transforming conventional driving into a safer, more efficient, and autonomous experience. The implementation of sensors in autonomous vehicles has emerged as a fundamental pillar for realizing this transformative vision. These sensory devices play a crucial role in collecting information from the surrounding environment, enabling autonomous vehicles to interpret and react intelligently to various situations.

The relevance of sensors is intrinsic to their ability to provide precise, real-time data about the vehicle's surroundings, ensuring continuous environmental perception. By equipping vehicles with advanced sensory systems, it becomes possible not only to detect obstacles but also to interpret traffic signals, identify pedestrians, and anticipate complex traffic conditions. This perception capability is essential to ensure the safety of occupants and other road users, as well as to optimize the operational efficiency of autonomous vehicles.

Among the various types of sensors used in this context, cameras, LiDAR, and RADAR stand out. Each of these devices presents distinct characteristics that complement each other to provide a comprehensive view of the environment. Cameras, for instance, capture visual information, while LiDAR uses lasers to measure distances with precision, and RADAR employs radio waves to detect objects and calculate their distances. These three types of sensors, in particular, will be discussed in subsequent subsections, allowing for a more in-depth analysis of their functionalities, advantages, and challenges inherent to their implementation in autonomous vehicles.

2.1.1 Camera

Cameras play an essential role in the visual perception capability of autonomous vehicles, as they are responsible for capturing images and videos that allow data processing systems to interpret and understand the topography of the surrounding environment. Crucial for obstacle detection, traffic sign recognition, and facilitating autonomous navigation, this sensory component plays a key role in the safety and operational effectiveness of these vehicles. In this chapter, three main types of cameras will be thoroughly addressed and explained: RGB, thermal, and stereoscopic cameras, highlighting their distinct characteristics, advantages, and challenges inherent to their implementation in autonomous vehicles.

The use of Red Green Blue (RGB) cameras in autonomous vehicles is widely spread due to their simplicity, low cost, and ability to capture color visual information. RGB cameras play a fundamental role in computer vision, providing an accurate representation of the surrounding environment. These cameras can detect nuances of color, textures, and details, allowing for precise object identification and a more sophisticated interpretation of the road scenario. Their cost-effectiveness makes them an economical choice for large-scale implementation, favoring their widespread adoption in autonomous systems.

The incorporation of thermal cameras in autonomous vehicles represents an innovative approach to environmental perception, particularly in challenging visibility conditions. Thermal cameras, sensitive to thermal radiation, offer a significant advantage in low-visibility environments, such as at night or in situations of heavy fog. The ability of these cameras to capture temperature-based images provides a complementary view to those obtained by traditional visual cameras. The usefulness of thermal cameras in detecting objects based on their heat, even in complete darkness, stands out as a key contribution to the safety and operational effectiveness of autonomous vehicles.

In the diagram presented in Figure 4, two images are shown, one captured by an RGB camera (on the left) and one captured by a thermal camera (on the right). It can be seen that, in the RGB image, the environment faces low-light conditions, making it impossible to see the background of the scene, with only the foreground of a moving pedestrian in front of the vehicle being discernible. In contrast, with the thermal camera, parked vehicles in the background can

be discerned, although without the precision of color detail. This analysis clearly highlights the disparity between the capabilities of these two cameras, emphasizing the complementarity of their use to optimize perception in autonomous vehicles.

Figure 4 – Comparison of images obtained from the RGB camera (left) and thermal camera (right) taken from the ForesightAuto page. ([DANZIGER, 2020](#)).



The application of stereoscopic cameras in autonomous vehicles is an essential approach to enhance the three-dimensional perception of the surrounding environment. These cameras operate in pairs, allowing the capture of images with disparity, which is then processed to calculate the distance of objects. This technique provides a deeper, more detailed view, significantly contributing to obstacle detection and the accurate interpretation of the environment's topography. The use of stereoscopic cameras, therefore, plays a crucial role in improving the visual perception capability of autonomous vehicles, strengthening safety and effectiveness in various driving scenarios.

However, it is crucial to consider the disadvantages associated with the use of cameras. Cameras can be sensitive to poor lighting conditions, negatively affecting their performance. Additionally, in adverse weather conditions, such as heavy rain or snow, the effectiveness of cameras may be compromised, posing a significant challenge for autonomous driving in diverse environments.

The complexity of image processing is also an important consideration. The high computational power required to interpret visual data can impact the overall performance of the autonomous system, necessitating efficient data processing solutions.

The mono RGB camera is often chosen due to its lower cost and ability to capture color visual information. Thermal cameras typically do not provide color information but are an excellent complement to RGB cameras. Stereoscopic cameras tend to be more expensive and operate at a lower frame rate and resolution, but their main advantage is providing depth information about the environment. However, by using a multi-camera system or combining

information provided by LiDAR, it is possible to achieve similar results to those obtained with stereoscopic cameras, making the mono RGB camera a more attractive option.

2.1.2 LiDAR

Providing a detailed three-dimensional view of the surroundings, LiDAR systems are essential for the environmental perception of autonomous vehicles, playing a crucial role. This technology relies on emitting laser light pulses and measuring the time it takes for them to return, allowing the creation of precise and up-to-date maps of the surrounding environment. The implementation of LiDAR is critical for obtaining accurate data about the terrain's topography and detecting objects in the surroundings, significantly contributing to the safety and operational effectiveness of autonomous vehicles.

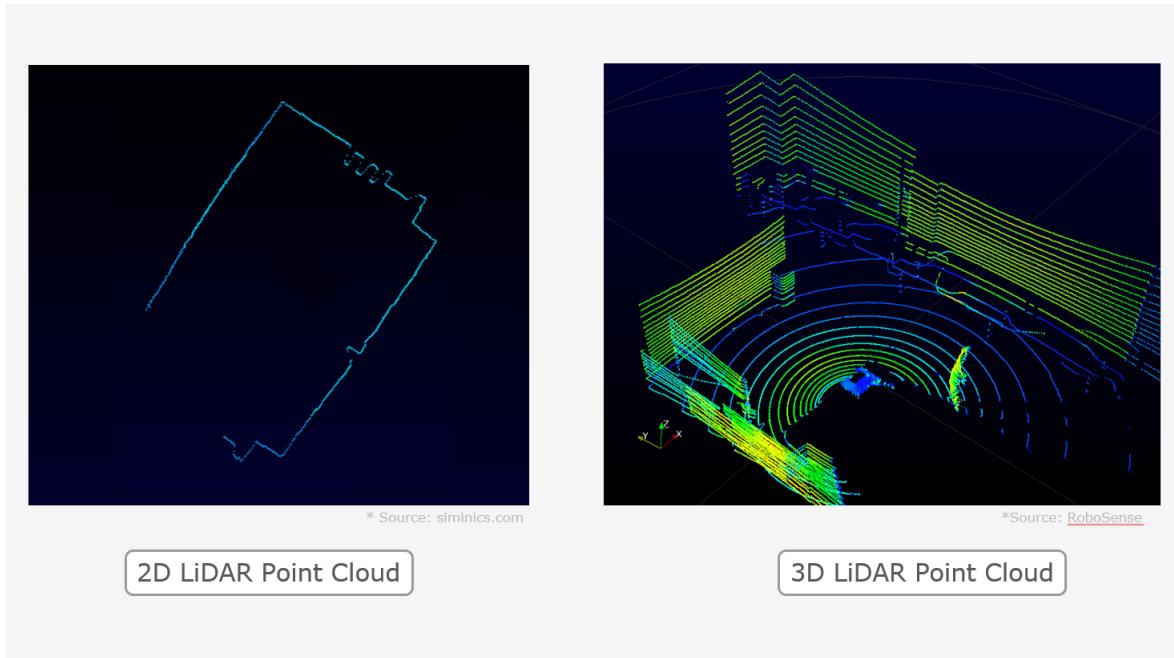
Within the LiDAR category, two main types stand out: solid-state LiDAR and mobile LiDAR. Solid-state LiDAR is characterized by the absence of moving parts in its internal components, providing robustness to the system and making it suitable for automotive environments, where durability and resistance to vibrations are essential. The viability and effectiveness of mobile LiDAR are supported by the ability of these systems to perform dynamic scans in real-time, making them particularly valuable for applications such as road and urban environment mapping.

The data format produced by LiDAR is crucial for subsequent interpretation. LiDAR systems generate point clouds that can be either two-dimensional (2D) or three-dimensional (3D), providing detailed information about the topography and position of objects in the environment. The sparsity of the data is also a significant characteristic, representing the density or dispersion of the points, serving as a valuable parameter for optimizing computational efficiency.

In Figure 5, two representations of point clouds obtained by LiDAR are presented. On the left, the point cloud resulting from a single-beam LiDAR (2D) is observed, while on the right, the corresponding representation generated by a multi-beam LiDAR (3D) is shown. The distinction between both is evident, highlighting that the 3D LiDAR provides a more detailed visualization. These additional details are particularly valuable in contexts related to autonomous driving, especially in perception tasks, where understanding the three-dimensional shape of obstacles is crucial.

Although LiDAR sensors are crucial for perception in autonomous vehicles, there are specific situations where they may encounter challenges or failures in object identification. Adverse weather conditions, such as heavy fog, rain, or snow, pose significant challenges for LiDAR sensors, as water or snow particles can scatter the laser light, impairing the accuracy of measurements. Additionally, highly reflective surfaces, such as mirrors or glass, can result in unwanted reflections, leading to incorrect interpretations of the environment. In dense urban environments, where multiple objects are close to each other, the phenomenon known as "occlusion" can occur, causing the LiDAR to fail to detect objects hidden behind closer obstacles.

Figure 5 – Comparison of 2D (left) and 3D (right) LiDAR Point Cloud Images (ROBOSENSE..., 2022).



Similarly, in cases of partial occlusion, where only parts of an object are visible, the LiDAR's ability to reconstruct the complete shape may be limited.

2.1.3 Radar

Radar sensors are essential components in object perception for autonomous vehicles, providing valuable information about the surrounding environment. These devices emit radio waves and record the time it takes for these waves to return after interacting with nearby objects. In this way, RADAR provides fundamental data on the distance, speed, and direction of objects, playing a crucial role in obstacle detection, collision prevention, and ensuring safe navigation.

The operation of RADAR is based on the principle of electromagnetic waves that propagate through space, reflecting off objects in the environment. By measuring the return time of these waves, RADAR calculates the distance to objects, enabling efficient and accurate perception.

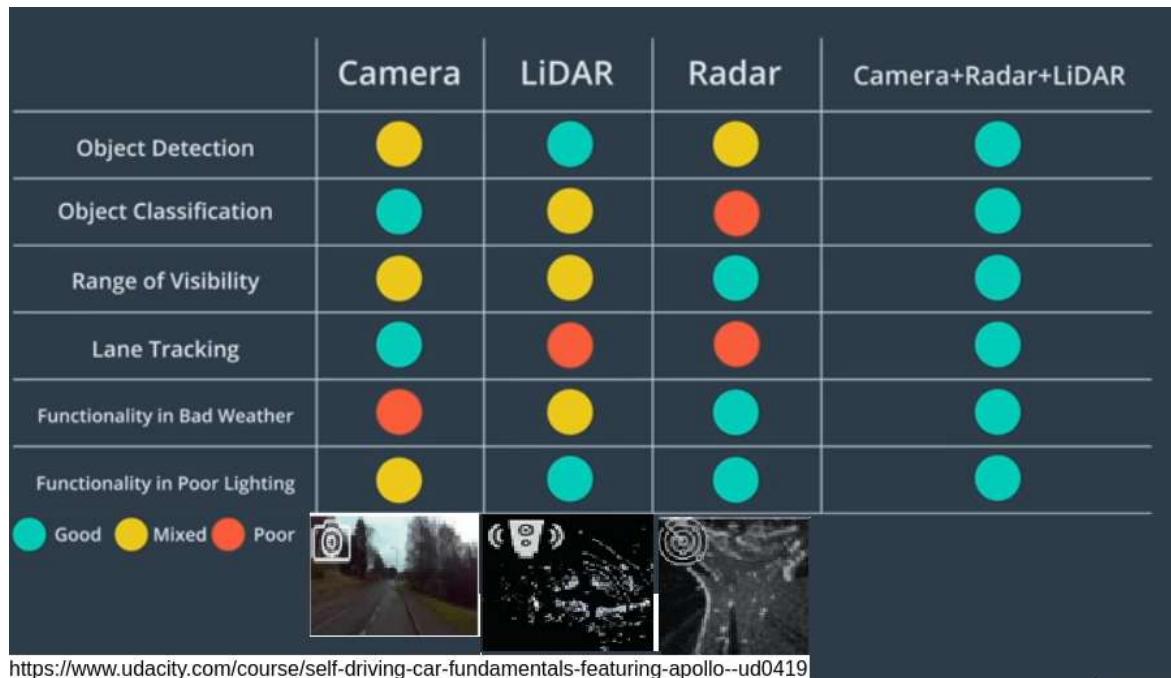
RADAR excels in various situations, being particularly effective in adverse weather conditions, such as fog, rain, or darkness, where other sensors may struggle. Additionally, its ability to penetrate visual obstacles, such as dense foliage, makes it a valuable choice for urban environments and areas with high traffic density.

However, RADAR presents limitations in high object-density situations, such as congested urban environments, where individual obstacle identification can be challenging. The accuracy of RADAR can also be affected by unwanted reflections, multiple reflections, and partial occlusions, resulting in a less reliable interpretation of the environment. Furthermore, objects with absorbent or poorly reflective surfaces may not be adequately detected.

Four formats stand out in processing this sensor, as pointed out by Srivastav and Mandal (2023): the RAD Tensor, a three-dimensional representation after the Fast Fourier Transform (FFT), provides a comprehensive view allowing direct inference of attributes; the Range-Azimuth Heatmap, a two-dimensional image obtained by compressing the Doppler dimension of the RAD Tensor; the Radar Point Cloud, a sparse three-dimensional representation valuable for detection and classification; and the Micro-Doppler Spectrogram, a two-dimensional representation that highlights distinctive movement features. By incorporating these formats into deep learning models, it is possible to richly explore RADAR information to enhance object detection and classification, contributing to the effectiveness and safety of autonomous driving (SRIVASTAV; MANDAL, 2023).

In summary, Figure 6 illustrates the overall concept of sensor quality for specific tasks under different conditions. The key conclusion is that, in each scenario, one sensor will perform well while others may not achieve the same level of accuracy. However, the combination of multiple sensors of different modalities provides both complementarity and redundancy, ensuring that sensor fusion consistently achieves excellent performance across all tasks and conditions.

Figure 6 – Main sensors used in autonomous vehicles and their performance under different conditions.
Adapted from (UDACITY, 2025).



2.2 Bird's-Eye View Representation

Visual representations play a crucial role in the perception of autonomous vehicles, providing essential information for decision-making. Among the various approaches, two commonly used representations are BEV and perspective. The perspective representation, captured from the vehicle's point of view, is effective in identifying details and understanding traffic rules. On the

other hand, BEV offers an overhead view, similar to that observed by a camera placed above the vehicle, standing out for its ability to provide a global view of the environment.

BEV stands out compared to perspective for perception tasks in autonomous vehicles due to several advantages. While perspective depicts the scene from the point of view of a camera mounted on the vehicle, BEV provides an orthographic top-down projection, eliminating the perspective distortion associated with perspective. The orthographic projection of BEV preserves the metric relationships and provides a more accurate representation of distances and sizes of objects in the scene. This is crucial for applications such as obstacle detection and trajectory planning, where accuracy in spatial information is essential.

Obtaining BEV representations from perspective is a beneficial and practically feasible strategy for vision systems in autonomous vehicles, despite the additional computational cost associated with the perspective-to-BEV transformation algorithm, which, under certain hardware configurations, may not be viable for real-time execution. The perspective representation, captured by cameras mounted on the vehicle, provides detailed information about the scene, including its height dimension. Classical methods, such as Inverse Perspective Transformation, rely on geometric strategies to generate BEV representations from multiple cameras. However, these methods often introduce distortions and are sensitive to calibration errors. With the advancement of deep learning techniques, modern approaches have emerged that aim to directly learn the transformation between perspective and BEV, training neural networks to overcome limitations related to scale variations, occlusions, and geometric inconsistencies. While these learning-based approaches improve robustness and adaptability, their computational demands must be carefully considered, especially for real-time deployment in resource-constrained environments.

In the context of LiDARs, the BEV representation is often obtained directly from the three-dimensional data, providing a detailed visualization of the spatial distribution of objects. The ability of BEV to offer a global, undistorted view, combined with its effectiveness in occlusion detection and accurate spatial representation, makes it a fundamental choice in autonomous vehicle perception systems. These features are essential for making safe and efficient decisions during navigation in complex and dynamic environments.

2.3 Multi-Sensor Fusion

As discussed in the previous sections, it is evident that each sensor has its own advantages and disadvantages, operating more effectively under specific conditions. In certain scenarios, one sensor may fail or provide a limited representation of the environment, while other sensors may offer more comprehensive and detailed readings. For example, an RGB camera is capable of capturing colour information, a characteristic that may be absent in LiDAR sensors. However, LiDAR excels in providing detailed information about distance and the three-dimensional shape of objects. This complementarity between the RGB camera and LiDAR can be leveraged to

enrich environmental perception by combining colour information with spatial details.

Considering this complementarity, it becomes clear that the use of multiple sensors, each operating uniquely, is more efficient than relying solely on one sensor. This approach allows the sensors to work complementarily, combining different information synergistically to enhance tasks such as object detection in autonomous vehicles. Additionally, the use of cameras, LiDAR, and RADAR is not limited to object detection; these sensors play critical roles in various other tasks such as navigation, obstacle recognition, and trajectory planning.

By employing a variety of sensors, redundancy in environmental information is also achieved. This redundancy is crucial to ensure that critical information is captured, even in the event of individual sensor failures or occlusions. This robust and integrated approach, which explores the distinct capabilities of each sensor, significantly contributes to the reliability and safety of autonomous vehicle systems in dynamic and challenging environments.

In the pursuit of effective multi-sensor fusion strategies, a wide range of techniques and methods are available, categorized into three distinct groups as proposed by [HU et al. \(2020\)](#). These categories include Probability-Based Methods, Classification-Based Methods, and Inference-Based Methods. It is crucial to highlight that the choice between these methods should be carefully considered, as each approach is better suited for specific tasks and is therefore not universally applicable in all sensor fusion scenarios. The following are brief explanations of each of these approaches.

2.3.1 Probability Based Methods

These methods focus on modelling the uncertainty associated with measurements, treating each sensor as a source of probabilistic information. The Extended Kalman Filter ([CANAN; AKKAYA; ERGINTAV, 2004](#)) exemplifies the application of Bayesian principles, iteratively adjusting probability distributions as new observations are acquired. Bayesian Networks, in turn, offer an effective graphical representation of probabilistic relationships, suitable for integrating complex information from various sensors.

Monte Carlo-based methods, such as Markov Chain Monte Carlo ([SYKACEK; REZEK; ROBERTS, 2000](#)) and Sequential Monte Carlo ([VEMULA; DJURIC, 2005](#)), describe probability density through weighted samples, being highly flexible in dealing with non-linear and non-Gaussian problems in state space models. Although more computationally intensive, these methods provide flexibility in representing general probability distributions, making them valuable for multi-sensor data fusion.

2.3.2 Classification Based Methods

These methods stand out in the classification of patterns and features extracted from different sensory sources, providing a deeper understanding of the data. Machine learning

algorithms, such as Support Vector Machines (BANERJEE; DAS, 2012) and Neural Networks, play fundamental roles in this context. The fusion of results from these algorithms offers a more comprehensive and robust representation of the environment. Voting approaches and ensemble methods, such as Random Forests (LOUPPE, 2015), are common for combining results from diverse classifiers, providing a more resilient approach to noise and variations in the input data.

2.3.3 Inference Based Methods

The multi-sensory fusion based on inference stands out for its use of statistical and logical inference to combine heterogeneous information. The evidence theory, also known as Dempster-Shafer theory (HAMDA; HADJALI; LAGHA, 2023), is crucial in this context, allowing for handling the uncertainty and imprecision inherent in sensory data. Furthermore, fuzzy logic provides a flexible way to deal with uncertainty, allowing the representation of vague concepts. Data fusion through fuzzy logic (ROTH; SCHILLING, 1995) is especially useful when the boundaries between classes are fuzzy, enabling a smoother integration of sensory information. Inferential methods, such as Bayesian networks (SMAILI; NAJJAR; CHARPILLET, 2007), are employed to model the relationship between different variables and make inferences about the state of the environment.

Another perspective for classifying the methods lies in the point at which the fusion is executed, as pointed out by Alaba, Gurbuz and Ball (2022) and Fayyad *et al.* (2020). Thus, there are methods that perform fusion at an early stage, combining the information without prior processing into a single representation to be analysed. At an intermediate stage, each piece of information from the sensors is pre-processed individually before being integrated. Finally, at a more advanced stage, all processing and decision-making occur after fusion in an integrated manner. Figure 7 visually illustrates these methods. Although this thesis focuses on feature fusion methods, brief explanations of each of these approaches will be presented next.

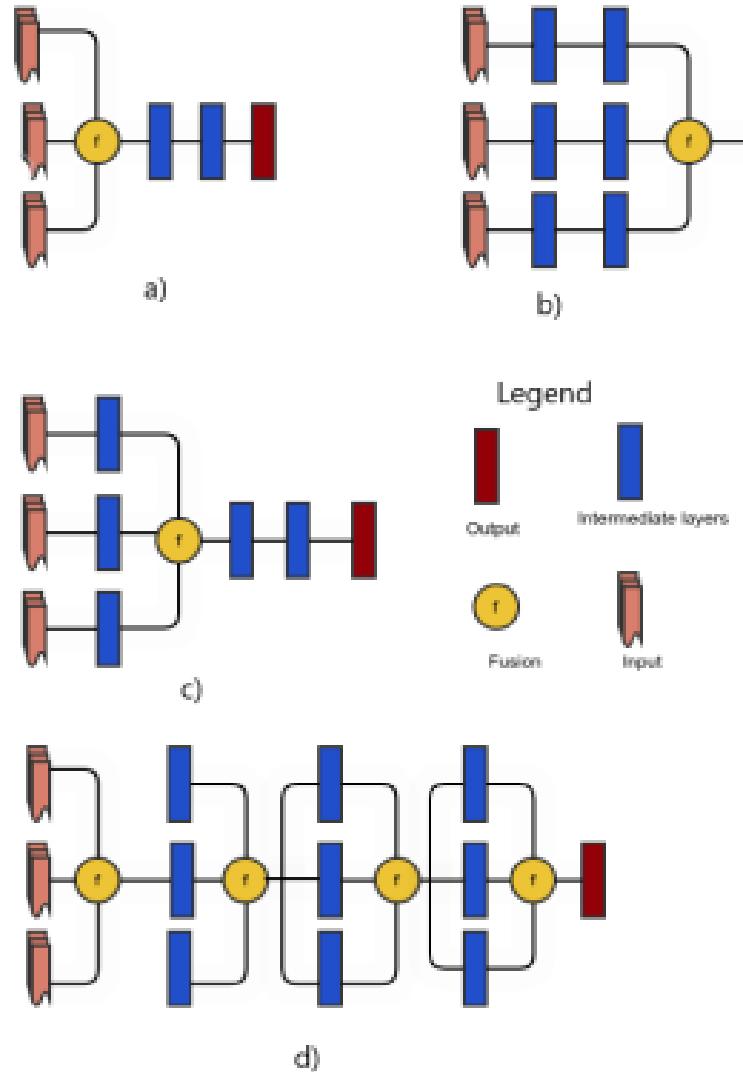
2.3.4 Data Fusion Methods

Data fusion is characterized by the grouping and direct combination of raw data, representing different sensors, at the initial stage of processing. This approach enables the creation of a unified and integrated representation, where each type of information is considered simultaneously, providing a richer and more detailed view of the scenario.

However, data fusion also presents challenges. Differences in sampling rates, alignment, and data representation between different sensors can complicate the fusion process. Synchronization and the conversion of data into a synchronized feature vector are critical aspects in dealing with these discrepancies.

Common methods of data fusion include concatenation and summation. Concatenation is often used to increase the width of feature maps, providing a variety of inputs and details. On

Figure 7 – Illustration of fusion methods categorized by stage: a) Data Fusion, b) Decision Fusion, and c) Feature Fusion ([ALABA; GURBUZ; BALL, 2022](#)).



the other hand, summation, based on the identity mapping technique, can reduce computational load but may result in information loss during the reduction of feature map width.

2.3.5 Feature Fuse Methods

Unlike data fusion, which combines raw data from the initial stage of processing, feature fusion occurs at an intermediate stage, after distinct features have been extracted from each source. This allows for more refined and specialized manipulation of the features, as each data set undergoes specific processes before fusion.

Several techniques are available to perform feature fusion. Concatenation is a common strategy, where features extracted from different sources are simply concatenated into a single vector. This method increases the dimensionality of the feature space, allowing the system to

capture a broader range of information.

Another approach is feature weighting, where weights are assigned to each set of features before fusion. These weights can be learned during model training, allowing for dynamic adaptation to the relative importance of each source.

2.3.6 Decision Fusion Methods

Unlike data fusion and feature fusion, which occur at the early or intermediate stages, decision fusion focuses on the combination of final results or decisions. This approach is particularly useful in scenarios where different information sources may generate partial decisions or predictions, and integrating these decisions can lead to a more reliable and comprehensive conclusion.

There are several techniques for performing decision fusion. A common approach is voting, where individual decisions are weighted according to their confidence or associated probability. This can be implemented through majority voting, weighted voting, or other decision aggregation methods.

Another technique is rule-based fusion, where specific decisions are combined based on predefined criteria. This may involve applying fusion rules, such as selecting the most frequent decision or considering the most confident decision.

2.4 Segmentation

Segmentation is a fundamental task in computer vision that involves partitioning an image, point cloud, or 3D data into distinct regions or objects based on visual or semantic characteristics, which can be understood as a pixel-level classification (LIU *et al.*, 2023). The goal is to assign a label to each pixel (in 2D data) or point/voxel (in 3D data) to understand and interpret scenes at a granular level. This fine-grained classification enables a more detailed and structured representation of the environment, facilitating downstream tasks such as object detection and scene understanding. Segmentation serves as a critical step in numerous applications, including autonomous vehicles (GAUTAM; MATHURIA; MEENA, 2022b), robotics (HURTADO; VALADA, 2024), and medical imaging (WANG *et al.*, 2022).

In contrast to object detection, which focuses on identifying and localizing objects within bounding boxes, segmentation provides a more refined understanding by delineating the precise boundaries of objects and regions. This capability is especially valuable in scenarios requiring detailed spatial awareness, such as obstacle avoidance in autonomous vehicles or tumour identification in medical imaging.

Segmentation can be categorized into three main types: semantic segmentation, instance segmentation, and panoptic segmentation. Each type serves a specific purpose, depending on

the requirements of the application, and addresses unique challenges in understanding scenes or objects.

Semantic segmentation focuses on assigning a class label to every pixel or point in an image or 3D space ([SEVAK *et al.*, 2017](#)). The objective is to group all elements belonging to the same class, such as roads, sidewalks, or buildings, into a unified representation. For example, in an urban scene, a semantic segmentation model would label all roads with one colour, all sidewalks with another, and buildings with a third. While semantic segmentation provides a high-level understanding of the environment, it does not differentiate between individual instances of the same class. For instance, two cars in a scene would be grouped under the same label, making it impossible to distinguish between them. This type of segmentation is particularly valuable in applications like autonomous driving, where the identification of general areas such as drivable regions is critical.

Instance segmentation builds upon semantic segmentation by distinguishing between different instances of the same class ([HAFIZ; BHAT, 2020](#)). While semantic segmentation groups all elements of a class together, instance segmentation assigns unique identifiers to each object within a class. For example, in an urban scene, instance segmentation would not only identify cars but also distinguish between individual cars in the scene, assigning each one a distinct label. This capability is crucial for tasks like tracking multiple objects or analysing crowded environments, where differentiating between objects of the same class is necessary. Instance segmentation is widely used in autonomous driving, robotics, and surveillance applications, where precise object-level understanding is essential.

Panoptic segmentation combines the strengths of both semantic and instance segmentation, providing a comprehensive understanding of a scene by segmenting all pixels into either a “thing” or a “stuff” category ([ELHARROUSS *et al.*, 2021](#)). “Things” are countable objects, such as cars, pedestrians, or bicycles, while “stuff” refers to amorphous regions, such as roads, sky, or vegetation. Panoptic segmentation not only labels individual instances of objects but also segments uncountable regions, offering a unified framework for obtaining a complete scene understanding. For example, in an urban scene, panoptic segmentation would identify individual cars and pedestrians while also labelling the surrounding roads and sidewalks. This holistic approach is particularly advantageous in tasks like autonomous driving and environmental mapping, where both object-level detail and scene context are crucial for decision-making.

Each type of segmentation plays a unique role in interpreting scenes and objects, addressing specific needs across a wide range of applications. The three types of segmentation can be visualized in Figure 8, which shows the original image alongside the results of semantic, instance, and panoptic segmentation.

Evaluating segmentation models requires robust and reliable metrics that measure their performance in capturing and labelling various regions or objects in an image or 3D scene. Among the most commonly used metrics in segmentation tasks are the mean Intersection over

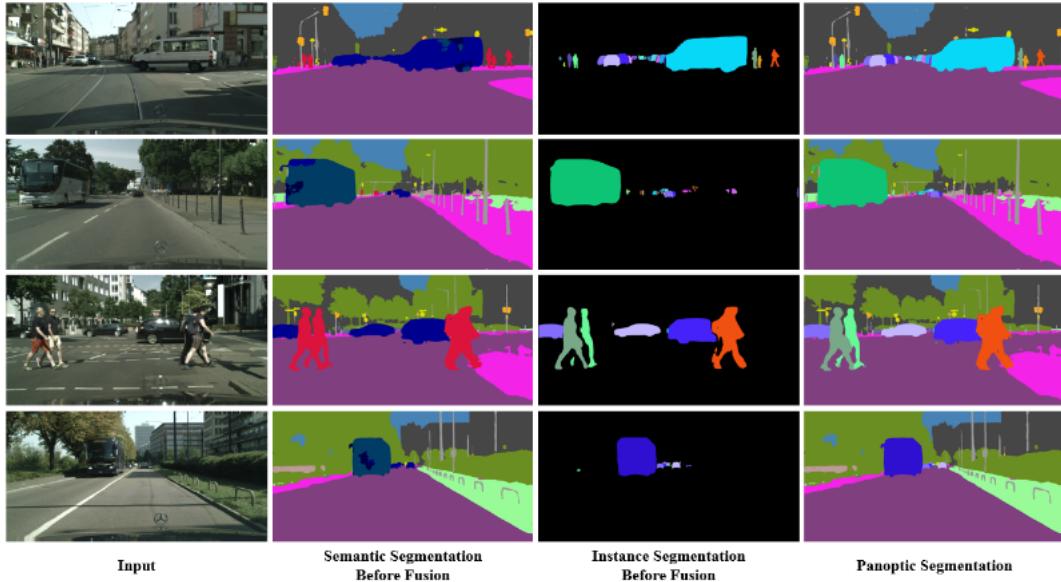


Figure 8 – Examples of the three types of segmentation. Image adapted from ([COSTEA; PETROVAI; NEDEVSCHI, 2018](#)).

Union (mIoU) and the Dice coefficient, which provide detailed insights into the accuracy and completeness of segmentation outputs. These metrics are widely used due to their ability to quantify the overlap and consistency between predicted and ground truth labels, making them essential for comparing models and optimizing performance.

The mIoU is one of the most fundamental metrics for segmentation evaluation. It quantifies the overlap between the predicted segmentation and the ground truth, offering a clear measurement of accuracy. For a single class, the IoU is calculated as:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|P \cap G|}{|P \cup G|}$$

where P represents the set of pixels or points predicted as belonging to a class, and G represents the corresponding ground truth set. The IoU ranges from 0 to 1, with higher values indicating better performance. A perfect IoU of 1 signifies complete alignment between the prediction and the ground truth, while lower values indicate discrepancies ([MINAEE *et al.*, 2020](#)).

The mIoU extends this calculation across all classes in the dataset and averages the IoU values, providing a single score that summarizes the model's performance. The formula for mIoU is given as:

$$mIoU = \frac{1}{N} \sum_{i=1}^N IoU_i$$

where N is the number of classes, and IoU_i is the IoU for class i . The mIoU is particularly valuable in segmentation tasks involving multiple object categories, as it accounts for the performance across all classes, ensuring that the model's effectiveness is evaluated comprehensively. This metric is favoured for its interpretability and ability to reflect both false positives and false

negatives, which are critical in applications like autonomous driving and medical imaging ([HE et al., 2024](#)).

The Dice coefficient, also known as the Sørensen-Dice index, is another widely used metric in segmentation. It measures the similarity between the predicted and ground truth regions and is defined as:

$$Dice = \frac{2 \cdot |P \cap G|}{|P| + |G|}$$

where $|P|$ is the size of the predicted region, $|G|$ is the size of the ground truth region, and $|P \cap G|$ is the size of their overlap. The Dice coefficient also ranges from 0 to 1, with higher values indicating better segmentation accuracy. A Dice score of 1 represents perfect overlap between the prediction and the ground truth.

The Dice coefficient emphasizes the overlap between the predicted and ground truth regions while being less sensitive to class imbalance compared to IoU. This property makes it particularly useful in medical imaging and other applications where certain classes may dominate the dataset. Additionally, the Dice coefficient is symmetric, meaning that interchanging the predicted and ground truth labels does not affect the score, ensuring consistency in evaluation.

Both mIoU and the Dice coefficient are widely used due to their complementary strengths in evaluating segmentation models. The mIoU is effective at penalizing both false positives and false negatives, making it suitable for applications requiring a balanced assessment of accuracy. On the other hand, the Dice coefficient is particularly advantageous in scenarios with class imbalance, as it provides a more forgiving evaluation of minority classes. This makes it invaluable in fields like medical imaging, where certain regions of interest may be significantly smaller than others.

Moreover, these metrics are interpretable and easily computable, making them ideal for comparing models and tracking improvements during training. By considering the overlap and consistency between predicted and ground truth regions, mIoU and Dice provide reliable indicators of a model's ability to perform precise and accurate segmentation. These metrics have become standard benchmarks in the segmentation literature and continue to guide advancements in the field.

According to Minaee et al. ([MINAEE et al., 2020](#)), the models used for segmentation can be categorized into distinct types based on their underlying architectural principles. This classification highlights the core contributions of each model architecture, such as the use of encoder-decoder frameworks, multi-scale analysis, skip connections, and advanced techniques like dilated convolutions. While segmentation models can also be grouped by their specific goals—such as semantic, instance, or panoptic segmentation—the architectural grouping offers a more consistent perspective given the diversity and volume of work in these areas. The following subsections provide a detailed explanation of each type of model used for segmentation, emphasizing their unique architectural characteristics and advancements.

2.4.1 Fully Convolutional Networks

Fully Convolutional Networks (FCNs) are foundational models in semantic segmentation that replace the fully connected layers in traditional Convolutional Neural Networks (CNN) with fully convolutional layers. This design allows FCNs to process images of arbitrary sizes and produce spatially aligned segmentation maps of the same resolution as the input. By integrating skip connections, FCNs combine fine-grained low-level features with high-level contextual features, improving accuracy in segmentation tasks. Notable works such as those by [Long, Shelhamer and Darrell \(2015\)](#) and [Liu, Rabinovich and Berg \(2015\)](#) demonstrated the applicability of FCNs across various domains, laying the groundwork for subsequent advancements in segmentation.

2.4.2 Convolutional Models With Graphical Models

Convolutional models combined with graphical models, such as Conditional Random Fields, aim to refine segmentation outputs by imposing spatial consistency. While convolutional layers excel at extracting local features, graphical models enforce global coherence by accounting for the relationships between neighbouring pixels or regions. This synergy enhances segmentation accuracy, particularly in scenarios with complex boundaries. Works like [Zheng et al. \(2015\)](#), [Chen et al. \(2016\)](#), [Lin et al. \(2016\)](#) and [Liu et al. \(2015\)](#) demonstrated the effectiveness of integrating CRFs into segmentation pipelines, setting a precedent for hybrid approaches.

2.4.3 Encoder-Decoder Based Models

Encoder-decoder architectures are widely used in segmentation due to their ability to capture both global context and fine-grained details. The encoder compresses the input into a compact representation, while the decoder reconstructs the segmentation map at the original resolution. Models like U-Net ([RONNEBERGER; FISCHER; BROX, 2015](#)) and its evolutions such as works by [Çiçek et al. \(2016\)](#), [Zhou et al. \(2018\)](#) and [Zhang, Liu and Wang \(2018\)](#) leverage symmetric skip connections, allowing the decoder to utilize spatial details from the encoder. These architectures have been particularly successful in medical imaging and satellite imagery, demonstrating robustness and flexibility across applications.

2.4.4 Multi-Scale and Pyramid Network Based Models

Multi-scale and pyramid network models are designed to capture features at various resolutions, enabling them to handle objects of different sizes effectively. Techniques such as the Pyramid Scene Parsing Network (PSPNet) ([ZHAO et al., 2017](#)) and Feature Pyramid Networks (FPN) ([LIN et al., 2017](#)) employ hierarchical pooling and multi-scale feature fusion to aggregate global and local context. These methods excel in urban scene segmentation, where objects like pedestrians and vehicles vary significantly in scale.

2.4.5 R-CNN Based Models

R-CNN-based models adapt the region proposal techniques used in object detection for segmentation tasks. By generating region proposals and applying segmentation algorithms within each region, these models achieve high accuracy, especially for instance segmentation. Mask R-CNN ([HE et al., 2018](#)) extends Faster R-CNN ([REN et al., 2016](#)) by adding a segmentation branch, making it a popular choice for tasks requiring precise instance-level delineation, such as autonomous driving and robotics.

2.4.6 Dilated Convolutional Models and DeepLab Family

Dilated convolutions, also known as atrous convolutions, allow networks to expand their receptive fields without increasing the number of parameters. This capability is central to the DeepLab family ([CHEN et al., 2017](#)) of models, which integrates dilated convolutions with pyramid pooling modules to capture both local and global context. DeepLabv3+ ([CHEN et al., 2018](#)) is a state-of-the-art model that combines encoder-decoder structures with dilated convolutions, achieving exceptional performance in semantic segmentation benchmarks.

2.4.7 Recurrent Neural Network Based Models

Recurrent Neural Networks (RNNs) are incorporated into segmentation pipelines to model sequential and spatial dependencies in the data. Variants such as Long Short-Term Memory (LSTM) networks are particularly effective in capturing temporal and spatial relationships, making them suitable for video segmentation. Works like those of [Visin et al. \(2016\)](#) and [Byeon et al. \(2015\)](#), [Liang et al. \(2016\)](#) demonstrate the potential of RNNs in improving segmentation consistency over time and space.

2.4.8 Generative Models and Adversarial Training

Generative models, including Generative Adversarial Networks (GANs), are employed in segmentation to improve the quality and realism of predictions. Adversarial training involves a generator producing segmentation maps and a discriminator evaluating their realism. This approach encourages the generator to produce more accurate outputs, as demonstrated by works such as [Luc et al. \(2016\)](#). GAN-based methods are particularly effective in applications like medical imaging, where data is limited, and realistic augmentations are essential.

According to [He et al. \(2024\)](#), 3D segmentation can be addressed using various approaches depending on the type of data representation and the specific requirements of the application. Unlike 2D segmentation, 3D segmentation involves understanding and labelling volumetric data or point clouds, providing a richer spatial and geometric context for applications such as autonomous navigation, medical imaging, and robotics. By analysing the three-dimensional structure of scenes, 3D segmentation enables a more precise identification and differentiation

of objects or regions in a given environment. The methods for 3D segmentation are often categorized based on how the data is represented and processed, such as using RGB-D images, projected 2D images, voxel grids, or raw point clouds. The following subsections provide a detailed exploration of each of these approaches, highlighting their characteristics, challenges, and notable contributions in the field.

2.4.9 *RGB-D Based Segmentation*

RGB-D based segmentation leverages RGB images combined with depth information to enhance segmentation tasks. Depth data provides additional geometric context that helps differentiate objects in a scene. The primary advantage of RGB-D segmentation lies in its ability to disambiguate objects that may appear similar in RGB space but differ in depth. This method has been widely applied in indoor scenes, where depth cameras like Kinect are commonly used. Several works have explored RGB-D segmentation, including [Gupta et al. \(2014\)](#), which introduced a multi-modal network combining RGB and depth features. Other notable approaches include works by [Qi et al. \(2017c\)](#) and [Hazirbas et al. \(2017\)](#), which proposed novel fusion strategies to integrate RGB and depth data effectively. These methods have been pivotal in advancing segmentation accuracy in complex environments.

2.4.10 *Projected Images Based Segmentation*

Projected images based segmentation focuses on transforming 3D point clouds or voxel grids into 2D images through projections. This approach enables the use of established 2D segmentation networks on 3D data, significantly reducing computational complexity. Common projection techniques include spherical, cylindrical, or BEV projections. Works like [Xu et al. \(2021\)](#) demonstrated the effectiveness of this method by projecting 3D data onto 2D planes for efficient processing. Similarly, [Guerry et al. \(2017\)](#) extended this idea by integrating multi-view projections to improve segmentation accuracy. These methods offer a practical trade-off between efficiency and performance, making them suitable for real-time applications such as autonomous driving.

2.4.11 *Voxel-Based Segmentation*

Voxel-based segmentation represents 3D data as a structured grid of voxels, enabling the application of CNNs designed for 3D volumetric data. This representation facilitates spatially consistent feature extraction but introduces challenges such as high memory consumption and computational overhead for large-scale datasets. Notable works in voxel-based segmentation include VoxNet by [Raj, Maturana and Scherer \(2015\)](#), which employs a 3D CNN for segmentation, and the work by [Graham, Engelcke and Maaten \(2017\)](#), which introduced sparse convolutions to address the memory efficiency issue. These methods are particularly effective in scenarios requiring detailed 3D representations, such as medical imaging and architectural modelling.

2.4.12 Point-Based Segmentation

Point-based segmentation operates directly on raw 3D point clouds without converting the data into intermediate representations like voxels or projections. This approach preserves the fine-grained details of the original 3D data, making it ideal for tasks that require high spatial resolution. PointNet by Qi *et al.* (2017a) is a seminal work in this domain, introducing a neural network architecture that processes unordered point clouds. Subsequent improvements, such as PointNet++ (QI *et al.*, 2017b), added hierarchical feature extraction to capture local structures. Other works, like those by Thomas *et al.* (2024), proposed advanced point-based networks incorporating attention mechanisms to further improve segmentation performance. These methods are widely used in autonomous navigation, robotics, and object recognition tasks.

2.5 3D Object Detection

3D object detection is a crucial area in the field of computer vision, playing an essential role in various applications such as autonomous vehicles, augmented reality, and industrial automation. Unlike 2D object detection, which focuses solely on location and classification in a single image, 3D object detection aims to understand the three-dimensional position of objects in space. This complex process involves not only identifying the presence of objects but also estimating their 3D coordinates and orientations.

There are various approaches to perform 3D object detection (ARNOLD *et al.*, 2019), and the choice of method often depends on the type of sensor available and the specifics of the application. A common approach is the use of individual sensors, such as LiDAR, RADAR, stereo cameras, or camera arrays. Each type of sensor provides unique information, and algorithms are tailored to extract specific features from each modality (QIAN; LAI; LI, 2022).

LiDAR sensors are frequently employed in 3D object detection due to their ability to measure distances with high precision. By emitting laser pulses and measuring the time it takes for them to return, LiDARs provide information about the scene's geometry. LiDAR-specific detection algorithms process 3D point clouds to identify objects and estimate their geometric properties.

RADAR, in turn, is effective in adverse conditions, such as fog or low visibility. Its ability to measure the reflectance of objects and detect motion makes it a valuable choice, especially in autonomous driving scenarios.

Stereo cameras, by capturing two simultaneous images of the same scene from slightly different angles, allow for depth information to be obtained. The disparity between the images is used to calculate the distances to objects in the three-dimensional scene. Algorithms specific to stereo cameras, often based on visual feature matching, are applied to perform 3D detection.

The use of camera arrays is also common, particularly in omnidirectional vision appli-

cations. By positioning cameras in different directions, it is possible to capture comprehensive information about the environment, enabling more complete object detection.

In addition to these approaches with individual sensors, 3D object detection can also benefit from data fusion from multiple sensors (MAO *et al.*, 2023). Combining LiDAR, cameras, and RADAR, for example, allows for the creation of more robust systems that are resilient to different environmental conditions. Data fusion can be achieved through techniques such as multimodal sensor fusion and extended Kalman filtering algorithms, providing a more comprehensive and reliable view of the three-dimensional environment. By integrating complementary information from different sensors, 3D object detection systems become better equipped to handle various challenges, contributing to safety and effectiveness in a range of applications, such as autonomous vehicles and industrial monitoring.

The 3D object detection process involves two main tasks: bounding box regression and classification. These tasks are essential for locating and identifying objects in the three-dimensional scene (QIAN; LAI; LI, 2022).

Bounding box regression is responsible for determining the precise location of the object in the scene, defining a bounding box that encompasses the object. This is typically expressed in three-dimensional coordinates, such as height, width, and depth. Specific algorithms, such as those used in CNN or machine learning algorithms, are trained to learn patterns in the data and perform this regression. In neural networks, dedicated regression layers at the end of the architecture are often used to estimate the bounding box coordinates.

Classification, on the other hand, refers to assigning a category or label to the detected object. Each object belongs to a specific class, such as car, pedestrian, or bicycle, and is associated with a unique identifier. Again, neural networks play a crucial role in this task, learning to distinguish visual and geometric features that characterize each object class. Classification layers typically precede or follow the regression layers in the network architecture.

The combination of these two tasks allows the detection system to locate and identify objects in a 3D environment. During training, algorithms are fed with labelled datasets, where the bounding box coordinates and object categories are provided. The algorithm adjusts its internal parameters to minimize the difference between predictions and the labels provided in the training data.

Accurate evaluation of 3D object detection algorithms is essential to measure the performance and reliability of these models. Two key metrics in this context are IoU and the Average Precision (AP) or mean Average Precision (mAP) (QIAN; LAI; LI, 2022).

IoU is a crucial metric that quantifies the overlap between the bounding box predicted by the model and the ground truth bounding box of the object. Expressed as the ratio between the intersection area and the union area, IoU is calculated by the formula $IoU = \frac{\text{Intersection Area}}{\text{Union Area}}$. An IoU of 1 indicates complete overlap, representing a perfect detection, while lower values

indicate less overlap.

AP is a widely used metric for evaluating performance in object detection tasks. It measures the detection precision at different confidence levels and is calculated from the precision-recall curve. AP is the average of precisions at all recall points and is especially relevant when detections have varying confidence scores associated with them.

mAP is an extension of AP that considers the variation in accuracy rates across different object categories. This is crucial in scenarios where the model needs to handle multiple object classes. mAP provides a more comprehensive view of the model's performance by aggregating the AP for each class and presenting a single score that summarizes the overall quality of the detections.

These metrics play a critical role in the objective evaluation of 3D object detection models. IoU highlights spatial accuracy, while AP and mAP provide insights into the model's ability to handle varying confidence levels and object categories. By considering these metrics together, researchers and professionals can gain a comprehensive understanding of the model's performance in challenging detection tasks in 3D environments.

According to the review by [Arnold et al. \(2019\)](#), 3D detection methods are categorized into three main groups: Image-Based Methods, Point Cloud-Based Methods, and Fusion Methods. Each category addresses specific challenges related to obtaining accurate information about the location and geometry of objects in the environment. In the following subsections, we will explore each category in detail, presenting relevant approaches that play a significant role in advancing these methods and in building robust and precise detection systems.

2.5.1 Image based methods

Image-based 3D detection methods focus on the precise estimation of 3D bounding boxes using information from monocular cameras. These methods face the challenge of extracting depth information from 2D images, as the z-dimension (depth) is not directly available. Here, we explore some notable approaches within this category.

The Mono3D method, proposed by [Chen et al. \(2016\)](#), stands out for its simple approach to proposing regions. Using context, semantics, projected shape features, and location priorities, the algorithm generates proposals through an exhaustive search in 3D space, which are then filtered by Non-Maximum Suppression. The next step involves using a Fast R-CNN model trained with 2D images to score and regress the 3D bounding boxes. Surprisingly, even without direct depth information, Mono3D outperforms previous methods.

The 3D Voxel Pattern (3DVP) approach, introduced by [Xiang et al. \(2015\)](#), innovates by incorporating visibility patterns into the model. This unique representation models appearance using RGB intensities, 3D shape as a set of voxels, and occlusion masks. The 3DVP patterns are obtained through clustering, and classifiers are trained for each pattern, providing a robust way

to distinguish visible, hidden, or truncated parts of objects. This technique proves effective in retrieving relevant information for 3D detection.

The Deep MANTA method ([CHABOT et al., 2017](#)) extends the capabilities of previous methods by adopting a many-task approach. This model uses a region-proposal network for 2D bounding box regression and localization, followed by 3D shape inference through matching with 3D models. By addressing multiple tasks simultaneously, such as vehicle position, part localization, and shape, Deep MANTA offers a comprehensive solution for 3D object detection. This approach stands out for its ability to consider multiple aspects simultaneously, providing a more complete view of the environment surrounding the autonomous vehicle.

2.5.2 Cloud Point Based Methods

3D detection methods based on point clouds can be categorized into three main subgroups: projection methods, volumetric methods, and point network-based methods.

Some methods based on Projection Based Methods, such as those proposed by [Li, Zhang and Xia \(2016\)](#), use cylindrical or spherical projections to transform 3D points into a 2D image. CNN can then be applied to these projections for object detection, followed by the regression of the 3D bounding box's dimensions and location. These approaches leverage the familiarity of object detection techniques in 2D images. Other methods, such as those presented in ([SIMON et al., 2018](#)), ([YU et al., 2017](#)), ([BELTRAN et al., 2018](#)), adopt the BEV projection to generate 3D proposals. The representation includes information about height, intensity, and point density. Networks like Faster R-CNN ([REN et al., 2016](#)) can be used for proposal generation and 3D bounding box regression, although some methods, such as Complex-YOLO ([SIMON et al., 2018](#)), adopt a single-stage detection approach for greater computational efficiency.

Some methods based on Volumetric Convolution Based Methods, such as those proposed by ([LI, 2017](#)), ([ENGELCKE et al., 2017](#)), adopt a volumetric approach, considering a 3D grid (voxel) representation of the scene. Fully CNN are applied directly to the volumetric representation. Although this approach explicitly encodes shape information, it is limited by efficiency, as most of the volume consists of empty cells, which drastically increases the computational cost.

In contrast to projection methods or volumetric representations, point-net-based methods, such as PointNet ([QI et al., 2017a](#)), treat the point cloud as direct input. These approaches aim to reduce information loss caused by projections or quantizations in 3D space. PointNet uses fully connected layers to perform point-wise transformations, aggregating global features through a max-pooling layer. Extensions like PointNet++ ([QI et al., 2017b](#)) enhance this architecture by incorporating hierarchical structures for progressively encoding complex features.

2.5.3 Fusion-Based Method

Given the individual limitations of modalities, fusion methods aim to combine information from images and point clouds. As discussed in the Multi-Sensor Fusion section, the fusion process can be implemented at the beginning, end, or deeply within the neural network architecture. Examples include MV3D ([CHEN et al., 2017](#)), which uses a deep fusion approach, and AVOD ([KU et al., 2018](#)), which proposes fusion at the beginning of the process. Information fusion provides an advantage by leveraging complementary data to improve 3D detection performance.

2.6 Attention Mechanisms

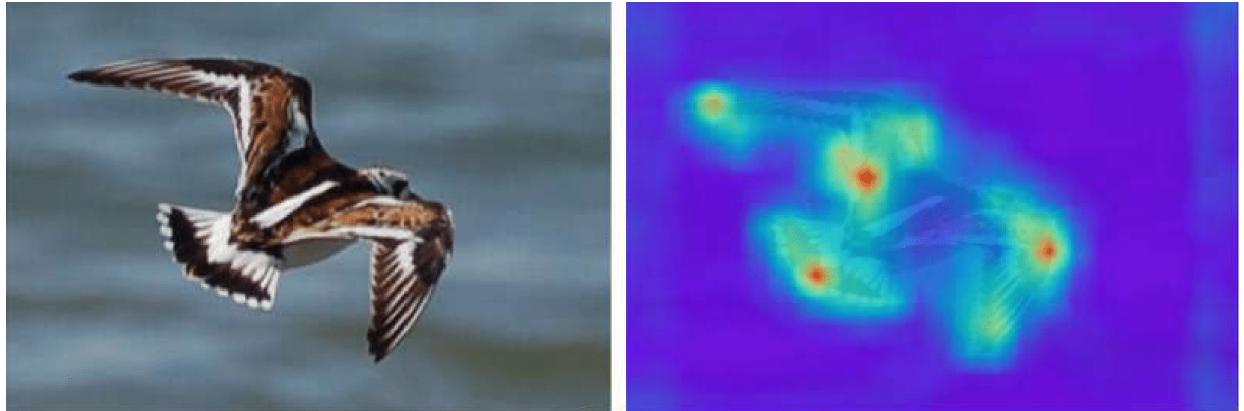
The rapid advancements in deep neural networks have driven the need for more sophisticated mechanisms capable of processing and interpreting complex data effectively. In this context, attention mechanisms have emerged as a cornerstone of modern neural architectures, enabling models to selectively focus on the most relevant parts of the input while ignoring less important details. This capability is particularly valuable for tasks that involve high-dimensional or complex data, such as natural language processing, computer vision, and sensor fusion.

Attention mechanisms operate under the premise that not all parts of an input contribute equally to solving a task. Inspired by the human visual system, these mechanisms dynamically assign weights to input elements, allowing the model to “attend” to regions or features that are most critical for the task at hand. This selective processing is especially beneficial when the input is large, such as high-resolution images or dense 3D point clouds.

An example of how the attention mechanism works can be seen in Figure 9. In the original image, a bird is present against a non-monochromatic background. If the goal is to identify objects such as animals, the attention mechanism aims to create a representation of the same dimensions as the original image, highlighting the regions that are most relevant to the task. The resulting image on the right illustrates this process, where warmer colours indicate areas of higher importance—primarily outlining the bird—while cooler colours represent less relevant regions, such as the background, which can be ignored.

The development of attention mechanisms marked a significant milestone in deep learning, particularly in addressing the challenges posed by long-range dependencies in data. The concept was first introduced by [Bahdanau, Cho and Bengio \(2016\)](#) in the context of machine translation. In traditional sequence-to-sequence (seq2seq) models, information from the input sequence was compressed into a single fixed-length vector, which often limited the model’s ability to handle long or complex sequences. [Bahdanau, Cho and Bengio \(2016\)](#) proposed an attention mechanism that allowed the decoder to dynamically focus on specific parts of the input sequence at each step, improving alignment and overall translation performance.

Figure 9 – Example of an image recalibrated using attention mechanisms (right) compared to the original (left). Adapted from ([WU et al., 2023](#)).



Building upon this foundational work, [Vaswani et al. \(2017\)](#). introduced the Transformer architecture, which popularized the concept of self-attention . Unlike RNNs or LSTMs, which process data sequentially, self-attention allows each element in the input sequence to attend to all other elements simultaneously. This innovation not only improved the model’s ability to capture long-range dependencies but also enabled efficient parallel processing, significantly reducing training time and computational overhead. The Transformer’s success has influenced advancements across multiple domains, including natural language processing (e.g., BERT ([DEVLIN et al., 2018](#))) and computer vision (e.g., Vision Transformers ([DOSOVITSKIY et al., 2021](#))).

Attention mechanisms can be categorized based on their operational behaviour and range of focus. Self-attention allows input elements to interact with one another and identify relationships within the same dataset, which is crucial for understanding contextual dependencies. It forms the backbone of Transformer-based architectures and has been successfully adapted to tasks such as spatial feature extraction in computer vision and sequence modeling in NLP ([VASWANI et al., 2017](#)). In contrast, cross-attention enables one set of inputs to attend to another, making it particularly effective for tasks involving multi-modal or multi-sensor data fusion. By aligning features from different modalities, such as LiDAR point clouds and camera images, cross-attention enhances the integration of complementary information, leading to more robust performance.

Attention mechanisms can also be divided into global attention ([LIU; SHAO; HOFFMANN, 2021](#)) and local attention ([AGUILERA-MARTOS et al., 2024](#)). Global attention considers all elements of the input data simultaneously, which is ideal for tasks requiring contextual awareness, such as machine translation and global feature analysis in images. However, this approach comes with high computational costs, especially for large-scale inputs. In contrast, local attention restricts the focus to a specific region of the input, reducing computational complexity while maintaining accuracy for tasks where spatial locality is essential, such as object detection and image segmentation.

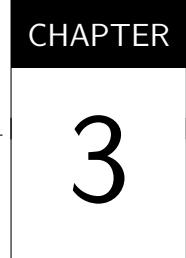
Another distinction lies between hard attention ([MNIH *et al.*, 2014](#)) and soft attention ([HERMANN *et al.*, 2015](#)). Hard attention selects specific regions or elements of the input while discarding the rest. This discrete approach is computationally efficient and interpretable but non-differentiable, often requiring reinforcement learning for optimization. Soft attention, on the other hand, assigns continuous weights to all input elements, enabling differentiability and seamless integration into gradient-based training pipelines. Soft attention is widely adopted in modern architectures, including Transformers, due to its flexibility and effectiveness.

Despite their widespread success, attention mechanisms face notable challenges and limitations. A key issue is the high computational cost associated with global attention, where the complexity grows quadratically with input size. This problem becomes particularly pronounced in tasks involving large-scale data, such as high-resolution images, long text sequences, or dense 3D point clouds. For example, processing LiDAR point clouds in autonomous driving applications requires handling millions of points, leading to significant resource usage and computational delays.

Scalability is another critical challenge when applying attention mechanisms to tasks with large and complex datasets. In applications such as 3D segmentation or multi-sensor fusion, datasets often contain unstructured and high-dimensional information, increasing memory consumption and processing time. This challenge is further compounded in real-time systems, where low-latency processing is crucial for performance.

To address these limitations, recent advancements have introduced more efficient forms of attention. Sparse attention, for example, selectively applies attention to a subset of input elements, reducing computational complexity while maintaining performance. Child *et al.* ([CHILD *et al.*, 2019](#)) introduced sparse Transformers that scale linearly or sub-linearly with input size, making them suitable for large-scale tasks. Another promising solution is deformable attention, proposed by Zhu *et al.* ([ZHU *et al.*, 2021](#)), which focuses on a sparse set of key positions in the input data. By dynamically learning these positions, deformable attention reduces computational overhead while preserving accuracy, making it particularly effective for dense data such as 3D point clouds.

In the scope of this thesis, attention mechanisms are explored for their potential in multi-sensory fusion. Although these mechanisms were not originally developed for sensor fusion, their ability to selectively highlight relevant features makes them well-suited for integrating multi-modal data in tasks such as 3D object detection and segmentation. The application of these methods to sensor fusion will be further elaborated in the following chapters, with a particular focus on their innovative adaptation to handle complex and heterogeneous inputs.



RELATED WORK

This chapter is organized as follows: the first section provides a detailed analysis of BEVFusion ([LIU et al., 2022](#)) and its contributions to the field; the second and third sections explore related studies that investigate attention mechanisms for multi-sensory fusion, both in specific contexts—such as segmentation and 3D object detection in autonomous vehicles—and in other application areas involving multiple sensors across various domains. The introduction to this chapter presents an overview of the relevant studies that underpin this research, with a focus on enhancing the efficiency of segmentation and 3D object detection through the implementation of attention mechanisms for multi-sensory fusion, using BEVFusion as the initial reference.

3.1 BEVFusion

BEVFusion is an efficient and generic multi-sensory fusion framework designed to address the complexity of 3D perception in autonomous driving environments ([LIU et al., 2022](#)). This innovative method unifies multimodal features into a shared BEV representation space, preserving both geometric and semantic information. This approach revolutionizes the traditional concept of sensor fusion, which previously relied on point-level fusion, and introduces a new way of integrating information from different sensory sources.

BEVFusion tackles the fundamental challenge of reconciling the visualization discrepancies between distinct sensor data, such as cameras and LiDAR, which operate in different visualization modalities. Traditionally, previous methods resorted to projecting LiDAR point clouds onto camera images for fusion ([LIU et al., 2022](#)). However, this approach presents significant limitations, as the projection from camera to LiDAR leads to semantic loss and geometric distortions.

In contrast, BEVFusion adopts the innovative approach of transforming all relevant features into the BEV space, which is conducive to nearly all perception tasks. This strategy

preserves both the geometric structure of LiDAR features and the semantic density of camera features. Additionally, BEVFusion effectively identifies and overcomes performance bottlenecks, such as latency in transforming camera views to BEV, through BEV pooling optimization techniques (LIU *et al.*, 2022).

The transformation from camera to BEV is a critical step in the sensor fusion process for 3D perception (LIU *et al.*, 2022). In the method proposed by BEVFusion, this transformation is efficiently addressed through preprocessing techniques and interval reduction, thereby optimizing the BEV feature aggregation operation.

Initially, during the preprocessing phase, each point in the camera feature point cloud is associated with a BEV grid. This step is crucial to ensure that both the 3D coordinates and the BEV grid index are pre-calculated for each point, enabling efficient reordering during inference (LIU *et al.*, 2022).

After the grid association, feature points that share the same BEV grid are sequentially organized into tensors. Then, during the interval reduction step, features within each BEV grid are aggregated using a symmetric function, such as mean, max, or sum. However, to overcome the inefficiencies of previous approaches, a specialized Graphics Processing Unit (GPU) kernel was developed to operate directly on the BEV grids, allowing for more efficient feature aggregation and significantly reducing process latency (LIU *et al.*, 2022).

These optimizations result in a camera-to-BEV transformation that is up to 40 times faster, reducing latency from over 500ms to just 12ms, representing a minimal fraction of the model's total execution time. Furthermore, this approach is highly scalable across different feature resolutions, which is crucial for the effective integration of multimodal sensory features into the shared BEV representation (LIU *et al.*, 2022).

After converting all sensory features into the shared BEV representation, BEVFusion proposes a simplified fusion process using an element-wise operator, such as concatenation. Although they reside in the same space, the LiDAR and camera BEV features may exhibit spatial misalignments due to inaccuracies in depth determination by the vision transformer (LIU *et al.*, 2022). To address this issue, the method adopts a convolution-based BEV encoder, integrating residual blocks to correct these local misalignments (LIU *et al.*, 2022). This strategy is essential to ensure the accuracy and consistency in fusing features from different sensors, enabling a more robust and cohesive representation of the information captured by the autonomous vehicle's perception system.

Experiments with BEVFusion, using the nuScenes dataset (CAESAR *et al.*, 2020), explored a wide variety of realistic driving scenarios and conditions. nuScenes offers a diverse collection of data captured in different cities, covering urban and suburban environments as well as various weather conditions and times of day. This allowed for the evaluation of the robustness and generalization of the models in different driving contexts, including challenging

situations such as heavy traffic, adverse weather conditions, and complex roadways. Additionally, the detailed 3D annotations provided by nuScenes enable an accurate assessment of model performance in tasks like object detection and segmentation. Thus, experiments with BEVFusion provided valuable insights into the framework's effectiveness and robustness across diverse autonomous driving environments.

3.2 Attention Mechanisms in Multi-Sensor Fusion for 3D Object Detection

BEVFusion has demonstrated remarkable results, significantly optimizing the time required to generate BEV representations and achieving state-of-the-art performance on several autonomous vehicle datasets for tasks such as 3D object detection and segmentation.

However, after acquiring the LiDAR and camera signals and converting them into feature space representations, BEVFusion combines them through concatenation and then processes them using a convolutional network. This approach presents an opportunity for refinement in the way this information is integrated and processed. This study aims to evaluate attention methods to enhance the fusion of LiDAR and camera signals in this context.

The concept of attention mechanisms has roots in various areas of computer science and artificial intelligence, but its specific application in neural networks and deep learning gained prominence primarily with the development of the Transformer, a machine learning model introduced in the paper "Attention is All You Need" ([VASWANI et al., 2017](#)).

Although the concept of attention has been explored in different contexts before, such as in language models, machine translation, and natural language processing, the *Transformer* was one of the first deep learning architectures to demonstrate the power and effectiveness of attention mechanisms across a wide range of tasks, including machine translation, text summarization, and computer vision tasks.

Due to its status as a constantly developing field within the scope of Deep Learning, there is a significant gap in the literature regarding studies dedicated to multi-sensor fusion using attention mechanisms. Nevertheless, some relevant works have explored these resources both in the context of perception for autonomous vehicles and in other application domains. This section will review studies that have investigated attention mechanisms as a means to address multi-sensor fusion in 3D detection contexts.

The Adaptive Fusion Transformer (AFTR), proposed in the paper "AFTR: A Robustness Multi-Sensor Fusion Model for 3D Object Detection Based on Adaptive Fusion Transformer" ([ZHANG et al., 2023](#)), introduces an end-to-end fusion framework that employs adaptive attention mechanisms to address the misalignment and feature diffusion issues commonly encountered in multi-sensor fusion. The key components of AFTR include the Adaptive Spatial Cross Atten-

tion (ASCA) mechanism and the Spatial Temporal Self-Attention (STSA) mechanism.

The Adaptive Spatial Cross Attention (ASCA) dynamically associates and interacts with data features from multiple sensors in 3D space, mitigating misalignment issues and reducing computational costs. By selectively interacting with relevant features using learned offsets, Adaptive Spatial Cross Attention (ASCA) improves alignment accuracy without the need for rigid projections. Additionally, Adaptive Spatial Cross Attention (ASCA) effectively integrates temporal information to counter misalignments induced by dynamic scenes, ensuring precise fusion across temporal frames.

The Spatial Temporal Self-Attention (STSA) further enhances alignment and robustness by incorporating temporal information and dynamically updating offsets to align object features across different timestamps. This mechanism enables the Adaptive Fusion Transformer (AFTR) to achieve state-of-the-art performance in 3D object detection tasks, demonstrating superior accuracy and robustness compared to existing fusion models.

In summary, the Adaptive Fusion Transformer (AFTR) represents a significant advancement in the field of multi-sensor fusion for autonomous driving systems, offering a robust and efficient solution to the challenges of data heterogeneity and dynamic scene perception. By addressing misalignment and feature diffusion problems, the Adaptive Fusion Transformer (AFTR) paves the way for more reliable and accurate autonomous driving systems in complex real-world environments.

The paper “MSF3DDETR: Multi-Sensor Fusion 3D Detection Transformer for Autonomous Driving” ([ERABATI; ARAUJO, 2022](#)) proposes a transformer architecture for the fusion of image and LiDAR data, aiming to improve detection accuracy. The MSF3DDETR is a single-stage, anchor-free, and post-processing-free model that takes multi-view images and LiDAR point clouds as input and predicts 3D bounding boxes.

The method proposed by MSF3DDETR consists of three main components: (1) a convolutional *Backbone* and *Neck*, (2) a *Transformer Head*, and (3) a *Loss* function. The convolutional *Backbone* extracts features from multi-view images using a shared ResNet, while LiDAR features are extracted using a base network such as SECOND or PointPillars. The novelty of MSF3DDETR lies in the *transformer head*, which employs a cross-attention block to fuse RGB and LiDAR features through an attention mechanism, leveraging learned object queries. These object queries interact within a multi-head self-attention block to refine the queries and predict 3D bounding box parameters. During training, the model is optimized using bipartite matching and set-to-set loss.

The paper “FUTR3D: A Unified Sensor Fusion Framework for 3D Detection” ([CHEN et al., 2023](#)) proposes the first unified end-to-end sensor fusion framework for 3D detection, called FUTR3D, which can be used in (almost) any sensor configuration. FUTR3D employs a *Query-Based Modality-Agnostic Feature Sampler* along with a *transformer decoder* and a

set-to-set loss for 3D detection, thus avoiding the use of *late fusion heuristics* and *post-processing tricks*.

The approach proposed by FUTR3D can be conceptually divided into four main parts. First, data from different sensor modalities are encoded by their *modality-specific feature encoders*. Then, a *Query-Based Modality-Agnostic Feature Sampler* (MAFS) is used to sample and aggregate features from all modalities according to the initial positions of the queries; this is the main innovative aspect of this work. Next, a *shared transformer decoder head* is employed to refine the *bounding box* predictions based on the fused features using an *iterative refine module*. Finally, the *loss* is based on *set-to-set matching* between predictions and *ground-truths*.

The MAFS operates in a *modality-agnostic* environment, meaning that it is designed to be independent of the input data modality. The input to the MAFS consists of a set of *object queries*, which are abstract representations of objects in a scene, and the features extracted from all available sensors in the perception system, such as LiDAR, cameras, and radar.

For each query, the MAFS samples features from all sensors. This sampling process is conducted adaptively, based on the location and orientation of the query in relation to the data provided by each sensor. In essence, the MAFS selects the relevant features from each sensor that are associated with the position and shape of the object represented by the query.

After sampling features from all sensors, the MAFS fuses these features into a unique and integrated representation of the object. This fusion process is carried out in a way that preserves the important information from each sensor while simultaneously leveraging the complementarities between the sensor modalities. For example, depth information provided by LiDAR can be combined with the visual features from cameras to improve the accuracy and robustness of object detection in challenging environments.

The output of the MAFS is a rich multi-modal representation of each object in the scene, encapsulating crucial information from all available sensors. This integrated representation is then used by the rest of the FUTR3D detection pipeline, including the *transformer decoder*, to make accurate and robust predictions about the location, orientation, and class of objects in the scene.

The paper “Cross Modal Transformer: Towards Fast and Robust 3D Object Detection” ([YAN et al., 2023](#)) presents the development of a robust 3D detector, called Cross Modal Transformer (CMT), for end-to-end multi-modal 3D detection. The CMT aims to overcome the challenges encountered in fusing information from different sensor modalities, such as images and point clouds, for accurate 3D object detection. The proposed method addresses the complexity of multi-modal integration without the need for explicit transformation of views, adopting a cross-attention mechanism in a transformer decoder. This approach allows the model to interact directly with image and point cloud *tokens*, producing precise 3D bounding box outputs.

The method proposed by CMT consists of three main components: a coordinate encoding module, a position-guided query generator, and a decoder with loss calculation. The coordinate encoding module is responsible for incorporating positional information into the multi-modal features, enabling effective fusion between different sensor modalities. To achieve this, 3D coordinates are implicitly encoded into the multi-modal *tokens*, avoiding the need for explicit feature alignment between views. The position-guided query generator initializes the queries with 3D reference points, which are then transformed into image and LiDAR spaces for relative coordinate encoding. These queries are then used to interact with the multi-modal *tokens* in the transformer decoder, generating class and 3D bounding box predictions. The decoder uses multiple layers to update the representations of the multi-modal *tokens* and predict the desired outputs. The loss is computed based on bipartite matching between the predictions and the ground-truth labels, using a combination of focal loss for classification and L1 loss for 3D bounding box regression.

Additionally, the CMT employs a training strategy called masked modal training to enhance its robustness. This strategy involves training the model with only a single input modality during certain iterations, ensuring that the model remains robust even in the absence of certain sensors. Experiments show that CMT maintains robust performance, even in extreme sensor failure scenarios, such as the absence of LiDAR data.

The paper “AutoAlign: Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection” ([CHEN et al., 2022](#)) proposes an adaptive and automated feature fusion strategy, enabling the model to dynamically align heterogeneous features in a data-driven manner.

The method proposed by AutoAlign is divided into three main stages: First, the Cross-Attention Feature Alignment (CAFA) module is introduced, allowing the adaptation of features between heterogeneous representations. Instead of a deterministic matching, CAFA enables each *voxel* to perceive the entire image and dynamically attend to 2D features at the pixel level based on learnable alignment maps.

Next, the Self-supervised Cross-modal Feature Interaction (SCFI) module is introduced to guide the learning of CAFA. This module utilizes the final predictions of the 3D detector as proposals, leveraging both image and point cloud features for accurate proposal generation. Moreover, the feature interaction between modalities at the instance level is enhanced through a self-supervised learning approach, which applies a similarity loss between paired 2D and 3D features to guide feature alignment. Finally, a joint training strategy for 2D-3D detection is proposed to regularize the features extracted from the image branch and improve detection accuracy.

The paper “DeepInteraction: 3D Object Detection via Modality Interaction” ([YANG et al., 2022](#)), unlike conventional strategies that fuse information from different modalities into a single feature map, the DeepInteraction method maintains modality-specific representations throughout the object detection process. This allows for interaction between modalities for

progressive information exchange and representation learning, using attention mechanisms to highlight relevant areas in both modalities.

DeepInteraction operates in two main stages: the encoder and the decoder. In the encoder, modality-specific representations (image and LiDAR) are learned independently. Instead of fusing these representations into a single feature map, the method keeps them separate to allow interaction between them. This is done by calculating the similarity between elements from both representations and adjusting their weights accordingly. In this way, the most relevant parts of one modality influence the representation of the other modality, and vice versa. This modality interaction process enables progressive information exchange, leading to a richer and more comprehensive representation of the data.

In the decoder, attention is used to improve 3D object predictions by highlighting important areas of the enhanced image or LiDAR representations. This allows the model to focus on relevant features for object detection, such as edges, textures, or specific patterns, thereby improving prediction accuracy.

An important aspect of the method is its ability to handle the complexities and nuances of three-dimensional data, such as the presence of multiple sensory modalities and the need to capture spatial and semantic relationships between objects.

The paper “LIFT: Learning 4D LiDAR Image Fusion Transformer for 3D Object Detection” ([ZENG et al., 2022](#)) addresses a significant challenge in 3D object detection for autonomous driving environments: the efficient fusion of sensor information over time. LiDAR and camera sensors are common in this context, offering complementary information, but fully exploiting these sequential data remains challenging. The paper proposes the LiDAR Image Fusion Transformer (LIFT) to model the mutual interaction between data from different sensors over time. LIFT learns to align sequential data from multiple modalities to achieve multimodal multi-frame information aggregation. Furthermore, it benefits from a data augmentation scheme between sensors and over time.

The proposed method consists of two main components: the Grid Feature Encoder and the 4D Sensor-Temporal Attention. The Grid Feature Encoder processes sequential data from different sensors into grid features. This is achieved by extracting pillar features to project LiDAR points and image features into BEV maps, and then applying point-wise attention to enhance feature representation. On the other hand, the 4D Sensor-Temporal Attention models the mutual correlations of sequential LiDAR and image data using the Transformer’s Self-Attention mechanism. This component uses 4D positional encoding to locate the *tokens* across sensors and time, and adopts a pyramidal context structure to expand the receptive field.

Attention plays a crucial role in the proposed method. In the Grid Feature Encoder, point-wise attention is used to enhance the representation of features within each pillar, allowing for dynamic aggregation of information between LiDAR and image modalities. In the 4D Sensor-

Temporal Attention, the attention mechanism is applied to model the mutual correlations between sequential LiDAR and image data over time, allowing the model to focus on the most relevant relationships for 3D object detection.

The paper “SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection” ([XIE et al., 2023](#)) addresses the following problem: existing 3D object detection methods often use dense representations of scenes, which can be inefficient and noisy, as objects occupy only a small portion of the 3D space. Therefore, the paper proposes the exclusive use of sparse candidates and representations to achieve more efficient and accurate detection.

The proposed method, called SparseFusion, operates in several stages. First, sparse candidates from each modality (LiDAR and camera) are acquired through modality-specific object detection. Then, the instance-level features generated by the camera branch are transformed into the LiDAR space of the instance-level features generated by the LiDAR branch. This is done using a simple, dedicated attention module. To mitigate negative transfer between modalities, geometric and semantic information transfer modules are applied before the parallel detection branches. Additionally, custom loss functions are designed for each module to ensure stable optimization.

In the sparse representation fusion method, the candidates from both modalities are concatenated, and then a self-attention layer is used to efficiently fuse the information from the two modalities. The authors argue that, although simple, the use of self-attention is innovative as it allows modality-specific detectors to encode the advantageous aspects of their respective inputs, while the self-attention module is capable of aggregating and preserving information from both modalities.

Furthermore, the paper proposes information transfer modules between modalities to mitigate negative transfer. The geometric transfer module projects LiDAR points onto depth maps from multiple views and uses them to generate depth-aware features for camera inputs. On the other hand, the semantic transfer module projects LiDAR points onto images, combining the resulting features with the features from the BEV representation.

The paper “Spatial Attention Frustum: A 3D Object Detection Method Focusing on Occluded Objects” ([WANG et al., 2020](#)) proposes the method called Spatial Attention Frustum (SAF), which aims to suppress irrelevant features and allocate limited computational resources to critical parts of the scene, providing greater relevance and facilitating processing for higher-level perceptual reasoning tasks.

SAF is designed to handle occluded objects, which usually have only part of their structure visible. To ensure the effectiveness of the method even in the presence of occluded objects with partial structures, the paper introduces a Local Feature Aggregation (LFA) module to capture more complex local features from the point cloud.

SAF uses a segmentation method based on monocular depth estimation and is guided by the object height, as spatial attention is directly related to the distance estimate of the objects. Additionally, the paper proposes a joint projection loss function between 2D and 3D bounding boxes to improve the overall accuracy of the method. To extract and process features, the method uses a fully convolutional network and an LFA module, which increases the receptive field of each point.

The fusion between LiDAR and camera is performed adaptively, allowing the model to determine which image information is relevant for object detection. To achieve this, a spatially modulated attention mechanism is used, where each object query acts only on the relevant regions of the image, improving the efficiency and robustness of the fusion process. Additionally, an image-guided query initialization strategy is introduced, which selects object queries based on both LiDAR and image features, increasing the model's ability to detect objects that are difficult to identify using just the point cloud.

The Transformer attention mechanism allows the model to adaptively determine where and what information should be extracted from the image, resulting in a robust and effective fusion strategy. Furthermore, an image-guided query initialization strategy is designed to handle objects that are difficult to detect in point clouds.

The paper “UniBEV: Multi-modal 3D Object Detection with Uniform BEV Encoders for Robustness against Missing Sensor Modalities” ([WANG et al., 2023](#)) aims to create well-aligned BEV feature maps from each available sensor modality. In contrast to previous BEV-based multi-modal detection methods, where each sensor modality follows a non-uniform approach to re-sample features from the sensor’s native coordinate systems to BEV features, UniBEV uses a uniform approach. This is achieved through a uniform deformable BEV encoder for all sensor modalities, facilitating alignment between the modalities.

UniBEV uses a set of learnable BEV query vectors with associated 3D spatial locations, shared across all modalities, to construct BEV features. These queries are designed for the native spatial coordinates of each sensor modality and are used to encode BEV features through deformable attention layers. Additionally, the method proposes a fusion module that uses channel-normalized weights to combine the BEV feature maps from different sensor modalities, ensuring that the number of channels in the fused features remains consistent, even when one modality is absent.

For object detection, UniBEV employs a common modality dropout training strategy, where during training, the BEV features of a randomly selected modality are discarded with a certain probability, thus simulating the absence of a sensor modality.

The paper “Unifying Voxel-based Representation with Transformer for 3D Object Detection” ([LI et al., 2022](#)) presents a new unified framework for 3D object detection from multimodal data, called UVTR. The main goal of this method is to unify multimodal representations in the

voxel space for accurate and robust 3D detection, both in single-modality and multi-modality scenarios.

To achieve this goal, UVTR introduces a modality-specific space, designed to represent different inputs in the voxel feature space. It then proposes modality-cross interactions, including knowledge transfer and modality fusion, to fully utilize inputs from different sensors. UVTR adopts a Transformer-based decoder to efficiently sample features from the unified space with learned positions, facilitating object-level interactions.

The modality-specific space is constructed differently for images and point clouds. In the case of images, the voxel space is constructed by sampling features from the image plane according to predicted depth scores and geometric constraints. For point clouds, the voxels are naturally generated from the precise position of the points.

A voxel encoder is then introduced for spatial interaction, establishing relationships between adjacent features in each voxel space. Afterward, modality-cross interaction is performed by transferring knowledge between modalities to optimize the model's features, guided by a modality rich in information toward a modality with less information. Furthermore, modality fusion is designed to better utilize all modalities both during training and inference, combining features from different modalities in a unified voxel space.

Finally, the Transformer-based decoder is employed for object-level interaction and final prediction. Inspired by the Deformable DETR, it uses reference positions to sample representative features, regardless of the spatial size of the 3D voxel spaces. Each object query interacts with the unified voxel features in each block of the decoder, allowing the model to make accurate and robust predictions.

3.3 Attention Mechanisms in Multi-Sensor Fusion for Segmentation

Although segmentation is a critical task for autonomous vehicles, few studies have explored sensor fusion or multimodal fusion with attention mechanism specifically for this purpose. This is primarily because segmentation tasks tend to focus heavily on vision, often neglecting the integration of other sensor modalities. Multimodal fusion has been applied in some cases, where data from different modalities, such as images, are combined to enhance segmentation performance. However, significant research in this area remains scarce, especially in applications related to autonomous vehicles. The following paragraphs will discuss the existing studies that leverage attention mechanisms to perform sensor fusion or multimodal fusion for segmentation tasks.

In the work by [Zhang et al. \(2025\)](#) proposes a two-stage attention-guided framework to enhance semantic segmentation through multimodal fusion. The method focuses on integrating

features from various modalities, such as RGB images and auxiliary data (e.g., depth, LiDAR), by leveraging attention mechanisms to refine and guide feature fusion at multiple levels.

The proposed framework consists of two primary components: a feature guidance module and a feature interaction module. These components employ cross-attention mechanisms to prioritize relevant features and reduce noise, dynamically capturing inter-modal relationships. Additionally, a lightweight decoding process ensures efficient representation learning for each modality without significantly increasing computational costs. The two-stage design combines early fusion of raw features and deep feature interaction, achieving a comprehensive integration of multimodal information.

In the work by [Zhuang et al. \(2021\)](#), a Perception-aware Multi-Sensor Fusion (PMF) scheme is proposed to enhance 3D LiDAR semantic segmentation by effectively integrating data from RGB images and LiDAR point clouds. The approach focuses on leveraging perceptual information from these two complementary modalities—appearance details from RGB images and spatio-depth characteristics from point clouds.

The PMF framework consists of three main components: (1) a perspective projection method to map sparse point clouds into the RGB image coordinate space, preserving image appearance information; (2) a two-stream network (TSNet) that separately processes RGB and LiDAR inputs, and incorporates residual-based fusion modules to combine their features effectively; and (3) perception-aware losses, which encourage the model to balance and leverage the perceptual strengths of both modalities. These losses measure perceptual differences between the modalities and supervise the network's learning by focusing on confident predictions from either stream.

A notable feature of this approach is its residual-based fusion mechanism, which dynamically combines image features into the LiDAR stream while retaining the structural information from the point clouds. The attention mechanism embedded within this module ensures that relevant information from both modalities is emphasized during the fusion process. This design allows PMF to handle challenges such as sparse LiDAR data and unreliable RGB inputs under adverse conditions.

The Efficient Perception-Aware Multi-Sensor Fusion (EPMF) method proposed by [Tan et al. \(2024\)](#) builds upon its predecessor, PMF, to address limitations in computational efficiency and scalability while maintaining robust performance in 3D semantic segmentation tasks. Both approaches utilize perception-aware multi-sensor fusion to integrate appearance information from RGB images and spatio-depth data from LiDAR point clouds. However, EPMF introduces significant advancements in data pre-processing and network architecture to optimize the fusion process.

EPMF employs a novel cross-modal alignment and cropping technique to mitigate the misalignment issues between RGB images and LiDAR point clouds. This process reduces unne-

essary computational overhead by aligning and cropping input data to only include overlapping regions, resulting in more compact and efficient inputs. Additionally, EPMF enhances the contextual module within the LiDAR stream by incorporating down-sampling operations and replacing standard convolutional layers with sparse invariant convolutional layers, tailored for the sparse nature of point cloud data.

Another notable improvement is the fusion strategy. Unlike PMF, which focuses on a residual-based fusion within the LiDAR domain, EPMF introduces a mechanism to integrate high-level LiDAR features directly into the camera stream. This adjustment boosts the performance of the camera stream without adding extra computational costs, enhancing the overall effectiveness of the perception-aware loss.

In the work by [Fooladgar and Kasaei \(2019\)](#), the Multi-Modal Attention Fusion Network (MMAF-Net) is introduced as a novel architecture for the semantic segmentation of RGB and depth data. The model aims to effectively integrate multi-modal inputs through an innovative attention-based fusion mechanism, enhancing segmentation accuracy while maintaining computational efficiency.

MMAF-Net employs an encoder-decoder structure with two dedicated encoders for RGB and depth data, generating intermediate feature maps for each modality. These features are combined in the decoder through a series of Attention-Based Fusion Blocks. The Attention-Based Fusion Blocks utilize two types of attention mechanisms: channel-wise attention, which emphasizes the most relevant channels, and spatial-wise attention, which highlights significant spatial regions within the feature maps. By combining these two approaches, the fusion process identifies and amplifies critical information, suppressing irrelevant features and improving segmentation performance.

The model is computationally efficient, employing long-range residual connections to recover information lost during the downsampling process. This design reduces the number of parameters and computational complexity, enabling MMAF-Net to achieve competitive results on benchmarks such as SUN-RGBD, NYU-V2, and Stanford-2D-3D-Semantic datasets. The attention mechanisms in MMAF-Net allow it to dynamically adapt to the complementary nature of RGB and depth data, making it particularly effective for tasks involving complex scenes and varied lighting conditions.

In the work by [Xu, Lu and Wang \(2021\)](#), the Attention Fusion Network (AFNet) for Semantic Segmentation of RGB-IR Images is proposed as an innovative approach for multi-spectral semantic segmentation. The model leverages a co-attention mechanism to fuse features from RGB and IR images, addressing the limitations of traditional fusion methods, such as simple summation or concatenation, which often fail to exploit contextual relationships and the complementary characteristics of multi-spectral data.

The proposed network employs an encoder-decoder structure. Two separate encoders,

based on modified ResNet architectures with dilated convolutions, extract feature maps from RGB and IR images. These encoders maintain high-resolution feature maps and capture detailed spatial information by removing downsampling in the last two blocks, which is crucial for accurately segmenting small objects.

The core innovation lies in the Attention Fusion Module, which implements a co-attention mechanism to guide the fusion of RGB and IR features. This module creates a symmetric structure where RGB features influence the fusion of IR features and vice versa. Using cosine similarity, attention matrices are generated to capture the spatial correlations between the two modalities. These matrices dynamically weight the feature maps, emphasizing relevant areas while suppressing irrelevant information. The attention-enhanced feature maps are then added back to the original feature arrays to complete the fusion process. This mechanism ensures a more effective integration of multi-spectral data, leveraging the complementary strengths of RGB and IR inputs.

The decoder restores the spatial resolution of the fused features using bilinear interpolation and convolutional layers, avoiding artifacts commonly introduced by deconvolution layers. This design results in outputs with high visual fidelity and accurate pixel-level classification.

AFNet demonstrates superior performance in semantic segmentation tasks involving RGB-IR data by enhancing the contextual representation and fully exploiting the complementary nature of the two modalities. The experimental results highlight its effectiveness in improving classification accuracy and localization, particularly in challenging environments such as low-light conditions.

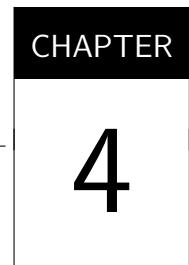
3.4 Final considerations

The analysis of related work highlights the increasing relevance of attention mechanisms in multi-sensor fusion for 3D object detection and segmentation. Several studies have explored spatial and channel attention to improve feature extraction and sensor fusion. For instance, the SAF method employs spatial attention to focus on occluded objects, enhancing the detection of partially visible structures. Similarly, UniBEV integrates deformable attention layers to align BEV features across different sensor modalities, ensuring robustness in multi-sensor fusion. These methods share common objectives with this thesis, as they aim to refine sensor fusion techniques to improve detection accuracy while addressing the challenges of occlusion and sensor misalignment.

For segmentation tasks, fewer studies have specifically explored attention mechanisms for multi-sensor fusion. Most segmentation approaches still focus predominantly on vision-based methods, often neglecting sensor fusion techniques. However, some relevant works have examined how attention mechanisms can refine feature selection in multimodal segmentation. For example, PMF has leveraged attention to optimize the integration of RGB and LiDAR data,

balancing the strengths of each modality.

The reviewed studies reinforce the potential of attention-based methods for multi-sensor fusion and highlight their applicability to BEV representations. While existing works have successfully improved object detection and segmentation, they often require high computational resources. This thesis builds upon these advancements by integrating attention mechanisms into BEVFusion, focusing on optimizing both performance and computational efficiency, which is critical for real-world deployment in autonomous vehicles.



CONDUCTED RESEARCH

This chapter outlines the details of the present research work. First, the dataset used for conducting experiments and evaluations is discussed. Next, the implementations and modifications made to the original BEVFusion code are described, aiming to integrate attention mechanisms to enhance 3D perception. Additionally, the software and hardware resources utilized in this investigation are presented.

4.1 Modifications to BEVFusion

The implementation of the BEVFusion method ([LIU et al., 2022](#)) is publicly available as open-source code, enabling other researchers to reproduce experiments and make modifications as needed. The original authors provided scripts that allow the model to be trained using only camera data, only LiDAR data, or a combination of both, for both detection and segmentation tasks. Additionally, pretrained weights are available for individual sensors for detection and segmentation, as well as for the combined sensor model, which uses these individual pre-trained models as starting points. These resources facilitate experimentation and provide deeper insights into the BEVFusion method.

However, the original model was trained using eight Nvidia A100 GPUs, each with 80 GB of Video Random Access Memory (VRAM), which allowed for a larger batch size during training. In contrast, the available resources in the lab consisted of a single RTX 4090 GPU with 24 GB of VRAM. This required modifications to the original code to use a significantly smaller batch size and reduce the learning rate during training. As a result, these adjustments led to results that differed from those reported by the original authors.

The requirement for a GPU with 24 GB of VRAM for training the original BEVFusion model arises primarily from the use of the *Swin Transformer* ([LIU et al., 2021](#)) backbones and the voxel size employed for processing LiDAR data. This thesis proposes adapting the

BEVFusion model by replacing the backbones and adjusting the voxel size to enable training and inference on GPUs with 12 GB of VRAM, a more modest and commonly available configuration in mid-range GPUs. Among the alternatives to the *Swin Transformer* are ResNet ([HE et al., 2015](#)) and MobileNet ([HOWARD et al., 2017](#)) architectures. The goal of these modifications is to reduce computational complexity and memory requirements without significantly compromising model performance.

In the original BEVFusion implementation, pretrained models for each sensor modality were utilized as backbones. Following this strategy, the ResNet50 architecture was selected as the image backbone because a pretrained model on the dataset used in BEVFusion experiments is publicly available. The LiDAR backbone remains unchanged, continuing to utilize VoxelNet. This adjustment significantly reduced memory consumption, allowing the model to be trained with a batch size of 2. Additionally, the training duration was reduced from 20 to 7 days for segmentation tasks and from 7 to 2 days for 3D object detection.

These training times were obtained using BEVFusion's default training configuration, which estimates 20 epochs for training a segmentation model and 6 epochs for training a 3D object detection model. It is important to note that training parameters such as the number of epochs, learning rate, and scheduling can be adjusted to improve performance, as done by the original BEVFusion authors. However, optimizing these hyperparameters requires extensive experimentation to determine the best configuration, which is beyond the scope of this work. Therefore, the same configuration provided in BEVFusion's official repository was used, with only minimal modifications suggested by the original authors to ensure compatibility with the hardware used in this study.

The original BEVFusion implementation processes each sensor modality through a specific backbone to extract features. These features are subsequently concatenated and processed by a convolutional layer. In this thesis, we propose a modification to this stage of BEVFusion processing. After extracting features from the sensors, the feature vectors are concatenated and passed through an attention mechanism module. This module identifies the most relevant parts of the unified features, generating an importance map that is multiplied by the original features before being processed by the convolutional layer.

During the training phase, dropout regularization is applied to the concatenated features, randomly zeroing out a portion of them with a certain probability. This strategy aims to reduce the model's susceptibility to overfitting. An additional potential benefit of applying dropout is improved robustness to sensor failures or less favourable sensor conditions. In this work, the probability of an input element being zeroed out by dropout was set to 0.25. However, it is important to note that this is a hyperparameter that can be tuned to optimize performance. Due to hardware limitations, a fixed value was used across all experiments.

Figure 10a illustrates the fusion process in the original BEVFusion implementation, where features extracted by each sensor's backbone are concatenated and directly processed by a

convolutional layer. In contrast, Figure 10b shows the proposed modification, where an attention mechanism is introduced to identify the most relevant parts of the unified features. This generates an importance map that is multiplied with the original features before being processed by the convolutional layer. This modification is designed to enhance the fusion process by focusing on key features, thereby improving model robustness and performance.

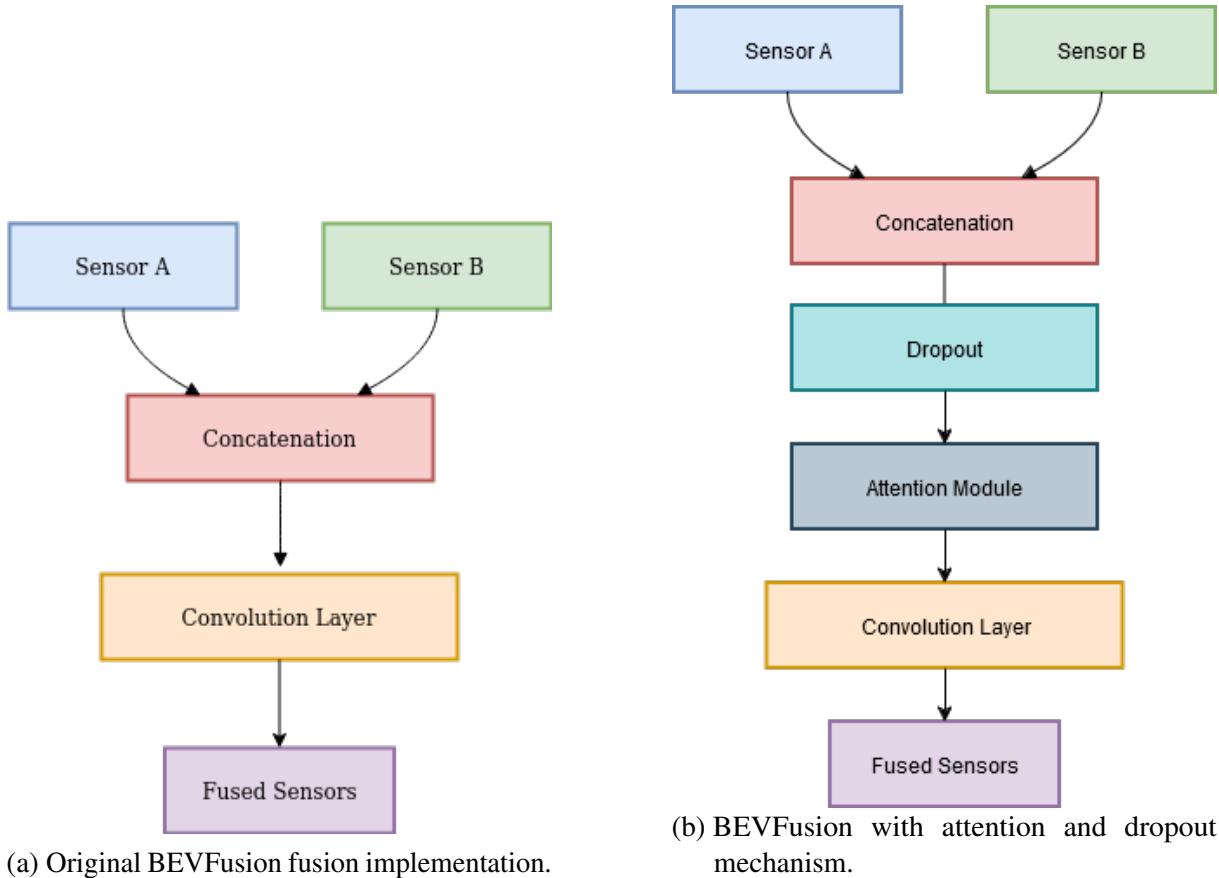


Figure 10 – Figure 10a illustrates the original BEVFusion fusion implementation, while Figure 10b shows the modified BEVFusion with the addition of attention and dropout mechanisms.

Figure 11 highlights the components of the original BEVFusion processing pipeline that were modified in this thesis. The image backbone, marked in blue, was replaced with ResNet50, a residual network with simpler convolutional layers, as an alternative to the Swin Transformer. While the Swin Transformer is a powerful backbone, it is computationally expensive. The sensor fusion process, marked in red, was enhanced by introducing dropout regularization and an attention mechanism, as described in the previous paragraph. The rest of the BEVFusion model remains unchanged and is implemented as provided in the official project repository by the original authors.

The attention mechanisms employed in this study for feature fusion include Spatial Squeeze and Channel Excitation, Channel Squeeze and Spatial Excitation, Concurrent Spatial and Channel Squeeze-and-Excitation, Channel Attention, Spatial Attention, and the Convolutional Attention Block Module (CBAM). These attention modules, along with their implementations,

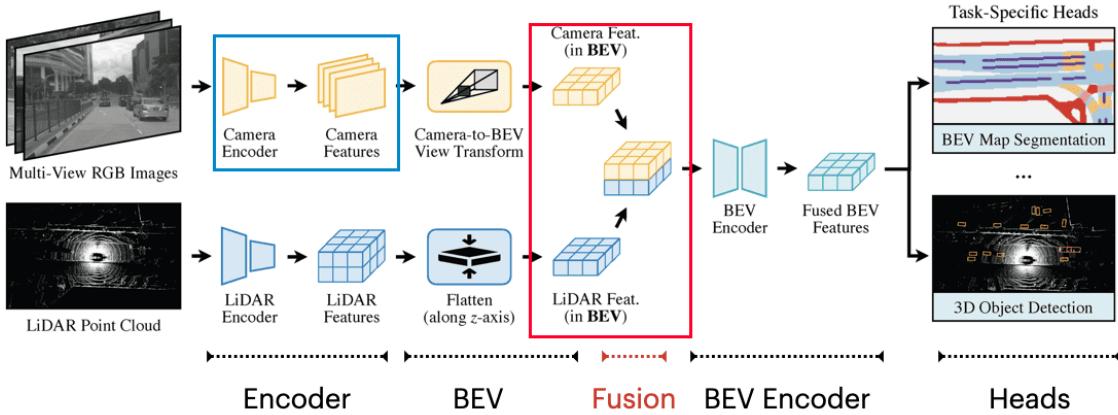


Figure 11 – Modifications to the BEVFusion pipeline. The image backbone (blue) was replaced with ResNet50, and dropout and an attention mechanism (red) were added to the sensor fusion process. Other components remain unchanged.

are detailed in the following subsections.

4.1.1 Spatial Squeeze and Channel Excitation

The Spatial Squeeze and Channel Excitation (cSE) mechanism, introduced in ([HU et al., 2019](#)) as Squeeze and Excitation (SE) blocks, is a channel attention method designed to improve the representational power of CNN. Unlike traditional convolutional layers that process spatial and channel information together, SE blocks explicitly model interdependencies between channels, enabling dynamic recalibration of channel-wise feature responses. This recalibration enhances the network's ability to emphasize informative features while suppressing less useful ones.

An SE block operates in two main stages: **Squeeze** and **Excitation**.

In the **Squeeze** stage, global spatial information is aggregated using global average pooling. For an input feature map $U \in \mathbb{R}^{C \times H \times W}$, where C , H , and W represent the number of channels, height, and width, respectively, the channel descriptor $z \in \mathbb{R}^C$ is computed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U(i, j),$$

where z_c represents the aggregated information for the c -th channel.

In the **Excitation** stage, a gating mechanism generates channel-specific weights. This involves passing z_c through two fully connected layers with a ReLU activation and a sigmoid function:

$$s_{cSE}(U) = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z_c)),$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the weights of the two fully connected layers, r is the reduction ratio, and σ represents the sigmoid activation.

The output feature map $\hat{U}_c \in \mathbb{R}^{C \times H \times W}$ is obtained by rescaling the input feature map U using the generated channel weights s_{cSE} :

$$\hat{U}_c = s_{cSE}(U) \cdot U,$$

SE blocks can be seamlessly integrated into existing CNNs architectures, such as ResNet and Inception, as demonstrated by [Hu et al. \(2019\)](#). By replacing or augmenting standard layers with SE blocks, significant performance gains have been achieved across a variety of tasks, including image classification, object detection, and segmentation, with minimal additional computational cost. Notably, SE-ResNet-50 surpassed the original ResNet-50 by achieving a lower top-5 error on the ImageNet dataset, demonstrating the effectiveness of the recalibration strategy. These improvements underline the potential of channel attention mechanisms like SE blocks to enhance the feature modelling capabilities of CNNs.

4.1.2 Channel Squeeze and Spatial Excitation

The Channel Squeeze and Spatial Excitation (sSE) mechanism, introduced by [Roy, Navab and Wachinger \(2018\)](#), focuses on recalibrating feature maps along the spatial dimension. Unlike the traditional channel-wise approach, sSE emphasizes the relative importance of spatial locations, which is particularly effective for fine-grained image segmentation tasks.

For an input feature map $U \in \mathbb{R}^{H \times W \times C}$, where H , W , and C represent the height, width, and number of channels, respectively, the sSE block first performs a channel squeeze. This is achieved by applying a 1x1 convolution:

$$s_{sSE}(U) = q = W_{sq} \cdot U,$$

where $W_{sq} \in \mathbb{R}^{1 \times 1 \times C \times 1}$ is the weight of the convolution operation, and $q \in \mathbb{R}^{H \times W}$ is the resulting projection tensor. Each element $q_{i,j}$ in q represents a linearly combined representation of all channels at the spatial location (i, j) .

Next, a sigmoid activation σ rescales the activations to the range $[0, 1]$. The recalibrated feature map \hat{U}_c is obtained by multiplying the spatial weights with the input tensor:

$$\hat{U}_c = \sigma(s_{sSE}(U)) \cdot U$$

4.1.3 Concurrent Spatial and Channel Squeeze and Excitation

The Concurrent Spatial and Channel Squeeze and Excitation (csSE) mechanism combines the benefits of both channel-wise (cSE) and spatial (sSE) recalibration. This approach, also introduced by [Roy, Navab and Wachinger \(2018\)](#), applies both recalibration mechanisms concurrently and combines their outputs through an element-wise addition.

Given an input feature map $U \in \mathbb{R}^{H \times W \times C}$, the recalibrated outputs of the cSE and sSE blocks are computed as:

$$\hat{U}_{cSE} = s_{cSE}(U) \cdot U, \quad \hat{U}_{sSE} = s_{sSE}(U) \cdot U.$$

The final recalibrated feature map \hat{U}_{scSE} is obtained by combining these outputs:

$$\hat{U}_{scSE} = \hat{U}_{cSE} + \hat{U}_{sSE}.$$

In this approach, a specific spatial location (i, j) and channel c are emphasized only if both recalibration mechanisms assign high importance to them. This concurrent recalibration enhances the feature map's relevance for tasks like image segmentation, encouraging the network to learn more meaningful spatial and channel-wise representations.

4.1.4 Channel Attention

The Channel Attention (CA) mechanism stands out by focusing on specific features within each channel of a feature representation. This technique becomes crucial in identifying important semantic attributes, significantly improving the effectiveness and efficiency of convolutional models.

In the work by [Chen et al. \(2017\)](#), an innovative architecture is proposed that integrates spatial attention, channel-wise attention, and multi-layer attention in convolutional networks for image captioning. The Channel Attention mechanism is applied to select semantic attributes in response to a sentence context, emphasizing relevant features for prediction. Two approaches, Channel-Spatial and Spatial-Channel, incorporate Channel Attention and Spatial Attention in different orders.

In the work by [Zhu et al. \(2022\)](#), the Channel Interaction Unit is introduced in a model for detecting lung nodules. The CIU uses channel-wise attention to capture local interactions between different channels, enhancing nodule detection and optimizing information, highlighting the effectiveness of Channel Attention in specific tasks.

The work of [Woo et al. \(2018\)](#) presents an innovative attention module that focuses on the primary channel and spatial dimensions. Using a channel-wise attention approach, CBAM highlights relevant features in each channel, significantly improving feature representation. With sequential attention modules for both channels and spatial dimensions, the model enhances its ability to learn "what" and "where" to focus.

In the work of [Yan et al. \(2021\)](#), the Channel-wise Attention-based Depth Network (CADepth-Net) is proposed for monocular depth estimation. By integrating channel-wise attention modules, CADepth-Net performs information aggregation and feature recalibration, emphasizing important details at different scales and reinforcing the applicability of channel-wise attention in various tasks.

In this study, a method analogous to the Channel Attention module proposed by [Woo et al. \(2018\)](#) is adopted. Channel Attention enhances feature representations by focusing on “what” is important within a feature map. This mechanism relies on pooling operations to distill spatial context into channel-wise descriptors, which are then used to compute attention weights dynamically. The primary steps include mean and max-pooling operations, followed by a shared Multi-Layer Perceptron (MLP) with ReLU activation, and normalization via a sigmoid function.

The spatial context descriptors are calculated as:

$$z_{avg} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U(i, j), \quad z_{max} = \max_{i,j} U(i, j),$$

where U represents the input feature map of dimensions $\mathbb{R}^{C \times H \times W}$, and z_{avg} and z_{max} are channel-wise descriptors obtained via average and max-pooling, respectively.

These descriptors are fed into a shared MLP to compute intermediate activations, which are then combined:

$$s_{CA}(U) = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z_{avg})) + \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z_{max})),$$

where W_1 and W_2 are learnable parameters of the MLP, and σ represents the sigmoid activation function.

Finally, the attention weights s_{CA} are applied to the original feature map U via element-wise multiplication, yielding the refined feature map:

$$\hat{U} = s_{CA}(U) \cdot U.$$

This process enables the model to emphasize relevant features while suppressing less critical ones, improving overall representational efficiency. Similar principles are utilized in the Spatial Attention mechanism, which instead focuses on “where” the significant information lies, emphasizing critical spatial regions to complement the channel-based focus. Both approaches work synergistically to refine feature importance across tasks like object detection and semantic segmentation.

4.1.5 Spatial Attention

Spatial Attention (SA) plays a crucial role in CNN, allowing the model to focus on specific regions of an image by assigning differentiated weights to relevant features. This mechanism is particularly valuable in tasks where the location of features is of great importance, such as object detection and semantic segmentation.

A prominent approach to incorporating spatial attention in CNN is through the convolutional block attention module, introduced by [Woo et al. \(2018\)](#). in their study on the subject. The

CBAM was designed to easily integrate into existing convolutional neural network architectures, offering a dedicated part for spatial attention.

In CBAM, spatial attention is computed based on the relationship between spatial features. Unlike channel attention, which focuses on “what” is informative, spatial attention focuses on “where” the relevant information is located, complementing channel attention. This approach is described in detail in the work of [Woo et al. \(2018\)](#).

To compute spatial attention, CBAM uses pooling operations, both average and max pooling, along the channel axis to generate efficient feature descriptors. These descriptors are then convolved to produce a spatial attention map that highlights the important regions of the image, improving the model’s focus on spatially significant areas.

The spatial context descriptors are calculated as:

$$z_{avg} = \frac{1}{C} \sum_{c=1}^C U(c), \quad z_{max} = \max_c U(c),$$

where $U \in \mathbb{R}^{C \times H \times W}$ is the input feature map, and z_{avg} and z_{max} are spatial descriptors obtained via average and max pooling across the channel axis.

The pooled descriptors are concatenated along the channel dimension and passed through a convolutional layer to compute the spatial attention map:

$$s_{SA}(U) = \sigma(f^{7 \times 7}([z_{avg}; z_{max}])),$$

where $f^{7 \times 7}$ represents a convolution operation with a 7×7 kernel, and σ is the sigmoid function that scales the attention values between 0 and 1.

The spatial attention map s_{SA} is then applied to the input feature map U via element-wise multiplication to generate the refined feature map:

$$\hat{U} = s_{SA}(U) \cdot U.$$

This integration of spatial attention into CNN enables the model to focus on relevant regions of the image, enhancing its ability to capture fine-grained spatial details and improving performance across various computer vision tasks.

4.1.6 CBAM

The CBAM, introduced in the work by [Woo et al. \(2018\)](#), is a lightweight and efficient attention mechanism designed to refine feature representations in CNNs. CBAM applies sequential attention along two dimensions: channel and spatial. By learning “what” features are important through channel attention and “where” to focus using spatial attention, CBAM enhances feature refinement in an end-to-end trainable manner.

CBAM first computes a channel attention map, $s_{CA} \in \mathbb{R}^{C \times 1 \times 1}$, for an input feature map $U \in \mathbb{R}^{C \times H \times W}$. Spatial information is aggregated using average pooling and max pooling to generate descriptors. The refined feature map U_c' is then obtained by applying the channel attention map. Next, the spatial attention is computed using the refined feature map, resulting in \hat{U}_c , which highlights more relevant information and leads to a more refined feature map.

$$U_c' = \sigma(s_{CA}(U)) \cdot U.$$

$$\hat{U}_c = \sigma(s_{SA}(U_c')) \cdot U_c'.$$

As shown in Figure 12, the CBAM module integrates seamlessly into CNNs with negligible computational overhead. Extensive experiments in (WOO *et al.*, 2018) demonstrate its effectiveness across tasks such as image classification, object detection, and segmentation.

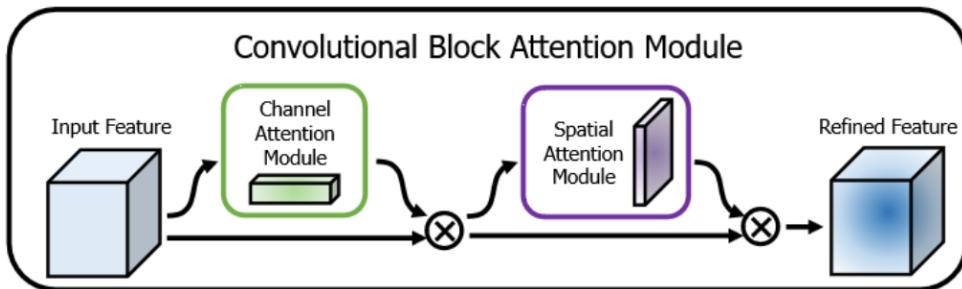


Figure 12 – Diagram of the CBAM module, showing sequential application of channel and spatial attention. Adapted from (WOO *et al.*, 2018).

Channel and Spatial Attention mechanisms have distinct yet complementary roles in feature recalibration. Channel Attention enhances feature maps by weighting each channel according to its relevance, effectively emphasizing the “what” aspect of the data. In contrast, Spatial Attention focuses on the “where” by highlighting important regions in the feature map. Combining these two approaches enables the model to dynamically capture both channel-specific and spatially distributed information, thereby improving performance in 3D object detection and semantic segmentation with minimal additional computational overhead.

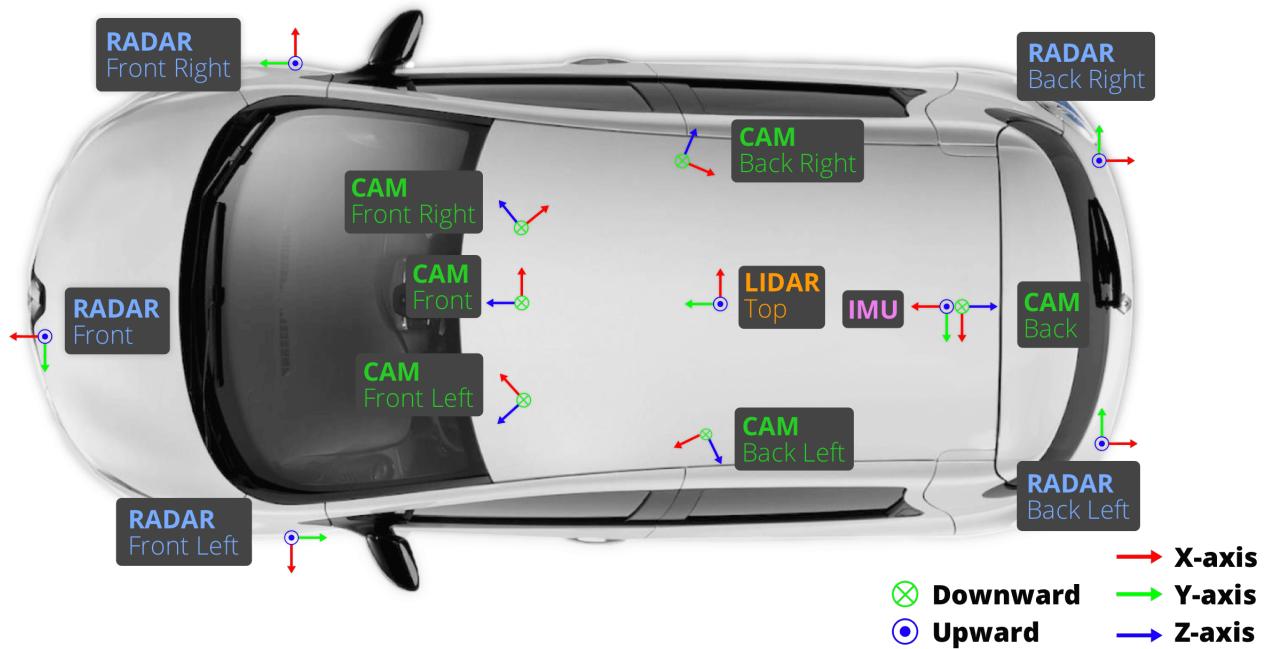
Attention mechanisms based on channel and spatial dimensions also offer the benefit of lower computational complexity compared to transformer-based approaches. Transformer models, which rely on self-attention, resulting in significantly higher memory and processing requirements. Despite this, many recent studies are pursuing transformer-based attention for multi-sensor fusion to leverage their capability to capture long-range dependencies and complex feature interactions. However, the increased computational cost makes them less practical for real-time applications or deployments on resource-constrained hardware.

4.2 nuScenes Dataset

The advancement of autonomous vehicle technology has driven the demand for robust and diverse datasets to train and evaluate perception and control algorithms. In this context, the nuScenes *Dataset* ([CAESAR et al., 2020](#)) emerges as a significant contribution to the autonomous vehicle research community. Developed by nuTonomy, a subsidiary of Aptiv, the nuScenes *Dataset* provides a wide range of sensory data and contextual information collected in real-world urban environments.

The nuScenes *Dataset* consists of a comprehensive collection of data captured by various sensors, including cameras, LiDAR, RADAR, GPS, and IMU, mounted on data collection vehicles, as shown in Figure 13. These sensors offer detailed information about the surrounding environment, including high-resolution images, three-dimensional point clouds, and data on location and orientation.

Figure 13 – Sensor configuration used to generate the nuScenes Dataset ([CAESAR et al., 2020](#)).



For the nuScenes dataset, approximately 15 hours of driving data were collected in Boston and Singapore by third-party entities. The complete dataset includes information from the Seaport district and One North in Boston, as well as Queenstown and Holland Village in Singapore. The driving routes were carefully selected by the data collectors to capture challenging scenarios, ensuring diversity in locations, times of day, and weather conditions. To achieve a balanced class distribution, additional scenes featuring rare classes, such as bicycles, were included. Based on these criteria, 1,000 scenes, each lasting 20 seconds, were manually selected. An example of some images from the dataset can be seen in Figure 14.

The nuScenes Dataset is accompanied by a comprehensive set of annotations, providing

Figure 14 – Example of adverse conditions in the nuScenes dataset (CAESAR *et al.*, 2020).



detailed information about objects in the environment, such as cars, pedestrians, bicycles, and traffic signs. These annotations are crucial for object detection, tracking, and prediction tasks, allowing perception algorithms to understand and interact with the surrounding environment in a precise and effective manner.

The applications of the nuScenes *Dataset* are vast and span a variety of fields, including academic research, autonomous vehicle technology development, simulation, and artificial intelligence algorithm training. Researchers and engineers can use the nuScenes *Dataset* to develop and test perception algorithms, trajectory planning, and autonomous vehicle control in a safe, controlled environment before deploying them in real-world scenarios.

nuScenes provides its own metric for detection tasks, known as the nuScenes Dataset Score (NDS), which was developed to address the limitations of conventional object detection metrics, such as mAP with an IoU threshold. These metrics fail to capture all aspects of nuScenes detection tasks, such as speed estimation and attributes, as well as the coupling of location, size, and orientation estimates. To overcome these limitations, the NDS metric proposes consolidating different types of errors into a single scalar score.

The formula for calculating the NDS is given by:

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

where mAP is the Mean Average Precision, the main metric for this task, True Positive (TP) is the set of the five True Positive metrics, and mean True Positive (mTP) is the mean True Positive metric, representing a set of metrics that are not fundamental for the 3D object detection task but can be useful in specific scenarios, such as assessing a sensor's ability to accurately measure the velocity of surrounding moving objects.

Half of the NDS is based on detection performance in terms of the mean precision for each class, while the other half evaluates the quality of detections in terms of the mean True

Positives. This enables a comprehensive evaluation of the detection models' capabilities in relation to different aspects of the nuScenes scenes. Additionally, to prevent the error metrics from exceeding 1, each metric is bounded between 0 and 1 in the NDS calculation formula.

The TP metrics used by the nuScenes dataset are calculated independently for each class and represent the average of the cumulative average at each recall level achieved above 10%. If a recall of 10% is not reached for a given class, all TP errors for that class are defined as 1. The defined TP metrics are:

The evaluation metrics include the Average Translation Error (ATE), which represents the Euclidean distance of the center in 2D, measured in meters. The Average Scale Error (ASE) is computed as $1 - IoU$ after aligning the centers and orientation. The Average Orientation Error (AOE) is defined as the smallest yaw angle difference between the prediction and the ground truth, measured in radians; it is evaluated over 360 degrees for all classes except barriers, where it is considered only over 180 degrees, and for cones, it is ignored. The Average Velocity Error (AVE) corresponds to the absolute velocity error in m/s, but it is not considered for barriers and cones. Lastly, the siglaAAEAverage Attribute Error is calculated as $1 - acc$, where acc represents the attribute classification accuracy, and similar to AVE, the attribute error for barriers and cones is ignored. All errors are greater than 0, but note that for translation and velocity errors, the errors have no upper limit and can take any positive value.

4.3 Materials and Resources

For the development of this thesis, we used the source code of BEVFusion, available online on GitHub¹. BEVFusion was entirely developed using mmdetection3d, an open-source object detection toolbox based on PyTorch 1.8. On the BEVFusion page, there was a means to replicate the original development environment in terms of software through a Dockerfile, which allowed the creation of a Docker container with all the necessary packages and dependencies to run the original algorithm.

As for the hardware, we will use a desktop computer equipped with a Ryzen 7 2700 processor running at 3.2 GHz, 64 GB of DDR4 RAM at 3000 MHz, and an Nvidia RTX 4090 graphics card. The task explored in this thesis is heavily GPU-dependent, making a more powerful processor than the one mentioned less critical. However, the GPU configuration is modest compared to the one used in BEVFusion (where 8 Nvidia A100 GPUs were employed), which will result in considerably longer training times and will require modifications. Thus, testing a wide range of specifications or parameters on the models will not be feasible due to the time required to train each one.

¹ <<https://github.com/mit-han-lab/bevfusion>>

4.4 Evaluation

The performance evaluation of the proposed models was conducted on the nuScenes validation set, considering each of the relevant metrics. The most important metrics for comparison were mAP and NDS, which were used for 3D Object Detection tasks. The primary metric for the segmentation task was mIoU.

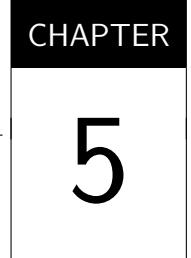
The mAP metric is widely used in object detection tasks, including 3D object detection. It calculates the AP across different recall levels for each class, taking into account both the precision and recall of the model. For 3D object detection, mAP is computed by considering the precision of the predicted bounding boxes, comparing them to the ground truth annotations. A higher mAP value indicates that the model is better at detecting objects accurately and with fewer false positives. In the context of nuScenes, mAP is computed for each object class, considering the 3D bounding box predictions.

For the segmentation task, mIoU is the primary metric used. It measures the overlap between the predicted segmentation mask and the ground truth mask for each region of interest. Specifically, mIoU is calculated as the intersection of the predicted and ground truth areas divided by their union. A higher mIoU score indicates better performance in segmenting the relevant regions. For this thesis, the regions of interest considered in the segmentation task include drivable area, pedestrian crossing, walkway, stop line, car park area, and divider.

Only the models developed in this thesis were included, including the BEVFusion model trained with the necessary adaptations to run on a single 24GB GPU.

For the LiDAR modality, the backbone used was the one provided by the BEVFusion author, while for the image modality, the ResNet50 backbone was obtained from another source but pre-trained on the nuScenes dataset. The nuScenes dataset considered only 10 classes for the object detection task, following the same criterion adopted by BEVFusion: car, truck, construction vehicle, bus, trailer, barrier, motorcycle, bicycle, pedestrian, and traffic cone.

For the segmentation task, the following regions of interest will be considered: drivable area, pedestrian crossing, walkway, stop line, car park area, and divider. These regions are particularly important for understanding the environment and will be used to evaluate the performance of the segmentation models.



RESULTS

In this chapter, we present the results obtained for the evaluated models in both 3D object detection and segmentation tasks. For each task, the models were specifically trained to perform only that task. In other words, a model trained for 3D object detection only performs 3D detection, while a segmentation model was exclusively trained for segmentation. Although both models share the same backbone and attention modules, the distinction lies in the specific task head appended to the end of the BEVFusion processing pipeline.

The original BEVFusion code, provided by the authors in the official repository, was used to train the models. However, some modifications were necessary to enable training and execution on the hardware available for this study.

For 3D object detection, all models were trained for six epochs with a learning rate of 10^{-5} and a batch size of 2. This contrasts with the configuration used by the original authors, who employed a learning rate of 10^{-4} and a batch size of 4. Consequently, the results obtained in this study differ from those reported by the original authors. The image backbone used was a ResNet50 pre-trained on the NuScenes dataset. Under these conditions, each model required slightly more than two days to train, with VRAM usage averaging around 20 GB.

For segmentation, the models were trained for 20 epochs using a ResNet50 backbone pre-trained on the NuScenes dataset. The batch size was reduced from 4 to 2, but the learning rate remained consistent with the original implementation provided by the authors. Each model took approximately seven days to train under these conditions, with VRAM usage averaging around 22 GB.

For all models incorporating dropout, a probability of 25% was applied to randomly zero out elements in the input. No hyperparameter tuning was performed due to the significant time required for training.

Table 1 presents the results obtained for the 3D object detection task. The abbreviations used in the table are defined as follows: CA refers to the use of the Channel Attention Module,

cSE stands for Spatial Squeeze And Channel Excitation, SA denotes Spatial Attention, sSE indicates Spatial Squeeze and Excitation, and scSE represents the combined Spatial and Channel Squeeze and Excitation.

Among the evaluated models, the cSE approach demonstrated the highest performance in the two most critical metrics for this task: mAP and NDS. Specifically, the cSE model achieved an mAP of 0.6606, representing a performance gain of **0.732%** compared to the baseline BEVFusion model, which achieved an mAP of 0.6558. Following the cSE approach, the CA model, another channel-focused mechanism, achieved the second-best mAP among the evaluated methods.

In the NDS metric, the cSE model also achieved the highest score, with 0.6980, marking an improvement of 0.374% over the baseline BEVFusion, which scored 0.6954. Similar to the mAP results, the Channel Attention model ranked second in NDS performance, further validating the effectiveness of channel-based methods.

Compared to the original BEVFusion model trained with the SwinTransformer, as reported by its authors, a mAP of 0.6852 and an NDS of 0.7138 were achieved, representing performance differences of 3.6% and 2.2%, respectively. It is worth noting that in related studies ([HONORATO; WOLF, 2024](#)), the BEVFusion model trained with the SwinTransformer using a setup similar to this work produced comparable results to those obtained here. This suggests that the slightly lower performance observed in this study may be attributed to the limitations imposed by using a lower batch size of and a higher learning rate.

Regarding the FPS metric, most models, including cSE, maintained an average inference rate of 7.9 frames per second. However, Channel Attention and CBAM, which also employs channel attention, fell slightly below this average. Nevertheless, their FPS values remain comparable and within acceptable limits relative to the other models. The FPS calculation was performed by running inference on 2000 images and then averaging the results.

Table 1 – Performance metrics for different models. Metrics with \uparrow indicate higher is better, and those with \downarrow indicate lower is better.

Model	mAP (\uparrow)	mATE (\downarrow)	mASE (\downarrow)	mAOE (\downarrow)	mAVE (\downarrow)	mAAE (\downarrow)	NDS (\uparrow)	FPS (img/s) (\uparrow)
BEVFusion	0.6558	0.2931	0.2559	0.3297	0.2607	0.1861	0.6954	7.9
CA	0.6593	0.2921	0.2542	0.3277	0.2637	0.1859	0.6973	7.6
cSE	0.6606	0.2903	0.2544	0.3264	0.2663	0.1859	0.6980	7.9
SA	0.6594	0.2919	0.2538	0.3377	0.2635	0.1863	0.6964	7.9
sSE	0.6584	0.2927	0.2546	0.3317	0.2684	0.1851	0.6960	7.9
CBAM	0.6597	0.2907	0.2549	0.3391	0.2644	0.1860	0.6963	7.5
scSE	0.6591	0.2925	0.2546	0.3327	0.2603	0.1849	0.6970	7.9

In Figures [15](#), [16](#), and [17](#), a comparison is presented between the BEVFusion model and the cSE variant across three different scenes. Each figure is organized into three rows: the top row shows the ground truth annotations, the middle row illustrates the results obtained using the BEVFusion model with cSE for multi-sensor fusion, and the bottom row presents the outputs

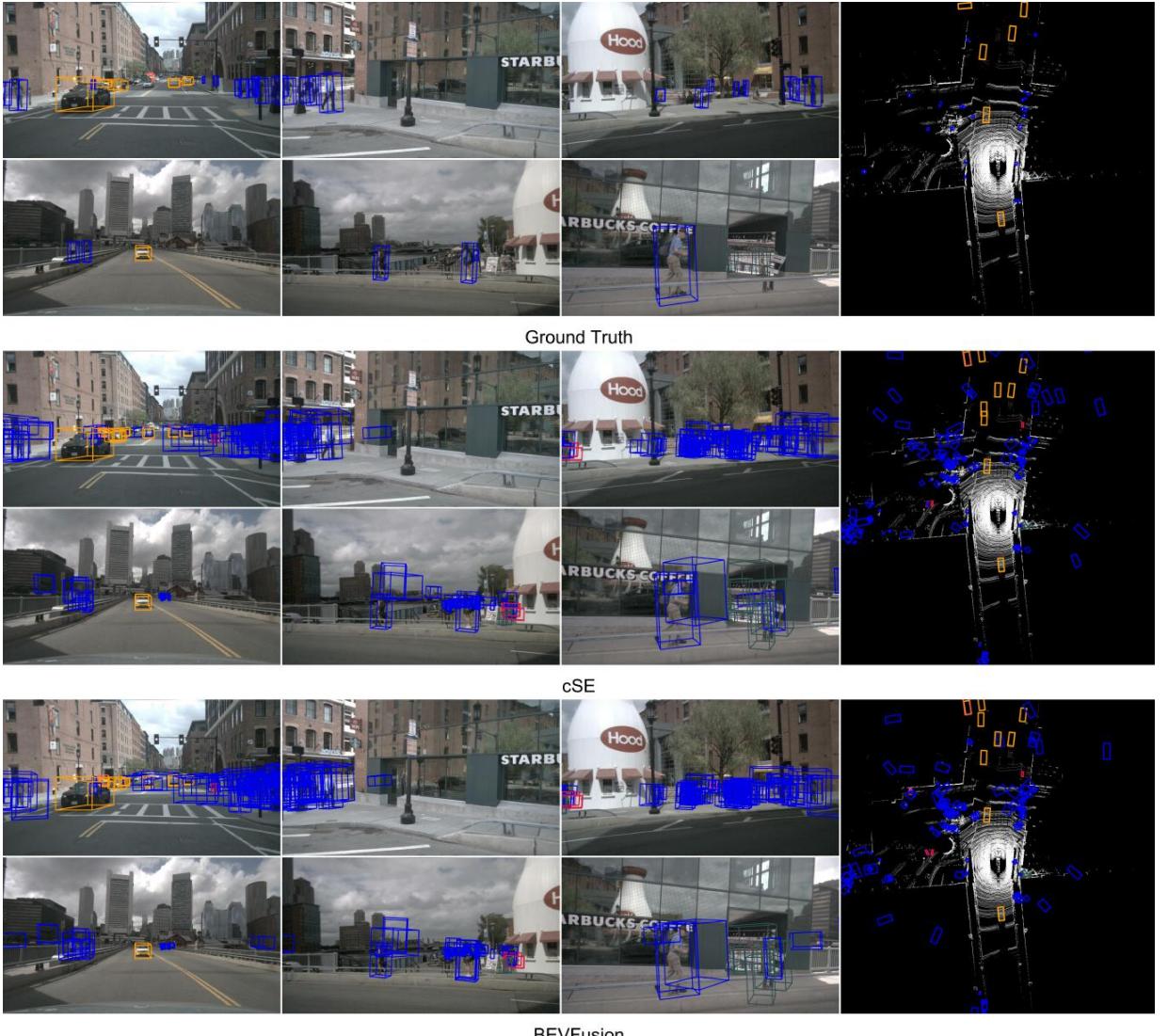


Figure 15 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference.

of the original BEVFusion model. On the left side of each image, the results for the six camera views are displayed, while the right side shows the corresponding LiDAR outputs.

Both models exhibit certain limitations when compared to the ground truth annotations. Common issues include false positives, where bounding boxes are assigned to non-existent objects, and instances where multiple bounding boxes are allocated to a single object. These challenges underscore the complexity of accurately detecting objects in multi-modal environments. It is important to note that the IoU threshold used to determine valid bounding boxes was provided by the original BEVFusion authors in the official repository. Additionally, the non-maximum suppression algorithm employed to reduce redundant bounding boxes was also sourced from their implementation.

Visually, the differences between the two models are not immediately apparent, as both exhibit similar patterns of inaccuracies. This observation aligns with the quantitative results,

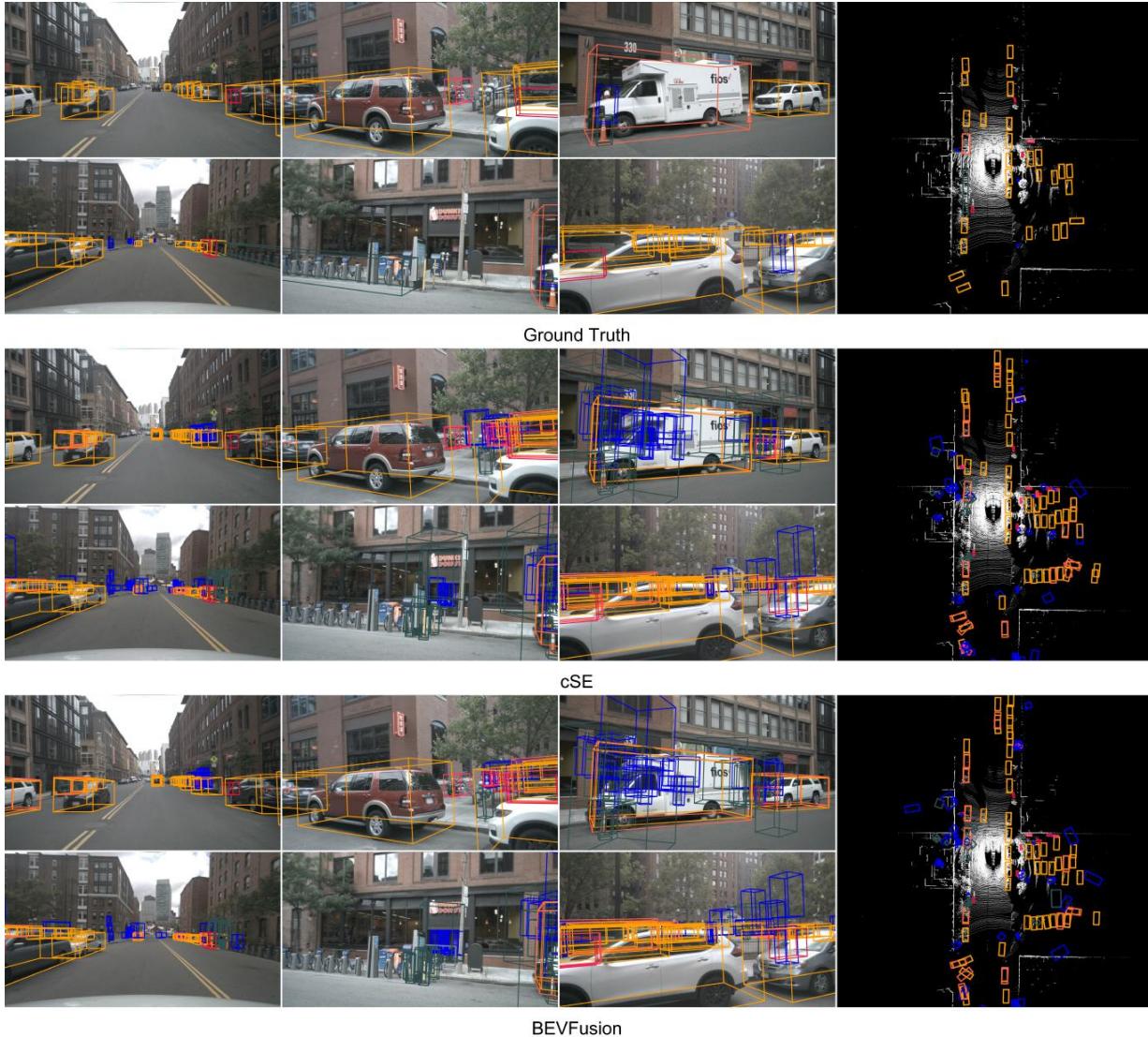


Figure 16 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference.

where the performance gap between the two models was relatively small.

Table 2 summarizes the results comparing training time, training configurations, and performance metrics using the original BEVFusion backbone, Swin Transformer, in two scenarios. The first scenario presents the results reported by the authors, where training was conducted on their hardware, although the exact training time was not disclosed. The second scenario evaluates BEVFusion using our hardware, comparing Swin Transformer with ResNet50 enhanced by the cSE attention mechanism, which achieved the best performance for this task.

Additionally, we observe that when using the RTX 4090—a high-performance GPU—there was no difference in FPS between the models with ResNet50 and Swin Transformer. However, when employing a less powerful GPU, such as the RTX 2070, a difference of 0.4 imgs/s FPS was observed, corresponding to a **17.39%** performance improvement. This suggests that in computationally constrained environments, ResNet-based models offer superior efficiency at the

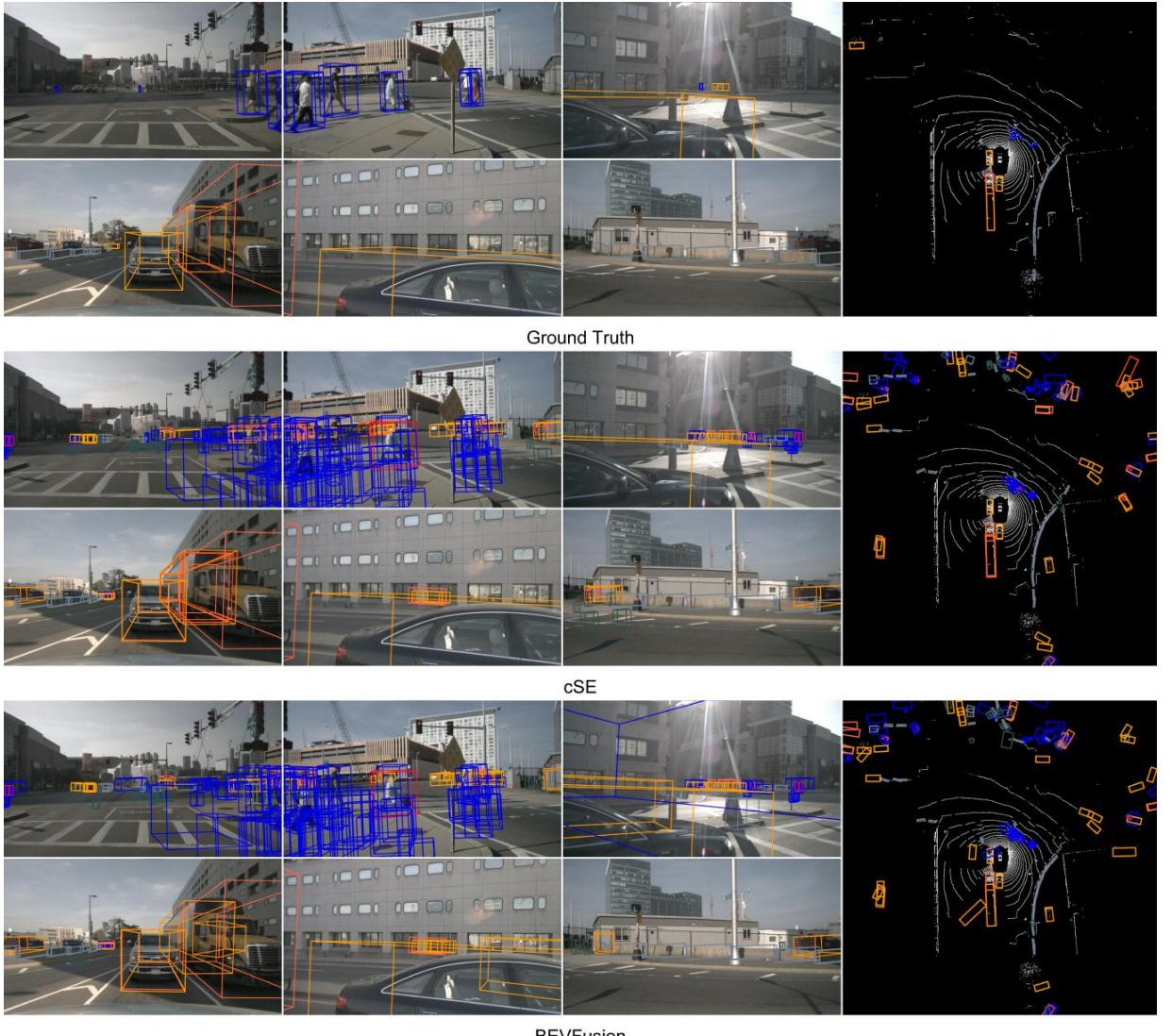


Figure 17 – Visual comparison of 3D object detection results between cSE and BEVFusion, using the Ground Truth as reference..

cost of a minor trade-off in key 3D object detection metrics.

As observed, there is a significant difference between the authors’ BEVFusion results and ours. However, when comparing BEVFusion trained on our hardware, the difference is less pronounced. This suggests that tuning the training parameters could potentially lead to improved performance.

Table 2 – Configuration and performance of models during training.

Model	Backbone	mAP	NDS	Learning Rate	Batch Size	VRAM	Training Time	FPS (4090)	FPS (2070)
BEVFusion	Swin Transformer	0.6852	0.7138	10^{-4}	4	80 GB	-	7.9	2.3
BEVFusion	Swin Transformer	0.6542	0.6922	10^{-5}	1	18 GB	6 days	7.9	2.3
BEVFusion	ResNet50	0.6558	0.6954	10^{-5}	2	20 GB	2 days	7.9	2.7
cSE	ResNet50	0.6606	0.6980	10^{-5}	2	20 GB	2 days	7.9	2.7

Table 3 summarizes the performance of the evaluated models in the segmentation task. As observed, the cSE model consistently achieves the best results across most categories, except

for Car Parking Area and Divider, where the CA model outperforms.

In terms of overall performance, the cSE model achieves the highest mIoU, with an average of 0.5868, closely followed by the Channel Attention model, which achieves an mIoU of 0.5854. This represents a **14.12%** improvement over the baseline BEVFusion model, which achieved a mean mIoU of 0.5142. When compared to the original BEVFusion model reported by its authors, trained using the Swin Transformer backbone and achieving an mIoU of 0.6295, the performance difference is approximately 7%. However, when the original BEVFusion model with Swin Transformer was trained under our configuration, with the batch size reduced from 4 to 1, it achieved an mIoU of 0.5932, reducing the performance gap to only 1.31%.

Table 3 – Average mIoU Metrics by Category.

Model	Drivable Area	Ped Crossing	Walkway	Stop Line	Car park Area	Divider	Mean
BEVFusion	0.7972	0.5069	0.5953	0.3782	0.3589	0.4487	0.5142
CA	0.8241	0.5509	0.6383	0.4394	0.5600	0.4997	0.5854
cSE	0.8331	0.5518	0.6454	0.4472	0.5492	0.4940	0.5868
SA	0.7963	0.5120	0.5944	0.3803	0.4036	0.4548	0.5236
sSE	0.7947	0.5183	0.6051	0.3834	0.3834	0.4605	0.5242
CBAM	0.8136	0.5443	0.6326	0.4234	0.5095	0.4905	0.5690
scSE	0.8288	0.5464	0.6436	0.4358	0.5091	0.4853	0.5749

The comparison of segmentation results between the cSE model and BEVFusion is shown in Figure 18, with the Ground Truth serving as the reference. The columns are organized as follows: the first column displays the Ground Truth, the second shows the results of the cSE model, and the third presents the results obtained with the BEVFusion model. As can be seen, the results are not perfect for either BEVFusion or cSE, but the cSE model is visibly superior.

The main weakness of both methods lies in their ability to segment distant details from the vehicle. This is evident in the first figure (from top to bottom), where the cross street is not detailed by either method, and in the second figure, where the upper area, which represents a drivable region, is not well marked by BEVFusion and shows some flaws in cSE.

In the third figure, from top to bottom, on the upper-right side, BEVFusion fails to segment the area correctly but performs better on the left side. In contrast, cSE segments the right side well but leaves the left side incomplete.

In the fourth figure, BEVFusion marks a drivable region where none exists according to the reference image. cSE also has a small failure in this area, but it is much less noticeable. The key observation here is that both methods struggle to mark the yellow details present in the reference image correctly.

Finally, in the last image, BEVFusion fails to segment the right boundary and does not mark the yellow details found in the reference image. Meanwhile, cSE, although it does not capture all the details fully, manages to mark most of them along with the red regions. However, the purple details are not correctly represented by either method.

It can be observed that, although neither model is perfect, the cSE approach produces results that are significantly closer to the Ground Truth compared to BEVFusion. The cSE model demonstrates greater consistency and precision in the segmented areas. In contrast, the BEVFusion model exhibits more noticeable errors, particularly in boundary regions and intersections.

Table 4 summarizes the efficiency metrics for segmentation. Using SwinTransformer, a batch size of 1 consumed 20 GB of VRAM, while with ResNet50, it was possible to use a batch size of 2 with a VRAM consumption of 22 GB. This suggests that, with a batch size of 1, the estimated memory usage would be around 11 GB, representing a **45%** reduction in VRAM consumption. Additionally, the training time was reduced from 20 days to 7 days, marking a **65%** decrease. Regarding the main metric, mean Intersection over Union (mIoU), models using ResNet50 performed slightly worse, with a minor performance drop of only **1.08%** for the CA attention mechanism compared to the BEVFusion baseline provided by the original authors. However, when compared to the BEVFusion model trained with a batch size of 4, the performance gap increased to **6.78%**, highlighting the significant impact of training configuration on the final model performance.

Table 4 – Configuration and performance of models for segmentation.

Model	Backbone	mIoU	Learning Rate	Batch Size	VRAM	Training Time
BEVFusion	Swin Transformer	0.6295	10^{-4}	4	80 GB	-
BEVFusion	Swin Transformer	0.5932	10^{-4}	1	20 GB	20 days
BEVFusion	ResNet50	0.5142	10^{-4}	2	22 GB	7 days
cSE	ResNet50	0.5868	10^{-4}	2	22 GB	7 days

In general, the cSE model consistently achieved the best performance across both tasks. Notably, for the segmentation task, both the cSE and CA mechanisms yielded the highest results. This variation in performance among channel-based attention mechanisms can likely be attributed to the concatenation of sensor data being performed along the channel axis. Channel-based attention enhances interdependencies between features at the channel level, effectively identifying which channels are most relevant to the problem. This capability is a key factor contributing to the superior performance observed in these models.

The channel-based attention methods demonstrated significantly better performance than spatial attention methods, most likely because the attention mechanism is applied in the feature space extracted from the sensors using deep learning models. In this space, "what" information is present is more important than "where" it is located. It is possible that if the same attention mechanisms were applied directly to the raw sensor signals, without prior feature extraction, spatial attention would yield better results. Furthermore, combining both attention mechanisms, as seen in hybrid approaches like CBAM or scSE, could potentially enhance performance beyond what was achieved using only channel or spatial attention individually.

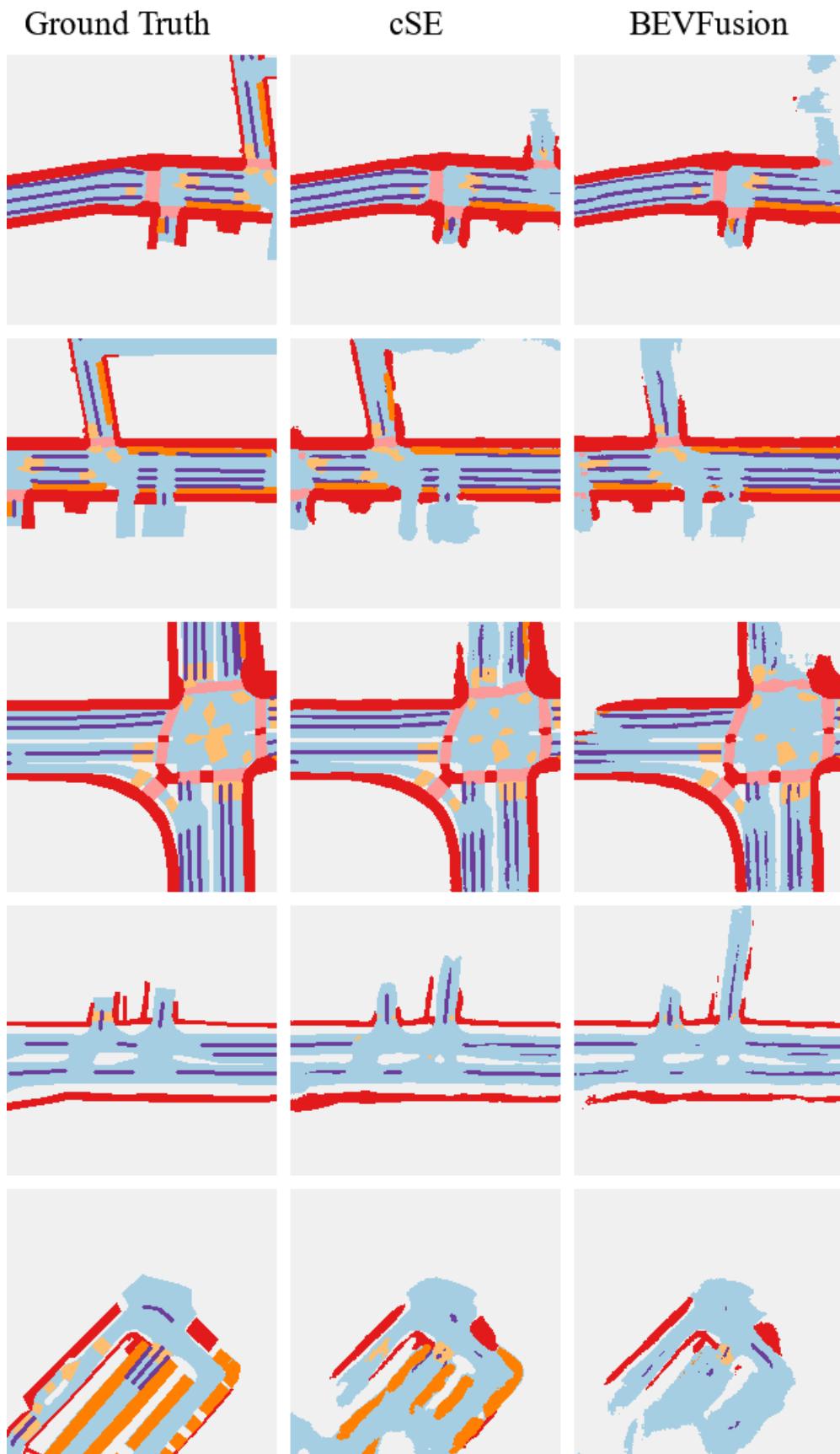
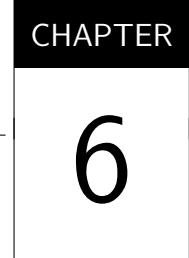


Figure 18 – Visual comparison of segmentation results between cSE and BEVFusion, using the Ground Truth as reference.



CONCLUSION

This thesis addresses the rapidly evolving field of autonomous vehicles, with a specific focus on visual perception. This domain is fundamental, as many other tasks in autonomous driving rely heavily on accurate perception. The challenge tackled in this study pertains to multi-sensory fusion. The objective was to optimize the fusion method of a state-of-the-art model while simultaneously reducing its computational cost without compromising performance in 3D object detection and semantic segmentation tasks.

In this study, several attention mechanisms were explored, including Channel Squeeze and Spatial Excitation, Spatial Squeeze and Channel Excitation, Concurrent Channel and Spatial Squeeze and Excitation, Channel Attention, Spatial Attention, and CBAM. While these mechanisms were originally designed to enhance the performance of CNN in tasks like classification and segmentation, their innovative application here lies in their adaptation for sensor fusion.

Additionally, dropout regularization was applied before feeding data into the attention modules. This approach forces the model to become more robust during training, reducing overfitting and enhancing performance when dealing with sensor failures.

This thesis builds upon the original implementation of BEVFusion provided by its authors. However, due to hardware limitations encountered during this study, several adjustments were necessary. The first modification, aligned with the goals of this thesis, was to replace the original image backbone, Swin Transformer, with the simpler ResNet50. Furthermore, given the 24GB VRAM limitation of the available GPU, the batch size had to be reduced from 4 to 2. For the 3D object detection task, the learning rate was decreased by a factor of 10, whereas for the segmentation task, the original learning rate was retained.

Experimental results on the NuScenes dataset, the same used by the original authors, demonstrated a performance improvement of 0.732% in mAP and 0.374% in NDS for the 3D object detection task in the best-performing model. For the semantic segmentation task, there was a significant 14.12% performance improvement compared to the base BEVFusion model,

using ResNet50 as the image backbone.

The Channel Squeeze and Spatial Excitation mechanism emerged as the best-performing model across both tasks. Particularly in the segmentation task, the top two models were those employing channel attention mechanisms. This suggests that channel-based attention is particularly effective in scenarios where sensor fusion occurs along the channel axis. In this configuration, data from different sensors are concatenated along the channel dimension. Channel attention mechanisms likely enhance the model's ability to identify the most relevant information from each channel. This is consistent with how these mechanisms generate attention maps, selectively highlighting specific channels corresponding to camera and LiDAR inputs, which improves segmentation performance.

An interesting avenue for future research, inspired by this finding, is the application of channel-wise dropout. After sensor fusion, the data dimensionality is $R^{80 \times 80 \times 336}$. By applying dropout to zero out entire channels, the model could potentially become even more robust, learning to handle scenarios where data from specific channels are entirely missing.

Regarding computational efficiency, the ResNet50-based model achieved a **65%** reduction in processing time, a **45%** decrease in memory consumption, and a **17.91%** increase in FPS. While there was some performance loss, the efficiency gains outweigh this trade-off. An interesting finding is that model performance is highly influenced by training configurations. A model trained on an RTX 4090 with a batch size of 1 was outperformed by the same model trained on an A100 with 80 GB of VRAM using a batch size of 4.

When using the RTX 4090, there was no difference in FPS between the Swin Transformer and ResNet50 models. However, on a less powerful GPU, such as the RTX 2070, the FPS difference was 0.4 imgs/s, representing a 17.39% performance gain. This suggests that in resource-constrained environments, ResNet50 offers a more efficient alternative while maintaining comparable detection performance.

In conclusion, this thesis successfully achieved the goal of reducing the computational cost of the model while attaining a substantial performance gain compared to the base BEVFusion model with a less computationally intensive backbone. Compared to the original backbone, SwinTransformer, the best-performing model in this study achieved results only 1.31% below its performance. This is a positive outcome, considering a 3.3-fold reduction in training time and improved compatibility with less powerful hardware for inference.

In future work, we can explore additional attention mechanisms, such as local channel attention and global channel attention, as well as ways to combine them. Additionally, we can investigate the use of attention mechanisms in models with Swin Transformer. Another possible research direction is to analyze the differences between spatial attention mechanisms and channel attention mechanisms across different scenarios and applications. For instance, we could apply attention mechanisms directly at the sensor level, without any prior preprocessing, to assess

whether spatial attention performs better in this setting.

BIBLIOGRAPHY

- AGUILERA-MARTOS, I.; HERRERA-POYATOS, A.; LUENGO, J.; HERRERA, F. **Local Attention Mechanism: Boosting the Transformer Architecture for Long-Sequence Time Series Forecasting**. 2024. Available: <<https://arxiv.org/abs/2410.03805>>. Citation on page 51.
- ALABA, S.; GURBUZ, A.; BALL, J. A comprehensive survey of deep learning multisensor fusion-based 3d object detection for autonomous driving: Methods, challenges, open issues, and future directions. Institute of Electrical and Electronics Engineers (IEEE), Aug. 2022. Available: <<http://dx.doi.org/10.36227/techrxiv.20443107.v2>>. Citations on pages 13, 37, and 38.
- ARNOLD, E.; AL-JARRAH, O. Y.; DIANATI, S. F. M.; OXTOBY, D.; MOUZAKITIS, A. A survey on 3d object detection methods for autonomous driving applications. **IEEE Transactions on Intelligent Transportation Systems**, v. 20, n. 10, p. 3782–3795, 2019. Citations on pages 46 and 48.
- BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. 2016. Available: <<https://arxiv.org/abs/1409.0473>>. Citation on page 50.
- BANERJEE, T. P.; DAS, S. Multi-sensor data fusion using support vector machine for motor fault detection. **Information Sciences**, v. 217, p. 96–107, 2012. ISSN 0020-0255. Available: <<https://www.sciencedirect.com/science/article/pii/S0020025512004185>>. Citation on page 37.
- BELTRAN, J.; GUINDEL, C.; MORENO, F. M.; CRUZADO, D.; GARCIA, F.; ESCALERA, A. de la. **BirdNet: a 3D Object Detection Framework from LiDAR information**. 2018. Citation on page 49.
- BYEON, W.; BREUEL, T. M.; RAUE, F.; LIWICKI, M. Scene labeling with lstm recurrent neural networks. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 3547–3555. Citation on page 44.
- CAESAR, H.; BANKITI, V.; LANG, A. H.; VORA, S.; LIONG, V. E.; XU, Q.; KRISHNAN, A.; PAN, Y.; BALDAN, G.; BEIJBOM, O. **nuScenes: A multimodal dataset for autonomous driving**. 2020. Citations on pages 13, 54, 76, and 77.
- CANAN, S.; AKKAYA, R.; ERGINTAV, S. Extended kalman filter sensor fusion and application to mobile robot. In: **Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004**. [S.l.: s.n.], 2004. p. 771–774. Citation on page 36.
- CHABOT, F.; CHAOUCH, M.; RABARISOA, J.; TEULIÈRE, C.; CHATEAU, T. **Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image**. 2017. Citation on page 49.
- CHEN, L.; ZHANG, H.; XIAO, J.; NIE, L.; SHAO, J.; LIU, W.; CHUA, T.-S. **SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning**. 2017. Citation on page 72.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. **Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs**. 2016. Available: <<https://arxiv.org/abs/1412.7062>>. Citation on page 43.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. **DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs**. 2017. Available: <<https://arxiv.org/abs/1606.00915>>. Citation on page 44.

CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. **Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation**. 2018. Available: <<https://arxiv.org/abs/1802.02611>>. Citation on page 44.

CHEN, X.; KUNDU, K.; ZHANG, Z.; MA, H.; FIDLER, S.; URTASUN, R. Monocular 3d object detection for autonomous driving. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 2147–2156. Citation on page 48.

CHEN, X.; MA, H.; WAN, J.; LI, B.; XIA, T. **Multi-View 3D Object Detection Network for Autonomous Driving**. 2017. Citation on page 50.

CHEN, X.; ZHANG, T.; WANG, Y.; WANG, Y.; ZHAO, H. **FUTR3D: A Unified Sensor Fusion Framework for 3D Detection**. 2023. Citation on page 56.

CHEN, Z.; LI, Z.; ZHANG, S.; FANG, L.; JIANG, Q.; ZHAO, F.; ZHOU, B.; ZHAO, H. **AutoAlign: Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection**. 2022. Citation on page 58.

CHILD, R.; GRAY, S.; RADFORD, A.; SUTSKEVER, I. **Generating Long Sequences with Sparse Transformers**. 2019. Available: <<https://arxiv.org/abs/1904.10509>>. Citation on page 52.

COSTEA, A. D.; PETROVAI, A.; NEDEVSCHI, S. Fusion scheme for semantic and instance-level segmentation. In: **2018 21st International Conference on Intelligent Transportation Systems (ITSC)**. [S.l.: s.n.], 2018. p. 3469–3475. Citations on pages 13 and 41.

DANZIGER, M. **Autonomous Vehicles Need Thermal Cameras**. 2020. Acessado em: 22/01/2024. Available: <<https://www.foresightauto.com/autonomous-vehicles-need-thermal-cameras/>>. Citations on pages 13 and 31.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citation on page 51.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021. Available: <<https://arxiv.org/abs/2010.11929>>. Citation on page 51.

ELHARROUSS, O.; AL-MAADEED, S.; SUBRAMANIAN, N.; OTTAKATH, N.; AL-MAADEED, N.; HIMEUR, Y. **Panoptic Segmentation: A Review**. 2021. Available: <<https://arxiv.org/abs/2111.10250>>. Citation on page 40.

- ENGELCKE, M.; RAO, D.; WANG, D. Z.; TONG, C. H.; POSNER, I. **Vote3Deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks**. 2017. Citation on page 49.
- ERABATI, G. K.; ARAUJO, H. **MSF3DDETR: Multi-Sensor Fusion 3D Detection Transformer for Autonomous Driving**. 2022. Citation on page 56.
- FAYYAD, J.; JARADAT, M. A.; GRUYER, D.; NAJJARAN, H. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. **Sensors**, v. 20, n. 15, 2020. ISSN 1424-8220. Available: <<https://www.mdpi.com/1424-8220/20/15/4220>>. Citation on page 37.
- FOOLADGAR, F.; KASAEI, S. **Multi-Modal Attention-based Fusion Model for Semantic Segmentation of RGB-Depth Images**. 2019. Available: <<https://arxiv.org/abs/1912.11691>>. Citation on page 64.
- GAUTAM, S.; MATHURIA, T.; MEENA, S. Image segmentation for self-driving car. In: **2022 2nd International Conference on Intelligent Technologies (CONIT)**. [S.l.: s.n.], 2022. p. 1–6. Citations on pages 13 and 24.
- _____. Image segmentation for self-driving car. In: **2022 2nd International Conference on Intelligent Technologies (CONIT)**. [S.l.: s.n.], 2022. p. 1–6. Citation on page 39.
- GRAHAM, B.; ENGELCKE, M.; MAATEN, L. van der. **3D Semantic Segmentation with Submanifold Sparse Convolutional Networks**. 2017. Available: <<https://arxiv.org/abs/1711.10275>>. Citation on page 45.
- GUERRY, J.; BOULCH, A.; SAUX, B. L.; MORAS, J.; PLYER, A.; FILLIAT, D. Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In: **2017 IEEE International Conference on Computer Vision Workshops (ICCVW)**. [S.l.: s.n.], 2017. p. 669–678. Citation on page 45.
- GUPTA, S.; GIRSHICK, R.; ARBELÁEZ, P.; MALIK, J. **Learning Rich Features from RGB-D Images for Object Detection and Segmentation**. 2014. Available: <<https://arxiv.org/abs/1407.5736>>. Citation on page 45.
- HAFIZ, A. M.; BHAT, G. M. A survey on instance segmentation: state of the art. **International Journal of Multimedia Information Retrieval**, Springer Science and Business Media LLC, v. 9, n. 3, p. 171–189, Jul. 2020. ISSN 2192-662X. Available: <<http://dx.doi.org/10.1007/s13735-020-00195-x>>. Citation on page 40.
- HAMDA, N. E. I.; HADJALI, A.; LAGHA, M. Multisensor data fusion in iot environments in dempster–shafer theory setting: An improved evidence distance-based approach. **Sensors**, v. 23, n. 11, 2023. ISSN 1424-8220. Available: <<https://www.mdpi.com/1424-8220/23/11/5141>>. Citation on page 37.
- HAZIRBAS, C.; MA, L.; DOMOKOS, C.; CREMERS, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: LAI, S.-H.; LEPESTIT, V.; NISHINO, K.; SATO, Y. (Ed.). **Computer Vision – ACCV 2016**. Cham: Springer International Publishing, 2017. p. 213–228. ISBN 978-3-319-54181-5. Citation on page 45.
- HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. **Mask R-CNN**. 2018. Available: <<https://arxiv.org/abs/1703.06870>>. Citation on page 44.

HE, K.; ZHANG, X.; REN, S.; SUN, J. **Deep Residual Learning for Image Recognition**. 2015. Citation on page 68.

HE, Y.; YU, H.; LIU, X.; YANG, Z.; SUN, W.; ANWAR, S.; MIAN, A. **Deep Learning Based 3D Segmentation: A Survey**. 2024. Available: <<https://arxiv.org/abs/2103.05423>>. Citations on pages 42 and 44.

HERMANN, K. M.; KOČISKÝ, T.; GREFENSTETTE, E.; ESPEHOLT, L.; KAY, W.; SULEYMAN, M.; BLUNSMON, P. **Teaching Machines to Read and Comprehend**. 2015. Available: <<https://arxiv.org/abs/1506.03340>>. Citation on page 52.

HONDA, H.; UCHIDA, Y. **CLRerNet: Improving Confidence of Lane Detection with LaneIoU**. 2023. Available: <<https://arxiv.org/abs/2305.08366>>. Citations on pages 13 and 25.

HONORATO, E. S.; UCHIDA, M. A. S.; SILVA, T. H. S.; WOLF, D. F. Out-of-distribution object detection in autonomous vehicles with yolo model. In: **2024 Latin American Robotics Symposium (LARS)**. [S.l.: s.n.], 2024. p. 1–6. Citation on page 103.

HONORATO, E. S.; UCHIDA, M. A. S.; TRAINA, A. J. M.; WOLF, D. F. Improving u-net with attention mechanism for medical image segmentation applications. In: **2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)**. [S.l.: s.n.], 2025. p. 1–6. Citation on page 103.

HONORATO, E. S.; WOLF, D. F. Enhancing 3d object detection in autonomous vehicles: Multi-sensor fusion with attention mechanisms. In: **2024 Latin American Robotics Symposium (LARS)**. [S.l.: s.n.], 2024. p. 1–6. Citations on pages 82 and 103.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDRETTI, M.; ADAM, H. **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017. Citation on page 68.

HU, J.; SHEN, L.; ALBANIE, S.; SUN, G.; WU, E. **Squeeze-and-Excitation Networks**. 2019. Available: <<https://arxiv.org/abs/1709.01507>>. Citations on pages 70 and 71.

HU, J. wen; ZHENG, C. W. Bo-yin; ZHAO, C. hui; HOU, X. lei; PAN, Z. X. Q. Asurvey onmulti-sensor fusion based obstacle detection for intelligent ground vehicles in off-road environments. **Frontiers of Information Technology Electronic Engineering**, Front. Inform. Technol. Electron. Eng., v. 21, n. 5, p. 675, 2020. Available: <https://journal.hep.com.cn/ckcest/fitee/EN/abstract/article_27529.shtml>. Citation on page 36.

HURTADO, J. V.; VALADA, A. **Semantic Scene Segmentation for Robotics**. 2024. Available: <<https://arxiv.org/abs/2401.07589>>. Citation on page 39.

KU, J.; MOZIFIAN, M.; LEE, J.; HARAKEH, A.; WASLANDER, S. **Joint 3D Proposal Generation and Object Detection from View Aggregation**. 2018. Citation on page 50.

LI, B. **3D Fully Convolutional Network for Vehicle Detection in Point Cloud**. 2017. Citation on page 49.

LI, B.; ZHANG, T.; XIA, T. **Vehicle Detection from 3D Lidar Using Fully Convolutional Network**. 2016. Citation on page 49.

LI, Y.; CHEN, Y.; QI, X.; LI, Z.; SUN, J.; JIA, J. **Unifying Voxel-based Representation with Transformer for 3D Object Detection**. 2022. Citation on page 61.

- LIANG, X.; SHEN, X.; FENG, J.; LIN, L.; YAN, S. **Semantic Object Parsing with Graph LSTM**. 2016. Available: <<https://arxiv.org/abs/1603.07063>>. Citation on page 44.
- LIN, G.; SHEN, C.; HENGEL, A. van dan; REID, I. **Efficient piecewise training of deep structured models for semantic segmentation**. 2016. Available: <<https://arxiv.org/abs/1504.01013>>. Citation on page 43.
- LIN, T.-Y.; DOLLÁR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. **Feature Pyramid Networks for Object Detection**. 2017. Available: <<https://arxiv.org/abs/1612.03144>>. Citation on page 43.
- LIU, D.; ZHANG, D.; WANG, L.; WANG, J. Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism. **Frontiers in Neuroscience**, v. 17, 2023. ISSN 1662-453X. Available: <<https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1291674>>. Citation on page 39.
- LIU, W.; RABINOVICH, A.; BERG, A. C. **ParseNet: Looking Wider to See Better**. 2015. Available: <<https://arxiv.org/abs/1506.04579>>. Citation on page 43.
- LIU, Y.; SHAO, Z.; HOFFMANN, N. **Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions**. 2021. Available: <<https://arxiv.org/abs/2112.05561>>. Citation on page 51.
- LIU, Z.; LI, X.; LUO, P.; LOY, C. C.; TANG, X. **Semantic Image Segmentation via Deep Parsing Network**. 2015. Available: <<https://arxiv.org/abs/1509.02634>>. Citation on page 43.
- LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**. 2021. Citation on page 67.
- LIU, Z.; TANG, H.; AMINI, A.; YANG, X.; MAO, H.; RUS, D.; HAN, S. **BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation**. 2022. Citations on pages 26, 53, 54, and 67.
- LONG, J.; SHELHAMER, E.; DARRELL, T. **Fully Convolutional Networks for Semantic Segmentation**. 2015. Available: <<https://arxiv.org/abs/1411.4038>>. Citation on page 43.
- LOUPPE, G. **Understanding Random Forests: From Theory to Practice**. 2015. Citation on page 37.
- LUC, P.; COUPRIE, C.; CHINTALA, S.; VERBEEK, J. **Semantic Segmentation using Adversarial Networks**. 2016. Available: <<https://arxiv.org/abs/1611.08408>>. Citation on page 44.
- MAO, J.; SHI, S.; WANG, X.; LI, H. **3D Object Detection for Autonomous Driving: A Comprehensive Survey**. 2023. Citation on page 47.
- MINAEE, S.; BOYKOV, Y.; PORIKLI, F.; PLAZA, A.; KEHTARNAVAZ, N.; TERZOPOULOS, D. **Image Segmentation Using Deep Learning: A Survey**. 2020. Available: <<https://arxiv.org/abs/2001.05566>>. Citations on pages 41 and 42.
- MNIH, V.; HEESS, N.; GRAVES, A.; KAVUKCUOGLU, K. **Recurrent Models of Visual Attention**. 2014. Available: <<https://arxiv.org/abs/1406.6247>>. Citation on page 52.

QI, C. R.; SU, H.; MO, K.; GUIBAS, L. J. **PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation**. 2017. Available: <<https://arxiv.org/abs/1612.00593>>. Citations on pages 46 and 49.

QI, C. R.; YI, L.; SU, H.; GUIBAS, L. J. **PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space**. 2017. Available: <<https://arxiv.org/abs/1706.02413>>. Citations on pages 46 and 49.

QI, X.; LIAO, R.; JIA, J.; FIDLER, S.; URTASUN, R. 3d graph neural networks for rgbd semantic segmentation. In: **2017 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. p. 5209–5218. Citation on page 45.

QIAN, R.; LAI, X.; LI, X. 3d object detection for autonomous driving: A survey. **Pattern Recognition**, Elsevier BV, v. 130, p. 108796, Oct. 2022. ISSN 0031-3203. Available: <<http://dx.doi.org/10.1016/j.patcog.2022.108796>>. Citations on pages 46 and 47.

RAJ, A.; MATURANA, D.; SCHERER, S. **Multi-Scale Convolutional Architecture for Semantic Segmentation**. Pittsburgh, PA, 2015. Citation on page 45.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. 2016. Available: <<https://arxiv.org/abs/1506.01497>>. Citations on pages 44 and 49.

ROBOSENSE Tech Show. 2022. Accessed in: 22/01/2024. Available: <<https://www.robosense.ai/en/tech-show-93>>. Citations on pages 13 and 33.

RONNEBERGER, O.; FISCHER, P.; BROX, T. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. 2015. Available: <<https://arxiv.org/abs/1505.04597>>. Citation on page 43.

ROTH, H.; SCHILLING, K. Sensor data fusion for control of mobile robots using fuzzy logic. **IFAC Proceedings Volumes**, v. 28, n. 24, p. 317–322, 1995. ISSN 1474-6670. 3rd IFAC/IFIP/IFORS Workshop on Intelligent Manufacturing Systems 1995 (IMS '95), Bucharest, Romania, 24-26 October. Available: <<https://www.sciencedirect.com/science/article/pii/S1474667017465694>>. Citation on page 37.

ROY, A. G.; NAVAB, N.; WACHINGER, C. **Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks**. 2018. Available: <<https://arxiv.org/abs/1803.02579>>. Citation on page 71.

SEVAK, J. S.; KAPADIA, A. D.; CHAVDA, J. B.; SHAH, A.; RAHEVAR, M. Survey on semantic image segmentation techniques. In: **2017 International Conference on Intelligent Sustainable Systems (ICISS)**. [S.l.: s.n.], 2017. p. 306–313. Citation on page 40.

SIMON, M.; MILZ, S.; AMENDE, K.; GROSS, H.-M. **Complex-YOLO: Real-time 3D Object Detection on Point Clouds**. 2018. Citation on page 49.

SINDAGI, V. A.; ZHOU, Y.; TUZEL, O. **MVX-Net: Multimodal VoxelNet for 3D Object Detection**. 2019. Available: <<https://arxiv.org/abs/1904.01649>>. Citations on pages 13 and 24.

SMAILI, C.; NAJJAR, M. E. E.; CHARPILLET, F. Multi-sensor fusion method using dynamic bayesian network for precise vehicle localization and road matching. In: **19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)**. [S.l.: s.n.], 2007. v. 1, p. 146–151. Citation on page 37.

- SRIVASTAV, A.; MANDAL, S. **Radars for Autonomous Driving: A Review of Deep Learning Methods and Challenges**. 2023. Citation on page 34.
- SYKACEK, P.; REZEK, I.; ROBERTS, S. Markov chain monte carlo methods for bayesian sensor fusion. **Dept. Eng. Sci., Univ. Oxford, Oxford, UK, Tech. Rep. PARG-00-10.[Online]**. Available: <http://www.robots.ox.ac.uk/~parg>, 2000. Citation on page 36.
- TAN, M.; ZHUANG, Z.; CHEN, S.; LI, R.; JIA, K.; WANG, Q.; LI, Y. Epmf: Efficient perception-aware multi-sensor fusion for 3d semantic segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 46, n. 12, p. 8258–8273, 2024. Citation on page 63.
- THOMAS, H.; TSAI, Y.-H. H.; BARFOOT, T. D.; ZHANG, J. **KPConvX: Modernizing Kernel Point Convolution with Kernel Attention**. 2024. Available: <<https://arxiv.org/abs/2405.13194>>. Citation on page 46.
- UDACITY. **Self-Driving Car Fundamentals: Featuring Apollo**. 2025. Acessado em: 22/02/2025. Available: <<https://www.udacity.com/course/self-driving-car-fundamentals-featuring-apollo--ud0419>>. Citations on pages 13 and 34.
- VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, ; POLOSUKHIN, I. Attention is all you need. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. Citations on pages 26, 51, and 55.
- VEMULA, M.; DJURIC, P. Multisensor fusion for target tracking using sequential monte carlo methods. In: **IEEE/SP 13th Workshop on Statistical Signal Processing, 2005**. [S.l.: s.n.], 2005. p. 1304–1309. Citation on page 36.
- VISIN, F.; ROMERO, A.; CHO, K.; MATTEUCCI, M.; CICCONE, M.; KASTNER, K.; BEN-GIO, Y.; COURVILLE, A. Reseg: A recurrent neural network-based model for semantic segmentation. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2016. p. 426–433. Citation on page 44.
- WANG, J.; ZHU, M.; WANG, B.; SUN, D.; WEI, H.; LIU, C.; NIE, H. Kda3d: Key-point densification and multi-attention guidance for 3d object detection. **Remote Sensing**, v. 12, n. 11, 2020. ISSN 2072-4292. Available: <<https://www.mdpi.com/2072-4292/12/11/1895>>. Citation on page 60.
- WANG, R.; LEI, T.; CUI, R.; ZHANG, B.; MENG, H.; NANDI, A. K. Medical image segmentation using deep learning: A survey. **IET Image Processing**, Institution of Engineering and Technology (IET), v. 16, n. 5, p. 1243–1267, Jan. 2022. ISSN 1751-9667. Available: <<http://dx.doi.org/10.1049/ijpr2.12419>>. Citation on page 39.
- WANG, S.; CAESAR, H.; NAN, L.; KOOLIJ, J. F. P. **UniBEV: Multi-modal 3D Object Detection with Uniform BEV Encoders for Robustness against Missing Sensor Modalities**. 2023. Citation on page 61.
- WOO, S.; PARK, J.; LEE, J.-Y.; KWEON, I. S. **CBAM: Convolutional Block Attention Module**. 2018. Citations on pages 13, 72, 73, 74, and 75.
- WU, Z.; WANG, M.; SUN, W.; LI, Y.; XU, T.; WANG, F.; HUANG, K. Cat: Learning to collaborate channel and spatial attention from multi-information fusion. **IET Computer Vision**, v. 17, n. 3, p. 309–318, 2023. Available: <<https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12166>>. Citations on pages 13 and 51.

XIANG, Y.; CHOI, W.; LIN, Y.; SAVARESE, S. Data-driven 3d voxel patterns for object category recognition. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 1903–1911. Citation on page 48.

XIE, Y.; XU, C.; RAKOTOSAONA, M.-J.; RIM, P.; TOMBARI, F.; KEUTZER, K.; TOMIZUKA, M.; ZHAN, W. **SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection**. 2023. Citation on page 60.

XU, C.; WU, B.; WANG, Z.; ZHAN, W.; VAJDA, P.; KEUTZER, K.; TOMIZUKA, M. **SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation**. 2021. Available: <<https://arxiv.org/abs/2004.01803>>. Citation on page 45.

XU, J.; LU, K.; WANG, H. Attention fusion network for multi-spectral semantic segmentation. **Pattern Recognition Letters**, v. 146, p. 179–184, 2021. ISSN 0167-8655. Available: <<https://www.sciencedirect.com/science/article/pii/S0167865521001021>>. Citation on page 64.

YAN, J.; LIU, Y.; SUN, J.; JIA, F.; LI, S.; WANG, T.; ZHANG, X. **Cross Modal Transformer: Towards Fast and Robust 3D Object Detection**. 2023. Citation on page 57.

YAN, J.; ZHAO, H.; BU, P.; JIN, Y. **Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation**. 2021. Citation on page 72.

YANG, Z.; CHEN, J.; MIAO, Z.; LI, W.; ZHU, X.; ZHANG, L. **DeepInteraction: 3D Object Detection via Modality Interaction**. 2022. Citation on page 58.

YU, S.-L.; WESTFECHTEL, T.; HAMADA, R.; OHNO, K.; TADOKORO, S. 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)**. [S.l.: s.n.], 2017. p. 102–109. Citation on page 49.

ZENG, Y.; ZHANG, D.; WANG, C.; MIAO, Z.; LIU, T.; ZHAN, X.; HAO, D.; MA, C. Lift: Learning 4d lidar image fusion transformer for 3d object detection. In: **2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2022. p. 17151–17160. Citation on page 59.

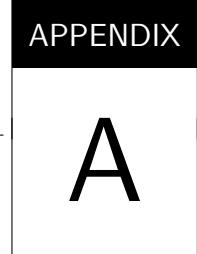
ZHANG, Y.; LIU, K.; BAO, H.; QIAN, X.; WANG, Z.; YE, S.; WANG, W. Aftr: A robustness multi-sensor fusion model for 3d object detection based on adaptive fusion transformer. **Sensors**, v. 23, n. 20, 2023. ISSN 1424-8220. Available: <<https://www.mdpi.com/1424-8220/23/20/8400>>. Citation on page 55.

ZHANG, Z.; LIU, Q.; WANG, Y. Road extraction by deep residual u-net. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 15, n. 5, p. 749–753, May 2018. ISSN 1558-0571. Available: <<http://dx.doi.org/10.1109/LGRS.2018.2802944>>. Citation on page 43.

ZHANG, Z.; WANG, W.; ZHU, L.; TANG, Z. Tag-fusion: Two-stage attention guided multi-modal fusion network for semantic segmentation. **Digital Signal Processing**, v. 156, p. 104807, 2025. ISSN 1051-2004. Available: <<https://www.sciencedirect.com/science/article/pii/S1051200424004329>>. Citation on page 62.

ZHAO, H.; SHI, J.; QI, X.; WANG, X.; JIA, J. **Pyramid Scene Parsing Network**. 2017. Available: <<https://arxiv.org/abs/1612.01105>>. Citation on page 43.

- ZHENG, S.; JAYASUMANA, S.; ROMERA-PAREDES, B.; VINEET, V.; SU, Z.; DU, D.; HUANG, C.; TORR, P. H. S. Conditional random fields as recurrent neural networks. In: **2015 IEEE International Conference on Computer Vision (ICCV)**. IEEE, 2015. Available: <<http://dx.doi.org/10.1109/ICCV.2015.179>>. Citation on page 43.
- ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH, N.; LIANG, J. **UNet++: A Nested U-Net Architecture for Medical Image Segmentation**. 2018. Available: <<https://arxiv.org/abs/1807.10165>>. Citation on page 43.
- ZHU, X.; SU, W.; LU, L.; LI, B.; WANG, X.; DAI, J. **Deformable DETR: Deformable Transformers for End-to-End Object Detection**. 2021. Available: <<https://arxiv.org/abs/2010.04159>>. Citation on page 52.
- ZHU, X.; WANG, X.; SHI, Y.; REN, S.; WANG, W. Channel-wise attention mechanism in the 3d convolutional network for lung nodule detection. **Electronics**, v. 11, n. 10, 2022. ISSN 2079-9292. Available: <<https://www.mdpi.com/2079-9292/11/10/1600>>. Citation on page 72.
- ZHUANG, Z.; LI, R.; JIA, K.; WANG, Q.; LI, Y.; TAN, M. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: **2021 IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2021. p. 16260–16270. Citation on page 63.
- ÇIÇEK ÖZGÜN; ABDULKADIR, A.; LIENKAMP, S. S.; BROX, T.; RONNEBERGER, O. **3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation**. 2016. Available: <<https://arxiv.org/abs/1606.06650>>. Citation on page 43.



PUBLISHED WORKS

During the course of this master's research, four papers were produced, three of which were accepted for publication.

The first paper, "Enhancing 3D Object Detection in Autonomous Vehicles: Multi-Sensor Fusion with Attention Mechanisms" ([HONORATO; WOLF, 2024](#)), presents results obtained using the original BEVFusion backbone combined with attention mechanisms that demand greater computational resources. These results were excluded from the final version of this thesis due to the extended training time required.

The second paper, "Out-of-Distribution Object Detection in Autonomous Vehicles with YOLO Model" ([HONORATO *et al.*, 2024](#)), also focuses on autonomous vehicles but is not directly related to the main topic of this thesis. It addresses Out-of-Distribution (OOD) detection in object recognition, driven by a curiosity to explore a relatively underexplored area with strong potential to improve autonomous perception systems.

The third paper, "Improving U-Net with Attention Mechanism for Medical Image Segmentation Applications" ([HONORATO *et al.*, 2025](#)), investigates the application of the same attention mechanisms employed in this work to enhance segmentation performance in medical imaging tasks. This paper was recognized with the Best Student Paper Award at the conference.

The details and abstracts of the three published works are presented below:

Enhancing 3D Object Detection in Autonomous Vehicles: Multi-Sensor Fusion with Attention Mechanisms

Authors: Eduardo Sperle Honorato; Denis Fernando Wolf

Published in: *2024 Latin American Robotics Symposium (LARS)*

Abstract: In the realm of Autonomous Vehicles (AVs), effective 3D object detection is

paramount for ensuring safe navigation in complex environments. The integration of data from multiple sensors, such as cameras and LiDAR, presents challenges in accurately perceiving the surrounding environment. In this paper, we propose several enhancements to the BEVFusion model, a state-of-the-art approach for fusing camera and LiDAR data for 3D object detection in AVs. Specifically, we investigate the integration of attention mechanisms to improve sensor fusion within the BEVFusion framework. Through extensive experiments on the nuScenes and nuScenes mini datasets, the best-performing model from our proposed approaches achieved a relative improvement of 1.2% in mAP and 0.6% in NDS compared to the baseline model. These findings highlight the effectiveness of our attention-based fusion strategy in enhancing detection accuracy, making it a robust solution for real-world autonomous driving scenarios.

Out-of-Distribution Object Detection in Autonomous Vehicles With YOLO Model

Authors: Eduardo Sperle Honorato; Mariana Aya Suzuki Uchida; Thiago Henrique Segreto Silva; Denis Fernando Wolf

Published in: *2024 Latin American Robotics Symposium (LARS)*

Abstract: This paper addresses the challenge of detecting Out-of-Distribution (OOD) objects in autonomous vehicles, focusing on identifying and localizing objects absent from the training data. We propose a novel method that leverages the existing object detection model, YOLOv5, to detect OOD instances without requiring model retraining or additional datasets. Our approach computes dissimilarity scores from class confidence outputs to effectively distinguish OOD objects in cropped images. Experiments on popular autonomous vehicle 2D object detection datasets demonstrate that, in a straightforward scenario, our method significantly reduces the False Positive Rate at 95% True Positive Rate while maintaining a comparable Area Under the Receiver Operating Characteristic curve (AUROC) to baseline models. In more challenging scenarios, our method outperforms competitors, demonstrating superior robustness. Key contributions include the proposed OOD detection method and the methodology for identifying OOD object instances.

Improving U-Net with Attention Mechanism for Medical Image Segmentation Applications

Authors: Eduardo Sperle Honorato; Mariana Aya Suzuki Uchida; Agma Juci Machado Traina; Denis Ferando Wolf **Abstract:** Medical image segmentation plays a vital role in numerous applications and has gained significant attention since the introduction of the U-Net model, which enabled convolutional neural networks to achieve high performance with manageable

computational costs. Recently, attention mechanisms have emerged as a promising approach to enhance model performance by emphasizing relevant features while suppressing irrelevant ones. This study explores the integration of channel and spatial attention mechanisms into the U-Net architecture, evaluating their impact on segmentation performance and computational cost. Experiments conducted on six public medical imaging datasets demonstrated performance improvements, with Intersection over Union (IoU) gains ranging from 1.62% to 33.66% compared to the original U-Net. These results highlight the potential of attention mechanisms to significantly improve the efficiency and effectiveness of medical image segmentation models.

