Universidade de Aveiro
2023

Lucas
Rodrigues Dal'Col

Multi-Modal End-to-End Holistic Perception for Autonomous Driving

Perceção Holística Multimodal Ponta-a-Ponta para Condução Autónoma

Universidade de Aveiro
2023

Lucas Rodrigues Dal'Col

Multi-Modal End-to-End Holistic Perception for Autonomous Driving

Perceção Holística Multimodal Ponta-a-Ponta para Condução Autónoma

Proposta de tese apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Mecânica, realizado sob orientação cientıfica do Doutor Miguel Armando Riem de Oliveira, Professor Auxiliar do Departamento de Engenharia Mecânica da Universidade de Aveiro, e do Doutor Vítor Manuel Ferreira dos Santos, Professor Associado com Agregação do Departamento de Engenharia Mecânica da Universidade de Aveiro.

O júri / The jury

Presidente / President

Vogais / Committee

Prof. Doutor Miguel Armando Riem de Oliveira
Professor Auxiliar da Universidade de Aveiro (orientador)

Prof. Doutor Vítor Manuel Ferreira dos Santos
Professor Associado com Agregação da Universidade de Aveiro (co-orientador)

Agradecimentos /
Acknowledgements

Keywords Motion Prediction, Holistic Perception, End-To-End Perception, Autonomous Driving

Abstract Autonomous driving systems require accurate perception algorithms to navigate safely through traffic. End-to-end perception approaches jointly learn perception tasks, transforming raw sensor data directly into object detection with motion predictions. To improve trajectory predictions, recent approaches attempted to model interactions between actors (dynamic objects), whose behavior depends on each other and their interplay with the scene. However, current models only consider interactions within the same object class. Modeling interactions between actors from different classes improves perception accuracy and enables a more holistic understanding of the scene by preventing misinterpretation of object behavior. In this context, this Ph.D. proposal aims to develop a multi-modal end-to-end holistic perception approach capable of modeling both inter-class and intra-class interactions between actors, and their interplay with the scene. Perception systems still face challenges in achieving high accuracy and robustness, with a complete holistic perception of the scene yet to be achieved despite recent progress.

Palavras-chave Previsão de Trajetórias, Perceção Holística, Perceção Ponta-a-Ponta, Condução Autónoma

Resumo Os sistemas de condução autónoma requerem algoritmos de perceção precisos para garantir a segurança da navegação autónoma no trânsito. As abordagens de perceção ponta-a-ponta aprendem tarefas de perceção em conjunto, transformando dados brutos do sensor diretamente na deteção de objetos com previsões de trajetória. Para melhorar as previsões de trajetória, abordagens recentes tentaram modelar interações entre atores (objetos dinâmicos), cujo comportamento depende um do outro e da sua interação com a cena. No entanto, os modelos atuais consideram apenas interações dentro da mesma classe de objeto. A modelação de interações entre atores de diferentes classes melhora a precisão da perceção e permite uma compreensão mais holística da cena, evitando a má interpretação do comportamento do objeto. Neste contexto, esta proposta de doutoramento visa desenvolver uma abordagem multimodal de perceção holística ponta-a-ponta, capaz de modelar interações entre atores da mesma classe e de diferentes classes, e também a sua interação com a cena. Os sistemas de perceção ainda enfrentam desafios para alcançar alta precisão e robustez, com uma perceção holística completa da cena ainda por atingir, apesar do progresso recente.

Contents

Intentionally blank page.

List of Tables

Intentionally blank page.

List of Figures

Intentionally blank page.

Chapter 1

Introduction

Autonomous vehicles are becoming a reality in our society, and research and development in autonomous driving are growing significantly. These vehicles hold the promise of preventing accidents, reducing traffic congestion and emissions, transporting people and goods, and reducing driving-related stress [1]. However, there are social concerns that need to be addressed before autonomous driving can be fully accepted by society. Governments must implement rules and regulations to ensure the safety of users. Additionally, the main responsibility for autonomous vehicles needs to be defined, and confidence and trust in these systems must be established [2]. Despite the great potential of autonomous driving in terms of social and technological impact, these systems are still far from being fully developed, as there are still many unsolved challenges. This is due to the complexity of autonomous vehicles, which require the ability to perceive, predict, decide, plan, and control the car in complex and uncontrolled environments.

The Society of Automotive Engineers (SAE) has defined six levels of driving automation to categorize autonomous systems [3]. These levels range from Level 0 (no driving automation) to Level 5 (fully driving automation). Figure 1.1 illustrates this categorization. In 2023, Level 2 automation is prevalent in brand new cars from all manufacturers, incorporating driver support features like lane centering and adaptive cruise control simultaneously. Companies such as Tesla and Volvo have developed Level 3 automation, which includes features such as traffic jam chauffeur but still requires a human driver to take control when necessary. Level 4 automation represents the emergence of robot taxis, a technological breakthrough achieved by companies such as Waymo, Cruise, and Motional. Robot taxis operate without a human driver, and the presence of pedals and steering wheels may vary. However, Level 5 automation, where complete autonomy is achieved under any environmental condition without any human intervention, has not yet been achieved. In conclusion, Level 4 robot taxis are starting to become a reality, and research and development in autonomous driving are currently focused at this level, although it is not yet fully established and accessible to the general public. Therefore, it is evident that autonomous driving still faces numerous unsolved challenges.

Regarding the approaches used in autonomous driving, there are two main methods: the modular approach and the end-to-end approach [4]. Figure 1.2 depicts the architecture of these two approaches. The modular approach is usually summarized in six core modules: object detection/tracking and motion prediction (perception), localization, assessment, behavior prediction, planning, and control. The major advantage of this approach is that these individual modules divide the challenging task of autonomous

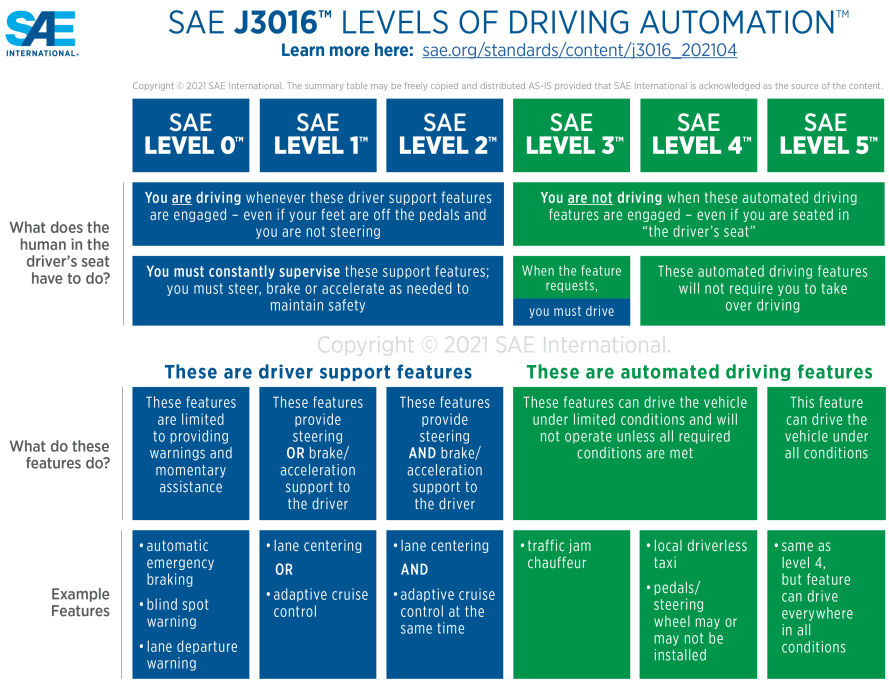

Figure 1.1: Levels of driving automation according to the SAE [3].

driving into an easier-to-solve set of problems. Recently, there has been a surge of interest in end-to-end approaches, which involve a single module generating continuous control of the steering wheel and pedals directly from sensory input.

Autonomous driving systems are usually developed using a modular approach [5]. Among its components, the perception module plays a critical role in accurately assessing the vehicle's surroundings to ensure safe navigation through traffic [2]. The perception module comprises three main tasks: object detection, object tracking, and motion prediction [6]. Object detection aims to identify objects of interest in the environment, object tracking involves tracking these objects over time, and motion prediction is designed to anticipate the movement of these objects to avoid collisions. These systems need to be accurate, robust and operate in real-time [7]. Accuracy refers to providing precise information about the driving environment, while robustness refers to effective functioning in adverse weather conditions or when sensors degrade. Real-time operation is crucial, especially when driving at high speeds. To achieve these goals, autonomous vehicles are equipped with sensors of different modalities, including cameras, LiDARs, Radars, and others.

The remainder of this Ph.D. proposal is organized as follows: Chapter 2 provides background information on perception for autonomous driving; Chapter 3 offers an overview of the state of the art in end-to-end perception for autonomous driving, identifying the research gap that this Ph.D. proposal aims to address; Chapter 4 presents the main objective of this Ph.D. proposal, along with corresponding research questions and specific objectives; Chapter 5 specifies the tasks related to this Ph.D. proposal, as well as the acknowledged risks and respective contingency plans. Finally, Chapter 6 outlines the hosting conditions for this Ph.D. proposal.

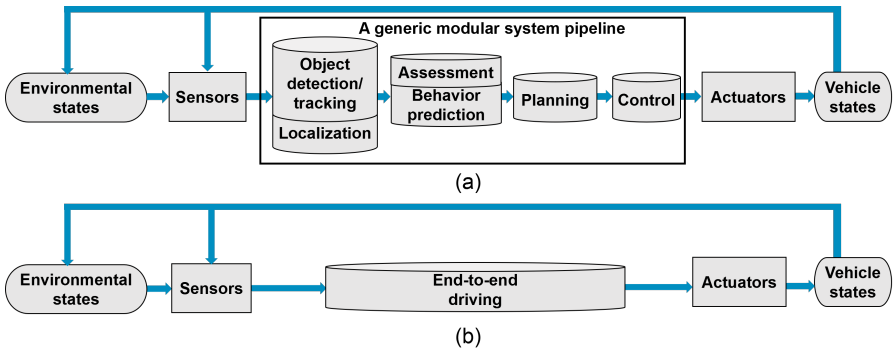

Figure 1.2: Architecture of the main approaches of autonomous driving: (a) modular approach and (b) end-to-end approach [4].

Intentionally blank page.

Chapter 2

Background

This chapter provides background information on perception for autonomous driving. Firstly, we briefly summarize typical sensor modalities used in autonomous vehicles and their functioning. Then, we present the three most commonly used datasets for benchmarking perception state-of-the-art algorithms.

2.1 Sensors

The sensors most commonly employed to perceive the surroundings of an autonomous vehicle are RGB Cameras, LiDARs, and Radars. These sensors fall under the category of exteroceptive sensors, which means they perceive the environment around them. Autonomous vehicles use several sensor modalities to enhance redundancy, which increases robustness and reliability. Sensors are generally classified as active or passive. Active sensors emit signals for measurements, while passive sensors do not emit any signals. Figure 2.1 illustrates examples of sensors used for perception, along with some applications for each sensor.

RGB Cameras

RGB cameras capture images that provide detailed texture information about the scene in a passive manner. However, this sensor modality is sensitive to lighting and weather conditions and does not directly provide depth information. RGB cameras are primarily used for 2D computer vision tasks such as object detection and semantic segmentation.

LiDARs

LiDAR is an active sensor that measures the distance between the sensor and the objects surrounding the vehicle. It emits laser beams and calculates the time taken for the reflected light to return to the receiver. 3D LiDARs generate 3D point clouds, which consist of a set of 3D points that provide precise depth information. However, LiDARs encounter challenges such as occlusions, sparsity, and sensitivity to weather conditions like fog and snow. This sensor modality is predominantly used for 3D computer vision tasks like object detection and localization.

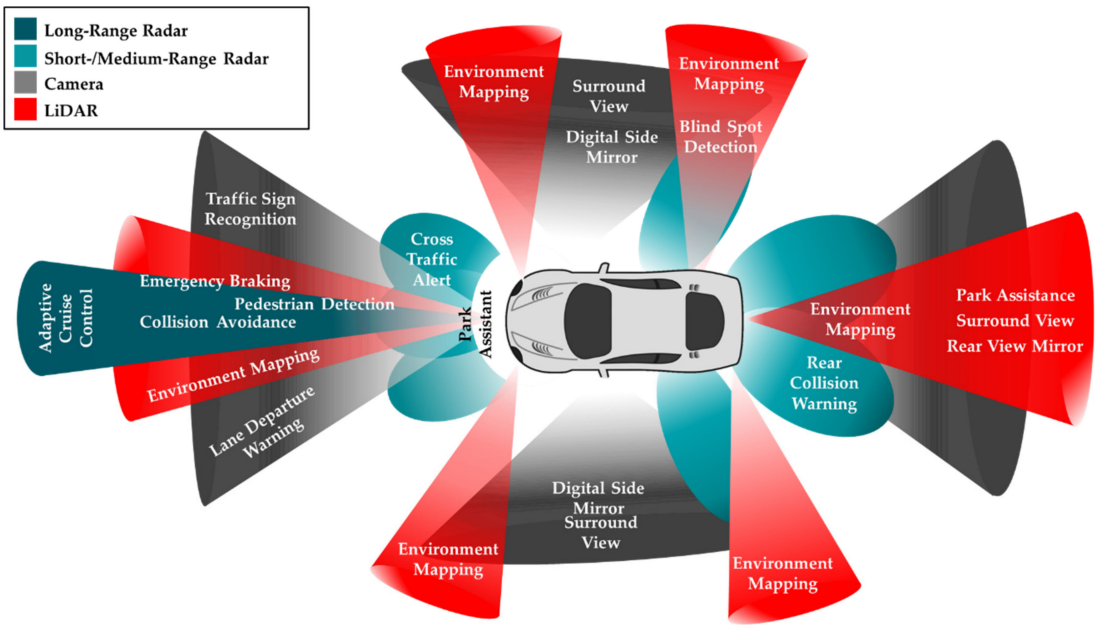

Figure 2.1: Example of sensors used for perception in autonomous driving, with their coverage and applications [8].

Radars

Radar is an active sensor that emits radio waves towards an obstacle and measures the time it takes for the reflected signal to return. It also estimates the radial velocity of the object using the Doppler effect. Radars are robust to various lighting and weather conditions. However, they face challenges in classifying objects due to their low resolution. Radars are commonly applied in adaptive cruise control and traffic jam assistance systems.

Sensor Fusion

As described above, each sensor modality has its own advantages and disadvantages. Fuse the information from these sensors can mitigate these disadvantages by leveraging high sensor redundancy. This fusion is called sensor fusion and is crucial for various perception tasks as it ensures robustness and reliability. In other words, sensor fusion combines the information acquired from each sensor to reduce detection uncertainties and overcome the limitations of individual sensors when operating independently [8]. The fusion between RGB camera and LiDAR is the most frequently used combination of sensor modalities. Point clouds from LiDAR sensors provide precise depth information but suffer from occlusions, sparsity, and noise, while RGB cameras offer color and high-resolution data but lack depth information. Table 2.1 summarizes the strengths and weaknesses of each sensor modality for several important factors in perception for autonomous driving [8].

Table 2.1: A comparison between sensors employed in autonomous vehicles: Camera, LiDAR, and Radar. The factors are important characteristics for perception in autonomous driving. The "✓" symbol indicates that the sensor operates competently under the specific factor. The "∼" symbol indicates that the sensor performs reasonably well under the specific factor. The "×" symbol indicates that the sensor does not operate well under the specific factor relative to the other sensors [8].

Factors	**Camera**	**LiDAR**	**Radar**	**Fusion**
Range	∼	∼	✓	✓
Resolution	✓	∼	×	✓
Distance Accuracy	∼	✓	✓	✓
Velocity	∼	×	✓	✓
Color Perception	✓	×	×	✓
Object Detection	∼	✓	✓	✓
Object Classification	✓	∼	×	✓
Lane Detection	✓	×	×	✓
Obstacle Edge Detection	✓	✓	×	✓
Illumination Conditions	×	✓	✓	✓
Weather Conditions	×	∼	✓	✓

2.2 Datasets

Several large-scale datasets are available to benchmark state-of-the-art algorithms for perception in autonomous driving. These datasets are continuously updated to enhance diversity and size. The three most commonly used datasets in the scientific community are KITTI [9], nuScenes [10] and Waymo Open Dataset [11]. In the following lines, we introduce these three datasets and provide information on their diversity and size.

KITTI

The KITTI dataset, introduced in 2012, serves as a benchmark for various perception tasks, including stereo, optical flow, visual odometry/SLAM, and 3D object detection. This dataset is the pioneer dataset for benchmark on perception. The KITTI autonomous vehicle is equipped with four high-resolution video cameras, a Velodyne laser scanner, and a localization system. This dataset contains annotations of 200K bounding boxes across 15K frames, with 7481 training samples and 7518 test samples. It also includes corresponding point clouds, comprising a total of 80K labeled objects. The dataset offers 50 scenes across eight classes in Karlsruhe - Germany. The benchmark uses three difficulty levels (easy, moderate, and hard) based on the height of 2D bounding boxes, occlusion level, and truncation. KITTI has significantly influenced the scientific community in data acquisition and benchmark. However, it was recorded only during sunny days, resulting in limited diversity in terms of lighting and weather conditions. Figure 2.2 presents an example of the KITTI dataset.

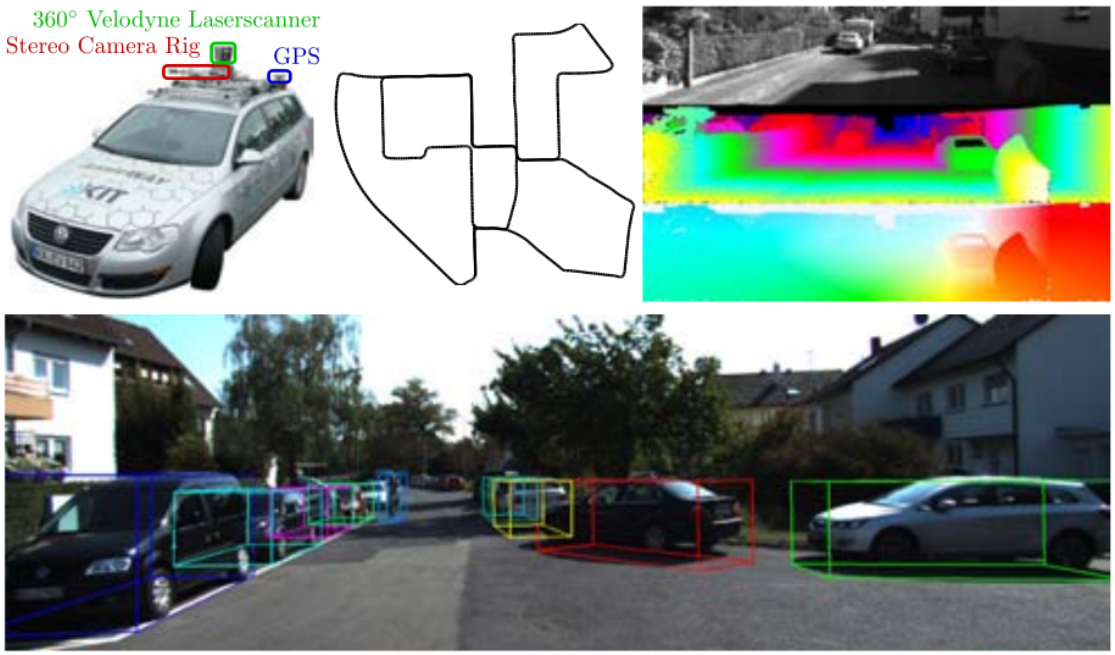

Figure 2.2: Example of the KITTI dataset with sensors from their autonomous vehicle (top-left), trajectory from their visual odometry benchmark (top-center), disparity and optical flow map (top-right) and 3D object labels (bottom) [9].

nuScenes

The nuScenes dataset, introduced in 2019 by Motional, is a large-scale dataset with 3D object annotations. It is the first dataset with full multimodality sensor suite, including one LiDAR, five Radars, and six Cameras. The dataset includes 1000 scenes, each lasting 20 seconds, captured in Boston - USA, and Singapore, comprising 1.4M camera images and 390K LiDAR sweeps. The dataset provides 1.4M manually annotated 3D bounding boxes for 23 object classes and 1.1B manually annotated LiDAR points for 32 classes. In terms of diversity, nuScenes includes scenes with various weather conditions (rainy, foggy, snowy, etc.) and lighting conditions (daytime and nightime). Figure 2.3 shows an example of the nuScenes dataset.

Waymo Open Dataset

The Waymo Open Dataset, introduced in 2020 by Waymo, is the largest dataset for perception in autonomous driving. Waymo's autonomous vehicles are equipped with one mid-range LiDAR, four short-range LiDARs, and five Cameras. This dataset contains synchronized LiDAR and Camera data for 2030 scenes, each lasting 20 seconds, captured in Phoenix - USA, and San Francisco - USA. It includes annotations of 112M bounding boxes across 200K frames for four object classes. Similar to nuScenes, the Waymo Open Dataset offers scenes with diverse weather and lighting conditions. Figure 2.4 illustrates an example of the Waymo Open Dataset.

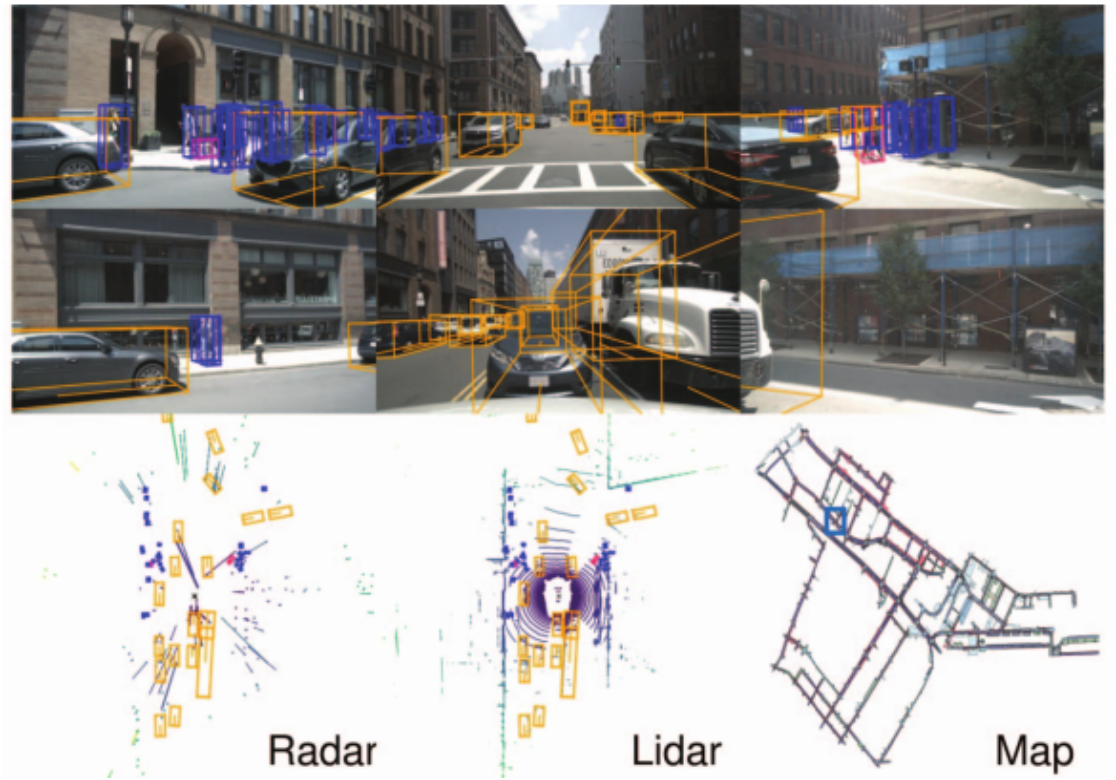

Figure 2.3: Example of the nuScenes dataset. There are 6 different camera views (top), lidar (bottom-center) and radar (bottom-left) data and the human annotated semantic map (bottom-right) [10].

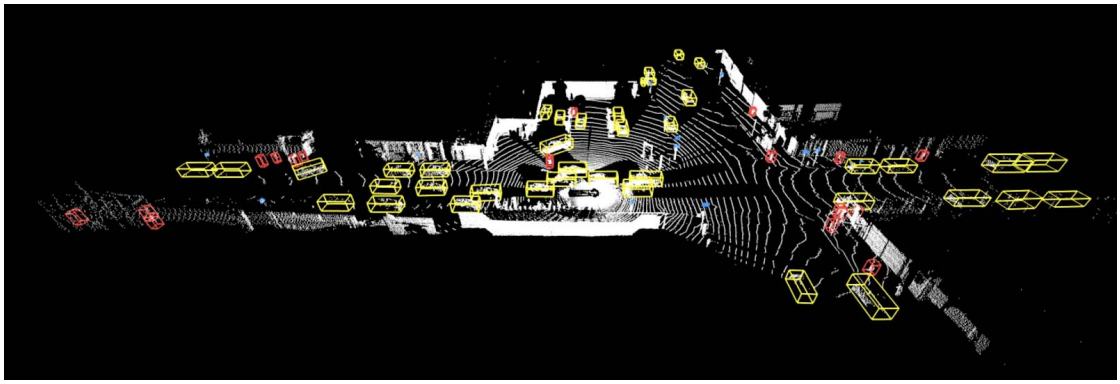

Figure 2.4: Example of Waymo Open Dataset for LiDAR labelling [11].

Intentionally blank page.

Chapter 3

State of the Art

The perception module of autonomous driving is divided into three tasks: object detection, object tracking, and motion prediction [6], as discussed in Chapter 1. These tasks are typically learned independently and executed sequentially, with each task using the output of the previous task as input. Such strategy makes each task easier to design and implement. However, computation is not shared among these tasks, sacrificing the potential advantage of joint optimization. Furthermore, uncertainty is rarely propagated, leading to information loss between these tasks [12]. Recently, end-to-end perception approaches have been introduced to jointly learn object detection, object tracking, and motion prediction within a single neural network [6], [12]–[19]. These approaches share computation among all tasks, enabling highly efficient algorithms for real-time operation, which is crucial for autonomous driving systems with low latency requirements. This is extremely important because high latency can be fatal in autonomous driving systems. Additionally, these approaches have the advantage of direct access to raw data among all tasks, enabling the object tracking and motion prediction tasks to leverage the object detection task through shared knowledge. In this way, end-to-end perception approaches reduce the detection of false negatives when dealing with occluded and far away objects, and the detection of false positives by accumulating evidence over time [13].

Nevertheless, most existing approaches predict the trajectory of individual actors (dynamic objects) independently, based solely on their past trajectory, without considering the interactions between actors. This strategy undermines the accuracy of the predicted trajectories, as the evolution of a trajectory is highly dependent on the behavior of other actors and environmental factors, such as lanes, traffic signs, and traffic states. For instance, a driver adjusts the vehicle's speed to maintain a safe distance from the vehicle ahead, yields to others with the right of way at intersections, and slows down when a traffic light turns yellow [20]. To accurately predict the motion of a vehicle on the road, it is necessary to not only understand its past motion but also analyze the interactions between actors and consider the surrounding context. This is particularly challenging in autonomous driving, as autonomous vehicles will coexist with human drivers for many years, who can perform a very diverse set of maneuvers, including compromising behaviors [21].

Several end-to-end perception approaches have attempted to address the challenge of actor interactions and human intention prediction. IntentNet [20] was the first to develop this joint approach, outperforming state-of-the-art end-to-end perception approaches and

demonstrating the importance of modeling these interactions. Their system receives as input point clouds and high-definition maps (HD maps) to predict discrete high-level behavior and continuous long-term trajectories. Their system is based on a fully convolutional neural network that outputs detection scores for vehicle and background classes, high-level action probabilities corresponding to discrete intention (e.g. keep lane, turn left, turn right, right change lane, left change lane, among others), and bounding boxes in the current and future frames representing the intended motion. SpAGNN [21] improved the state-of-the-art by exploiting graph neural networks (GNNs) to model the interaction between vehicles in the scene. As input, their system receives point clouds and HD maps. They use a convolutional neural network (CNN) to perform object detection and estimate their initial states, and then use a GNN to iteratively update the state of the actors through a message-passing process. The authors of [22] proposed a transformer-like [23] module integrated with a recurrent neural network (RNN), which they called Interaction Transformer. This approach captures the spatial-temporal dependencies between vehicles using point clouds, RGB images, and HD maps as input. STITNet [12] is a spatial-temporal-interactive network that jointly performs object detection and motion prediction of pedestrians, using only point clouds. They proposed a graph layer that models the interaction between pedestrians in order to enhance the motion prediction task. JFP [24] is the most recent approach in this context and focuses on output representations to model the joint probability of multiple objects. They proposed using a pairwise graphical model with a dynamic interactive graph, leading to consistent trajectories prediction for all vehicles presented in the scene. These approaches demonstrated the importance of modeling the interactions between actors. However, these approaches have focused only on a single category of objects, such as cars or pedestrians. A holistic perception foresees perceiving the entire contextual environment, including other categories of dynamic objects. The interaction across different object categories (e.g. vehicles, pedestrians, and cyclists, among others) has not been investigated so far, leaving a gap in the literature.

The presented approaches are advancing the state of the art, covering several possibilities, and making more robust end-to-end perception approaches. However, all these approaches designed their object detection task as an object-level approach. This means that the objects are represented by 2D or 3D bounding boxes, which has some limitations. For example, these representations approximate the shape of objects and often include background or other objects within the bounding boxes. The exact location of the object is not accurately described, which is particularly challenging in high-traffic environments. Another weakness is the difficulty in detecting objects that are not included in the training set, as these methods rely on global shape and texture information from Regions of Interest (RoI) specific to each object category. This information is difficult to generalize to categories of objects that have never been present in the training set [25]. Pixel-level representations, on the other hand, are more accurate since they cover most of the shortcomings of object-level representations. The shape information of the objects is conserved, and the exact location of the object is more accurately described. The ability to detect small and distant objects increases, and it is easier to generalize to classes not defined in the training set [26]. This is possible, since these approaches effectively decompose RoIs into grid cells, extracting local information shared by many object categories. These methods also make predictions for all occupied cells [27]. In this context, several approaches have been proposed to perform end-to-end perception at pixel-level.

MotionNet [27] laid the groundwork and proposed a spatio-temporal pyramid network (STPN) based on three main heads: i) cell classification to perceive the category of each cell; ii) motion prediction to predict the future trajectory of each cell; and iii) state estimation to estimate the motion status of each cell, i.e., static or moving, providing auxiliary information for motion prediction. Other approaches of end-to-end perception at pixel-level [26], [28]–[31] used MotionNet as a backbone network to explore different point cloud representations, such as range-view and birds-eye-view, and the addition of different sensor modalities. Birds-eye-view representation offers advantages in end-to-end models by maintaining object size regardless of range, simplifying object detection and motion forecasting, and enabling effective fusion of historical LiDAR data and high-definition map features [32]. However, this representation discretizes LiDAR data into voxels, leading to the loss of fine-grained information for detecting smaller objects. On the other hand, range-view methods [33] operate in the dense, non-quantized LiDAR data, providing access to full sensor information and strong detection performance, particularly for smaller objects. Range-view is also suitable for fusing data from sensors that natively capture range-view data, such as LiDAR and camera. However, range-view methods require learning the transformation from range-view to birds-eye-view and handling variations in object size with range, making the problem more complex and requiring a larger dataset for competitiveness with birds-eye-view based methods [34]. Given these advantages and disadvantages, the fusion of both birds-eye-view and range-view representations becomes important to leverage the strengths of each approach and overcome their limitations, enabling comprehensive perception of the scene and accurate detection across various object sizes and ranges [35].

Most end-to-end perception approaches rely only on point clouds and HD maps, and only a few of them exploit the benefits of sensor fusion with RGB images [22], [28], [29], [34], [36], [37]. As discussed in Chapter 2, point clouds contain precise depth, physical, and metric information about the environment, whereas RGB images provide color and high-resolution data, but no depth information [29]. HD maps provide detailed and accurate information about the environment, including road geometry, lane markings, traffic signs, and landmarks, leveraging the environment information provided to the perception system. Sensor fusion integrates data from multiple sensing modalities to reduce the number of detection uncertainties and overcome the shortcomings of individual sensors [8], [38]. This allows the development of robust models that accurately perceive the surroundings under various environmental conditions. However, the benefits of sensor fusion between point clouds, RGB images, and HD maps remain largely unexplored.

The interactions between actors from different classes have not been studied so far and remain a research gap that this Ph.D. proposal aims to address. Modeling these interactions is an essential step toward achieving more accurate and reliable perception systems. These interactions can significantly affect the behavior and trajectory of each actor. For example, the presence of a cyclist in the vicinity of a car can significantly affect the behavior of the car, and in turn, the trajectory of the car can influence the trajectory of the cyclist. Therefore, failing to model these interactions can lead to inaccurate predictions and potential safety hazards. In this context, this Ph.D. proposal aims to develop a complete end-to-end holistic perception approach capable of modeling the interactions between actors from the same and different classes and their interplay with the scene. Additionally, integration of this end-to-end holistic perception in a pixel-

level representation appears to offer several advantages that can leverage the perception systems to another level of robustness and accuracy. This novel research line would model the interaction between actors for various known and unknown object categories, in which their interplay with other actors (from the same category or not) and with the contextual environment may differ depending on their category and the current environment. Furthermore, exploring the benefits of sensor fusion between RGB images, point clouds and HD maps also seems to be very promising, allowing the perception system to work properly in a large set of environments with diverse weather and lighting conditions. By addressing these challenges, the proposed research aims to advance the state of the art in perception for autonomous driving and contribute to the development of more accurate, robust, and reliable perception systems.

Chapter 4

Objectives

The main objective of this Ph.D. proposal is to develop a multi-modal end-to-end holistic perception approach for autonomous driving. Its primary contribution is the complete holistic perception of the scene, considering the interactions between actors from the same and different classes (cars, buses, trucks, pedestrians, and cyclists), and their interplay with the scene (lanes, traffic signs, and traffic states), to accurately detect and predict their future motions. The joint detection and motion prediction of each actor in the scene will be based on three mutual components: i) processing its past motion; ii) modeling the interactions with other actors from the same and different classes; and iii) processing the contextual information about the scene. Based on this research objective, the proposed research question is: "How can interactions between actors from different categories of objects be effectively modeled into an end-to-end perception approach, in order to accurately detect and predict their future motions and improve the overall holistic perception of the scene?" In summary, this Ph.D. proposal can be divided into four research specific objectives and derived research questions:

Explore perception datasets: We intend to use the two most used large-scale datasets nuScenes [10] and Waymo Open Dataset [11], which provide thousands of real-word labeled scenes with diverse weather and lighting conditions, and offer benchmarks with standard metrics. Based on this specific objective, the derived research question is: "How can large-scale perception datasets such as nuScenes and Waymo Open Dataset be effectively utilized to train and evaluate the performance of the developed end-to-end holistic perception architectures?"

Development of end-to-end holistic perception architectures: End-to-end perception architectures all rely on multi-task learning [13]. Multi-task learning consists in learning tasks simultaneously through shared knowledge. It is the appropriate machine learning paradigm, as it is able to jointly optimize and learn the tasks of end-to-end perception. Our objective is to explore and combine multi-task learning [13] with RNNs and Transformers that are suitable to capture temporal dependencies between different time steps and spatial dependencies between actors, respectively [22]. Based on this specific objective, the derived research question is: "How can multi-task learning, combined with RNNs and Transformers, be leveraged to capture temporal and spatial dependencies, in order to model the interactions between actors from the same and different classes, as well as their interplay with the scene?"

Explore pixel-level architectures: The end-to-end perception algorithms that model the interactions between actors are all based on object-level [27]. Our objective is to organically combine the advantages of a holistic perception approach with pixel-level perception. In this context, we propose to explore several pixel-level architectures to serve as a backbone for the main end-to-end holistic perception architecture. Based on this specific objective, the derived research question is: "Which pixel-level architectures can be explored and integrated into the main end-to-end holistic perception architecture to capitalize on the advantages of both holistic perception and pixel-level representation?"

Explore sensor fusion techniques: Most end-to-end perception approaches receive as input only point clouds [36]. Point clouds are usually transformed into birds-eye-view and range-view representations, which are suitable for fusing with HD maps and RGB images, respectively [34]. Therefore, our objective is to explore the benefits of combining these four data modalities. Based on this specific objective, the derived research question is: "Which sensor fusion techniques can be effectively employed to combine birds-eye-view and range-view representations of point clouds, HD maps, and RGB images, to enhance the overall perception system and enable robust perception in diverse environmental conditions?"

Chapter 5

Work Plan

The detailed description of this Ph.D. proposal consists of eight tasks to be completed over a four-year period starting from 01/10/2023. We outline each task of the work plan in detail and also present the contingency plans. Figure 5.1 presents the timeline of the proposed work plan.

Task 1 - State-of-the-art review: In this task, all relevant publications in the field of end-to-end perception will be continuously analyzed throughout the entire time span of the Ph.D. The goal is to conduct a narrative review of the literature, to stay up-to-date with the new methods in this area, and may adapt the research statement to explore new opportunities and ideas in the field. At the end of this task, we expect to have a comprehensive understanding of the current state-of-the-art. The contingency plan includes considering alternative research directions, such as investigating perception approaches in other domains or exploring emerging technologies that could be applied to end-to-end perception.

Task 2 - Perception datasets: Recently, all perception algorithms have relied on neural networks. Neural networks are data-driven approaches and require large-scale datasets. Accordingly, the goal of this task is to explore nuScenes [10] and Waymo Open Dataset [11], which are real-world datasets with thousands of labeled scenes captured by a full sensor suite from a real autonomous vehicle. The intention is to explore how to use their data to develop and evaluate our perception algorithm. Additionally, we will study the standard metrics used to evaluate the perception tasks, such as mean average precision, average multiple object tracking accuracy, average displacement error, and trajectory collision rate, among others. These metrics are used in the datasets' benchmarks, where our algorithm can be compared with state-of-the-art perception algorithms. At the end of this task, we expect to understand how to effectively use these datasets to compare our algorithm with the state-of-the-art methods. The contingency plan involves considering alternative datasets or supplementary data sources if necessary to ensure a comprehensive evaluation and validation of our perception algorithm.

Task 3 - End-to-end holistic perception architectures: We intend to explore and implement state-of-the-art end-to-end perception algorithms that model a holistic perception of the scene. From these algorithms, we gather several ideas on how to model

the interactions between actors and their interplay with the scene. The most common architectures used to model these interactions are RNNs and transformers [22], combined with the multi-task learning paradigm for end-to-end perception [13]. A comparison and evaluation of their advantages and disadvantages will be performed to select the best architecture or combination of architectures for the holistic perception model. The goal is to develop an end-to-end holistic perception algorithm that captures the complex interplay between multiple actors across different object categories and their interactions with the scene. At the end of this task, we expect to prove the importance of modeling inter-class interactions between actors by improving the accuracy and reliability of the perception system, which is the key objective of this Ph.D. proposal. In case of misselection of the architecture, the contingency plan includes investigating other relevant algorithms or approaches that can effectively capture the complex interplay between multiple actors and their interactions.

Task 4 - End-to-end perception architectures at pixel-level: In this task, we intend to explore several pixel-level approaches to combine them as a backbone architecture into our end-to-end holistic perception architecture. The goal of this task is to organically combine the benefits of pixel-level representation with holistic perception. The advantages of pixel-level representation include preserving the shape of objects, describing more accurately the exact location of objects, improving the detection of small and distant objects, and facilitating generalization to unseen object categories [26]. At the end of this task, we expect to demonstrate how the incorporation of pixel-level representation can contribute to a more comprehensive and effective perception system. In case the initial exploration of pixel-level approaches does not yield satisfactory results, the contingency plan includes simplifying the developed perception algorithm to an object-level representation.

Task 5 - Sensor fusion techniques: Most end-to-end perception approaches receive as input point clouds and HD maps, and only a few of them explored the benefits of sensor fusion with RGB images [22], [28], [29], [34], [36]. In this context, the goal is to explore sensor fusion techniques to aggregate the benefits of fusing both birds-eye-view and range-view point cloud representations with RGB images and HD maps, and then integrate this technique into our perception algorithm. One advantage of birds-eye-view is that the size of objects remains constant regardless of range. However, this representation loses information necessary to detect smaller objects due to the discretization of the point cloud into voxels. On the other hand, range-view is the native representation of point clouds, providing strong detection performance for smaller objects. However, in this representation, the size of the objects varies with range [34]. Fusing these four data modalities can overcome the shortcomings of individual sensors working independently and combine the advantages of both point cloud representations. At the end of this task, we expect that these four sensing modalities and the selected sensor fusion technique leverage the robustness and accuracy of our perception algorithm. In case the initial sensor fusion technique does not yield the desired results, the contingency plan includes considering alternative sensor fusion methods or variations to ensure optimal integration and performance. In the worst-case scenario, we consider the dropout of one or more data representations to reduce the complexity of the system.

Task 6 - Domain adaptation for AtlasCar 2 in Aveiro: AtlasCar 2 [39] is an autonomous vehicle developed by the Atlas Project at the University of Aveiro. We intend to explore domain adaptation to adapt our perception algorithm to receive as input data from the sensor suite of the AtlasCar 2, which may differ from the sensor suite (and respective calibration) of the autonomous vehicle used to produce the public datasets. The goal is to test and validate our perception algorithm with an autonomous vehicle available at the hosting institution. At the end of this task, we expect to validate the generality of the perception algorithm in a real-case study. An impacting risk is the shutdown of the Atlas Project, and consequently the non-availability of the AtlasCar 2. In this case, the contingency plan includes using the CARLA simulator [40] to mitigate this risk. Using CARLA, we can implement Domain Adaptation for several sensor suites and setups, not only the one available in AtlasCar 2 and the one from the datasets.

Task 7 - Writing of articles: The goal is to publish our state-of-the-art contributions in international journals and conferences (milestone 1). The international conferences that we aim to publish are: IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), IEEE International Conference on Computer Vision (ICCV), and IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Concerning the international journals, we aim to publish in: IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Neural Networks and Learning Systems. This task is attached to the development tasks and we expect to contribute to the state-of-the-art in perception for autonomous driving by publishing in these top-tier conferences and journals.

Task 8 - Writing of the thesis: This task is concerned with writing the Ph.D. thesis (milestone 2). An introduction contextualization and a systematic literature review will be presented, as well as a summary of the research questions that the Ph.D. aimed to answer. Then, the document will present all relevant contributions of the Ph.D. alongside the methodologies, results, and conclusions.

General risks and contingency plans: The perception algorithms that will be developed are based on neural networks. As such, it is a necessity to have powerful hardware capable of training these networks. In the Laboratory for Automation and Robotics (LAR) at the University of Aveiro, there is a server called DeepLar with several GPUs that is mainly used to train neural networks. A general impact risk is the malfunctioning of the hardware available to develop and train the perception algorithms. To mitigate this risk, the supervision team has several contacts and partners in other institutions, where we can request the usage of their hardware. Another option is the usage of servers provided by Google CoLaboratory and Kaggle to train neural networks.

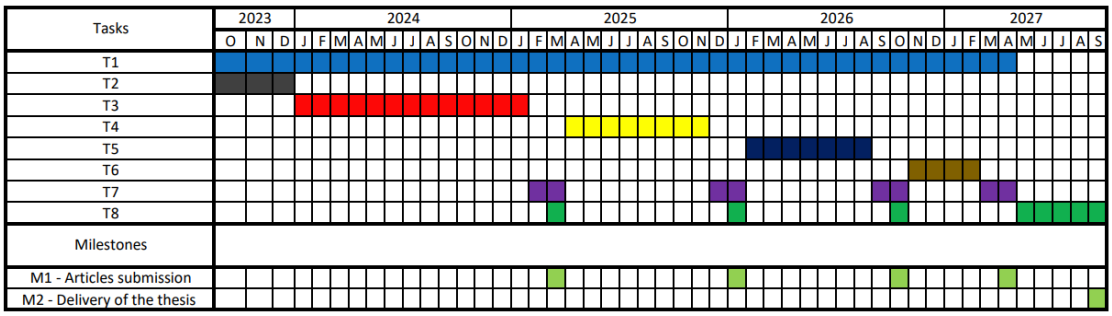

Figure 5.1: Gantt chart of the Ph.D. proposal.

Chapter 6

Hosting Conditions

This Ph.D. proposal will be hosted by the **University of Aveiro (UA)**, which is renowned for its excellence in autonomous driving research. The UA offers access to state-of-the-art resources such as the AtlasCar 2, a full-size autonomous vehicle from the Atlas Project at the Department of Mechanical Engineering (DEM), equipped with several sensors to deal with real-world scenarios. Additionally, the UA offers access to DeepLar, a server with several GPUs used to train neural networks. The Institute of Electronics and Informatics Engineering of Aveiro (IEETA), a research unit of the UA, and its Intelligent Robotics and Systems (IRIS) research group provide expertise in intelligent mobile robotics, with a particular focus on topics such as autonomous driving, perception, control, and artificial intelligence, among others. The strong multidisciplinary experience of these research groups (Atlas Project and IRIS) in perception for autonomous driving will provide a platform for the discussion of ideas and knowledge sharing with other researchers working in the same field. The interaction with these research groups and the access to hardware resources will be critical in achieving the main objectives of the Ph.D. proposal.

The foreign institution that will support this Ph.D. is the **Computer Vision Center (CVC), at the Autonomous University of Barcelona (UAB)**. The Advanced Driver Assistance Systems (ADAS) group of CVC focuses on applying Machine Learning techniques to problems related to autonomous driving, such as perception. They also developed an autonomous vehicle called Elektra (http://adas.cvc.uab.es/elektra/), demonstrating their expertise in autonomous driving. The opportunity to discuss ideas with other researchers who have expertise in this field will be valuable.

The supervisor, Prof. Miguel Oliveira, has wide experience in autonomous driving and robotics. His Ph.D. thesis entitled "Automatic Information and Safety Systems for Driving Assistance" overlaps with the proposed Ph.D. research. He is the creator of ATOM, a calibration framework for robotics systems using the Atomic Transformations Optimization Method, which demonstrates his expertise in the field of robotics. He has published several articles in high-impact journals and international conferences in related areas, such as autonomous driving, sensor calibration, and color correction. As such, his supervision and expertise are essential for this Ph.D.

Co-supervisor Prof. Vítor Santos has vast experience in autonomous driving and deep learning, having published several articles in top journals and conferences. He supervised six former Ph.D. students, including the supervisor Prof. Miguel Oliveira. He is the coordinator of the Atlas Project and one of the founders of the Portuguese

Robotics Open and the Portuguese Society of Robotics. Therefore, his supervision and expertise are significant for the proposed Ph.D.

Co-supervisor Prof. Antonio López has extensive experience in autonomous driving, computer vision, and deep learning. He is the Principal Investigator of the ADAS group at CVC and one of the developers of CARLA Simulator, widely used to develop perception algorithms for autonomous driving. He has published several articles in high-impact journals and conferences. Therefore, his expertise in computer vision and perception algorithms using deep learning is critical for this Ph.D.

Bibliography

[1] T. J. Crayton and B. M. Meier, "Autonomous vehicles: Developing a public health research agenda to frame the future of transportation policy," *Journal of Transport Health*, vol. 6, no. February, pp. 245–252, Sep. 2017, ISSN: 22141405. DOI: 10.1016/j.jth.2017.04.004. [Online]. Available: http://dx.doi.org/10.1016/j.jth.2017.04.004%20https://linkinghub.elsevier.com/retrieve/pii/S2214140517300014.

[2] R. Qian, X. Lai, and X. Li, "3D Object Detection for Autonomous Driving: A Survey," *Pattern Recognition*, vol. 130, p. 108 796, Oct. 2022, ISSN: 00313203. DOI: 10.1016/j.patcog.2022.108796. arXiv: 2106.10823. [Online]. Available: https://doi.org/10.1016/j.patcog.2022.108796%20http://arxiv.org/abs/2106.10823%20http://dx.doi.org/10.1016/j.patcog.2022.108796%20https://linkinghub.elsevier.com/retrieve/pii/S0031320322002771.

[3] SAE International, *SAE levels of Driving Automation™ refined for clarity and international audience*, https://www.sae.org/blog/sae-j3016-update, Online; accessed 09 January 2023, 2021.

[4] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, Jun. 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2983149. arXiv: 1906.05113. [Online]. Available: https://ieeexplore.ieee.org/document/9046805/.

[5] A. Kendall, J. Hawke, D. Janz, *et al.*, "Learning to Drive in a Day," in *2019 International Conference on Robotics and Automation (ICRA)*, vol. 2019-May, IEEE, May 2019, pp. 8248–8254, ISBN: 978-1-5386-6027-0. DOI: 10.1109/ICRA.2019.8793742. arXiv: 1807.00412. [Online]. Available: https://ieeexplore.ieee.org/document/8793742/.

[6] M. Liang, B. Yang, W. Zeng, *et al.*, "PnPNet: End-to-End Perception and Prediction With Tracking in the Loop," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 11 550–11 559, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.01157. [Online]. Available: https://ieeexplore.ieee.org/document/9157054/.

[7] D. Feng, C. Haase-Schutz, L. Rosenbaum, *et al.*, "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021, ISSN: 1524-9050. DOI: 10.1109/TITS.2020.

2972974. arXiv: 1902.07830. [Online]. Available: https://ieeexplore.ieee.org/document/9000872/.

[8] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review.," *Sensors (Basel, Switzerland)*, vol. 21, no. 6, p. 2140, Mar. 2021, ISSN: 1424-8220. DOI: 10.3390/s21062140. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2140%20http://www.ncbi.nlm.nih.gov/pubmed/33803889%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8003231.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2012, pp. 3354–3361, ISBN: 978-1-4673-1228-8. DOI: 10.1109/CVPR.2012.6248074. [Online]. Available: http://ieeexplore.ieee.org/document/6248074/.

[10] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 11 618–11 628, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.01164. arXiv: 1903.11027. [Online]. Available: https://ieeexplore.ieee.org/document/9156412/.

[11] P. Sun, H. Kretzschmar, X. Dotiwalla, *et al.*, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 2443–2451, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.00252. arXiv: 1912.04838. [Online]. Available: https://ieeexplore.ieee.org/document/9156973/.

[12] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "STINet: Spatio-Temporal-Interactive Network for Pedestrian Detection and Trajectory Prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 11 343–11 352, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.01136. arXiv: 2005.04255. [Online]. Available: https://ieeexplore.ieee.org/document/9156474/.

[13] W. Luo, B. Yang, and R. Urtasun, "Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 3569–3577, ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00376. arXiv: 2012.12395. [Online]. Available: https://ieeexplore.ieee.org/document/8578474/.

[14] W. Zeng, W. Luo, S. Suo, *et al.*, "End-To-End Interpretable Neural Motion Planner," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June, IEEE, Jun. 2019, pp. 8652–8661, ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00886. arXiv: 2101.06679. [Online]. Available: https://ieeexplore.ieee.org/document/8954347/.

[15] F. Duffhauss and S. A. Baur, "PillarFlowNet: A Real-time Deep Multitask Network for LiDAR-based 3D Object Detection and Scene Flow Estimation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,

IEEE, Oct. 2020, pp. 10 734–10 741, ISBN: 978-1-7281-6212-6. DOI: 10.1109/IROS45743.2020.9341002. [Online]. Available: https://ieeexplore.ieee.org/document/9341002/.

[16] G. P. Meyer, J. Charland, S. Pandey, *et al.*, "LaserFlow: Efficient and Probabilistic Object Detection and Motion Forecasting," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 526–533, Apr. 2021, ISSN: 2377-3766. DOI: 10.1109/LRA.2020.3047793. arXiv: 2003.05982. [Online]. Available: https://ieeexplore.ieee.org/document/9310205/.

[17] S. Ye, H. Yao, W. Wang, Y. Fu, and Z. Pan, "SDAPNet: End-to-End Multitask Simultaneous Detection and Prediction Network," in *2021 International Joint Conference on Neural Networks (IJCNN)*, vol. 2021-July, IEEE, Jul. 2021, pp. 1–8, ISBN: 978-1-6654-3900-8. DOI: 10.1109/IJCNN52387.2021.9533290. [Online]. Available: https://ieeexplore.ieee.org/document/9533290/.

[18] Z. Chen, Y. Wang, X. Liu, and X. Wang, "FS-GRU: Continuous Perception and Prediction with inter Frame Feature Sharing," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, Oct. 2022, pp. 517–522, ISBN: 978-1-6654-6880-0. DOI: 10.1109/ITSC55140.2022.9922356. [Online]. Available: https://ieeexplore.ieee.org/document/9922356/.

[19] Y. Zhang, Y. Ye, Z. Xiang, and J. Gu, "SDP-Net: Scene Flow Based Real-Time Object Detection and Prediction from Sequential 3D Point Clouds," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12622 LNCS, 2021, pp. 140–157, ISBN: 9783030695248. DOI: 10.1007/978-3-030-69525-5_9. [Online]. Available: http://link.springer.com/10.1007/978-3-030-69525-5_9.

[20] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to Predict Intention from Raw Sensor Data," *2018 2nd Annual Conference on Robot Learning, CoRL 2018, Zurich, Switzerland, 29-31 October 2018, Proceedings*, vol. 87, pp. 947–956, Jan. 2018. arXiv: 2101.07907. [Online]. Available: http://arxiv.org/abs/2101.07907%20http://proceedings.mlr.press/v87/casas18a.html.

[21] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2020, pp. 9491–9497, ISBN: 978-1-7281-7395-5. DOI: 10.1109/ICRA40945.2020.9196697. arXiv: 1910.08233. [Online]. Available: https://ieeexplore.ieee.org/document/9196697/.

[22] L. L. Li, B. Yang, M. Liang, *et al.*, "End-to-end Contextual Perception and Prediction with Interaction Transformer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Oct. 2020, pp. 5784–5791, ISBN: 978-1-7281-6212-6. DOI: 10.1109/IROS45743.2020.9341392. arXiv: 2008.05927. [Online]. Available: https://ieeexplore.ieee.org/document/9341392/.

[23] Vaswani, S. Ashish, P. Noam, *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2016, pp. 47–82. DOI: 10.1007/978-3-319-29409-4_3. [Online]. Available: http://link.springer.com/10.1007/978-3-319-29409-4_3.

[24] W. Luo, C. Park, A. Cornman, B. Sapp, and D. Anguelov, "JFP: Joint Future Prediction with Interactive Multi-Agent Modeling for Autonomous Driving," in *6th Conference on Robot Learning (CoRL 2022)*, Dec. 2022, pp. 1–11. arXiv: 2212.08710. [Online]. Available: http://arxiv.org/abs/2212.08710.

[25] D. Zhou, J. Fang, X. Song, *et al.*, "Joint 3D Instance Segmentation and Object Detection for Autonomous Driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 1836–1846, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.00191. [Online]. Available: https://ieeexplore.ieee.org/document/9156967/.

[26] Y. H. Khalil and H. T. Mouftah, "End-to-End Multi-View Fusion for Enhanced Perception and Motion Prediction," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, vol. 2021-Septe, IEEE, Sep. 2021, pp. 1–6, ISBN: 978-1-6654-1368-8. DOI: 10.1109/VTC2021-Fall52928.2021.9625271. [Online]. Available: https://ieeexplore.ieee.org/document/9625271/.

[27] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird's Eye View Maps," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 11 382–11 392, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.01140. arXiv: 2003.06754. [Online]. Available: https://ieeexplore.ieee.org/document/9157538/.

[28] Y. H. Khalil and H. T. Mouftah, "LiCaNext: Incorporating Sequential Range Residuals for Additional Advancement in Joint Perception and Motion Prediction," *IEEE Access*, vol. 9, pp. 146 244–146 255, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3123169. [Online]. Available: https://ieeexplore.ieee.org/document/9585703/.

[29] Y. H. Khalil and H. T. Mouftah, "LiCaNet: Further Enhancement of Joint Perception and Motion Prediction Based on Multi-Modal Fusion," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, no. July 2021, pp. 222–235, 2022, ISSN: 2687-7813. DOI: 10.1109/OJITS.2022.3160888. [Online]. Available: https://ieeexplore.ieee.org/document/9738812/.

[30] Y. H. Khalil and H. T. Mouftah, "LidNet: Boosting Perception and Motion Prediction from a Sequence of LIDAR Point Clouds for Autonomous Driving," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, IEEE, Dec. 2022, pp. 3533–3538, ISBN: 978-1-6654-3540-6. DOI: 10.1109/GLOBECOM48099.2022.10001152. [Online]. Available: https://ieeexplore.ieee.org/document/10001152/.

[31] Z. Wei, X. Qi, Z. Bai, *et al.*, "Spatiotemporal Transformer Attention Network for 3D Voxel Level Joint Segmentation and Motion Prediction in Point Cloud," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, vol. 2022-June, IEEE, Jun. 2022, pp. 1381–1386, ISBN: 978-1-6654-8821-1. DOI: 10.1109/IV51971.2022.9827310. [Online]. Available: https://ieeexplore.ieee.org/document/9827310/.

[32] N. Djuric, H. Cui, Z. Su, *et al.*, "MultiXNet: Multiclass Multistage Multimodal Motion Prediction," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, vol. 2021-July, IEEE, Jul. 2021, pp. 435–442, ISBN: 978-1-7281-5394-0. DOI: 10.1109/

IV48863.2021.9575718. arXiv: 2006.02000. [Online]. Available: https://ieeexplore.ieee.org/document/9575718/.

[33] A. Laddha, S. Gautam, G. P. Meyer, C. Vallespi-Gonzalez, and C. K. Wellington, "RV-FuseNet: Range View Based Fusion of Time-Series LiDAR Data for Joint 3D Object Detection and Motion Forecasting," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, Sep. 2021, pp. 7060–7066, ISBN: 978-1-6654-1714-3. DOI: 10.1109/IROS51168.2021.9636083. arXiv: 2005.10863. [Online]. Available: https://ieeexplore.ieee.org/document/9636083/.

[34] S. Fadadu, S. Pandey, D. Hegde, *et al.*, "Multi-View Fusion of Sensor Data for Improved Perception and Prediction in Autonomous Driving," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Jan. 2022, pp. 3292–3300, ISBN: 978-1-6654-0915-5. DOI: 10.1109/WACV51458.2022.00335. arXiv: 2008.11901. [Online]. Available: https://ieeexplore.ieee.org/document/9706809/.

[35] A. Laddha, S. Gautam, S. Palombo, S. Pandey, and C. Vallespi-Gonzalez, "MV-FuseNet: Improving End-to-End Object Detection and Motion Forecasting through Multi-View Fusion of LiDAR Data," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2021, pp. 2859–2868, ISBN: 978-1-6654-4899-4. DOI: 10.1109/CVPRW53098.2021.00321. arXiv: 2104.10772. [Online]. Available: https://ieeexplore.ieee.org/document/9522936/.

[36] A. Mohta, F.-C. Chou, B. C. Becker, C. Vallespi-Gonzalez, and N. Djuric, "Investigating the Effect of Sensor Modalities in Multi-Sensor Detection-Prediction Models," in *Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Jan. 2021. arXiv: 2101.03279. [Online]. Available: http://arxiv.org/abs/2101.03279.

[37] J. Chen, Z. Xu, and M. Tomizuka, "End-to-end autonomous driving perception with sequential latent representation learning," *IEEE International Conference on Intelligent Robots and Systems*, pp. 1999–2006, 2020, ISSN: 21530866. DOI: 10.1109/IROS45743.2020.9341020. arXiv: 2003.12464.

[38] M. Shah, Z. Huang, A. Laddha, *et al.*, "LiRaNet: End-to-End Trajectory Prediction using Spatio-Temporal Radar Fusion," no. CoRL, Oct. 2020. arXiv: 2010.00731. [Online]. Available: http://arxiv.org/abs/2010.00731.

[39] V. Santos, J. Almeida, E. Ávila, *et al.*, "ATLASCAR - Technologies for a computer assisted driving system on board a common automobile," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2008, pp. 1421–1427, 2010. DOI: 10.1109/ITSC.2010.5625031.

[40] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *1st Conference on Robot Learning (CoRL 2017), Mountain View, United States.*, Nov. 2017, pp. 1–16. arXiv: 1711.03938. [Online]. Available: http://arxiv.org/abs/1711.03938.

Abstract

Autonomous driving systems require accurate perception algorithms to navigate safely through traffic. End-to-end perception approaches jointly learn perception tasks, transforming raw sensor data directly into object detection with motion predictions. To improve trajectory predictions, recent approaches attempted to model interactions between actors (dynamic objects), whose behavior depends on each other and their interplay with the scene. However, current models only consider interactions within the same object class. Modeling interactions between actors from different classes improves perception accuracy and enables a more holistic understanding of the scene by preventing misinterpretation of object behavior. In this context, this Ph.D. proposal aims to develop a multi-modal end-to-end holistic perception approach capable of modeling both inter-class and intra-class interactions between actors, and their interplay with the scene. Perception systems still face challenges in achieving high accuracy and robustness, with a complete holistic perception of the scene yet to be achieved despite recent progress.

Keywords

Object Detection, Motion Prediction, Holistic Perception, End-To-End Perception, Deep Learning, Autonomous Driving.

Introduction and State of the Art

Autonomous driving is an emerging technology that holds the promise of revolutionizing transportation. Autonomous driving systems are usually developed using a modular approach [1, 2]. One of the most critical components is the perception module, which is responsible for accurately assessing the environment surrounding the vehicle to enable safe navigation through traffic [3]. The perception module must be accurate, robust, and operate in real-time [4].

The perception module is divided into three tasks: object detection, object tracking, and motion prediction [5]. These tasks are usually learned independently and executed sequentially. Information is not shared among these tasks, and uncertainty is rarely propagated between them, leading to information loss [6, 7]. Recently, end-to-end perception approaches jointly learn and optimize all these tasks within a single neural network, using multi-task learning [5–12]. These approaches are efficient for real-time operations by sharing computational resources between all tasks. The three tasks directly access raw data, leveraging the shared knowledge between them. These approaches have yielded promising results, reducing the detection of false negatives for distant and occluded objects, and false positives by accumulating evidence over time [5, 6, 13, 14].

Most of these approaches predict the trajectory of an actor (dynamic object) independently based on its past trajectory, without taking into account the interactions between actors. This strategy undermines the accuracy of the predicted trajectories, as the evolution of a trajectory is highly dependent on the behavior of other actors and and environmental factors. To address this limitation, recent approaches attempted to model the interactions between actors and their interplay with the scene [7,14–17]. These approaches model spatial and temporal dependencies using different neural network architectures, such as graph neural networks (GNN) [16], recurrent neural networks (RNN) [17], convolutional neural networks (CNN) [14], and transformers [17]. These interaction models improved the accuracy of perception systems, demonstrating their importance in predicting motion. However, these approaches focus only on a single category of objects, such as cars or pedestrians, and inter-class interactions are not considered.

The interactions between actors from different classes have not been studied so far and remain a research gap that this Ph.D. proposal aims to address. Modeling these interactions is an essential step toward achieving more accurate and reliable perception systems. These interactions can significantly affect the behavior and trajectory of each actor. For example, the presence of a

cyclist in the vicinity of a car can significantly affect the behavior of the car and, in turn, the trajectory of the car can influence the trajectory of the cyclist. Therefore, failing to model these interactions can lead to inaccurate predictions and potential safety hazards. In this context, this Ph.D. proposal aims to develop a complete end-to-end holistic perception approach capable of modeling the interactions between actors from the same and different classes and their interplay with the scene. Additionally, the goal is to explore the benefits of sensor fusion between different data modalities [17–22], allowing the system to work properly in a large set of environments with diverse weather and lighting conditions [23].

Objectives

The main objective of this Ph.D. proposal is to develop a multi-modal end-to-end holistic perception approach for autonomous driving. Its primary contribution is the complete holistic perception of the scene, considering the interactions between actors from the same and different classes (cars, buses, trucks, pedestrians, and cyclists), and their interplay with the scene, to accurately detect and predict their future motions. The detection and motion prediction of each actor in the scene will be based on three joint components: i) processing its past motion; ii) modeling the interaction with other actors from the same and different classes; and iii) processing the contextual information about the scene (lanes, traffic signs, and traffic states). Based on this objective, the proposed research question is: how can interactions between actors from different categories of objects be effectively modeled into an end-to-end perception approach, in order to accurately detect and predict their future motions and improve the overall holistic perception of the scene? In summary, this Ph.D. proposal can be divided into three specific objectives:

Explore perception datasets: We intend to use the two most used large-scale datasets nuScenes [24] and Waymo Open Dataset [25], which provide thousands of real-word labeled scenes with diverse weather and lighting conditions, and offer benchmarks with standard metrics.

Development of end-to-end holistic perception architectures: Our objective is to explore and combine multi-task learning [6] with RNNs and Transformers that are suitable to capture temporal dependencies between different time steps and spatial dependencies between actors, respectively [17].

Explore sensor fusion techniques: Most end-to-end perception approaches receive as input only point clouds [18]. Point clouds are usually transformed into birds-eye-view and range-view representations, which are suitable for fusing with HD maps and RGB images, respectively [21]. Therefore, our objective is to combine the benefits of fusing these four data modalities.

Detailed Description

The detailed description of this Ph.D. proposal consists of seven tasks, each with a corresponding methodology and implementation plan, to be completed over a four-year period starting from 01/12/2023. In this section, we outline each task in detail and also present the risks involved and respective mitigation plans.

Task 1 - State-of-the-art review: To present this Ph.D. proposal, we performed a systematic review of the literature that culminated in the research statement and objectives defined in the previous section. In this task, all relevant publications in the field of end-to-end perception approaches for autonomous driving will be continuously analyzed throughout the entire time span of the Ph.D. The goal is to conduct a narrative review of the literature to stay up-to-date with the

new methods in this area and may adapt the research statement to explore new opportunities and ideas in the field.

Task 2 - Perception datasets: Recently, all perception algorithms have relied on neural networks. Neural networks are data-driven approaches and have the need for large-scale datasets with thousands of examples (image frames and/or point cloud sweeps). Accordingly, the goal of this task is to explore nuScenes [24] and Waymo Open Dataset [25], which are real-world datasets with thousands of labeled scenes captured by a full sensor suite from a real autonomous vehicle. The intention is to explore how to use their data to develop our perception algorithm. Additionally, we will study the standard metrics used to evaluate the three tasks of perception algorithms, such as mean average precision (mAP), average multiple object tracking accuracy (AMOTA), average displacement error (ADE), and trajectory collision rate (TCR), among others. These metrics are used in the datasets' benchmarks, where our algorithm can be compared with state-of-the-art perception algorithms.

Task 3 - End-to-end holistic perception architectures: We intend to explore and implement state-of-the-art end-to-end perception algorithms that model a holistic perception of the scene. From these algorithms, we gather several ideas on how to model the interactions between actors and their interplay with the scene. The most common architectures used to model these interactions are recurrent neural networks (RNNs) and transformers [17], combined with the multi-task learning paradigm for end-to-end perception [6]. A comparison and evaluation of their advantages and disadvantages will be performed to decide the best architecture or combination of architectures for the holistic perception model. The goal is to develop an end-to-end holistic perception algorithm, that captures the complex interplay between multiple actors across different object categories and their interactions with the environment. At the end of this task, we expect to prove the importance of modeling interactions between actors from different classes by improving the accuracy and reliability of the perception system, which is the key objective of this Ph.D. proposal.

Task 4 - Sensor fusion techniques: Most end-to-end perception approaches receive as input point clouds and HD maps, and only a few of them explored the benefits of sensor fusion with RGB images [17–21]. Point clouds are usually transformed into birds-eye-view representation, which is suitable for fusing with HD maps. However, range-view is also a very commonly used point cloud representation, which is appropriate for fusing with RGB images [21]. Fusing data modalities from different sensors can overcome the shortcomings of individual sensors working independently [23]. Within this context, we aim to explore sensor fusion techniques for these four data modalities and incorporate them into our perception algorithm to leverage its robustness and accuracy.

Task 5 - Domain adaptation for AtlasCar 2 in Aveiro: AtlasCar 2 [26] is an autonomous vehicle developed by the Atlas Project at the University of Aveiro. We intend to explore domain adaptation to adapt our perception algorithm to receive as input data from the sensor suite of the AtlasCar 2, which may differ from the sensor suite (and respective calibration) of the autonomous vehicle used to produce the public datasets. In this context, the goal is to test and validate our perception algorithm with an autonomous vehicle available at the hosting institution.

Task 6 - Publication in international journals and conferences: During the development tasks, the goal is to contribute to the state-of-the-art in the field of perception for autonomous driving, publishing these contributions in international journals and conferences (milestone 1). The international conferences that we aim to publish are the following: IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), IEEE International Conference on Computer Vision (ICCV), IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), and IEEE International Conference on Robotics and Automation (ICRA). Concerning the international

journals, our aim is to publish in: IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Neural Networks and Learning Systems.

Task 7 - Writing of the thesis: This task is concerned with writing the Ph.D. thesis (milestone 2). An introduction contextualization and a systematic literature review will be presented, as well as a summary of the research questions that the Ph.D. aimed to answer. Then, the document will present all relevant contributions of the Ph.D. alongside the methodologies, results, and conclusions. This task will be done mutually with the other tasks, refining the writing at the end of each development task, and reserving 5 months for the final writing.

Acknowledged risks and mitigation plans: The main impacting risk is the shutdown of the Atlas Project, and consequently the non-availability of the AtlasCar 2 for task 5. In this case, we plan to use the CARLA Simulator [27] from the foreign hosting institution CVC to mitigate this risk. Using CARLA, we can implement Domain Adaptation for several sensor suites and setups, not only the one available in Atlas Car and the one from the datasets. Another impacting risk is the malfunctioning of the hardware (DeepLar server - DEM/UA) available to develop and train the perception algorithms. To mitigate this risk, the supervision team has several contacts and partners in other institutions (including CVC), where we can request the usage of their hardware in case of malfunctioning of our hardware.

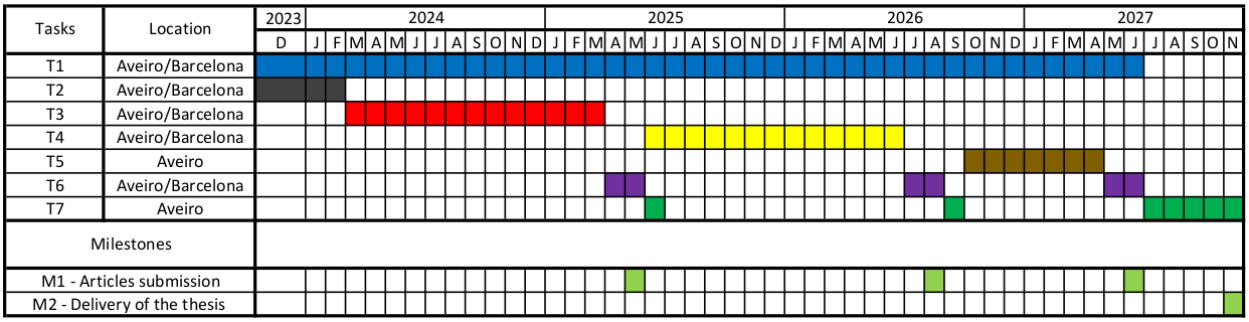

Figure 1: Gantt chart of the Ph.D. proposal.

Hosting Conditions

This Ph.D. proposal will be hosted by the University of Aveiro (UA), which is renowned for its excellence in autonomous driving research. The UA offers access to state-of-the-art resources such as the AtlasCar 2, a full-size autonomous vehicle from the Atlas Project at the Department of Mechanical Engineering (DEM), equipped with several sensors to deal with real-world scenarios. Additionally, the UA offers access to DeepLar, a server with several GPUs used to train neural networks. The Institute of Electronics and Informatics Engineering of Aveiro (IEETA), a research unit of the UA, and its Intelligent Robotics and Systems (IRIS) research group provide expertise in intelligent mobile robotics, with a particular focus on topics such as autonomous driving, perception, control, and artificial intelligence, among others. The strong multidisciplinary experience of these research groups (Atlas Project and IRIS) in perception for autonomous driving will provide a platform for the discussion of ideas and knowledge sharing with other researchers working in the same field. The interaction with these research groups and the access to hardware resources will be critical in achieving the main objectives of the Ph.D. proposal.

The foreign institution that will support this Ph.D. is the Computer Vision Center (CVC), at the Autonomous University of Barcelona (UAB). The Advanced Driver Assistance Systems (ADAS) group of CVC focuses on applying Machine Learning techniques to problems related

to autonomous driving, such as perception. They also developed an autonomous vehicle called Elektra (http://adas.cvc.uab.es/elektra/), demonstrating their expertise in autonomous driving. The opportunity to discuss ideas with other researchers who have expertise in this field will be valuable.

The supervisor, Prof. Miguel Oliveira, has wide experience in autonomous driving and robotics. His Ph.D. thesis, "Automatic Information and Safety Systems for Driving Assistance," overlaps with the proposed Ph.D. research. He is the creator of ATOM, a calibration framework for robotics systems using the Atomic Transformations Optimization Method, which demonstrates his expertise in the field of robotics. He has published several articles in high-impact journals and international conferences in related areas, such as autonomous driving, sensor calibration, and color correction. As such, his supervision is essential for this Ph.D.

Co-supervisor Prof. Vítor Santos has vast experience in autonomous driving and deep learning, having published several articles in top journals and conferences. He supervised six former Ph.D. students, including the supervisor Prof. Miguel Oliveira. He is the coordinator of the Atlas Project and one of the founders of the Portuguese Robotics Open and the Portuguese Society of Robotics. Therefore, his supervision is significant for the proposed Ph.D.

The co-supervisor, Prof. Antonio López, has extensive experience in autonomous driving, computer vision and deep learning. He is the Principal Investigator of the ADAS group at CVC and one of the developers of CARLA Simulator, widely used to develop perception algorithms for autonomous driving. He has published several articles in high-impact journals and conferences. Therefore, his expertise in computer vision and perception algorithms using deep learning is critical for this Ph.D.

References

[1] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to Drive in a Day," in *2019 International Conference on Robotics and Automation (ICRA)*, vol. 2019-May, pp. 8248–8254, IEEE, may 2019.

[2] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58443–58469, jun 2020.

[3] R. Qian, X. Lai, and X. Li, "3D Object Detection for Autonomous Driving: A Survey," *Pattern Recognition*, vol. 130, p. 108796, oct 2022.

[4] D. Feng, C. Haase-Schutz, L. Rosenbaum, H. Hertlein, C. Glaser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 1341–1360, mar 2021.

[5] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun, "PnPNet: End-to-End Perception and Prediction With Tracking in the Loop," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11550–11559, IEEE, jun 2020.

[6] W. Luo, B. Yang, and R. Urtasun, "Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3569–3577, IEEE, jun 2018.

[7] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "STINet: Spatio-Temporal-Interactive Network for Pedestrian Detection and Trajectory Prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11343–11352, IEEE, jun 2020.

[8] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-To-End Interpretable Neural Motion Planner," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June, pp. 8652–8661, IEEE, jun 2019.

[9] F. Duffhauss and S. A. Baur, "PillarFlowNet: A Real-time Deep Multitask Network for LiDAR-based 3D Object Detection and Scene Flow Estimation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10734–10741, IEEE, oct 2020.

[10] G. P. Meyer, J. Charland, S. Pandey, A. Laddha, S. Gautam, C. Vallespi-Gonzalez, and C. K. Wellington, "LaserFlow: Efficient and Probabilistic Object Detection and Motion Forecasting," *IEEE Robotics and Automation Letters*, vol. 6, pp. 526–533, apr 2021.

[11] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird's Eye View Maps," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11382–11392, IEEE, jun 2020.

[12] Y. H. Khalil and H. T. Mouftah, "LidNet: Boosting Perception and Motion Prediction from a Sequence of LIDAR Point Clouds for Autonomous Driving," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 3533–3538, IEEE, dec 2022.

[13] S. Ye, H. Yao, W. Wang, Y. Fu, and Z. Pan, "SDAPNet: End-to-End Multi-task Simultaneous Detection and Prediction Network," in *2021 International Joint Conference on Neural Networks (IJCNN)*, vol. 2021-July, pp. 1–8, IEEE, jul 2021.

[14] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to Predict Intention from Raw Sensor Data," *2018 2nd Annual Conference on Robot Learning, CoRL 2018, Zurich, Switzerland, 29-31 October 2018, Proceedings*, vol. 87, pp. 947—-956, jan 2018.

[15] W. Luo, C. Park, A. Cornman, B. Sapp, and D. Anguelov, "JFP: Joint Future Prediction with Interactive Multi-Agent Modeling for Autonomous Driving," in *6th Conference on Robot Learning (CoRL 2022)*, pp. 1–11, dec 2022.

[16] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9491–9497, IEEE, may 2020.

[17] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun, "End-to-end Contextual Perception and Prediction with Interaction Transformer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5784–5791, IEEE, oct 2020.

[18] A. Mohta, F.-C. Chou, B. C. Becker, C. Vallespi-Gonzalez, and N. Djuric, "Investigating the Effect of Sensor Modalities in Multi-Sensor Detection-Prediction Models," in *Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, jan 2021.

[19] Y. H. Khalil and H. T. Mouftah, "LiCaNext: Incorporating Sequential Range Residuals for Additional Advancement in Joint Perception and Motion Prediction," *IEEE Access*, vol. 9, pp. 146244–146255, 2021.

[20] Y. H. Khalil and H. T. Mouftah, "LiCaNet: Further Enhancement of Joint Perception and Motion Prediction Based on Multi-Modal Fusion," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, no. July 2021, pp. 222–235, 2022.

[21] S. Fadadu, S. Pandey, D. Hegde, Y. Shi, F.-C. Chou, N. Djuric, and C. Vallespi-Gonzalez, "Multi-View Fusion of Sensor Data for Improved Perception and Prediction in Autonomous Driving," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3292–3300, IEEE, jan 2022.

[22] Y. H. Khalil and H. T. Mouftah, "End-to-End Multi-View Fusion for Enhanced Perception and Motion Prediction," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, vol. 2021-Septe, pp. 1–6, IEEE, sep 2021.

[23] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review.," *Sensors (Basel, Switzerland)*, vol. 21, p. 2140, mar 2021.

[24] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, no. March 2019, pp. 11618–11628, IEEE, jun 2020.

[25] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, IEEE, jun 2020.

[26] V. Santos, J. Almeida, E. Ávila, D. Gameiro, M. Oliveira, R. Pascoal, R. Sabino, and P. Stein, "ATLASCAR - Technologies for a computer assisted driving system on board a common automobile," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2008, pp. 1421–1427, 2010.

[27] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *1st Conference on Robot Learning (CoRL 2017), Mountain View, United States.*, pp. 1–16, nov 2017.

RELATÓRIO FINAL DE ATIVIDADES – Bolsa de Doutoramento FCT

Resumo

Eu, Lucas Rodrigues Dal'Col, melhor identificado no âmbito do contrato de bolsa de investigação celebrado com a FCT – Fundação para a Ciência e Tecnologia, com o identificador DOI **10.54499/2023.02251.BD**, venho, por este meio, submeter o **Relatório Final de Atividades** relativo aos trabalhos desenvolvidos no contexto da tese de doutoramento financiada pela FCT, intitulada: "***Multi-Modal End-to-End Holistic Perception for Autonomous Driving***".

A investigação foi conduzida no Instituto de Engenharia Eletrónica e Informática da Universidade de Aveiro (IEETA), no âmbito do Programa Doutoral em Engenharia Mecânica (PDEM) do Departamento de Engenharia Mecânica (DEM) da Universidade de Aveiro (UA) e, parcialmente, no Computer Vision Center (CVC) da Universitat Autònoma de Barcelona (UAB), em Espanha. O trabalho foi orientado cientificamente pelo Doutor Miguel Armando Riem de Oliveira, Professor Auxiliar do DEM-UA, pelo Doutor Vítor Manuel Ferreira dos Santos, Professor Associado com Agregação do DEM-UA, e pelo Doutor Antonio Manuel López Peña, Investigador Principal do Autonomous Driving Lab – CVC-UAB.

O presente relatório tem como objetivo apresentar à FCT uma síntese das atividades desenvolvidas no decurso do plano de trabalhos definido até ao momento do cancelamento da bolsa de doutoramento por motivos profissionais. Durante o período de vigência da bolsa, as atividades decorreram com sucesso, em conformidade com o plano inicialmente aprovado.

Contribuições

No contexto do estado da arte na área de *Joint Perception and Prediction for Autonomous Driving*, o trabalho desenvolvido resultou na seguinte contribuição principal:

- Foi realizado um estudo aprofundado e sistemático do estado da arte na área de *joint perception and prediction* aplicada à condução autónoma. Este trabalho constitui o **primeiro levantamento sistemático sobre o tema**, propondo uma taxonomia que categoriza as abordagens existentes com base na representação de entrada, na modelação do contexto da cena e na representação de saída, destacando as suas principais contribuições e limitações. Além disso, foram realizadas análises qualitativas e quantitativas dos métodos existentes, bem como discutidas potenciais direções futuras de investigação com base nas lacunas identificadas no estado da arte.
- Este trabalho culminou na publicação de um artigo de revisão científica intitulado *"Joint Perception and Prediction for Autonomous Driving: A Survey"*, na revista *IEEE Transactions on Intelligent Transportation Systems*, com **fator de impacto de 8.4**. Trata-se de uma revista de elevado prestígio e exigência editorial, reconhecida como uma das mais

relevantes na área de condução autónoma, o que reforça a importância e o impacto científico deste contributo.

- Sendo o primeiro levantamento sistemático publicado sobre esta temática, o artigo tem como objetivo servir de referência para investigadores e profissionais da área, facilitando a compreensão do paradigma de *joint perception and prediction* e auxiliando na seleção informada e fundamentada de métodos adequados aos respetivos casos de estudo.

Publicações

A publicação resultante do trabalho desenvolvido no âmbito da bolsa encontra-se referenciada como segue:

- **L. Dal'Col**, M. Oliveira and V. Santos, "Joint Perception and Prediction for Autonomous Driving: A Survey," in *IEEE Transactions on Intelligent Transportation Systems*, doi: 10.1109/TITS.2025.3607743.

Aveiro, 31 de Outubro de 2025

Assinado por: **Lucas Rodrigues Dal'Col**
Num. de Identificação: 18028231
Data: 2025.10.31 16:54:47 +0000

(Lucas Rodrigues Dal'Col - 10.54499/2023.02251.BD)