Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

# Predictive Techniques for Scene Understanding by using Deep Learning in Autonomous Driving

PhD. Thesis Presented by

**Carlos Gómez Huélamo**

2023

Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

# Predictive Techniques for Scene Understanding by using Deep Learning in Autonomous Driving

PhD. Thesis Presented by

**Carlos Gómez Huélamo**

Advisors

**Dr. Luis Miguel Bergasa Pascual**

**Dr. Rafael Barea Navarro**

Alcalá de Henares, TBD

**A mi Madre, allá donde esté ..... .**

*"En este vasto mundo*
*navegáis en pos de un sueño,*
*surcando el ancho mar*
*que se extiende frente a vosotros.*
*El puerto de destino es el mañana*
*cada día más incierto.*
*Encontrad el camino,*
*cumplid vuestros sueños,*
*estáis todos en el mismo barco*
*y vuestra bandera es la libertad"*
Opening 3 de One Piece
Autor: The Babystars

# Acknowledgements

Esta Tesis Doctoral supone el culmen a cuatro años (Abril 2019 - Abril 2023) realmente duros, cargado de emociones, triunfos, pandemias, estafas y tropiezos, todo a partes iguales. Este es probablemente (aunque como diría Sean Connery interpretando a James Bond en 1983, *Never Say Never Again*) mi último gran documento individual, académicamente hablando.

Durante mi etapa universitaria (2013 hasta el momento, 2023) he tenido ciertos momentos puntuales en los que he sentido un salto cualitativo como profesional: El primero fue en el segundo cuatrimestre de segundo de carrera, cuando las cosas se pusieron tensas con Control II e Informática Industrial. Vaya sudores. El segundo probablemente fue con el fallecimiento de mi madre durante mi ERASMUS+ en Irlanda. Duros y oscuros momentos, alejado de mis seres queridos. El tercer momento llega en segundo de máster, durante mi querido ERASMUS+ en Finlandia, donde compagino una estancia preciosa en Tampere con el máster y un pre-inicio de doctorado. Me equivoqué al empezar tan pronto con la beca, "queriendo cobrar" cuanto antes, en vez de terminar tranquilamente el TFM y plantear la tesis, pero eso no lo sabría hasta tiempo después. Pero no es hasta la tesis donde empezaron los quebraderos de cabeza reales. Continuamente altibajos, mala planificación por mi parte, momentos puntuales donde me equivoqué rotundamente al empecinarme en soldar una estructura compleja para nuestro vehículo sin ayuda, no estudiar PyTorch tras el congreso WAF 2018 tras la sugerencia de mi tutor, no enfocarme en técnica individual hasta bien entrado el doctorado, no querer hacer nada hasta que no tuviese la teoría perfectamente asimilada, tener demasiado respeto a la Inteligencia Artificial y escurrir el bulto de mi tesis en un compañero mientras yo me dedicaba a integrar y corregir los bugs del grupo que para mí *era lo fácil*. Mal. Todo mal. Pero todo cambió tras mi segunda estancia, en Estados Unidos, cuando tras llorar por no entender el camino a seguir, nadie que me ayudara, decidí crear mi propio camino, con paciencia y fé, práctica y error compaginado con lectura de artículos, para mejorar mi confianza y autoestima, y finalmente logré empezar a entender lo que era el Deep Learning. Gracias a todos mis errores, desventuras y discusiones, a día de hoy, excepto momentos inevitables, me encuentro con muchísima capacidad para atacar y gestionar prácticamente cualquier problema, consultar documentación y organizarme, aunque esta sigue siendo mi tarea

pendiente.

Cada año, desde hace ya varios, mi primera publicación en Instagram viene seguida de la frase "Trabaja duro en silencio y deja que tu éxito haga todo el ruido". Filosofía Kaizen, de mejora y aprendizaje continuo, para así cada día entender el mundo un poquito mejor. Si toda la dedicación y estudio que he depositado en este trabajo sirven para algo en mi futuro, sé que todo el esfuerzo habrá merecido la pena.

Después de este particular monólogo, a lo cual soy muy propenso y de lo cual mis amigos y compañeros no cesan en su empeño de recordádmelo, debo, como no puede ser de otra manera, dar paso a los agradecimientos.

En primer lugar, me gustaría agradecer a mis profesores del grupo RobeSafe, especialmente a mis tutores Luis Miguel Bergasa Pascual y Rafael Barea Navarro, por ofrecerme estar en el grupo (así como aguantarme) durante todos estos años e intentar que tuviésemos el mejor *experimental setup* y *roadmap* en el laboratorio, aunque no fuese siempre sencillo. Sin dudas considero realmente interesante la temática propuesta en esta tesis doctoral, predicción de agentes en el contexto de vehículo inteligente, ya que entra en el plano filosófico sobre cómo razonar el futuro de los objetos y cómo podría afectar a la capacidad ejecutiva del agente que deba tomar una decisión. Mi mente hace tiempo que cambió y me fijo siempre que conduzco de todo lo que intento reproducir con mis estudios. Habrá que seguir esta tendencia muy de cerca en los próximos años, porque personalmente considero que sus aplicaciones son fantásticas.

A mis tutores en las estancias de doctorado, Christoph Stiller y Eduardo Molinos en el Karslruhe Institute of Technology (KIT, Alemania) y Wei Zhan y Masayoshi Tomizuka en la University of California, Berkeley (UCB, Estados Unidos). Se suele decir que unas veces se gana y otras se aprende, y yo en estas estancias quizás aprendí demasiado ... No obstante, me guardo grandísimos momentos (admirar las secuoyas gigantes o hacer mi primera escalada en roca entre ellos) y amigos, como Su Shaoshu o Frank Bieder, con los que aún guardo un cierto contacto.

A mis compañeros, mejor dicho, amigos, de laboratorio: Javier Araluce, Rodrigo, Felipe (chavalín), Santiago, Miguel Antunes, Miguel Eduardo, antiguos compañeros como Javier del Egido, Óscar, Alejandro, Eduardo, Roberto y Pablo Alcantarilla, y a nuevos becarios como Fabio, Navil y Pablo. Gracias de corazón por estar ahí, en nuestras charlas sobre tecnología, empleos, el camino correcto a seguir y la vida en general.

Especial mención vuelvo a hacer a Miguel Eduardo y mi compañero Marcos Conde, cuya apoyo intenso ayudó a centrar mi camino en técnica y escritura. Muchísimas gracias por todo lo que aprendí a vuestro lado. Especial mención a mi querido profesor Ángel Álvarez, por sus valiosos consejos para afrontar mi carrera profesional y mi vida en general de la mejor forma posible. Eres muy grande.

A mis amigos de la universidad, especialmente a Rocío, Juan Carlos, Esther, Sergio, Pablo, Rubén y Adrián Rocandio. Aún me acuerdo de cuando empezamos con la carrera y como el paso inexorable del tiempo nos moldea a conveniencia. Os deseo lo mejor en vuestro futuro.

A mis buenos amigos Samuel y Adrián, con quien gran parte de mi vida he compartido. Con especial cariño guardo las interminables charlas sobre la vida y el futuro después de Karate, de comer, de cenar, en el coche, siempre quejándonos de la hora que marcaba el reloj al final de tan interminables conversaciones.

A mi familia, uno de los pilares de mi vida. A mi padre Juan Antonio y a mi madre Petra, que en paz descanse, les debo todo lo que soy y es por ello por lo que les estaré siempre agradecido. Querido padre, gracias por ser tan Genaro, arisco y pesado. Siempre has sido mi ejemplo a seguir, aunque cuando tenga 42 años me sigas regañando por subir con las zapatillas puestas. Querida madre, no se muere quien se va, sólo se muere el que se olvida, y tú nunca caerás en el olvido. A mi querida hermana-calili-chessmaster Silvia, con quien tantas regañinas he tenido, pero el cariño que nos tenemos las supera a todas. A mi perrita Nuka (a.k.a. Dragón, ChupaChups cuando le cortamos el pelo o Nuki-Nuki), cuyos paseos matutinos son probablemente el ingrediente secreto para la elaboración de esta tesis, dando rienda suelta a mi cabeza para imaginar nuevas propuestas mientras miraba el cielo azul. A mi querido *experimental setup* que me ha acompañado durante toda mi vida académica: silla de madera de la cocina, ratón de Hello Kitty tomado prestado de mi hermana y mi querido portátil táctil. Sois la base de todo mi trabajo.

Al resto de la familia, amigos, compañeros, entrenadores y profesores, gracias por todo.

Y, por último, la persona más importante de mi vida ahora mismo. Mi querida Marta, la persona más maravillosa y buena que conozco. Hemos compartido risas, lloros, besos y abrazos. Nunca me cansaré de repetirte lo suave que tienes la piel tras darte un beso en la mejilla y después hacerte de rabiar. Espero que esta situación esté dentro de un while cuya condición sea *True*.

*"Te quiero más que ayer, pero menos que mañana. Hoy, y siempre"*

Mi querido lector, disculpa mi monólogo de agradecimientos, es mi forma de ser y la cual tengo por bandera, aunque creo que ha quedado bonito. Podría decir mil anécdotas más de mi doctorado, pero como diría Aragorn, legítimo Rey de Gondor, enfrente de la mismísima Puerta Negra: *Hoy no es ese día.*

Vamos a la lectura importante, que empiece el *Rock and Roll* !!

# Resumen

TBD

**Palabras clave:** Autonomous Driving, Deep Learning, Motion Prediction, Scene Understanding.

# Abstract

TBD

**Keywords:** Autonomous Driving, Deep Learning, Motion Prediction, Scene Understanding.

# Contents

# List of Figures

# List of Tables

# List of Source Code

# List of Acronyms

| | |
|---|---|
| AD | Autonomous Driving. |
| ADS | Autonomous Driving Stack. |
| AI | Artificial Intelligence. |
| BEV | Bird's Eye View. |
| DL | Deep Learning. |
| ITS | Intelligent Transportation Systems. |
| MP | Motion Prediction. |
| SOTA | State-of-the-Art. |

# Chapter 1

# Introduction

*Aaay, el oro, la fama, el poder.*
*Todo lo tuvo el hombre que en su día se autoproclamó*
*el rey de los piratas, ¡GOLD ROGER!*
*Mas sus últimas palabras no fueron muy afortunadas:*
*"¿¡MI TESORO!? Lo dejé todo allí, buscadlo si queréis,*
*ojalá se le atragante al rufián que lo encuentre.*

Opening 1 de One Piece: "We are"
Autor original: Hiroshi Kitadani

## 1.1. Motivation

Autonomous Driving (AD) have held the attention of technology enthusiasts and futurists for some time as evidenced by the continuous research and development in Intelligent Transportation Systems (ITS) over the past decades, being one of the emerging technologies of the *Fourth Industrial Revolution*, and particularly of the Industry 4.0.

The concept *Fourth Industrial Revolution* or Industry 4.0 was first introduced by Klaus Schwab , CEO (Chief Executive Officer) of the World Economic Forum, in a 2015 article in Foreign Affairs (American magazine of international relations and United States foreign policy). A technological revolution can be defined as a period in which one or more technologies are replaced by other kinds of technologies in a short amount of time. Hence, it is an era of accelerated technological progress featured by Researching, Development and Innovation whose rapid application and diffusion cause an abrupt change in society. In particular, the *Fourth Industrial Revolution* conceptualizes rapid change to industries, technology, processes and societal patterns in the 21st century due to increasing inter-connectivity and smart automation. This industrial revolution focuses on operational efficiency, being the following four themes which summarize it:

- Decentralized decisions: Ability of cyber physical systems to make decisions on their own and to perform their tasks as autonomously as possible.
- Information transparency: Provide operators with comprehensive information to make decisions. Inter-connectivity allows operators to gather large amounts of information and data from all points in the manufacturing process in order to identify key areas or aspects that can benefit from improvement to enhance functionality.
- Technical assistance: Ability to assist humans with unsafe or difficult tasks and technological facility of systems to help humans in problem-solving and decision-making.
- Interconnection: Ability of machines, sensors, devices and people to communicate and conect with each other via the Internet of Things (IoT) or the Internet of People (IoP).

Based on the aforementioned principles, this revolution is expected to be marked by breakthroughs in emerging technologies in fields such as nanotechnology, quantum computing, 3D printing, Internet of Things (IoT), fifth-generation wireless technologies (5G), Robotics, Computer Vision (CV), Artificial Intelligence (AI) or the scope of this PhD thesis, Autonomous Driving Stacks (ADSs). The sum of all these advances are resulting in machines that can potentially see, hear and what is more important, think, moving more deftly than humans.

An ADS, also referred in the literature as Intelligent Vehicle (IV), driverless car or autonomous car, is a vehicle tan can sense its surrounding and moving safely with little or even no human input. These ADSs must combine a variety of sensors to understand the traffic scenario, like RADAR (RAdio Detection A Ranging), LiDAR (Light Detection and Ranging), cameras, Inertial Measurement Unit (IMU), wheel odometry, GNSS (Global Navigation Satellite System) or ultrasonic sensors, and detect, track and predict (which is the main purpose of this thesis) the most relevant obstacles around the ego-vehicle. Then, advanced control and planning systems process this sensory information in combination with a predefined global route to calculate the corresponding control commands to drive throughout the environment, ensuring a safe driving.

The dream of seeing fleets of ADSs efficiently delivering goods and people to their destination has fueled billions of dollars and captured consumer's imaginations in investment in recent years. Nevertheless, according to the "Autonomous driving's future: Convenient and connected" report, published by the global management consulting firm McKinsey & Company in January 2023, even after some setbacks have pushed out timelines for AD launches and delayed customer adoption, the transportation community still broadly agrees that AD has the potential to transform consumer behaviour, transportation and

society at large. AD is considered as one of the solutions to the before mentioned problems and one of the greatest challenges of the automotive industry today.

Statistics show that 69 % of the population in the European Union (EU), including associated states, lives in urban areas. According to the World Health Organization, nearly one third of the world population will live in cities by 2030, leading to an overpopulation in most of them. Aware of this problem, the Transport White Paper published by the European Commission in 2011 indicated that new forms of mobility ought to be proposed so as to provide sustainable solutions for people and goods safely. For example, regarding safety, it sets the ambitious goal of halving the overall number of road deaths in the EU between 2010 and 2020. Nevertheless, this goal does not seem to be easy since only in 2014 more than 25,700 people died on the roads in the EU, many of them caused by an improper behaviour of the driver on the road. A similar study made by the National Highway Traffic Safety Administration (NHTSA, tranportation organization of the United States) reported in 2015 that around 94 % of traffic accidents happen because of human error. In that sense, the existence of reliable and economically affordable ADSs are expected to create a huge impact on society affecting social, demographic, environmental and economic aspects. It can produce substantial value for the auto industry, drivers and society, making driving safer, more convenient and more enjoyable. While the human driver or not could select whether to drive, in autonomous mode hours on the road previously spent driving could be used to work, watch a funny movie or even to video call a friend. For employees with long commutes, AD might shorten the workday, increasing worker productivity. Since workers, specially those related to digital jobs or related fields, may perform their jobs from an ADS, they could more easily move further away from the office, which, in turn, could attract more people to suburbs and rural areas. Besides this, it is estimated to cause a reduction in road deaths, reduce fuel consumption and harmful emission associated and improve traffic flow, as well as an improvement in the overall driver comfort and mobility in groups with impaired faculties, such as disable or elderly people, providing them with mobility options that go beyond car-sharing services or public transportation. Other industrial applications of autonomous vehicles are agriculture, retail, manufacturing, commercial and freight transport or mining.

## 1.2. Historical Context

ADSs have become a challenge for auto competitions and technology companies, which has derived in an intense competition. Though today companies such as Mercedes, Ford or Tesla are racing to build ADSs for a radically changing consumer world, the research

and development of autonomous robots is not new.

In 1500, centuries before the invention of the automobile, Leonardo da Vinci designed a cart that could move without being pulled or pushed. In 1868, Robert Whitehead invented a torpedo that could propel itself underwater in order to be a game-changer for naval fleets all over the world. In terms of robotic solutions for intelligent mobility, the study was started in the 1920s, being the concept of Autonomous Car defined in Futurama, an exhibit at the 1939 New York Wolrd's Fair. General Motors created the exhibit to display its vision of what the world would look like in 20 years, including an automated highway system that would guide ADS. By 1958, General Motors made this concept a reality (at least as a proof of concept) being the car's front end embedded with sensors to detect the current flowing through a wire embedded in the road. The first semi-automated car was developed in 1977 by Japan's Tsukuba Mechanical Engineering Laboratory. The vehicle reached speeds up to 30 km/h with the support of an elevated rail.


Figure 1.1: Stanley, 2005 DARPA Grand Challenge winner
Source: *Stanford university*

Nevertheless, the first truly autonomous cars appeared in the 1980s with Carnegie Mellon University's Navlab and ALV projects funded by the USA company DARPA (Defense Advanced Research Projects Agency) in 1984 and EUREKA Prometheus project (1987) developed by Mercedes-Benz and Bundeswehr University Munich's. By 1985, the ALV project had shown self-driving speeds on two-lane roads of 31 km/h with obstacle avoidance added in 1986 and off-road driving in day and night conditions by 1987. Furthermore, from the 1960s through the second DARPA Grand Challenge in 2005 (212 km off-road course near the California-Nevada state line, surpassed by all but one of the 23 finalists), automated vehicle research in the United States was primarily funded by DARPA, the US Army and US Navy, yielding rapid advances

in terms of speed, car control, sensor systems and driving competence in more complex conditions. This caused a boost in the development of autonomous prototypes by companies and research organizations, most of them from the United States. Figure 1.1 shows Stanley, the 2005 DARPA Gran Challenge winner, from Stanford university.

Even though self-driving cars have not yet displaced conventional cars, there can be found several examples of how it has become a hot topic for powerful companies such as Delphi Automotive Systems, Audi, BMW, Tesla, Mercedes-Benz or Waymo.

In 2005 Delphi broke the Navlab's record achievement (driving 4,584 km while remaining 98 % of the time autonomously) by piloting an Audi, improved with Delphi technology, over 5,472 km through 15 states while remaining in self-driving mode 99 % of the time. Moreover, in 2005 the USA states of Michigan, Virginia, California, Florida, Nevada and the capital, Washington D.C., allowed the testing of automated cars on public roads.

In 2017, Audi stated that its A8 car prototype would be automated at speeds up to 60 km/h by using its perception system named "Audi AI". Also, in 2017 Waymo (self-driving technology development company subsidiary of Alphabet Inc) started a limited trial of a self-driving taxi service in Phoenix, Arizona.

Figure 1.2 shows the total number of autonomous test miles and miles per disengagement in California (Dec 2019 - Nov 2020) by some of the most important AD technology development companies around the world. The concept disengagement is quite useful to assess the quality of an ADS, defined as the deactivation of the autonomous mode when a failure of the autonomous technology is detected or when a safe operation requires that the autonomous vehicle test driver disengages the autonomous mode, resulting in control being seized by the human driver.

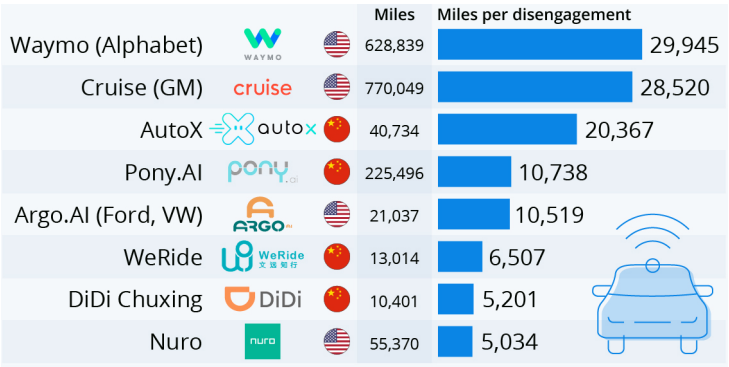


Figure 1.2: Number of autonomous test miles and miles per disengagement (Dec 2019 - Nov 2020)
Source: *DMV California, via The Last Driver License Holder*

At the moment of writing this thesis (2023), many vehicles on the road are considered to be semi-autonomous due to safety features like braking systems, assisted parking, lane boundaries detection or predict the long-term behaviour of the users around the vehicle to execute the most optimal action in a safely way. Regarding this, the Society of Automotive Engineers (SAE) published the concept of autonomy levels in 2014, as part of its "Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems" report. Figure 1.3 illustrates the six levels of autonomy (the higher the level, the more autonomous the car is), where it can be appreciated that Level Zero means "No Automation", being the acceleration, braking and steering controlled by a human driver at all times, and Level Five represents Full Automation, where there is a full-time automation of all driving tasks on any road, under any conditions, whether there is a human on board or not.

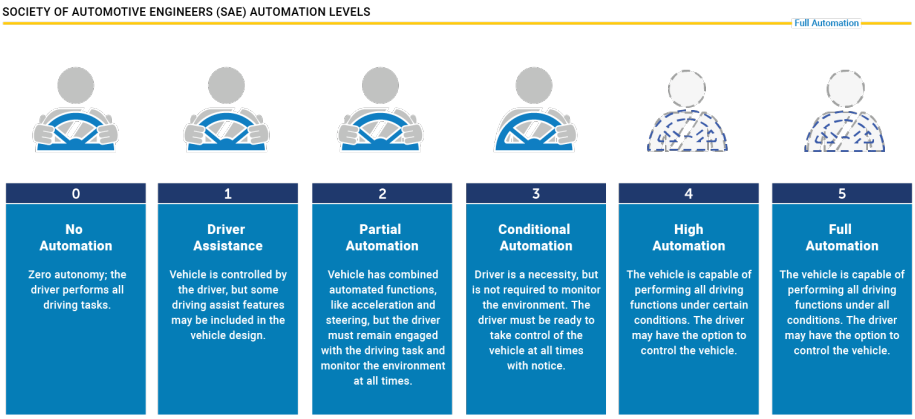


Figure 1.3: Society of Automotive Engineers (SAE) automation levels
Source: *NHTSA (National Highway Traffic Safety Administration)*

In that sense, today most vehicles only included basic Advanced Driver Assistance Systems (ADAS), but major advancements in AD capabilities are on the horizon. According to a 2021 McKinsey consumer survey, growing demand for AD systems could create billions of dollars in revenue. Based on a consumer interest in AD features and commercial solutions available on the market today, ADAS and AD could generate between $300 and %400 billions in the passenger car market by 2035. Figure 1.4 illustrates an interesting study reporting the revenues of ADAS and AD from Level 1 (Driver Assistance) to Level 4 (High Automation). As expected, Level 5 is excluded from this study due to the huge difficulties the automotive companies would have to face to adapt their systems under totally different environmental conditions.

## 1.3. Autonomous Driving architecture

To sum up what commented above, increasing the level of autonomous navigation in mobile robots (from agriculture to public and private transport) are expected to create

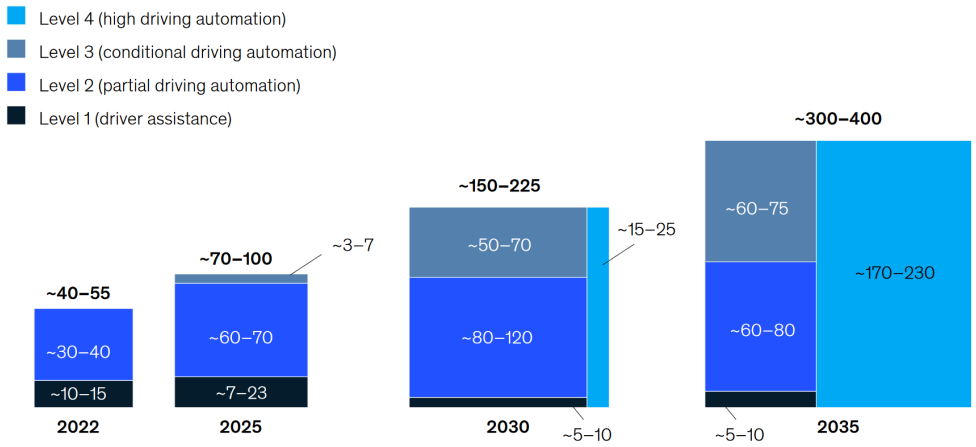


Figure 1.4: Advanced Driver Assistance systems (ADAS) and Autonomous Driving (AD) revenues in $ billion
Source: *McKinsey Center for Future Mobility*

tangible business benefits to those users and companies employing them. However, designing an autonomous navigation system does not seem to be an easy task. In the State-of-the-Art (SOTA) we can distinguish two main kind of software architectures: End-to-End and modular. Figure 1.5 illustrates the entire AD architecture starting from sensing, all the way to longitudinal (throttle/brake) and lateral (steering angle) control of the vehicle, which are the commanded signals that feed the low-level electronic system that moves the vehicle, like a drive-by-wire system [1]. End-to-End are considered black-box models, where a single neural network performs the driving task (throttle/steering/brake) from raw sensor data, in such a way the error be may vanished since intermediate representations are jointly optimized, but these are not very interpretable. On the other hand, modular architectures (considered as glass models as counterpart to End-to-End approaches) separate the driving task into individually programmed or trained modules. This solution is more interpretable, since the know-how of a research group or company is easily transferred, they allow parallel development, being the standard solution in industrial research, but the error is propagated, where intermediate representations can led to suboptimal performance. For example, incorrect object detection can lead to low-quality tracking and motion prediction.

Considering the features of the research group and main projects (Techs4AgeCar, AI-VATAR) where this thesis has been developed, we integrate our algorithms in a software modular approach. An example of modular An example of modular approach is shown in Figure 1.6. Despite the fact in the literature some authors include certain modules in different layers, specially motion prediction algorithms, which are usually classified as a perception algorithm, but sometimes is included as part of the planning or decision-making layers, we can hierarchically break down (from raw data to the driving task) a standard AD architecture into the following software layers:

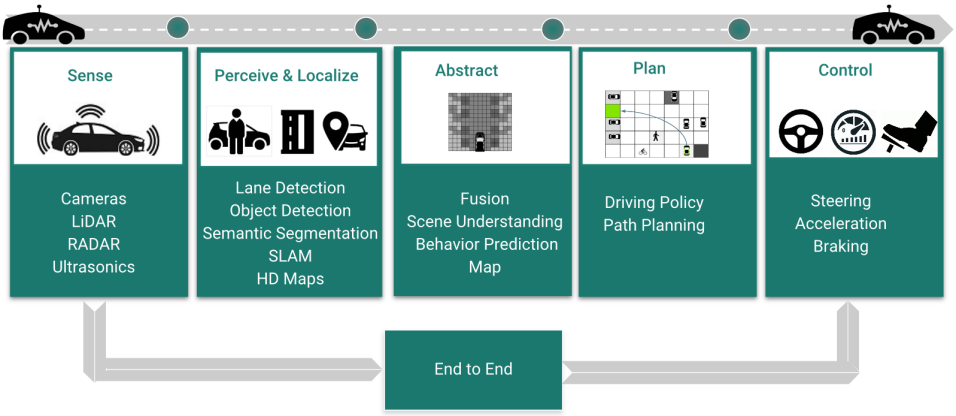


Figure 1.5: Autonomous Driving Stack (ADS) modular vs end-to-end pipeline
Source: *Vrunet: Multi-task learning model for intent prediction of vulnerable road users* [2]

- Localization layer: Positions and locates the vehicle on a map with real-time and centimetric accuracy approach. The main source of information is a robust differential-GNSS, though IMUs, wheel odometry and even cameras are commonly employed.

- Perception layer: Understand the environment around the ego-vehicle thanks to the information collected by the sensors. If defined as multi-stage, the perception layer first detects the most relevant obstacles, then track them over time to finalize long-term predict with plausible predictions. In order to perform object detection, LiDAR, camera and RADAR are the main sensors that provide the corresponding raw data. Additionally, HD map information is frequently used in the motion prediction tasks by most SOTA algorithms.

- Mapping layer: Responsible for creating a topologic, semantic and geographical modeling of the environment through which the vehicle drives, being the HD Map graph the most common source of information.

- Planning layer: This layer is comprised of three components: route, behavioural and trajectory planner. The route planner computes the most optimal (in terms of distance, time and so forth and so on) global route from some predefined start and goal. It uses the localization and mapping output. On the other hand, the behaviour planner, also referred as decision-making layer by some authors, it performs high-level decision-making of driving behaviours such as lane changes or progress through intersections, mostly focused on the previously computed global route and current localization. It can be seen as an atomization of the global route in different behaviors to reach the goal. Finally, the trajectory planner, also known as local planner, generates a time schedule for how to follow a path given constraints such as position, velocity and acceleration in order to meet the previously decided behaviour and taking into account the prediction from the perception layer, avoiding obstacles in optimal direction and speed conditions.

- Control layer: Once the local plan is calculated, the control layer is responsible for generating the commands that are sent to the actuators. It receives as input some waypoints from the calculations made in the trajectory planner. Once these waypoints are received, most authors perform spline interpolations and a velocity profile that ensures a smooth and continous trajectory.

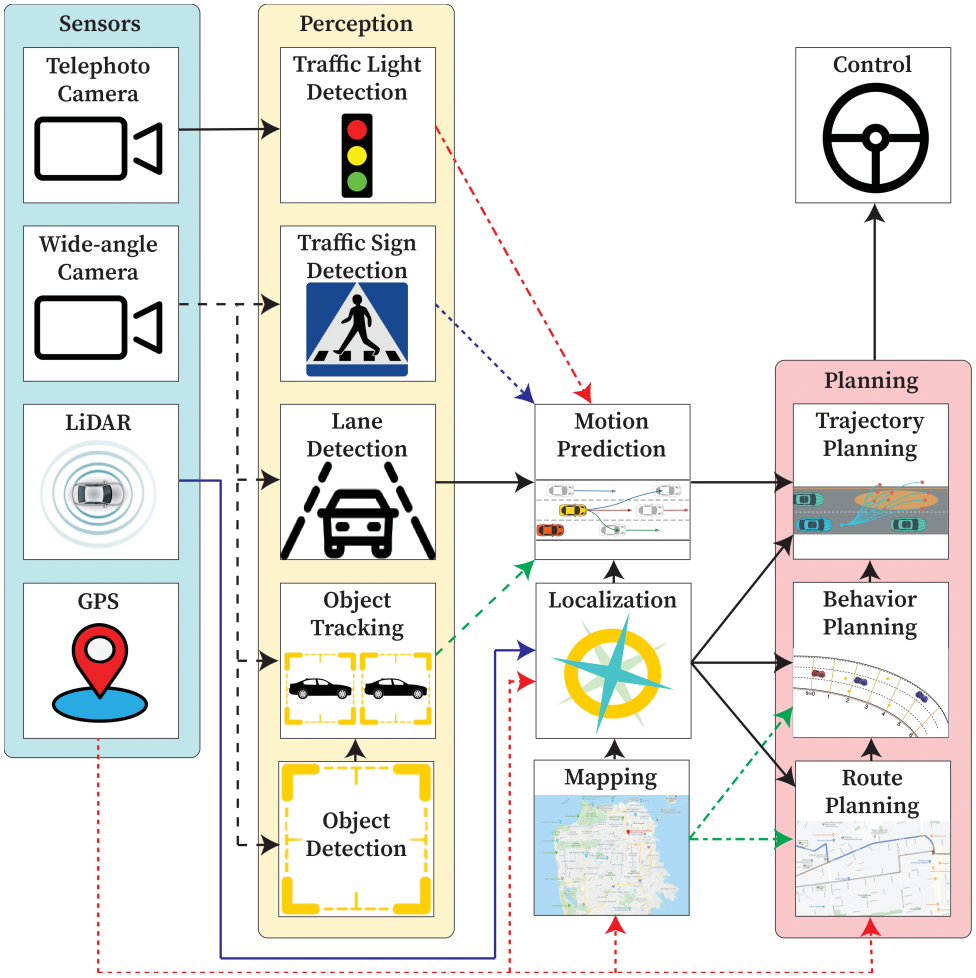


Figure 1.6: Autonomous Driving Stack (ADS) modular pipeline
Source: *Pylot: A modular platform for exploring latency-accuracy tradeoffs in autonomous vehicles* [3]

## 1.4. Problem statement

As commented in previous sections, in order to operate efficiently and safely in highly dynamic, complex and interactive driving scenarios, ADS need to smartly reason like human beings via predicting future motions of surrounding traffic participants during navigation. Nevertheless, achieving accurate and robust Motion Prediction (MP) in one of the most difficult and interesting challenges to achieve full-autonomy, since it is equivalent to a bridge between the former stages of the perception layer, where the scene is understood detecting and tracking static and dynamic objects of the environment, and

the planning a control layer, where the driving commands are sent to the vehicle. Here are some of the most important challenges:

1. Heterogeneity of traffic participants. Traffic partipants (specially those which are dynamic) can be roughly classified as cyclists, pedestrians or other vehicles. The prediction model should be capable of differentiating the motion patterns of heterogeneous traffic participants, in such a way fine-grained classification (detection module) is quite beneficial to include additional metadata along with the past observations.

2. Complexity of road structure. Road structures are highly diverse and complex, specially in highways and urban areas, which noticeably affect the motion behaviours of traffic participants.

3. Variable number of interactive agents. The prediction model must deal with a number of associated traffic participants within a certain area that can vary from time to time, such as intersections or roundabouts. Then, while driving, a comprehensive representation of the scene must be able to accommodate an arbitrary number of involved traffic participants.

4. Multimodality of driving behaviours. In real-world, despite we know the behaviour our vehicle will carry out, the motion patterns of other traffic participants can be considered inherently multimodal since there is usually more than one reasonable option for a driver to choose, specially in intersections, when the number of lanes increases or even in the same lane with different velocity profiles (constant velocity, sudden break, sudden acceleration). In that sense, a robust and reliable MP model is expected to be human-like and capture different plausible motion modalities where an agent can travel in the prediction horizon.

5. Complex interpendencies among traffic participants and road infrastructure. Agent-Agent, Agent-Road and Road-Road interpendencies are of great importance for MP and interaction modeling, even more taking into account the complexity of road structures and heterogeneity of traffic participants aforementioned. As expected, an agent future trajectory will be affected not only by its own past trajectory and driving objectives (given by the behaviour planner) but also by other surrounding agents past trajectories, traffic rules and physical constraints.

## 1.5. Objectives and Structure of this work

The main scope of this thesis is to study the SOTA and development of novel and efficient interaction-aware Deep Learning based MP models, focusing on long-term (from 3 to 6 s) prediction horizon and AD, where traffic participants can range from trucks

to pedestrians, instead of models focused on pedestrian trajectory prediction. The main inputs that will be used throughout this work are the physical (map) information and historical states (that may include agent position, velocity, orientation, object type and category) of traffic participants in Bird's Eye View (BEV), assuming these objects have been previously tracked by our ego-vehicle (also referred as the autonomous car). Though the evaluation of these methods will be done using a single target agent, as proposed by some of the most important prediction datasets, like Argoverse 1 [4] and Argoverse 2 [5], some of the proposed methods will be trained considering multi-agent. In this thesis, the solutions to the aforementioned challenges will be discussed and investigated progressively. In order to achieve the main scope, the following objectives will be met:

1. Research of SOTA MP, focused on Deep Learning (DL) and the AD paradigm.
2. Propose of several MP architectures, studying the progressive incorporation of DL mechanisms and different sources of information and metadata, achieving SOTA accuracy while reducing in millions of parameters previous models as well as inference time.
3. Validate the proposed models in downstream applications, such as decision-making or behavioural planning, taking into account former stages of the perception layer (detection and tracking) instead of static files (benchmarks) in hyper-realistic simulation, as a preliminary stage before implementing it in a real-world vehicle.

The organization of this document has been done as follows:

- Chapter 2 reviews the most important features and methods of physics-based and learning-based MP methods. The physics-based methods are reviewed according to a taxonomy similar to existing reviews. The learning-based methods are reviewed based on two classification criteria: scene representation and trajectory decoding.
- Chapter 3 presents a technical background, mostly focused on DL mechanisms to deal with temporal sequences and interactions, to deeply understand the proposed methods.
- Chapter 4 illustrates the different prediction models developed in the thesis using different validation environments, from unimodal physic-based prediction to the final model of the thesis which takes into account agents interactions, map information and past observations using a novel scene representation with heuristic proposals, graph-based encoding, DL-based goal proposals and motion refinement.
- Chapter 5 addresses the integration of the final model of the thesis with upstream and downstream modules to contribute the entire pipeline and closed-loop for AD.
- Chapter 6 summarizes the thesis and provides some promising directions for future work in the areas of MP and validation.

# Chapter 2

# Related Works

One of the crucial tasks that ADSs must face during navigation, specially in arbitrarily complex urban scenarios, is to predict the behaviour of dynamic obstacles [4], [6]. In a similar way to humans that pay more attention to close obstacles and upcoming turns, rather than considering the obstacles far away, the perception layer of an ADS must focus more on the salient regions of the scene, and the more relevant agents to predict the future behaviour of each traffic participant before conducting a maneuver, such as lane changing or accelerating.

Most traditional predictions methods [7], which usually only consider physics-related factors (like the velocity and acceleration of the target vehicle that is going to be predicted) and road-related factors (prediction as close as possible to the road centerline), are only suitable for **short-time** prediction tasks [7] and simple traffic scenarios, such as constant velocity (CV) in a highway or a curve (Constant Turn Rate Velocity, CTRV) where a single path is allowed, i.e. multiple choices computation are not required. Recently, MP methods based on DL have become increasingly popular since they are able not only to take into account these above-mentioned factors but also consider interaction-related factors (like agent-agent [8], agent-map [9] and map-map **liang2020learninggraph**) in such a way the algorithm can adapt to more complex traffic scenarios (intersections, sudden breaks and accelerations, etc.). It must be consider that multimodal, specially in the field of vehicle motion prediction, does not refer necessarily to different directions (e.g. turn to the left, turn to the right, continue forward in an intersection), but it may refer to different predictions in the same direction that model a sudden positive or negative acceleration, so as to imitate a realistic human behaviour in complex situations. As expected, neither classical nor Machine Learning (ML) methods can model these situations [7].

In order to classify DL based MP methods, we distinguish several important features: Motion history, Social information (agent interactions), Map information (road encoding), how the model returns the output trajectory and its corresponding distribu-

Table 2.1: Main state-of-the-art methods for Motion Prediction. Main categories are Encoder (splitted into motion history, social info (agent interactions) and map info (physical information)), Decoder, Output representation and Distribution over future trajectories.

| **Method** | **Encoder** | | | **Decoder** | **C** |
|---|---|---|---|---|---|
| | Motion history | Social info | Map info | | |
| SocialLSTM **alahi2016social** | LSTM | spatial pooling | – | LSTM | |
| SocialGan [8] | LSTM | maxpool | – | LSTM | |
| Jean **mercat2020multiattentmotion** | LSTM | attention | – | LSTM | |
| TNT **zhao2020tnt** | polyline | maxpool, attention | polyline | MLP | |
| LaneGCN **liang2020learninggraph** | 1D-conv | GNN | GNN | MLP | |
| WIMP [10] | LSTM | GNN+attention | polyline | LSTM | |
| VectorNet **gao2020vectornet** | polyline | maxpool, attention | polyline | MLP | |
| SceneTransformer **ngiam2021scene** | attention | attention | polyline | attention | |
| HOME **gilles2021home** | raster | attention | raster | conv | |
| GOHOME **gilles2021gohome** | 1D-conv+GRU | GNN | GNN | MLP | |
| MP3 **casas2021mp3** | raster | conv | raster | conv | cost |
| ExploringGAN **gomez2022exploring** | LSTM | attention | polyline | LSTM | |
| Multimodal **cui2019multimodal** | raster | conv | raster | conv | |
| MultiPath **chai2019multipath** | raster | conv | raster | MLP | |
| MultiPath++ **varadarajan2021multipath++** | LSTM | RNNs+maxpool | polyline | MLP | con |
| Trajectron++[6] | LSTM | RNNs+attention | raster | GRU | c |
| CRAT-PRED**schmidt2022crat** | LSTM | GNN+attention | – | MLP | |
| **Ours - Social baseline** | LSTM | GNN+attention | – | LSTM | |
| **Ours - Map baseline** | LSTM | GNN+attention | polyline | LSTM | |

tion. Table 2.1 summarizes several SOTA methods, inspired in the survey proposed by **varadarajan2021multipath++**.

- **Motion history**: Most methods encode the sequence of past observed states using 1D-convolution **liang2020learninggraph mercat2020multiattentmotion**, able to model spatial information, or via a recurrent net **gomez2022exploring alahi2016social** (LSTM, GRU), which are more useful to handle temporal information. Other methods that use a raster version of the whole scenario represent the agent states rendered as a stack of binary mask images depicting agent oriented bounding boxes **gilles2021home**. On the other hand, other approaches encode the past history of the agents in a similar way to the road components of the scene given a set of vectors or polylines **zhao2020tnt**, **gao2020vectornet** that can model the high-order interactions among all components, or even employing attention to combine features across road elements and agent interactions **ngiam2021scene**.

- **Social information**: In complex scenarios, motion history encoding of a particular target agent is not sufficient to represent the latent space of the traffic situation, but the algorithm must deal with a dynamic set of neighbouring agents around the target agent. Common techniques are aggregating neighbour motion history with a permutation-invariant set operator: soft attention **ngiam2021scene**, **gomez2022exploring**, a combination of soft attention and

RNN **varadarajan2021multipath++** / GNN **schmidt2022crat** or social pooling **alahi2016social**, **gupta2018sgan**. Raster based approaches rely on 2D convolutions **chai2019multipath casas2021mp3** over the spatial grid to implicitly capture agent interactions in such a way long-term interactions are dependent on the neural network receptive fields.

- **Map information**: High-fidelity maps **can2022maps** have been widely adopted to provide offline information (also known as physical context) to complement the online information provided by the sensor suite of the vehicle and its corresponding algorithms. Recent learning-based approaches **mahjourian2022occupancy**, **ivanovic2021heterogeneous**, [9], which present the benefit of having probabilistic interpretations of different behaviour hypotheses, require to build a representation to encode the trajectory and map information. Map information is probably the feature with the clearest dichotomy: raster vs vector treatment. The raster approach encodes the world around the particular target agent as a stack of images (generally from a top-down orthographic view, also known as Bird's Eye View). This world encoding may include from agent state history, agent interactions and usually the road configuration, integrated all this different-sources information as a multi-channel image **gilles2021home**, in such a way the user can use an off-the-shelf Convolutional Neural Network (CNN) based pipeline in order to leverage this powerful information. Nevertheless, this representation has several downsides: constrained field of view, difficulty in modeling long-range interactions and even difficulty in representing continuous physical states due to the inherent world to image (pixel) discretization. On the other hand, the polyline approach may describe curves, such as lanes, boundaries, intersections and crosswalks, as piecewise linear segments, which usually represents a more compact and efficient representation than using CNNs due to the sparse nature of road networks. Some state-of-the-art algorithms not only describe the world around a particular agent as a set-of-polylines [10] **zhao2020tnt** in an agent-centric coordinate system, but they also leverage the road network connectivity structure **liang2020learninggraph zeng2021lanercnn** treating road lanes as a set of nodes (waypoints) and edges (connections between waypoints) in a graph neural network so as to include the topological and semantic information of the map.

- **Decoder**: Pioneering works of DL based MP usually adopt the autoencoder architecture, where the decoder is often represented by a recurrent network (GRU, LSTM, etc., specially designed to handle temporal information) to generate future trajectories in an autoregressive way, or by CNNs **gilles2021home gilles2021gohome** / MLP **liang2020learninggraph schmidt2022crat** using the non-autoregressive strategy. The method may use an autoregressive strategy where the pipeline generates tokens (in this case, positions or relative displacements) in a sequential manner, in such a way the new output is dependent on the previously generated output, whilst

MLP **schmidt2022crat**, CNN **gilles2021home** or transformer **ngiam2021scene** based strategies usually follow a non-autoregressive strategy, where from a latent space the whole future trajectory is predicted.

- **Output**: The most popular model output representation is a sequence of states (absolute positions) or state differences (relative displacements for any dimension considered). The spacetime trajectory may be intrinsically represented as a continous polynomial representation or a sequence of sample points. Other works **gilles2021home gilles2021gohome** first predict a heatmap and then decode the corresponding output trajectories after sampling points from the heatmap, whilst **casas2021mp3 zeng2019end** learn a cost function evaluator of trajectories that are enumerated heuristically instead of being generated by a learned model.

- **Trajectory Distribution**: The choice of output trajectory distributions has several approaches on downstream applications. Regardless the agent to be predicted is described as a (non-)holonomic **triggs1993motion** platform, an intrinsic property of the motion prediction problem is that the agent must follow one of a diverse set of possible future trajectories. A popular choice to represent a multimodal prediction are Gaussian Mixture Models (GMMs) due to their compact parameterized form, where mode collapse (associated frequently to GMMs) is addressed through the use of trajectory anchors **chai2019multipath** or training tricks **cui2019multimodal**. Other approaches model a discrete distribution via a collection of trajectory samples extracted from a latent space and decoded by the model **rhinehart2018r2p2** or over a set of trajectories (fixed or a priori learned) **liang2020learninggraph**.

After classifying main SOTA methods, we conclude this section presenting the main characteristics of our baseline approaches. We make use of LSTM to encode the past motion history, GNN in combination with soft-attention the compute social interactions, a set-of-polylines to represent the most important map information and LSTM to decode the trajectories from the latent space. The output multimodal prediction is represented by a set of states with their corresponding confidences indicating the most plausible modes.

*Llegaré a ser el mejor, El mejor que habrá jamás*
*Mi causa es ser su entrenador, Tras poderlos capturar.*
*Viajaré a cualquier lugar, Llegaré a cualquier rincón*
*Y al fin podré desentrañar, El poder de su interior.*
*¡Pokémon! Hazte con todos (solos tú y yo),*
*Es mi destino, mi misión*
*¡Pokémon! Tú eres mi amigo fiel,*
*Nos debemos defender.*

Opening 1 de Pokémon: "Gotta catch 'em all!"
Autor original: Jason Paige

## 2.1. Introduction

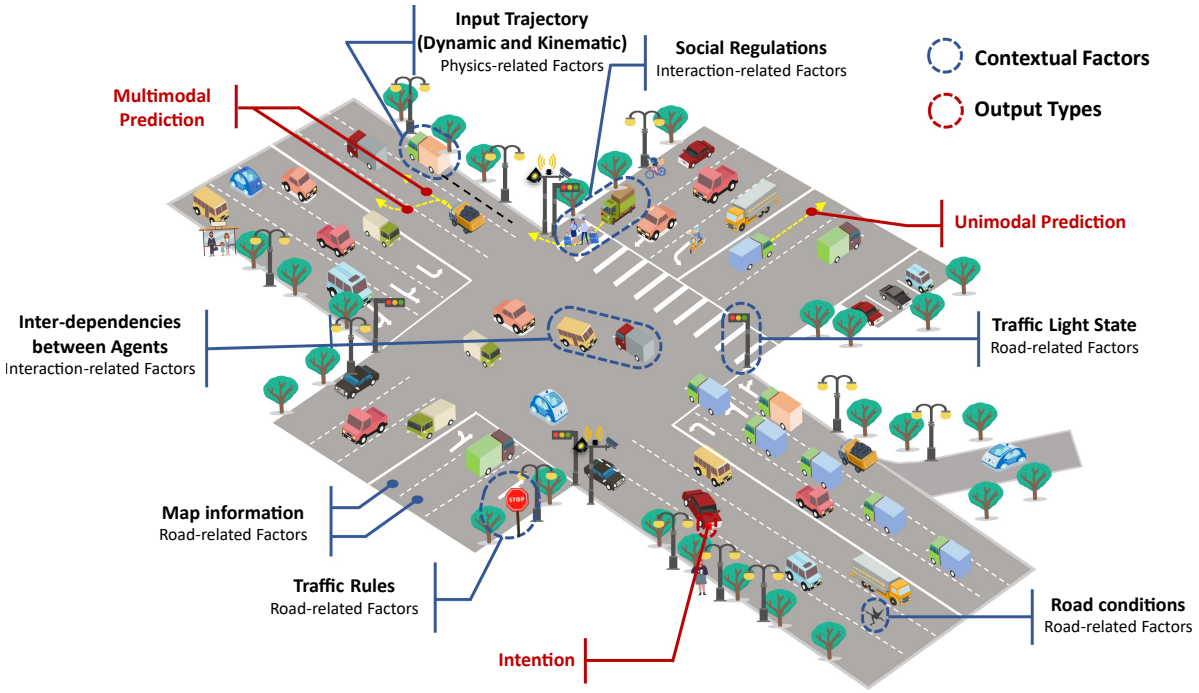


Figure 2.1: Contextual factors and output types in Vehicle Motion Prediction

## 2.2. Problem Formulation of Motion Prediction

Given a sequence of past trajectories $a_P = [a_{-T'+1}, a_{-T'+2}, ..., a_0]$ for an agent, we aim to predict its future steps $a_F = [a_1, a_2, ..., a_T]$ up to a fixed time step $T$. Running in a specific traffic scenario, each actor will interact with static HD maps $m$ and the other dynamic actors, meeting the corresponding traffic and social rules. Therefore, the probabilistic distribution that we want to capture is $p(a_F|m, a_P, a_P^O)$, where $a_P^O$ denotes the other actors' observed states. One important thing to note is that this thesis is focused on non-conditional motion prediction, different to Conditional Behaviour Prediction (CBP). Most existing works focus on a passive prediction scheme, where the future states of a particular agent are predicted given its past information, other surrounding agents information and interactions as well as the physical context. However, when using such a MP model, downstream planning modules, specially the behaviour planning (also referred as decision-making, as stated in Section 1.3), determine the ego-vehicle (our vehicle) action according to the predicted trajectories in a passive manner, that is, without modifying the output of the prediction model. Nevertheless, to ensure safety under various predicted trajectories of the surrounding agents, our vehicle must overly conservative with inefficient maneuvers, especially un highly interactive scenarios, because passive MP models ignore the fact that the future actions of an agent can influence the future actions of other agents, what is the most realistic situation. To this end, researchers started to investigate a more

coherent interactive prediction and planning framework which relies on predicting the surrounding agents future trajectories conditioned on the ego-vehicle future actions [11] [12] [10]. Under such frameworks, the ADS can reason

surrounding agents' future trajectories conditioned on the ego agent's future actions [3], [4], [5], [6], [7], [8], [9], [10]. Under such frameworks, the autonomous agents can reason over potential actions while considering their influence on surrounding agents. It can then induce more efficient and less conservative maneuvers in interactive scenes. Some of these prior works merely demonstrated that their models are able to support conditional prediction from the perspective of architecture [4], [6]. Another line of works focused on

More interestingly, some existing works formulated an alternative prediction task to evaluate the prediction module in a self-contained way [8], [7], [10]. We follow [10] to refer to this task as conditional behavior prediction (CBP). In the CBP task, the future trajectories of the target agents are predicted conditioned on the ground-truth future trajectory of an assigned ego agent. Standard prediction metrics are adopted to quantify the performance. It allows us to leverage large-scale naturalistic traffic datasets to develop and validate a conditional prediction model before closed-loop testings. In those works, a model that can achieve the smallest prediction error after granted the additional future information of the ego agent is considered the best. However, we can only evaluate the prediction accuracy given the actual future trajectory of the ego agent with a static offline dataset. It is impossible to quantify the model performance when it is queried by an arbitrary plan of the ego agent. Therefore, we should be careful when interpreting the evaluation results. In particular, we argue that it is risky to train and evaluate the model for conditional inference. In the current CBP task, the prediction model essentially learns the posterior distribution of future trajectories conditioned on the future trajectory of the ego agent. In this way, the ego agent's future trajectory is treated as an observation. Since the actual ego agents in the offline dataset make decisions according to the states of the surrounding agents, the surrounding agents under CBP are implicitly assumed to get additional hints on the future behavior of the ego agents. With such an unrealistic assumption, it is natural to consider the CBP model with the lowest prediction error as the best option for the CBP task. However, the surrounding agents in the real world are not informed of the planned trajectories of the ego agents. Consequently, as illustrated in Fig. 1, there will be a discrepancy between: 1) what an autonomous agent is informed by querying a CBP model with a potential plan; and 2) how the others will actually react if the agent executes the plan. As we will show later, this discrepancy may lead to overly confident anticipation on the ego agent's influence on the surroundings, resulting in potential safety hazards. This discrepancy is formally captured in the theory of causality [11] by the difference between observation and intervention. With an intervention to a set of random vari- ables, we enforce the value of a random variable without treating it as the consequence of other random variables. The resulting distribution of the remaining random variables under the intervention is consistent with what will actually happen if

we have the privilege to manipulate the target random variable as desired. Consequently, we argue that we should build the prediction model to approximate the future trajectory distribution under the intervention of enforcing the ego agent's future trajectory. We refer to this new task as the interventional behavior prediction (IBP) task. In IBP, we still want to train and evaluate the model with an offline dataset. The task setting is essentially the same as CBP, except for learning an interventional distribution instead of a conditional one. The remaining issue is then how to properly evaluate an IBP model with an offline dataset. Without knowing the ground-truth distribution under intervention, we can only compare the model's output against the ground-truth future trajectories for evaluation. However, such evaluation metrics are naturally biased toward a CBP model. The dataset is collected without intervention. The ego agent in the dataset follows an internal reactive policy. Therefore, the distribution of ground-truth labels given the same input essentially follows a conditional distribution. As a result, a CBP model will always outperform an IBP model if prediction accuracy is the only evaluation metric with an offline dataset. To this end, we propose to verify the inherent temporal independence of a prediction model before comparing the prediction performance to ensure a proper evaluation for the IBP task. Under the interventional distribution, the predicted states of the target agents at earlier timesteps should be independent from the ego agent's states at latter timesteps.

The output of our model is $A_F = \{a_F^k\}_{k\in[0,K-1]} = \{(a_1^k, a_2^k, ..., a_T^k)\}_{k\in[0,K-1]}$ for each actor, while motion forecasting tasks and subsequent decision modules usually expect us to output a set of trajectories. TNT **zhao2020tnt**-like methods' distribution can be approximated as

$$\sum_{\tau\in T(m,a_P,a_P^O)} p(\tau|m, a_P, a_P^O)p(a_F|\tau, m, a_P, a_P^O) \tag{2.1}$$

where $T(m, a_P, a_P^O)$ is the space of candidate goals depending on the driving context. However, the map space $m$ is large, and the goal space $T(m, a_P, a_P^O)$ requires careful design. In that sense, some methods expect to accurately predict the actor motion by extracting good features. For example, LaneGCN **liang2020learninggraph** tries to approximate $p(a_F|m, a_P, a_P^O)$ by modeling $p(a_F|M_{a_0}, a_P, a_P^O)$, where $M_{a_0}$ is a "local" map features that is related to the actor state $a_0$ at final observed step $t = 0$. To extract $M_{a_0}$, they use $a_0$ as an anchor to retrieve its surrounding map elements and aggregate their features. As stated by **wang2022ganet**, computing the local map information is only a part of the solution, but also proposing preliminary guidance for the model in a heuristic way, as well as calculating the goal area maps information using DL, may be of great importance for accuracy trajectory prediction. Then, our future probability distribution is enhanced by these preliminary preprocessed proposals and predicted goals as anchors to explicitly aggregate their surrounding map features as goal areas.

## 2.3. Physic-based Motion Prediction

## 2.4. Deep Learning based Motion Prediction

## 2.5. Vehicle Motion Prediction

# Chapter 3

# Theoretical Background

*Desde que el mundo cambió, estamos mucho más unidos*
*con los Digimon, luchamos juntos contra el mal.*
*Algo extraño pasaba, Digievolucionaban,*
*en tamaño y color, ellos son los Digimon.*

Opening 1 de Digimon: "Butterfly"
Autor original: Kōji Wada

## 3.1. Kalman Filtering

## 3.2. Convolutional Neural Networks

## 3.3. Recurrent Neural Networks

## 3.4. Generative Adversarial Networks

It then applies a function to generate $\mathbf{x}' = G(\mathbf{z})$. The goal of the generator is to fool the discriminator to classify $\mathbf{x}' = G(\mathbf{z})$ as true data, *i.e.*, we want $D(G(\mathbf{z})) \approx 1$. In other words, for a given discriminator $D$, we update the parameters of the generator $G$ to maximize the cross-entropy loss when $y = 0$, *i.e.*,

$$\max_G\{-(1-y)\log(1-D(G(\mathbf{z})))\} = \max_G\{-\log(1-D(G(\mathbf{z})))\}.$$

If the generator does a perfect job, then $D(\mathbf{x}') \approx 1$, so the above loss is near 0, which results in the gradients that are too small to make good progress for the discriminator. So commonly, we minimize the following loss:

$$\min_G\{-y\log(D(G(\mathbf{z})))\} = \min_G\{-\log(D(G(\mathbf{z})))\},$$

which is just feeding $\mathbf{x}' = G(\mathbf{z})$ into the discriminator but giving label $y = 1$.

To sum up, $D$ and $G$ are playing a "minimax" game with the comprehensive objective function:

$$\min_D \max_G \{-E_{x\sim\text{Data}} \log D(\mathbf{x}) - E_{z\sim\text{Noise}} \log(1 - D(G(\mathbf{z})))\}.$$

## 3.5. Attention Mechanisms

## 3.6. Graph Neural Networks

## 3.7. Training losses

# Chapter 4

# Predictive Techniques for Scene Understanding

*Avanzad, sin temor a la oscuridad.*
*Luchad jinetes de Theoden.*
*Caerán las lanzas, se quebrarán los escudos.*
*Aún restará la espada.*
*Rojo será el día, hasta el nacer del sol.*
*Cabalgad, cabalgad, cabalgad hacia la desolación*
*y el fin del mundo. Muerte, muerte, muerte.*

Discurso de Theoden, Rey de Rohan
El Señor de los Anillos: El Retorno del Rey

# Chapter 5

# Applications in Autonomous Driving

*La fuerza de tus convicciones determina tu éxito,*
*no el número de tus seguidores.*

Reamus Lupin
Harry Potter y Las Reliquias de la Muerte, Parte 2

# Chapter 6

# Conclusions and Future Works

*El mundo no es todo alegría y color, es un lugar terrible y por muy duro que seas es capaz de arrodillarte a golpes y tenerte sometido a golpes permanente si no se lo impides; Ni tú ni yo ni nadie golpea mas fuerte que la vida. Pero no importa lo fuerte que golpeas, sino lo fuerte que pueden golpearte y los aguantas mientras avanzas, hay que soportar sin dejar de avanzar.*
*¡Así es como se gana!*
*Si tú sabes lo que vales, vé y consigue lo que mereces pero tendrás que soportar los golpes y no puedes estar diciendo que no estás donde querías llegar por culpa de él o de ella, eso lo hacen los cobardes y tú no lo eres.*
*TÚ ERES CAPAZ DE TODO.*

Discurso de Rocky a su hijo
Rocky Balboa

## 6.1. Conclusions

## 6.2. Future Works

# Bibliography


[1] J. F. Arango, L. M. Bergasa, P. A. Revenga, *et al.*, "Drive-by-wire development process based on ros for an autonomous electric vehicle", *Sensors*, vol. 20, no. 21, p. 6121, 2020.

[2] A. Ranga, F. Giruzzi, J. Bhanushali, *et al.*, "Vrunet: Multi-task learning model for intent prediction of vulnerable road users", *arXiv preprint arXiv:2007.05397*, 2020.

[3] I. Gog, S. Kalra, P. Schafhalter, M. A. Wright, J. E. Gonzalez, and I. Stoica, "Pylot: A modular platform for exploring latency-accuracy tradeoffs in autonomous vehicles", in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 8806–8813.

[4] M.-F. Chang, J. Lambert, P. Sangkloy, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[5] B. Wilson, W. Qi, T. Agarwal, *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting", *arXiv preprint arXiv:2301.00493*, 2023.

[6] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data", in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, Springer, 2020, pp. 683–700.

[7] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving", *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255–2264.

[9] S. Casas, W. Luo, and R. Urtasun, "Intentnet: Learning to predict intention from raw sensor data", in *Conference on Robot Learning*, PMLR, 2018, pp. 947–956.

[10] S. Khandelwal, W. Qi, J. Singh, A. Hartnett, and D. Ramanan, "What-if motion prediction for autonomous driving", *arXiv preprint arXiv:2008.10587*, 2020.

[11] C. Tang and R. R. Salakhutdinov, "Multiple futures prediction", *Advances in neural information processing systems*, vol. 32, 2019.

[12] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2821–2830.