Expert Systems With Applications 238 (2024) 121983

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Pedestrian and vehicle behaviour prediction in autonomous vehicle system — A review

Luiz G. Galvão, M. Nazmul Huda [*]

Department of Electronic and Electrical Engineering, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, London, United Kingdom

ARTICLE INFO

Keywords:
Deep learning
Autonomous vehicle
Pedestrian
Vehicles
Behaviour prediction

ABSTRACT

Autonomous vehicles (AV)s have become a trending topic nowadays since they have the potential to solve traffic problems, such as accidents and congestion. Although AV systems have greatly evolved, it still have their limitations. For example, Google reported that their AVs have been involved in several collisions and near misses. While most of these collisions and near misses were caused by third parties, the AVs should be able to predict and avoid them. Events like this show that there is still room for improvement in the AV system. This paper aims to present a review of the state-of-the-art algorithms proposed to enable AV behaviour prediction systems to predict trajectories and intentions for pedestrians and vehicles. This will be achieved by using information from previous literature review papers, recent works, and results obtained using well-known datasets.

1. Introduction

Road traffic accidents and congestion have posed significant challenges for many countries today. Road traffic accidents claim the lives of 1.35 million people annually and it is ranked 8th leading cause of death worldwide (WHO, 2018). In addition, it has been reported that road traffic accidents are responsible for 20 to 50 million non-fatal causalities, and 95% of these accidents are caused by human errors and imprudence. It reported in the UK that, in 2020 and 2021 there were 92,055 and 119,850 road traffic causalities, respectively, and 1676 of these causalities led to death (GOVUK, 2020, 2021). Congestion has a significant negative impact on society, affecting the economy, environment, public health and safety (Afrin & Yodo, 2020; Levy et al., 2010). Enforced legislation, advanced driving assistance system (ADAS), other methods of transportation and road improvements have been used to address these road traffic issues. However, it is predicted that the number of road users will double by 2050 and these current measures will not be sufficient (COLONNA, 2018). AVs are a trending topic nowadays and companies such as Waymo and Uber have already deployed several AVs on the roads to solve the aforementioned road traffic problems. Although AV systems have considerably evolved, they still have limitations such as efficiently and safely navigating in complex scenarios. This could be achieved by avoiding congestion, predicting, preventing, or mitigating any road traffic collisions. These are challenging tasks since the AVs have to share the roads with human road users, and as reported by World Health Organisation (WHO), most road traffic collisions are linked to human error and imprudence. Another major AV limitation is gaining public confidence that they are safe to ride. Authors (Petrović et al., 2020) investigated 300 traffic collisions in California (US) between 2015 and 2017 that involved AVs. They found that most of the collisions were caused by conventional drivers, who were following the AVs too close, and violated the right-of-way, traffic signals, and traffic signs. Google published a paper reporting the performance of their Waymo driver between 2019 and 2020, to show transparency and make the public more comfortable and confident with AVs. In the report, the Waymo driver drove 6.1 million miles and was involved in 47 road traffic collisions and near-miss events - these include both actual and counterfactual simulated events (Schwall et al., 2020). Most of the reported collisions were induced by humans where one or more road traffic rules were broken, such as violating the speed limit, driving on the wrong side of the road, not obeying the stop sign or the red traffic light signal, performing inappropriate lane change or junction merging, not yielding the right of way to the Waymo driver, and not yielding to the slowing down behaviour of the Waymo's driver. Although some of these cited events could not be avoided by the AV drivers, for example, the conventional drivers hitting the rear of the AV while it was stationary or slowing down, there were instances where they could have been. For instance, accidents caused by changing lane manoeuvre, merging from a junction, or making a turning manoeuvre, could have been avoided if the AV was able to make an accurate and longer prediction horizon of the trajectories and intention of the conventional drivers. This shows that there is still room for improvement in the AV system, mainly in the behaviour prediction

[*] Corresponding author.
E-mail addresses: luizg.galvao@brunel.ac.uk (L.G. Galvão), mdnazmul.huda@brunel.ac.uk (M.N. Huda).

https://doi.org/10.1016/j.eswa.2023.121983
Received 24 December 2022; Received in revised form 11 September 2023; Accepted 1 October 2023
Available online 12 October 2023

of other road users since it enables the AVs to make a risk assessment of the situation in order to take appropriate action. The goal of this paper is to review the most relevant works that aimed to predict the trajectories and intentions of vehicles and pedestrians.

There are several literature reviews covering both traditional and Deep Learning (DL) techniques to predict the behaviour of vehicles, for example, Lefèvre et al. (2014), Leon and Gavrilescu (2019), Shirazi and Morris (2016), Sivaraman and Trivedi (2013) and Mozaffari et al. (2020). Sivaraman and Trivedi (2013) briefly reviewed the behaviour prediction of vehicles but at that time this topic was fairly new and only traditional techniques were reviewed. Lefèvre et al. (2014) presented a survey and classified vehicle prediction behaviour algorithms into physics-based, manoeuvre-based, and interaction-aware-based algorithms. They concluded that a behaviour prediction algorithm needs to consider the interaction between vehicles as well as the scene context to have a longer prediction horizon. In addition, they reviewed the existing risk assessment methods for autonomous vehicles and concluded that a risk assessment module was highly dependent on the behaviour prediction algorithm. In this review, the authors only covered traditional techniques since DL techniques for vehicle behaviour prediction were still emerging at the time. Shirazi and Morris (2016) reviewed techniques used to analyse vehicles, drivers, and pedestrians' behaviour at road intersections. Only traditional techniques were analysed, however, the focus was not on the prediction behaviour of the vehicles. Leon and Gavrilescu (2019) reviewed methods used for vehicle tracking, behaviour prediction, and decision-making. Both traditional and DL techniques have been covered. The authors concluded that DL techniques had better results since they are more robust, flexible and have better generalisation ability. Mozaffari et al. (2020) performed a systematic and comparative review of the different DL methods used to predict vehicle trajectories and its intentions. They presented a more detailed taxonomy of the prediction behaviour algorithms compared to Lefèvre et al. (2014). They categorised the algorithms based on the type of input, the type of output, and the method of prediction. Although the review was extensive and very informative, the authors did not cover in detail what intention behaviour the works were trying to predict, for example, lane change, overtaking, or making a turn; and do not provide specific information on what datasets were used.

The following works have performed pedestrian behaviour prediction reviews, Chen, Ding, et al. (2020), Kong and Fu (2018), Ridel et al. (2018), Rudenko et al. (2020), Sharma et al. (2022) and Ahmed et al. (2019a). Kong and Fu (2018) presented traditional and DL techniques that were used to recognise and predict human action. Ahmed et al. (2019a) presented a survey on the detection and intention prediction of pedestrians and cyclists. A review on pedestrian behaviour was presented by Ridel et al. (2018), where they briefly described the traditional and DL techniques that were used. Chen, Li, et al. (2020) discussed the required architecture, the traditional and DL techniques to detect and predict pedestrian actions. Although, these works reviewed DL techniques, only a limited amount of works were considered. A detailed human trajectory prediction survey was done by Rudenko et al. (2020), where they reviewed a substantial amount of published works to propose a taxonomy, identify the available datasets and evaluation metrics, and the limitations of the current methods. However, the authors did not review methods used to predict pedestrian intentions. A comprehensive survey was done by Sharma et al. (2022) on pedestrian intention prediction for AV systems.

To the authors' knowledge, the work presented by Gulzar et al. (2021) is the only one that reviewed the behaviour prediction of both pedestrian and vehicle. The authors presented a novel taxonomy that unifies both pedestrian and vehicle behaviour prediction problems. However, the authors did not explore the evaluation metrics, datasets, features, and the results of the reviewed works.

Unlike the previously cited review works on both pedestrian and vehicle behaviour prediction, this paper:

- Presents a behaviour prediction general problem formulation.
- Presents the most used terminologies in the pedestrian and vehicle behaviour prediction domain.
- Reviews not only pedestrian or vehicle behaviour prediction algorithms, but both of them;
- Briefly presents the most important traditional techniques and focuses more on the DL techniques for pedestrian, and vehicles prediction algorithms;
- Summarises the key information extracted from the reviewed studies on predicting pedestrian and vehicle behaviour in tables. These tables report the methods employed, the problem that the algorithms are trying to solve, the datasets used, and the results acquired.
- Reviews works that have performed prediction behaviour of heterogeneous agent traffic.
- Introduce a general framework for a behaviour prediction system highlighting the system dependence on the AV's hardware and the perception module, and its typical outputs. In addition presents a risk assessment for a general behaviour prediction system.
- Identifies the requirements and challenges to design a pedestrian and vehicle behaviour prediction system for AV.
- Discusses whether the current techniques have met the previously mentioned requirements, and suggests future works.

Some of the commonly used terminologies in the pedestrian and vehicle behaviour prediction literature are listed below (Mozaffari et al., 2020).

- **Object behaviour:** means the object trajectories or intentions.
- **Object trajectory:** vectors with a sequence of data, typically comprised of tracking information that describes the path an object had followed.
- **Object intention:** is a course of actions that an object intends to perform to achieve its goal. In the vehicle domain, these courses of action are known as manoeuvres, such as, turning, changing lanes, stopping, cut-in/cut-out, etc. In the pedestrian domain, these intentions are crossing/non-crossing, stopping, etc.
- **Observation time horizon (OTH):** the time that an algorithm observes the past behaviours of an object to predict its future behaviour.
- **Prediction time horizon (PTH):** Most of the reviewed works use prediction horizon to refer to the time that an algorithm can predict an object's behaviour before it happens. However, in some works, the term 'prediction' is replaced with 'anticipation', and it is defined as the time that an algorithm can predict an object's behaviour before it begins. This paper adopts the term prediction and its first meaning.
- **Ego Vehicle (EV):** observes the others traffic agents using onboard sensors.
- **Target object:** the object that the EV is observing to predict its behaviour.
- **Surrounding objects:** the objects that may interact with and affect the behaviour of the target object.
- **Multi-modal behaviour:** means that an observed history of behaviours could lead to multiple several potential future behaviours.
- **Trajectory prediction:** means to predict the future motion of an object given a time frame of its and/or surrounding object's trajectories, contextual information, and interactions between the objects in the scene.
- **Intention prediction:** usually uses the same history information that trajectory prediction uses, however, the system aims to predict the future discrete action of the target object.
- **Interaction:** influences that one or more objects have on each other.

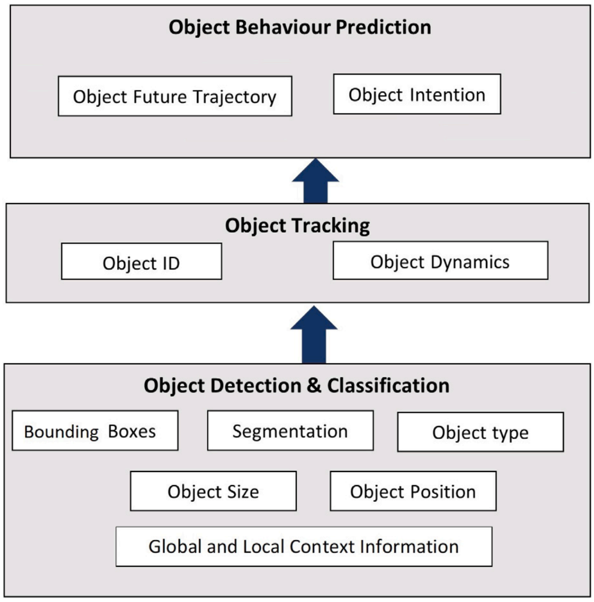

Fig. 1. Object behaviour prediction full pipeline process. The detection and classification stage outputs the object position, size, type, bounding box, segmentation, and, global and local context information. The object tracking stage outputs the ID for each detected object and its dynamics (e.g., speed). The output of the object behaviour prediction module can be the object's intention and its future trajectory.

A typical AV system architecture comprises perception, planning, and acting modules (Durrant-Whyte, 2001; Galvao et al., 2021; Pendleton et al., 2017; Siegwart et al., 2011). However, the Waymo driver system has an additional module called behaviour prediction, which comes before the planning module (Waymo, 2020). The perception module is responsible to inform what is around the AV, for example, static objects (e.g., traffic lights, traffic signs, road works, etc.) and non-static objects (e.g., pedestrians, vehicles, etc.), and traffic road contexts (e.g., road lanes, edges, curbs, pedestrian crossings, etc.). The behaviour prediction module is responsible to anticipate the behaviour (e.g., trajectories and intentions) of other traffic agents. The planning module takes the perceived and predicted information to decide what action the AV should take in order to achieve its final goal. Finally, the acting module performs the actual motion of the AV through actuators that control the steering wheel, accelerator, and brakes. This paper adopts the Waymo driver architecture, that behaviour prediction is a separate module in an AV system.

This paper is structured as follows: Section 2 presents a general problem formulation for pedestrian and vehicle behaviour prediction; Sections 3 and 4 present the most relevant algorithms used to predict the behaviour of vehicles and pedestrians, respectively; Section 5 presents behaviour prediction algorithms for heterogeneous road agents; and Section 6 discusses the findings.

2. Behaviour prediction general problem formulation

A full pipeline of an object behaviour prediction system, as depicted in Fig. 1, is composed of detection and classification, tracking, and prediction stages. This paper only reviews works that studied the behaviour prediction stage. Literature reviews on the detection and tracking can be found in the following works (Abbas et al., 2021; Ahmed et al., 2019b; Antonio & Romero, 2018; Dendorfer et al., 2021; Galvao et al., 2021; Ragesh & Rajesh, 2019; Shobha & Deepu, 2018).

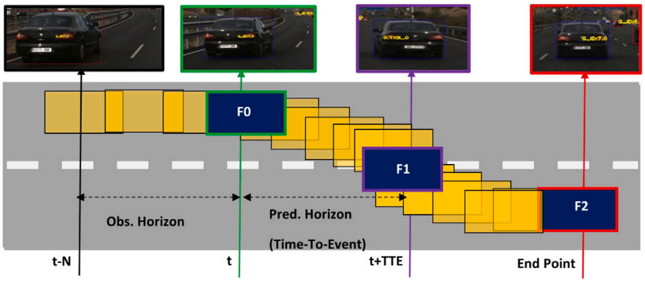

Fig. 2. An example of lane change prediction problem: F0 is where the vehicle manoeuvre starts, F1 is where the actual manoeuvre happens, and F2 is the end of the manoeuvre.

Based on the vehicle and pedestrian intention prediction problem formulation proposed by Achaji et al. (2022), Biparva et al. (2021), Bouhsain et al. (2020), Fernández-Llorca et al. (2020), Gazzeh and Douik (2022), Izquierdo et al. (2021), Kotseruba et al. (2020), Naik et al. (2022), Piccoli et al. (2020), Rasouli et al. (2019, 2020), Vitas et al. (2020), Yang, Zhang, et al. (2022), Yao et al. (2021b), Zeng (2022), Zhang, Angeloudis, and Demiris (2022), and Xue et al. (2018), a general intention prediction problem formulation is as follows: a sequence of feature vector $\{F_{t-OTH}, \ldots, F_t\}$ extracted from a given sequence of video frames $\{t - OTH, \ldots, t\}$ acquired from an image sensor is used by a model to determine the probability of the target agent intention $I_a^{t+n} \epsilon \{0, 1\}$, where t is the specific time of the last observed frame and n is the number of frames from the last observed frame to the final frame of the event, also known as time-to-event (TTE). The prediction intention estimation can be described by the equation

$$p(I_a | F_{t-T_{obs}:t}). \tag{1}$$

Based on the vehicle and pedestrian trajectory prediction problem formulation proposed by Altché and de La Fortelle (2017), Dai et al. (2019), Deo and Trivedi (2018a, 2018b), Kim et al. (2017), Lee, Choi, et al. (2017), Li et al. (2019a, 2019b), Mangalam et al. (2020), Messaoud et al. (2019), Mohamed et al. (2020), Sadeghian et al. (2019), Sun et al. (2020), Vemula et al. (2018), Xin et al. (2018), Xu et al. (2018), Zhang et al. (2019), Zhu et al. (2019), and Gupta et al. (2018), a general trajectory prediction problem could use the same sequence of feature vector used by intention prediction algorithm. However, in this case, the sequence's purpose is to predict the future path of the target agent, spanning up to the specified PTH. The trajectory prediction estimation can be described by the equation

$$p(FuturePath_{t:PTH} | F_{t-T_{obs}:t}). \tag{2}$$

Fig. 2 depicts an example of predicting a vehicle's lane change manoeuvre. Here, image sequences from $t - N$ to t are used to extract a sequence of feature vectors, which are subsequently used to make predictions. In this case, a successful lane change manoeuvre prediction occurs when the vehicle's intention is correctly recognised before reaching the F1 stage.

The differences among the reviewed problem formulation of vehicle and pedestrian behaviour prediction are:

- Some problem formulations are for trajectories and others for intention.
- The input features used may be different, for example, some authors have used only position and speed, while others have used local and global context vectors.
- Some considered top-view and others considered on-board view datasets.
- Authors have used different predictive models.

The listed items above are further discussed in the following Sections 3–5.

In the literature, some works use predicted intentions to improve the accuracy of the future trajectories, and other works use predicted trajectories to improve the accuracy of the predicted intention (Biparva

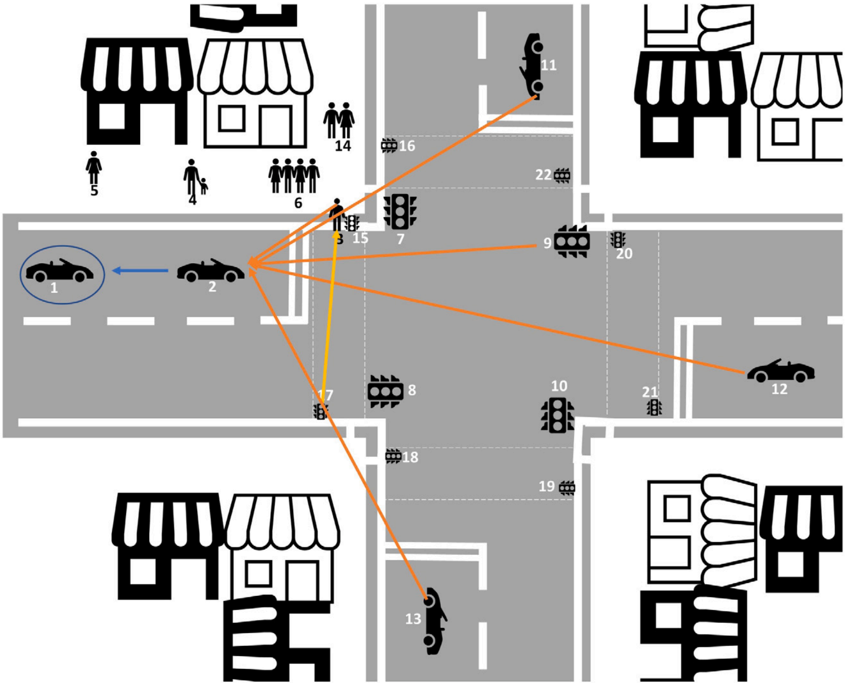

Fig. 3. General interactions among traffic agents and their environments. Object 1 is the target object (blue circled), the blue arrow shows the direct interaction between the target object and object 2; the orange arrows show the interaction between object 2 and objects 3, 9, 11, 12, and 13; the yellow arrow shows the interaction between object 17 and object 3.

et al., 2021; Mozaffari et al., 2020). These works will be discussed in the upcoming sections.

Before discussing the behaviour prediction of pedestrians and vehicles, it is important to understand their potential interactions. As depicted in Fig. 3, interactions among different traffic agents can cascade and get very challenging, for example, in order to predict the actions of object 1, it might be required to consider the actions of the:

- Object 2, since it can change direction and velocity.
- Object 3, since its action will affect the action of object 2.
- Object 17, since it will affect the action of object 3.
- Object 9, since it will affect the action of object 2.
- Object 13, since it can make a right turn, which will affect the action of object 2.
- Object 11 or 13, since they may break the law by not obeying the red traffic light.

Table 1
Motion, context and intention features that can be used to predict vehicle behaviour.

Information	Features
MOTION	Target Vehicle (TV): Lateral/longitudinal position, velocity, acceleration, yaw, yaw rate, and relative speed. TV-to-lane: lateral offset, and lateral speed. TV-to-Surrounding Vehicle (SV): distance from surrounding vehicles.
CONTEXT	Road: Lane marking, number of lanes, lane width, lane curvature, type of lines, entries, exits, left/right/forward arrows, crosswalks, traffic light, traffic signs, type of roads (urban, country, highway-motorway), bumps, road holes, road works, left/right-hand side traffic, and junctions. Vehicle: indicators, brake lights, warning lights, type of the vehicle, and sirens' light status. Other road agents: pedestrians, animals, cyclists, and trams. Environment: sunny, snowing, rainy, foggy, and dark.
INTENTION	Braking, turning left/right, lane keeping, left/right lane change, speeding, normal driving, aggressive driving, abnormal driving, merging, exiting, cutting in/out, and yielding.

3. Vehicle behaviour prediction

In the vehicle behaviour prediction domain, the literature often uses the terms prediction behaviour of drivers/vehicles or prediction behaviour of target/surrounding vehicles. The former usually means to predict the behaviour of the ego vehicle using its internal data, such as the steering angle, brake pedal position, velocity, speed, indicators status, etc. Berndt and Dietmayer (2009), Girma et al. (2020), Raimundo and Favio (2021), Xing et al. (2017). This approach is suitable for AV systems when considering vehicle-to-vehicle communication. The latter approach involves the ego vehicle using on-board sensors to gather information from the surrounding vehicles to predict their behaviour. In this review, only the latter approach is reviewed, as vehicle-to-vehicle communication is not yet available, and AVs would still share roads with conventional human drivers.

Vehicle behaviour prediction is a crucial component of the AV behaviour prediction system as it would enable the AV to perform risk assessment, plan future movements, and make appropriate decisions to avoid/mitigate the impact of collisions. Ideally, a vehicle behaviour prediction algorithm should be fast, cost-effective, accurate, generalise well in different traffic scenes, consider the interdependence between agents, and have a long prediction horizon. A long prediction horizon provides the AV with more time to make decisions and take appropriate actions. A typical vehicle behaviour prediction pipeline consists of multiple steps, starting with the detection of the target vehicle and the surrounding vehicles. This detection information is used to obtain tracking information. Subsequently, this tracking information is used as an observation feature to predict future trajectories. In order to enhance the quality and duration of predictions, context information of the traffic scene and the intention manoeuvre of other vehicles can be considered. Table 1 provides the type of motion, context, and intention information that has been and could be used by the researchers to predict vehicle behaviour.

Although vehicles have some characteristics that simplify their behaviour prediction, such as constrained movement due to their inertial property, having to obey traffic road rules, and navigating inside the road boundaries. It is still a challenging task since their behaviour is dependent on other vehicles' actions, traffic regulations, road geometry, and different driving environments (Lefèvre et al., 2014; Mozaffari et al., 2020). Moreover, vehicles have multi-modal behaviour, different types of vehicles might provide different motion information, and prediction can be affected if surrounding vehicles are occluded.

The main two sources of data used to predict the behaviour of vehicles are top-view and on-board sensors. Top-view data are captured from static sensors usually installed on tall buildings, while on-board sensors are captured from sensors installed on the EV. Top-view data have the advantage of providing more precise information since the acquired data have better quality, the vehicles surrounding the TV are captured, and vehicles are not easily occluded. However, it only covers a specific and fixed portion of the traffic scene, limiting the algorithm to generalise to other traffic scenarios. Top-view sensors are typically used in two types of traffic environments: highways-motorways and complex traffic scenes, such as busy urban areas and junctions. Highway-Motorway datasets can suffer imbalanced samples, where there are more instances of constant velocity behaviour than the specific manoeuvres of interest (Altché & de La Fortelle, 2017). On-board sensor data can capture different traffic scenarios, however, its data quality can be affected by noises, surrounding vehicles can be occluded, and in order to detect all the vehicles surrounding the EV and the TV, more than one sensor might be required (e.g., front, rear, and sides cameras.) (Izquierdo et al., 2021). On-board sensor data is particularly advantageous for AV applications because the algorithms that use them, could be directly integrated into AVs, which are already equipped with on-board sensors. Several sensors, such as cameras, radar, and LIDAR could be used to acquire both top-view and on-board data (Izquierdo et al., 2019; SIMulation, 2007; Zhou et al., 2020; Zyner et al., 2019). However, this research mainly focuses on works that have used camera sensors. For more information about the available datasets for vehicle behaviour prediction, please refer to Izquierdo et al. (2021). Table 2 summarise the most relevant vehicle trajectory and intention prediction works from 2009 to 2022. From the table, it is observed the following:

- **Shift to Deep Learning and NGSIM Dataset**: Up to 2016, the majority of the works used traditional techniques and their OWN datasets, however, after 2016 most of the works adopted DL techniques and used the NGSIM dataset.
- **Expanding Information Sources**: Vehicle behaviour prediction algorithms have evolved from using only motion information to incorporating additional sources, including manoeuvre, interaction, and driver-style information.
- **Limited Use of Other Datasets**: While the NGSIM dataset gained popularity, other datasets such as Apollo, KITTI, LISA, INTERACTION, HighD, and PREVENTION were rarely used.
- **Trajectory Prediction Dominance**: The majority of the research efforts was to predict trajectories. It was not until 2020 that more research began to address the prediction and recognition of vehicle intentions.
- **Focus on Lane Changing and Turning Manoeuvres**: Most research works focused on predicting the trajectories and intentions related to lane changing and turning manoeuvres. Other types of manoeuvres such as reversing, braking, and U-turns were seldomly used.
- **Evaluation Metrics**: The most common evaluation metric for trajectory prediction was the Root Mean Square Error (RMSE), while for intention prediction, accuracy was the predominant evaluation metric.

The following two subsections discuss the algorithms used to predict vehicle behaviour. The first covers the algorithms used to predict trajectories, and the latter, the algorithms used to detect and predict vehicle intention.

3.1. Trajectory prediction

As reported in Table 2, vehicle trajectory prediction has been achieved using one or more of the following approaches: physics-based, manoeuvre-based, or interaction-aware motion models (Lefèvre et al., 2014). Physics-based motion models were one of the first approaches to be proposed and it uses the principles of physics to predict vehicle motions. This approach is computationally efficient, meets real-time requirements, and does not require the dataset to be human-labelled. However, they are less suitable for complex scenarios like busy urban scenarios and junctions. This is because they do not take into account the TV intentions, the contextual information of the scene, or the interaction between the TV and the SVs. This lack of information limits the prediction horizon for the EV to less than 1 s (Lefèvre et al., 2014). In order to overcome the limitation of a short prediction horizon associated with the physics-based approach, manoeuvre-based approaches were introduced. In the manoeuvre-based approach, the EV uses the predicted intention of the TV to predict future trajectories. This increases both the trajectory prediction horizon and accuracy, as the predicted trajectory would match the predicted intention. However, if the predicted manoeuvre is incorrect, the whole predicted trajectory may also be inaccurate. The interaction-aware approach uses the trajectories and the intentions of both the TV and the SVs to predict the TV trajectory. This approach further extends the prediction horizon and improves the accuracy of the predicted trajectories. On the other hand, it comes with complexities in implementation, demands greater computational power, and raises questions about how to determine which vehicles should be considered as SVs, and not all SV might be reliably detected by the EV.

The previously cited approaches have been implemented using either traditional or DL techniques. Traditional techniques encompass Linear methods like KF and Switching Linear Dynamic Models, as well as Non-linear methods such as EKF, UKF, Switching Non-Linear Dynamic Models, Particle filters, Bayesian filtering, Monte Carlo simulation, Naive Bayes Classifiers, Dynamic Bayesian Networks, HMM, SVM, case-based reasoning, random decision Forest, Artificial Neural Network (ANN), SVM, and Gaussian Process NN (Biparva et al., 2021). Traditional techniques have the advantage of being fast to infer and not requiring an extensive dataset. However, they struggle to generalise well and have limited prediction horizons. Additionally, most traditional techniques do not inherently account for vehicle interactions and may require additional features. The DL techniques used in the literature were based on ANNs, Convolutional Neural Networks (CNN), Fully Connect Networks (FCN), Recurrent Neural Networks (RNN), Graph Convolutional Neural Networks (GCNN), Gated Recurrent Unit (GRU), or Long-short Term Memory (LSTM) (Biparva et al., 2021). The main advantage of DL techniques is their ability to implicitly extract the required features to predict vehicle behaviour. Some DL techniques even consider the interaction between vehicles by themselves, for instance, RNN and GCNNs. Yet, DL techniques may not address the multi-modal behaviour of vehicles as they tend to average the multiple possible modalities to minimise the regression error. They also require an extensive dataset to generalise well, take longer to train, may suffer gradient vanishing, and may not provide accurate trajectory prediction for longer time horizons.

The following paragraphs will discuss the most relevant DL algorithms used to predict vehicle trajectories.

Altché and de La Fortelle (2017) and Kim et al. (2017), to the authors' knowledge, were one of the first ones to use LSTM-RNN to predict the future trajectories of the surrounding vehicles by using their past trajectories as input feature. Park et al. (2018) predicted future trajectories using an encoder–decoder LSTM. The encoder encodes past trajectories of the surrounding vehicles, while the decoder decodes future trajectories in an Occupancy Grid Map (OGM). The authors also applied a beam search algorithm, to reduce the error propagation

Table 2
Relevant works for **vehicle** trajectory and intention prediction.

Work	Methods	Algorithm objectives	Dataset-results
PF+RBF Hermes et al. (2009)	Trajectory prototype. Particle Filter (PF) to track and generate motion hypothesis. RBF to classify trajectories. QRLCS to measure similarity between trajectories. **Evaluation**: RMSE.	Predict future trajectories of the ego and surrounding vehicles.	**OWN** See Table 3.
Lim et al. (2010)	Extended Kalman Filter **Evaluation**: Mean Distance Error.	Estimate Position and Velocity.	**OWN** Graphs.
Kasper et al. (2012)	Bayesian Networks. Occupancy Grid Map (OGM). **Evaluation**: FP, FN, and Accuracy.	Detection of lane change manoeuvre.	**OWN** Accuracy: 83.8%.
Kumar et al. (2013)	SVM. Bayesian Filter. **Evaluation**: Recall, Precision, and F1-score.	Predict lane change manoeuvres of the EV.	**OWN** Recall: 1 Precision: 0.8 F1-score: 0.9 APT: 0.97 s
Yoon and Kum (2016)	Target lane model to predict in which lane the target vehicle will go. 3rd Order Linear System to model trajectory. Auto encoder to cluster the available trajectories into 3 prototype trajectories. Multi-layer Perceptron (MLP) network to predict the target lane and the probability for each one of the prototype trajectories. OTH/PTH: (1 s, 2 s, 3 s, 4 s,5 s)/5 s. **Evaluation**: Prediction time and absolute error of lateral position.	Predict lane change of surrounding vehicles.	**NGSIM** Absolute error: 0.7 m.
Khosroshahi et al. (2016)	Features: linear changes, angular changes, and angular changes histogram. Multi-layer LSTM. **Evaluation**: Accuracy.	Classify manoeuvre intention at intersections.	KITTI 2 classes: 85%. 3 classes: 75%. 8 classes: 65%. 12 classes: 40%.
Dueholm et al. (2016)	Detection: DMP + Feature Pyramid + HOG Tracking: MDP + TLD. Trajectory: KF. **Evaluation**: Recall.	Predict future trajectories of the surrounding vehicles.	**OWN** Recall: 92%
Kim et al. (2017)	LSMT-RNN. OGM. Data-driven approach. PTH: 0.5 s, 1 s, and 2 s. Information: Position, the velocity of surrounding vehicles, and velocity and yaw rate of ego vehicle. **Evaluation**: Mean Absolute Error(MAE).	Predict the future position of the surrounding vehicle using OGM.	**OWN** MAE:1.51 for 2 s; 0.88 for 1 s; and 0.59 for 0.5 s.
Lee, Kwon, et al. (2017)	CNN. **Evaluation**: Accuracy.	Predict lane change manoeuvre.	OWN Accuracy: 89.87%
DESIRE Lee, Choi, et al. (2017)	Observation, sample generation, and rank refinement. CVAE + RNN (GRU) to predict multi-modal trajectories considering latent variables. IOC (based on Reinforcement Learning) to rank and refine the predicted trajectories. Spatial Grid-Based Pooling Layer to extract interaction feature. SCF to combine agents' interactions and scene context. OTH/PTH: 2 s/4 s. **Evaluation**: L2 distance error and miss rate.	Predict the future position of the surrounding vehicles considering static and dynamic scene context and interaction between agents.	SDD **KITTI** See Table 3.
Altché and de La Fortelle (2017)	LSTM encoder–decoder. **Evaluation**: average RMSE.	Predict the target vehicle's future position by considering surrounding vehicles.	NGSIM See Table 3.
Xing et al. (2017)	Two LSTM networks, one to encode past trajectories and predict intention manoeuvre, the other to encode past trajectories, and the predicted manoeuvre to decode future trajectories. **Evaluation**: lateral and longitudinal RMSE.	Predict vehicle trajectory using past trajectories and predicted manoeuvre intention.	NGSIM See Table 3.
Park et al. (2018)	LSTM encoder–decoder. OGM. Beam search algorithm. OTH/PTH: 3 s/2 s. **Evaluation**: MAE.	Predict the future position of the target and the surrounding vehicles.	**OWN** MAE (Grid): 1.27 for 2 s; 1.14 for 1.6 s; 0.99 for 1.2 s; 0.84 for 0.8 s; and 0.64 for 0.4 s.

(*continued on next page*)

Table 2 (*continued*).

Work	Methods	Algorithm objectives	Dataset-results
M-LSTM Deo and Trivedi (2018b)	Tracking history and Manoeuvres classification (Lane change, brake, and normal driving) to allow multi-modal prediction. LSTM encoder–decoder to encode tracked history motions and to decode multi-modal future motions. OTH/PTH: 3 s/5 s. **Evaluation**: RMSE.	Trajectory prediction of surrounding vehicles considering the interaction between traffic agents.	**NGSIM** See Table 3.
C-VGMM+VIM Deo et al. (2018)	HMM for manoeuvre recognition. IMM + VGMM to predict trajectories. Markov Random Field for vehicle interaction. PTH: 5 s **Evaluation**: Manoeuvre classification accuracy, mean and median error for the trajectory prediction.	Manoeuvre Intention (lane change, overtaking, cutting-in, drift into ego lane) and Trajectory Prediction.	**LISA-A** MAE overtakes and cut-ins: 2.49 for 5 s; 1.94 for 4 s; 1.39 for 3 s; 0.82 for 2 s; and 0.29 for 1 s. MAE stop-and-go: 2.17 for 5 s; 1.65 for 4 s; 1.14 for 3 s; 0.64 for 2 s; 0.20 for 1 s. Accuracy for overtakes and cut-ins: 55.89% Accuracy stop-and-go: 87.19% Time: 6FPS.
CS-LSTM Deo and Trivedi (2018a)	LSTM encoder–decoder to encode previous motion information and to decode future motion. Convolutional Social Pooling to learn agent's interdependence motions. Multi-modal prediction (6 classes: RLC, LLC, NLC, brake, and normal). OTH/PTH: 3 s / 5 s. **Evaluation**: RMSE and Negative log-likelihood (NLL).	Predict future motions of surrounding vehicles taking into consideration motion, spatial configuration, and interdependence between agents.	**NGSIM** See Table 3. Computation time: 0.29 s (reported by Li et al. (2019b)).
SA-LSTM Su et al. (2018)	Surrounding-Aware LSTM. OTH: 6, 9, and 12 frames. **Evaluation**: Accuracy.	Predict lane change manoeuvre and future trajectories.	**NGSIM** Avg. Accuracy: 86.19%.
MATF Zhao et al. (2019)	Hybrid Model (LSTM + CNN) LSTM to encode past trajectories for multiple agents. CNN to encode context information. MATF to fuse interaction, spatial structure, and context information. Conditional generative adversarial training to detect uncertainty in predicting manoeuvres. Environment: Highway-Motorway and pedestrian crowd scenes. OTH/PTH: 3 s / 5 s. **Evaluation**: RMSE.	Trajectory prediction by considering social interaction and scene context.	**NGSIM** See Table 3.
Benterki et al. (2019)	Features: local position, velocity, acceleration, distance to lane markings, yaw angle and rate, lateral velocity, and acceleration. ANN and SVM. **Evaluation**: Recall, Accuracy, Precision, and F1-score.	Predict lane change manoeuvres of the surrounding vehicles.	**NGSIM** **ANN** Accuracy: 98.8%. Prediction: 2.4 s. **SVM** Accuracy: 97.1%. Prediction: 1.9 s.
ST-LSTM Dai et al. (2019)	Spatio-temporal LSTM. Short-cut connections to avoid gradient vanishing. Weighted sum to integrate the outputs. Consider the 6 vehicles around the target vehicle. OTH/PTH: 3 s/6 s. **Evaluation**: RMSE.	Trajectory prediction by considering spatial and temporal information.	**NGSIM I-80** See Table 3.
GRIP Li et al. (2019b)	Fixed Graph Convolutional (10 blocks) Model to represent interactions between agents. Single LSTM encoder–decoder to make trajectory predictions. OTH/PTH: 3 s/5 s. Hardware: 4.0 GHz i7, 32GB memory, and NVIDIA Titan XP. **Evaluation**: RMSE.	Predict surrounding vehicle trajectories considering the interaction between them.	**NGSIM** See Table 3. Computation time: 0.05 s.
GRIP++ Li et al. (2019a)	Dynamic Graph Convolutional (3 blocks) Model to represent interactions between agents. Three GRU-RNN encoder–decoder to make trajectory predictions. OTH/PTH: 3 s/5 s. Hardware: 4.0 GHz i7, 32GB memory, and NVIDIA Titan XP. **Evaluation**: RMSE, WSADE, and WSFDE.	Predict surrounding vehicle trajectories considering the interaction between them.	**ApolloScape** WSADE: 1.2588. WSFDE: 2.3631. NGSIM See Table 3. Computation time: 0.02 s.
NLS-LSTM Messaoud et al. (2019)	Local and non-local social pooling. LSTM encoder–decoder. **Evaluation**: RMSE.	Predict vehicle trajectory using local and non-local social pooling.	**HighD** See Table 3 **NGSIM** See Table 3

(*continued on next page*)

Table 2 (*continued*).

Work	Methods	Algorithm objectives	Dataset-results
Benterki et al. (2020)	Hybrid Model ANN to classify manoeuvres. LSTM to predict trajectories. OTH: 3 s, 5 s, and 6 s. PTH: 1 s, 3 s, and 5 s. **Evaluation**: RMSE and classification accuracy.	Manoeuvre classification and trajectory prediction.	**NGSIM** See Table 3
Fernández-Llorca et al. (2020)	Two stream CNN (Disjoint). Spatio-temporal Multiplier Networks (ST) (cross-stream connections). ResNet-50 to extract both temporal and contextual information. OTH/PTH: 2 s/(1–2 s). 4 Sizes of RoI are used x1, x2, x3 and x4. Dense optical flow to extract movement context. **Evaluation**: Classification accuracy and Prediction Accuracy.	Recognition and prediction of lane change/keep manoeuvre using stacked visual cues from videos.	**PREVENTION** *Disjoint* Classification Accuracy:89.46%. Prediction Accuracy:91.02%. *ST* Classification Accuracy: 90.30%. Prediction Accuracy: 91.94%.
ARIMA-Bi-LSTM Zhang and Fu (2020)	Off-line Bi-LSTM. Online ARIMA + Bi-LSTM. PTH: 5 s. **Evaluation**: RMSE and Accuracy.	Predict trajectories and turning manoeuvres at intersections.	**NGSIM-LP** GS: lateral 0.032; long. 0.1093. TL: lateral 0.2719; long. 0.1592. TR: lateral 0.1168 long. 0.3954 Accuracy: 94.2% at 1 s, 93.5% at 2 s, and 74.5% at 3 s.
Izquierdo et al. (2021)	TSM to differ between target and surrounding vehicles. TIM to extract motion pattern. Greyscale image to extract context information. Compared various CNN models to detect and predict manoeuvres. OTH: 1 s. **Evaluation**: Accuracy, precision, recall, anticipation (s), and AUC.	Detection and prediction of lane change performed by surrounding vehicles. Present a baseline to compare human performance against automated systems. Briefly compared the available datasets.	**PREVENTION** **Manoeuvre Detection:** Accuracy: 82.7%. Anticipation:2.28 s. **Manoeuvre Prediction:** Accuracy: 83.4%. Prediction: 0.72 s.
Biparva et al. (2021)	4 action recognition models were evaluated: Two-stream CNN, Two-stream Inflated 3D CNN, STM network, and SlowFast Network. 4 Sizes of RoI. Dense optical flow to extract movement context. OTH:PTH: 2 s/(1–2 s). **Evaluation**: Accuracy (%).	Recognition and prediction of lane change/keep event using stacked visual cues from videos.	**PREVENTION** Accuracy for STM: 91.91% for 2 s; 86.51% for 1 s.
ST-Conv-LSTM Huang et al. (2021)	Spatial–temporal Convolutional LSTM. OTH/PTH: 2.4 s/1 s. **Evaluation**: Accuracy.	Predict lateral (lane change) and longitudinal (holding, sharp acceleration, deceleration, and stopping) intention.	**BDD100K** Accuracy: 57.9%.
IPTM-LSTM Zhang, Song, et al. (2021)	Intention encoder–decoder LSTM. Trajectory encoder–decoder LSTM. IPTM. **Evaluation**: Accuracy and RMSE.	Use intention to predict trajectory of travelling straight, turning left/right and braking.	**NGSIM-LP** Avg. Intention Accuracy: 90.94% RMSE: See Table 3 **INTERACTION** Avg. Intention Accuracy: 86.92%.
LSTM-GAN He et al. (2021)	LSTM + Generative Confrontation Network. **Evaluation**: Accuracy.	Predict vehicle turning intention.	**OWN** Accuracy: 90.9%.
Luan et al. (2022)	Game theory model to predict the intention of the driver. Recognise the vehicle behaviour using past vehicle state. Nash-optimisation function. **Evaluation**: Lateral position error, yaw rate error, probability error.	Predict the trajectory of lane change manoeuvres using driver style (aggressive or conservative) and behaviour recognition.	**NGSIM** Graphs.
AI-TP Zhang, Zhao, et al. (2022)	Approach: Data-driven. **Features**: Past trajectories. **Model(s)**: graph attention mechanism (AI-TP), ConvGRU, **Evaluation**: MSE.	Trajectory prediction.	**NGSIM** See Table 3

caused by the greedy strategy that the decoder LSTM uses to maximise the output probabilities.

Deo and Trivedi (2018b) presented a Manoeuvre-LSTM model that encodes motion and interaction of the surrounding vehicles to assign probabilities for each manoeuvre. The assigned probabilities enable multi-modal trajectory predictions. During that period, the algorithm achieved better RMSE results than the state-of-the-art algorithms, but the RMSE values for long PTH were still high. Although the algorithm considered the interaction between vehicles, it did not consider their inter-dependencies. In order to overcome this limitation, (Deo & Trivedi, 2018a) combined convolutional social pooling and encoder–decoder LSTM to predict manoeuvres and future trajectories. The convolution social pooling can learn the interaction and interdependence of the surrounding vehicles. The downside of the algorithm is that the social tensor of the convolutional social network was fixed to the defined spatial grid around the target vehicle, and it did not consider visual context information. The disadvantage of the last two algorithms is that the predicted trajectories are dependent on the manoeuvre classification performance. For example, Deo and Trivedi (2018a) compared their algorithm with and without considering manoeuvre intention

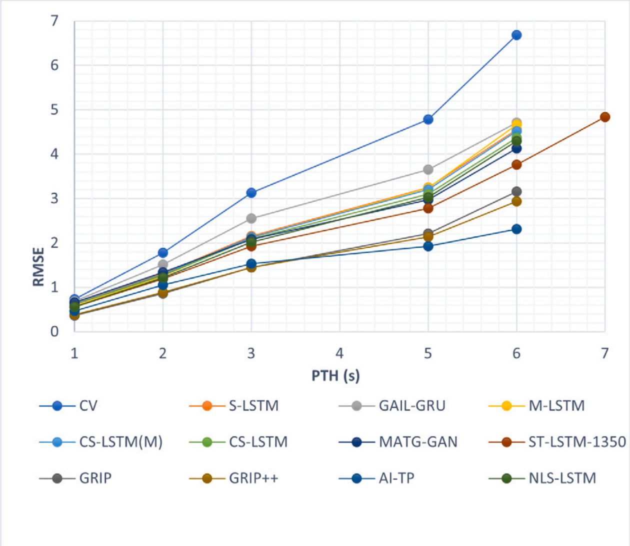

Fig. 4. Vehicle Trajectory Prediction Performance using the NGSIM dataset, with an OTH of 3 s, and PTH ranging from 1–5 s (See Table 3).

and they reported that the algorithm without manoeuvre had better performance.

Dai et al. (2019) claimed that the existing LSTM models suffered from vanishing gradients and were not able to learn spatial interactions between traffic agents. Therefore, they modified the conventional LSTM model by adding shortcut connections and treated spatial interaction between traffic agents as time series. Their model performed better than the M-LSTM (Deo & Trivedi, 2018b) model, which considered manoeuvre prediction information.

Li et al. (2019b) presented the Graph-based Interaction-aware Trajectory Prediction (GRIP) algorithm to predict future trajectories of the TV considering the SV information. GRIP used a GCNN to learn interaction patterns between the TV and SVs. The learnt patterns were then fed to an encoder–decoder LSTM model for predicting future trajectories. GRIP became the state-of-the-art algorithm and was one of the few works to report inference times. The disadvantage of GRIP is that it used a fixed graph structure to learn the interaction between agents, which may not be suitable for complex urban scenarios. In response, Li et al. (2019a) proposed GRIP++, an enhanced version that used both fixed and dynamic graphs to learn the interaction between agents. GRIP++ offered improved computational efficiency compared to the existing algorithms.

Benterki et al. (2020) proposed a hybrid method where they combined ANN and LSTM. ANN was first used to classify the target vehicle's manoeuvre (LLC, RLC, and NLC) using the following manually selected features: yaw, yaw rate, lateral velocity, and lateral acceleration. Subsequently, the LSTM used the vehicle's position and the predicted manoeuvre to predict future trajectories. While the authors tested their algorithm in a real vehicle scenario, only three tests were performed: two for right lane changes and one for left lane-change manoeuvre.

Luan et al. (2022) used vehicle behaviour and driver style to predict future trajectories. History trajectories of the surrounding vehicles were used to determine the type of driver, whether aggressive or conservative. Then, the predicted type of driver was used by a game theory model to predict the intention of the driver. Vehicle behaviour was recognised by using past vehicle state. A comprehensive trajectory was then predicted by feeding the predicted driver intention and the recognised vehicle behaviour into two Nash-optimisation functions. The authors claimed that with the inclusion of the type of driver information, the prediction of the vehicle trajectory was improved, however, their results could not be directly compared to state-of-the-art algorithms such as (Deo & Trivedi, 2018a; Li et al., 2019a, 2019b).

The previously cited works did not take into account the visual context of the scene, which is an important feature as it considers the constraints of the environment. Authors (Lee, Choi, et al., 2017) presented a Deep Stochastic Inverse Optimal Control RNN encoder–decoder (DESIRE) network that considers scene context. The DESIRE uses an RNN encoder to encode past trajectories, a Conditional Variational Auto-Encoder (CVAE) to enable multi-modal predictions, an RNN decoder to decode future trajectories, and a CNN to extract scene context information. In order to refine the predicted results, DESIRE applies Inverse Optimal Control (IOC) to the predicted trajectories and the extracted context information. The authors concluded that the model achieving the best results was the one that considered both scene context and vehicle interactions. Although their algorithm performs better than linear methods, it cannot be directly compared to other works in the literature, since they used different metrics and datasets. Zhao et al. (2019) aimed to predict future trajectories using interaction information between agents and the scene context. An LSTM network was used to encode multi-agent past trajectories, and a CNN was used to extract feature vectors from the scene context. The outputs of the LSTM and CNN were fused using a multi-agent tensor fusion (MATF) network, and the output of the MATF was then fed into an FCN to predict the future trajectories. While these last two cited works considered visual context and achieved good performance, they did not outperform algorithms that did not consider visual contexts, such as GRIP and ST-LSTM.

Table 3 and Fig. 4 report the results for most algorithms reviewed in this paper. Note that the graph only contains the works that have used the same dataset, OTH, PTH, and evaluation metrics. The following observations can be made from the table and the graph:

- Not all algorithms can be directly compared since they have used different datasets, metrics, OTH, and/or PTH. Additionally, some of the works combined the predicted lateral and longitudinal trajectories to calculate their metrics, while others calculated the metrics for lateral and longitudinal trajectories, separately.
- When comparing the algorithms that used the same dataset, metrics, OTH, and PTH, it is observed that GRIP has the best performance for PTHs of 1 s, 2 s, and 3 s; while AI-TP has the best performance for PTHs of 4 s and 5 s.
- With the exception of KITTI and LISA-A, all the other datasets are top-view cameras, and the most frequently used dataset is the NGSIM.
- The most used metric is RMSE.
- Most of the works adopted an OTH of 3 s and PTH of up to 5 s.
- It is noticed that the algorithms' performance worsens as the prediction horizon increases.

3.2. Intention recognition and prediction

The difference between intention recognition and prediction is that for intention recognition, the manoeuvre can be recognised without any anticipation, while for intention prediction, the manoeuvre event must be recognised before it happens. Generally, the researcher specifies the desired anticipation time and then the accuracy of the manoeuvre detection is calculated. The intention of a vehicle's manoeuvre can be recognised by using either prototype trajectories or the manoeuvre intention estimation method.

The literature assumes that there is a motion pattern for the different types of vehicle manoeuvres. Consequently, previous trajectory samples can be used to define a set of prototypes of trajectories which are then used to represent the different motion patterns. Vehicle manoeuvres are then predicted by using initially observed trajectories performed by the vehicle, and matching it to the best available motion patterns. However, this approach is computationally expensive because

Table 3
Results for the most relevant **vehicle** trajectory prediction works.

Work	Dataset	Metrics	Axis	Obs. Hor.	1 s	2 s	3 s	4 s	5 s	6 s
CV Deo and Trivedi (2018a)	NGSIM	RMSE	Both	3 s	0.73	1.78	3.13	4.78	6.68	-
S-LSTM Alahi et al. (2016)	NGSIM	RMSE	Both	3 s	0.65	1.31	2.16	3.25	4.55	-
GAIL-GRU Kuefler et al. (2017)	NGSIM	RMSE	Both	3 s	0.69	1.51	2.55	3.65	4.71	-
C-VGMM+VIM Deo et al. (2018)	NGSIM	RMSE	Both	3 s	0.66	1.56	2.75	4.24	5.99	-
M-LSTM Deo and Trivedi (2018b)	NGSIM	RMSE	Both	3 s	0.58	1.26	2.12	3.24	4.66	-
CS-LSTM(M) Deo and Trivedi (2018a)	NGSIM	RMSE	Both	3 s	0.62	1.29	2.13	3.20	4.52	-
CS-LSTM Deo and Trivedi (2018a)	NGSIM	RMSE	Both	3 s	0.61	1.27	2.09	3.10	4.37	-
MATF GAN Zhao et al. (2019)	NGSIM	RMSE	Both	3 s	0.66	1.34	2.08	2.97	4.13	-
ST-LSTM-1350 Dai et al. (2019) avg.	NGSIM	RMSE	Both	3 s	0.56	1.19	1.93	2.78	3.76	4.84
GRIP Li et al. (2019b)	NGSIM	RMSE	Both	3 s	**0.37**	**0.86**	**1.45**	2.21	3.16	-
GRIP++ Li et al. (2019a)	NGSIM	RMSE	Both	3 s	0.38	0.89	1.45	2.14	2.94	-
AI-TP Zhang, Zhao, et al. (2022)	NGSIM	RMSE	Both	3 s	0.47	0.1.05	1.53	**1.93**	**2.31**	-
NLS-LSTM Messaoud et al. (2019)	NGSIM HighD	RMSE	Both	3 s	0.560.20	1.22 0.57	2.02 1.14	3.03 1.90	4.30 2.91	- -
OGM-LSTM Kim et al. (2017)	NGSIM	RMSE	Lateral Longi.		0.56 3.05	1.24 6.70	- -	- -	- -	- -
Dual LSTM Xing et al. (2017)	NGSIM	RMSE	Lateral Longi.	5 s	0.15 0.47	0.26 1.39	0.38 2.57	0.45 4.04	0.49 5.77	- -
Altché and de La Fortelle (2017)	NGSIM	RMSE	Lateral Longi.		0.11 0.71	0.25 1.98	0.33 3.75	0.40 5.96	0.47 9.00	- -
ANN-LSTM Benterki et al. (2020)	NGSIM	RMSE	Lateral Longi.	3 s	0.043 0.122	- -	0.125 0.235	- -	0.235 0.264	- -
IPTM-LSTM Zhang, Song, et al. (2021)	NGSIM-LP	RMSE	Both	3 s	0.77	1.34	2.19	–	–	-
MATF GAN Zhao et al. (2019)	Massachusetts	RMSE	Both	3 s	0.75	1.4	2.0	2.7	–	-
PF+RBF Hermes et al. (2009)	OWN	RMSE	Both	–	0.7	1.4	5.0	–	–	-
CS-LSTM(M) Deo and Trivedi (2018a)	NGSIM	NLL	Both	3 s	0.58	2.14	3.03	3.68	4.22	-
C-VGMM+VIM Deo et al. (2018)	LISA-A	MAE	Both	3 s	0.24	0.69	1.18	1.66	2.18	-
DESIRE Lee, Choi, et al. (2017)	KITTI SDD	DE PE	Both	2 s	0.281.29	0.67 2.35	1.22 3.47	2.06 5.33	–	–

it requires a substantial number of sample trajectories to determine the numerous possible motion patterns.

In contrast, the manoeuvre intention estimation methods use vehicle motion and road context features to classify the different types of manoeuvres, for instance, stopping/non-stopping, turning left/right, etc. Although this method is less complex than calculating the numerous trajectory probabilities, a large training dataset is required to make the system robust to the different road scenarios. Another limitation is that the manoeuvre classes may not be sufficient to cover the real vehicle intention complexity. For instance, the system may predict a braking manoeuvre, but the braking can be normal or harsh. A proposed solution is to sub-categorise the manoeuvre, for example, normal/harsh stopping, normal/sharp right/left turn, however, this adds complexity to the dataset labelling (Mozaffari et al., 2020).

Intention prediction algorithms can also use predicted trajectories and interaction between vehicles to achieve better accuracy. Traditional methods used to predict the intention of vehicles are Heuristics, Bayesian Networks, HMM, and SVM. DL methods commonly used are RNN, LSTM, and action recognition models.

The following paragraphs will discuss the most relevant DL algorithms used to predict vehicle intention manoeuvre.

Khosroshahi et al. (2016) implemented a multi-layer LSTM network to classify manoeuvre intentions at complex intersections. They extracted samples representing manoeuvres intentions from the KITTI dataset to train and test the algorithm. The input features included linear and angular changes, as well as a histogram of angular changes of the vehicle trajectories. The authors performed experiments with different numbers of manoeuvre classes: 2 (straight or turning), 3 (straight, turning left/right), 8 and 12 classes. The algorithm performed well with 2 and 3 classes, but the accuracy significantly decreases with 8 and 12 classes.

Lee, Kwon, et al. (2017) transformed real-world images into a simplified version of Bird's Eye View (BEV) and fed them into a CNN to predict lane change behaviour. Zhang and Fu (2020) used an offline Bidirectional LSTM to learn driving behaviour and an online Auto-Regressive Integrated Moving Average (ARIMA) to learn past trajectories and predict future ones. The outputs of the offline Bi-LSTM and ARIMA were then fed into another Bi-LSTM to recognise turning behaviour as left-turn, right-turn, or going straight. The algorithm went through evaluation using the NGSIM Lankershim and Peachtree Street dataset, and was able to meet real-time requirements while achieving good accuracy recognition for the PTH of 1 s and 2 s. However, accuracy dropped when considering PTH of 3 s, and it only considered turning left/right and going straight manoeuvres. Whereas vehicles at intersections can perform more complex manoeuvres as reported by Khosroshahi et al. (2016). In addition, the dataset used was acquired from top-view sensors while AVs are equipped with on-board camera sensors. Benterki et al. (2019) compared two conventional methods to predict lane-change manoeuvre, ANN and SVM. They concluded that ANN and SVM have almost the same performance; however, ANN showed the best results.

Izquierdo et al. (2021) used CNN, action recognition, and prediction methods to recognise and predict lane-keeping/changing manoeuvres. Instead of using a sequence of images, they encoded context, interaction, and dynamic state information in a unique enriched image. The enriched image was created by extracting the red channel from a greyscale version of the original image, using a target selection method (TSM), and a temporal integration method (TIM). The authors also investigated the human performance in recognising and predicting lane changes. Their findings indicated that humans can detect 83.9% of the lane change events with an average anticipation of 1.66 s before the manoeuvre is completed. Only 3 out of 72 users were able to predict the lane change events before they started, with an average prediction horizon of 1.08 s. On the other hand, their best algorithm, which considers the trade-off between accuracy and anticipation, achieved 86.4% accuracy with an average anticipation of 2.09 s when considering TTE equal to 0. When TTE was set to 1 s, their algorithm achieved an anticipation of 2.69 s, a prediction of 0.72 s, and an average accuracy of 83.4%.

Fernández-Llorca et al. (2020) and Biparva et al. (2021) recognised and predicted lane-keeping/changing manoeuvres using video action recognition approaches. Biparva et al. (2021) used four types of video action recognition approaches: Two-stream CNN, Two-stream Inflated 3D CNN, spatio-temporal Multiplier Networks, and SlowFast Networks. All of the aforementioned networks used spatial and temporal information from a single image, a sequence of images, or a sequence of optical flow images for recognition and prediction tasks. Moreover, four sizes of RoI were used, denoted as x1, x2, x3 and x4, to consider the interaction between agents, and to extract contextual information around the target vehicle. The network with the best recognition performance was the SlowFast CNN achieving an accuracy of 90.98% with an OTH of 2 s before the TTE. Meanwhile, the network with the best prediction performance was the spatiotemporal multiplier, achieving an accuracy of 91.94% with an OTH of 2 s. The limitations of the previously cited works are as follows: the distribution of the manoeuvre classes was imbalanced, with more lane-keeping samples than lane-changing ones; the time required to recognise and predict a single instance was not provided; and some of the algorithms, such as the SlowFast network, was not able to complete its training due to the GPU memory limitation.

Furthermore, it was observed from the previous vehicle intention prediction works that the authors have selected a fixed PTH to predict the vehicle's intentions. The drawback of using a fixed PTH is that manoeuvre samples may vary in length. For instance, the lane-change manoeuvre performed by an aggressive driver will be shorter than a lane-changing manoeuvre performed by a normal driver.

4. Pedestrian behaviour prediction

At present, AV systems can effectively detect and track pedestrians, however, this alone is not enough to prevent potential collisions. In order to avoid a collision, AV systems must predict pedestrian behaviours. This section aims to provide a literature review of the challenges and techniques used over the years for addressing the pedestrian behaviour prediction.

Table 4
Features and information used to predict pedestrian intention.

Feature	Information
Bbox coordinates	Position, speed, height and width.
Bbox cropped image	Pedestrian appearance, local and surrounding context.
Full image	Global context and some interaction between different traffic objects.
Body Pose	Displacement, action, skeleton, and landmarks.
Ego vehicle position/speed	Interaction between pedestrian and ego vehicle. Pedestrian behaviour is affected by ego vehicle speed.

Pedestrian behaviour prediction has been applied in three main types of datasets: datasets that are recorded using drones, for example, ETH and UCY; datasets recorded from static cameras; and datasets recorded from a car dash cameras, for example, Daimler, JAAD, PIE or KITTI. Datasets from car cameras are more appropriate to train models for AV because they provide more realistic representation. However, when the car is in motion, it may affect the position of the pedestrian bounding box, and pedestrians can be easily occluded. Car cameras datasets can be categorised as either naturalistic or non-naturalistic, as discussed by Fang and López (2018). In non-naturalistic datasets, the pedestrian behaviours and intentions are performed by actors, whereas, in naturalistic datasets, behaviours and intentions are recorded from actual road traffic scenarios. Some of the features that have been used for predicting pedestrian behaviour are listed in Table 4. Pedestrian behaviour prediction has been heavily investigated in the past years and it has many challenges. For instance, pedestrians are highly dynamic, they can move in many directions and change them very quickly, and can be easily occluded by other objects. They can also become distracted by their own objects or external environments, their movements may be influenced by other traffic agents, and they can be difficult to detect in poor visibility conditions. As reported in Tables 5 and 7, researchers have proposed various methods and features to address these challenges over the years. From these tables, the following observations can be made:

- Until 2018, most of the works used traditional methods and their OWN dataset. Thereafter, most authors adopted DL techniques and used the ETH and UCY datasets for trajectory prediction, as well as JAAD and PIE datasets for intention prediction.
- Pedestrian behaviour prediction algorithms have evolved, from solely using motion information to using pedestrian appearance, body pose landmarks, local/global context, interactions between agents, and ego vehicle dynamics.
- Prior to 2018, the focus was predominantly on trajectory prediction, thereafter substantial research efforts have been dedicated to predict pedestrian intentions.
- Most of the intention prediction works were to predict the crossing intention.
- The most used evaluation metrics for intention prediction were accuracy, F1-score, precision, recall, Area Under the Curve (AUC), and Receiver Operating Characteristic Curve (ROC-AUC).
- The most used evaluation metrics for trajectory prediction were Average Displacement Error (ADE), Average Final Displacement Error (FDE), and MSE. Other metrics are Average Non-linear Displacement Error (ANDE), Mean Average Displacement (MAD), and Final Average Displacement (FAD).

The following subsections discuss some of the algorithms reported in Tables 5 and 7. The first subsection provides an in-depth exploration of trajectory prediction algorithms, while the subsequent subsection explore intention prediction algorithms.

Table 5
Relevant works for **pedestrian** trajectory prediction.

Work	Methods	Dataset/Results
Schneider and Gavrila (2013)	**Approach**: Dynamic. **Features**: constant velocity, acceleration, turn. **Models**: Recursive Bayesian filters – Compared EKF and IMM filters. **PTH**: $<$ 2 s. **Evaluation**: MLPE.	**Daimler** IMM has not shown significant performance over simpler models.
Keller and Gavrila (2013)	**Approach**: Dynamic. **Features**: optical flow. Compared the performance between GDPMs, PHTM, KF and IMMKF. Provided human performance on classifying pedestrian behaviour prediction. **Evaluation**: Mean Combined Longitudinal and Lateral RMSE.	**OWN** (on-board) GDPM and PHTM showed better accuracy, however, they are more computationally expensive. 10–50 cm Time Horizon 0.77 s.
Kooij et al. (2014)	**Approach**: Dynamic + Context. **Features**: Head orientation, distance between vehicle and pedestrian, distance between pedestrian and curb. **Models**: Dynamic Bayesian Filters (SLDS). **Evaluation**: Predictive log likelihood.	**OWN** (on-board) Outperforms state-of-art algorithm PHTM. Best result of −0.33 was achieved in the critical, vehicle-seen and stopping scenario using the full context information.
Social-LSTM Alahi et al. (2016)	**Approach**: Data driven. **Features**: Past trajectories. **Models**: Social pooling layer, and LSTM. **OTH/PTH**: 8 (3.2 s)/12 (4.8 s) frames. **Evaluation**: ADE, FDE, and AND.	**ETH and UCY** ADE/FDE/AND: 0.27/0.61/0.15.
Karasev et al. (2016)	**Approach**: Dynamic + Context. **Features**: pedestrian state (position, orientation, and speed), predicted goals, environment context (building, sidewalk, crosswalk, road and grass), dynamic environments such as traffic lights, and assumed rational behaviour for the agent. **Models**: Jump-Markov Process, and Rao-Blackwellized filter. **Evaluation**: L2 error, and Average prediction error.	**OWN** (on-board) for training and KITTI for evaluation. Displayed in a graph.
Rehder et al. (2018)	**Approach**: Data driven + Goal-directed. **Features**: visual cues, predicted pedestrian destinations, and trajectories. **Models**: RMDN, LSTM, topology network, and Markov Decision Process. **Evaluation**: Predicted probability distribution, Average accuracy of predicted destination, and prediction accuracy over time.	**OWN** (on-board) Outperformed IMM. Results were not clear, but from graph Prediction accuracy $10^{(-1)}$ for 1.5 s. Destination plays an important role when trying to predict pedestrian intention.
SR-LSTM Zhang et al. (2018)	**Approach**: Data driven and social behaviour. **Features**: trajectories and current state of the neighbours. **Model(s)**: SR-LSTM and attention mechanism. **Evaluation**: MAD, and FAD.	**ETH** and **UCY** MAD: 0.45; FAD: 0.94.
Social-GAN Gupta et al. (2018)	**Approach**: Data driven. **Features**: Past trajectories. **Model(s)**: GAN, Pooling Module, and LSTM. **PTH**: 8 and 12 metres. **Evaluation**: ADE and FDE.	**ETH, UCY** ADE: 0.39/0.58. FDE: 0.78/1.18.
Social attention Vemula et al. (2018)	**Approach**: Data driven. **Features**: Past trajectories. **Model(s)**: ST-Graph, LSTM, and Attention. **OTH/PTH**: 8 (3.2 s)/12 (4.8 s) time steps. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE: 0.30 m. FDE: 2.59 m.
SS-LSTM Xue et al. (2018)	**Approach**: Data driven. **Features**: Past trajectories, neighbour feature (occupancy maps: grip, circle and log), and individual information. **Model(s)**: CNN, and Hierarchical-LSTM. **OTH/PTH**: 8/12 frames. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE: 0.070 pixels. FDE: 0.133 pixels.
CIDNN Xu et al. (2018)	**Approach**: Data driven. **Features**: Past trajectories, and interactions. **Model(s)**: stacked-LSTM, and MLP. **OTH/PTH**: 5/5 frames. **Hardware**: Intel Xeon CPU E52643 4.40 and TITAN GPU. **Evaluation**: ADE.	**GC/ETH/UCY/CUHK/Subway** ADE: 0.012/0.09/0.12/0.008/0.016. Inference: 0.43 ms

(*continued on next page*)

Table 5 (*continued*).

Work	Methods	Dataset/Results
LSTM-Bayesian Bhattacharyya et al. (2018)	**Approach**: Data driven. **Features**: Bbox coordinates past trajectories and ego vehicle odometry. **Model(s)**: Two stream architecture, Bayesian RNN (LSTM), and CNN. **OTH/PTH**: 0.5/1 s. **Evaluation**: MSE in pixels and NLL.	**CityScapes(on-board)** MSE/NLL: 505/3.92.
DBN-SLDS Flohr et al. (2018)	**Approach**: Data driven. **Features**: context cues (VRU actions, and its static and dynamic environment). **Model(s)**: DBN and SLDS. **TTE** = [−15, 0] **PTH**:1 s. **Evaluation**: Prediction error.	**OWN (on-board, non-naturalistic** Graphs.
MX-LSTM Hasan et al. (2018)	**Approach**: Data driven. **Features**: Past trajectories, and head pose estimation. **Model(s)**: tracklets, vislets, VFO social pooling, and LSTM. **OTH/PTH**: 8/12 frames. **Evaluation**: MAD and FAD in metres.	**UCY** MAD/FAD: 0.49/1.12 m. **Towncentre** MAD/FAD: 1.15/2.30 m.
Scene-LSTM Manh and Alaghband (2018)	**Approach**: Data driven. **Features**: Past trajectories and scene divided into grid cells. **Model(s)**: Scene Data Filter, and Coupled-LSTM. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE, FDE and NDE.	**UCY and ETH** ADE/FDE/NDE: 0.7/0.7/0.9.
SoPhie Sadeghian et al. (2019)	**Approach**: Data driven. **Features**: Past trajectories, social interactions, and images of the scene. **Model(s)**: CNN, LSTM, GAN, Social and physical attention mechanism. **PTH**: 12 future timesteps. **Evaluation**: ADE and FDE.	**ETH, UCY** ADE: 0.54 m. FDE: 1.15 m. **SDD** ADE: 16.24 pixels. FDE: 29.38 pixels.
StarNet-DNN Zhu et al. (2019)	**Approach**: Data driven. **Features**: Past trajectories. **Model(s)**: StarNet DNN (Host and hub networks), and LSTM. **PTH**: 8 frames. **Hardware**: Tesla V100 GPU. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.30/0.57. Inference: 0.073 s.
PECNet Mangalam et al. (2020)	**Approach**: Data driven and goal directed. **Features**: Past trajectories and estimated end point destination. **Model(s)**: CVAE, attention mechanism, and social pooling. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.29/0.48 m. **SDD** ADE/FDE: 9.96/15.88 p.
ST-GCNN Mohamed et al. (2020)	**Approach**: Data driven. **Features**: Past trajectories and sequence of images. **Model(s)**: GCN, and TXP-CNN. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.44/0.75 m.
RSBG Sun et al. (2020)	**Approach**: Data driven. **Features**: Past trajectories and local context. **Model(s)**: GCN, CNN, and LSTM. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.48/0.99 m.
LVTA Xue et al. (2020)	**Approach**: Data driven. **Features**: Past trajectories and velocities. **Model(s)**: attention mechanism, and LSTM. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.46/0.92 m.
Holistic-LSTM Quan et al. (2021)	**Approach**: Data driven. **Features**: bbox past trajectories, crossing intention, pedestrian scale, depth estimation, and global scene dynamics (depth and optical flow). **Model(s)**: ConvLSTM, modified LSTM with more inputs, and attention mechanism. **OTH/PTH**: 0.5/1 s. **Evaluation**: MSE, CMSE, and CFMSE of the bbox coordinates.	**JAAD** MSE: 389. **PIE** MSE: 167. **S-KITTI** MSE: 525/1.5 s.

(*continued on next page*)

Table 5 (*continued*).

Work	Methods	Dataset/Results
Bi-TraP Yao et al. (2021a)	**Approach**: Data driven and Multi-modal goal estimation. **Features**: bbox past trajectories. **Model(s)**: CVAE, Gaussian distribution, GMM, and Bi-directional GRU. **OTH/PTH (JAAD/PIE)**: 0.5/1.5 s. **OTH/PTH (ETH/UCY)**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**JAAD** ADE: 1206. **PIE** ADE: 511. **ETH-UCY** ADE/FDE: 0.18/0.35.
BA-PTP Czech et al. (2022)	**Approach**: Data driven. **Features**: vehicle odometry, bbox, body, head orientation, and pose. **Model(s)**: attention mechanism and Bi-GRU, **OTH/PTH (PIE)**: 0.5/1.5 s. **OTH/PTH (ECP)**: 0.6/1.6 s. **Evaluation**: MSE, CMSE, and CFMSE.	**PIE** MSE/CMSE/CFMSE: 420/383/1513. **ECP-Intention** MSE/CMSE/CFMSE: 768/680/1966
SGNet Wang et al. (2022)	**Approach**: Data driven, and goal directed. **Features**: Past trajectories. **Model(s)**: Stepwise goal estimator, attention mechanism, GRU, and CVAE. **OTH/PTH (JAAD, PIE, HEV-I)**: 1.6/0.5,1.0,1.5 s. **OTH/PTH (ETH & UCY)**: 3.2/4.8 s. **OTH/PTH (NuScenes)**: 2/6 s. **Evaluation**: MSE, CMSE, CFMSE, ADE and FDE.	**JAAD** MSE/CMSE/CFMSE: 1049/996/4076 p (1.5 s). **PIE** MSE/CMSE/CFMSE: 442/413/1761 p (1.5 s). **ETH and UCY** ADE/FDE: 0.35/0.83 Euclidean space. **NuScenes** ADE/FDE: 1.32/2.50.
PTPGC Yang, Sun, et al. (2022)	**Approach**: Data driven. **Features**: Past trajectories, length of attributes, and number of pedestrians. **Model(s)**: Graph attention, convLSTM, and Temporal CNN. **OTH/PTH**: 3.2/4.8 s. **Evaluation**: ADE and FDE.	**ETH and UCY** ADE/FDE: 0.67/1.29.

4.1. *Trajectory prediction*

Both traditional and DL techniques have been used in order to predict pedestrian trajectories. Traditional techniques relies on hand-crafted functions, such as EKF, IMM, and social forces, to predict pedestrians' future trajectories. However, these functions have limitations in handling complex scenarios. To address this, several researchers adopted DL techniques such as: CNN, Generative Adversarial Network (GAN), GCNN, LSTM, GRU, CVAE, attention mechanism, and/or Multi-Layer Perceptron (MLP).

Although LSTM networks have many advantages, it struggles to learn dependencies between multiple correlated sequences. For this reason, Alahi et al. (2016) proposed a Social LSTM network to predict pedestrian trajectories. Social pooling layers were introduced to enable LSTM networks to share their hidden state. This enables the algorithm to learn interactions among pedestrians. Social-LSTM only considers motion features to model human interactions, however, Xu et al. (2018) argues that spatial position should also be considered. For this reason, they presented a model where MLP layers were used to encode location, and LSTM was used to encode motion for each neighbour. Both sets of encoded information were then used as input to a crowd interaction module to predict pedestrian displacement. In a different approach, Xue et al. (2020) used two LSTM layers to encode the pedestrian's location and velocity, along with a temporal attention mechanism to extract the most relevant features from the velocity and location inputs.

Humans are highly dynamic, which makes the task of predicting their trajectories more challenging. In response to this, Rehder et al. (2018) implemented a DNN that would first predict the future destinations of the pedestrians, and then predict their future trajectories. They have used CNN, LSTM and Mixture Density Network to predict potential destinations, and another CNN to plan and predict future trajectories based on these potential destinations. CVAE was used by Mangalam et al. (2020) to predict future endpoints, these then were subsequently used to predict multi-modal longer-term trajectories. They also presented a novel self-attention-based social pooling layers that extract relevant features from the neighbours using non-local attention. Yao et al. (2021a) also proposed a goal-direct method, where they combine CVAE and bi-directional GRU to encode past trajectories and decode multi-modal future trajectories. Goal-directed models have the disadvantage that only one goal is estimated over a long-term prediction. For this reason, if a pedestrian changes direction the estimated goal may be incorrect, and consequently affecting the estimated predicted trajectories. Wang et al. (2022) proposed a method where they model and estimate goals continuously by using RNNs.

While many studies relied on historical trajectories for predicting future ones, they often overlooked the current state of the pedestrian. In order to overcome this issue, Zhang et al. (2019) introduced a state refinement LSTM that considered both the current and previous state of the target pedestrian and the surrounding pedestrians. This state refinement module enables the network to incorporate interactions through a message-passing mechanism. It also uses a motion gate as an attention mechanism to focus on the most relevant features of the neighbours.

Previous research, when considering human-to-human interactions, would often take into account only nearby neighbours, even though more distant neighbours might also influence the behaviour of the target pedestrian. A GAN was presented by Gupta et al. (2018) that not only considers local neighbours but all neighbours in the scene. The GAN network comprises an LSTM generator to generate multi-potential trajectories, a pooling module to learn human-to-human interactions, and an LSTM discriminator to select acceptable trajectories from the generated ones. Similarly, Vemula et al. (2018) considered all the pedestrians in the scene using a spatio-temporal graph and LSTM. Additionally, they adopted an attention mechanism to learn the relevance of each agent, regardless of how far they are from each other. A star-like network was introduced by Zhu et al. (2019) to account for all agents in the scene. The network has a centralised hub network, which gathers motion information from all pedestrians in the scene, and a host network for each pedestrian. The host networks query the hub network for social information to predict trajectories. Graph attention and convolutional LSTM were also proposed by Yang, Sun, et al. (2022) to consider the surrounding neighbours.

Xue et al. (2018) emphasised the importance of considering scene layout when predicting pedestrian trajectories. As a result, they used three different LSTMs to learn information about individuals, social interactions, and scene layout. One LSTM used the trajectory of the target pedestrian as its input, another used an occupancy map as its input, and the final one used feature vectors extracted from the original image by a CNN as its input. Likewise, Manh and Alaghband (2018) took scene layout into account, where they used a two-level grid structure of the original image and trajectory information as inputs to a two-stream LSTM for predicting future trajectories. CNN, LSTM, attention mechanism, and GAN were used by Sadeghian et al. (2019) to predict trajectories using both past trajectories and scene context as inputs. The CNN extracted scene-related features, the LSTM extracted motion-related features, the attention mechanism extracted both the physical and position relevant features, and the GAN generated multiple trajectories and then selected the most suitable ones.

Mohamed et al. (2020) classified methods such as social pooling or the combination of hidden state features, used to model human interactions, as “aggregation methods”. They claimed that these types of methods have limitations in accurately modelling human interactions because the aggregation occurs within the feature space and does not directly model physical interactions. Furthermore, some of these aggregation methods, such as pooling layers, may overlook to capture important information. Given these considerations, the authors proposed a social spatio-temporal GCN (ST-GCN) to model interactions among pedestrians. The ST-GCN model’s output is subsequently used as input for a time extrapolate CNN to predict future trajectories.

The above works have not considered group-based interactions, which involve two or more individuals exhibiting similar movements, behaviours, or goals. A recursive social behaviour graph and GCN was implemented by Sun et al. (2020) to explore and learn group-based interactions. The authors also used CNN and LSTM to obtain an individual representation of each pedestrian in the scene. The individual representations, along with the learned group-based features, were combined and used by a decoder LSTM to predict future trajectories.

Bhattacharyya et al. (2018) claimed that they were the pioneers in using an on-board dataset to predict pedestrian behaviour. The authors used a two-stream LSTM architecture to encode bounding box coordinates, ego-vehicle odometry information, and feature vectors extracted from the original image by a CNN. Another work that used an on-board dataset is (Czech et al., 2022), in which the authors used a multi-stream RNN to individually encode bounding box coordinates, head orientation, body orientation, pose skeleton, and past trajectories. The encoded information from each stream is fused through an attention mechanism and subsequently input to an RNN decoder to predict future bounding boxes. The drawback of the latter two algorithms is that they did not consider social interaction among the agents.

Hasan et al. (2018) argues that head orientation and movement are correlated. Consequently, they proposed a two-stream LSTM to encode both trajectory and head orientation information. The two encoded information, were then merged using a View Frustum social pooling layer. The disadvantage of this method is that it is only suitable for top-view and BEV datasets.

Usually, when a system adopts LSTM networks and requires the use of multiple types of inputs, these inputs are first combined before being fed to LSTM cells. This practice is required because LSTM cells are designed to accept only a single input sequence, which can constrain their ability to capture relevant information from various input sources. Quan et al. (2021) adapted the conventional LSTM cell to accept four additional input sequences: vehicle speed, pedestrian intention, correlation among frames, and bounding box location. The vehicle speed was estimated by using optical flow and depth information; the pedestrian intention was estimated using convLSTM; and the correlation among frames was derived from optical flow images.

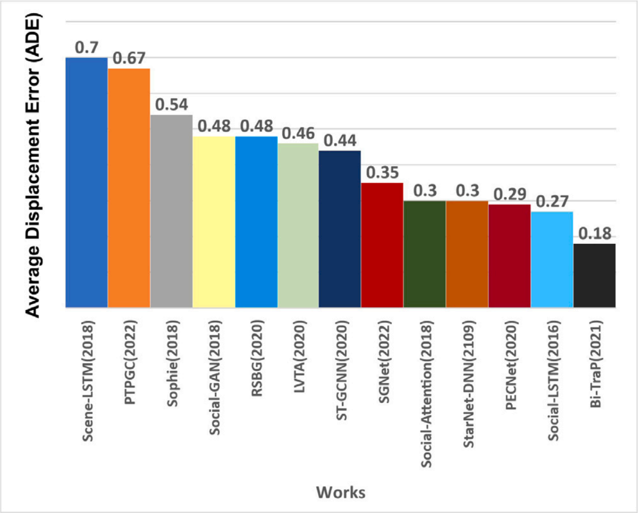

Fig. 5. **Pedestrian** Trajectory Prediction Performance using the ETH and UCY datasets, with an OTH of 3.2 s, a PTH of 4.8 s, and Average Displacement Error (ADE) in metres (See Table 6).

Table 6 and Fig. 5 report the results for the most relevant studies in pedestrian trajectory prediction. It is not possible to directly compare all of them since some of them have used different datasets, metrics, OTH, and PTH. However, when examining the results of the algorithms that used the same dataset, metrics, OTH, and PTH, the Bi-Trap (Yao et al., 2021a) algorithm outperformed others. Bi-Trap achieved ADE and FDE values of 0.18 m and 0.35 m, respectively.

4.2. Intention recognition and prediction

The difference between pedestrian intention recognition and prediction aligns with what was explained on Section 3. Recognition does not require anticipation, while prediction does. The main methods used to predict pedestrian intentions include CNN, GCNN, GRU, LSTM, attention mechanism, multi-tasking, and transformer networks.

CNN: Fang et al. (2017) and Fang and López (2018) used CNNs to extract human skeleton features and used SVM/RF classifier to predict if the pedestrian is crossing the road. Abdulrahim and Salam (2016) also used CNNs, along with depth information to learn 3D human body landmarks, including additional information such as the pedestrian shoulders, neck, and face. While CNNs can extract spatial features, their capability to capture temporal dependencies is limited. To overcome this limitation, Yang et al. (2021) implemented a 3D-CNN to extract spatio-temporal information. Additionally, Piccoli et al. (2020) proposed an alternative model called FuSSI-Net, designed to extract both spatio-temporal information. FuSSI-Net is a spatio-temporal Dense-net that takes a sequence of bounding boxes and skeleton features as inputs to predict crossing intention. Although these last two models can extract spatial and temporal information, they are limited to short-time horizon prediction and become computationally expensive as the input sequence length increases.

LSTM: Rasouli et al. (2019) used LSTM to encode local context, trajectories, and ego vehicle information. Subsequently, the encoded information was decoded to estimate the probability of a pedestrian crossing the road. Bouhsain et al. (2020) used bounding box coordinates and velocities features as inputs for a sequence-to-sequence LSTM, which was used to predict both the pedestrian intentions and the future position of the pedestrians’ bounding boxes. In a different approach, Lian et al. (2022), introduced a stacked-LSTM model, where appearance, context, and dynamic features of the pedestrian were used to predict crossing intentions. LSTM networks have the ability to learn and memorise features over the long term, as they capture

Table 6
Results for the most relevant **pedestrian** trajectory prediction works.

Work	Dataset	OTH	PTH	ADE	FDE	AND	MAD	FAD	MSE
Social-LSTM Alahi et al. (2016)	ETH & UCY	3.2 s	4.8 s.	0.27 m	0.61 m	0.15 m	–	–	–
Scene-LSTM Manh and Alaghband (2018)	ETH & UCY	3.2 s	4.8 s	0.7 m	0.7 m	0.9 m	–	–	–
Social-GAN Gupta et al. (2018)	ETH & UCY	3.2 s	4.8 s	0.48 m	0.98 m	–	–	–	-
Social-attention Vemula et al. (2018)	ETH & UCY	3.2 s	4.8 s	0.30 m	2.59 m	–	–	–	-
Sophie Sadeghian et al. (2019)	ETH & UCY SDD	3.2 s	4.8 s	0.54 m 16.24 pi	1.15 m 29.38 pi	–	– –	– –	– –
StarNet-DNN Zhu et al. (2019)	ETH & UCY	3.2 s	4.8 s	0.30 m	0.57 m	–	–	–	-
PECNet Mangalam et al. (2020)	ETH & UCY SDD	3.2 s	4.8 s	0.29 m 9.96 pi	0.48 m 15.88 pi	–	– –	– –	– –
ST-GCNN Mohamed et al. (2020)	ETH & UCY	3.2 s	4.8 s	0.44 m	0.75 m	–	–	–	–
RSBG Sun et al. (2020)	ETH & UCY	3.2 s	4.8 s	0.48 m	0.99 m	–	–	–	–
LVTA Xue et al. (2020)	ETH & UCY	3.2 s	4.8 s	0.46 m	0.92 m	–	–	–	–
Bi-TraP Yao et al. (2021a)	ETH & UCY JAAD PIE	3.2 s 0.5 s 0.5 s	4.8 s 1.5 s 1.5 s	**0.18 m** 1206 511	**0.35 m** – –	–	– –	– –	– – –
SGNet Wang et al. (2022)	ETH & UCY JAAD PIE NuScenes	3.2 s 1.6 s 1.6 s 2 s	4.8 s 1.5 s 1.5 s 6 s	0.35 m – – 1.32	0.83 — — 2.5	–	– –	– –	- 1049 442 -
SGNet Wang et al. (2022)	ETH & UCY	3.2 s	4.8 s	0.35 m	0.83 m	–	–	–	–
PTPGC Yang, Sun, et al. (2022)	ETH & UCY	3.2 s	4.8 s	0.67 m	1.29 m	–	–	–	–
SS-LSTM Xue et al. (2018)	ETH & UCY	3.2 s	4.8 s	0.070 npu	0.133 npu	–	–	–	-
SR-LSTM Zhang et al. (2018)	ETH & UCY	3.2 s	4.8 s	–	–	–	0.45	0.94	–
CIDNN Xu et al. (2018)	ETH & UCY	4 s	4 s	0.11	–	–	–	–	–
MX-LSTM Hasan et al. (2018)	UCY Towncentre	3.2 s	4.8 s	–	– –	– –	0.49 m 1.15 m	1.12 m 2.30 m	– –
Holistic-LSTM Quan et al. (2021)	JAAD PIE S-KITTI	0.5 s	1 s 1 s 1.5 s			–	– –	– –	389 167 525
BA-PTP Czech et al. (2022)	PIE ECP	0.5 s 0.6 s	1.5 s 1.6 s			–	– –	– –	420 768

long-distance dependencies (Chung et al., 2014). Nevertheless, they have limitations in extracting spatial features, managing dependencies among the extracted features, exhibiting longer training times, and assigning uniform attention to all inputs, even though some inputs can be more relevant than others (Sharma et al., 2022). Ahmed et al. (2023) used a 2D pose estimator in conjunction with LSTM to predict crossing behaviour of the pedestrian.

GRU: GRUs serve as an alternative to LSTMs, as they also learns temporal information. Kotseruba et al. (2020) used pedestrian appearance features, which were extracted using a VGG network, and ego vehicle velocity information as inputs for a GRU network to predict pedestrian intentions. Rasouli et al. (2020) used pedestrian appearance, global context, body pose, bounding boxes, and ego-vehicle speed features as inputs to a stacked GRU network to predict pedestrian crossing behaviour. These features were gradually integrated into the GRU network, starting with pedestrian appearance, followed by global context, body pose, bounding boxes, and concluding with the ego vehicle speed. GRUs offer the advantage of requiring less memory and being faster than LSTMs. However, they tend to be less accurate when handling with long input sequences (Chung et al., 2014).

GCN: A spatio-temporal GCN was presented by Zhang, Angeloudis, and Demiris (2022), where they used a sequence of skeleton features to predict crossing intentions. The skeleton joints were connected by nodes and edges to learn both spatial and temporal features. Cadena et al. (2022) used two GCNs, which took human body key points, local context, and ego speed information as inputs to predict crossing intentions. GCNs has the advantage of extracting interactions among the target pedestrian and its neighbours, considering both spatial and temporal dependencies (Sharma et al., 2022). In addition, GCNs can handle non-Euclidean data formats, such as scenarios where pedestrians are dispersed across a scene, which cannot be represented using a grid-like structure. However, they can only handle short-term sequences and do not perform well when applied to regression tasks.

Attention Mechanism: Lian et al. (2022) also used a self-attention mechanism to extract the most relevant information from the pedestrian appearance, the pedestrian's surroundings, and dynamic features. Rasouli et al. (2019) combined different attention mechanism layers at different locations of the network to investigate their impact on the model performance. Attention mechanism approaches enable networks like LSTM to focus more on the most relevant features, and less on redundant ones.

Table 7
Relevant works for **pedestrian** intention prediction.

Work	Methods	Problem	Dataset/Results
Schneider and Gavrila (2013)	Approach: Dynamic. **Features:** Recursive Bayesian filters – Compared EKF and IMM filters (constant velocity/acceleration/turn). PTH: < 2 s. **Evaluation**: MLPE.	Trajectory and intention prediction.	**Daimler** IMM has not shown significance performance over simpler models.
Keller and Gavrila (2013)	Approach: Dynamic. **Features**: Compared the performance between GDPMs using optical flow information, PHTM, KF and IMMKF. Provided human performance on classifying pedestrian behaviour prediction. **Evaluation**: Mean Combined Longitudinal and Lateral RMSE.	Trajectory and intention prediction.	**OWN** (on-board) GDPM and PHTM showed better accuracy, however, they are more computationally expensive. 10–50 cm Time Horizon 0.77 s.
Bonnin et al. (2014)	Approach: Dynamic + Context. **Features**: distance and time to curb, distance and time to ego lane, distance and time to zebra crossing, distance and time to collision point, difference of time to collision point, face, global and relative orientation. Single Neural Network as classifier to learn the different features. Inner-city and zebra model. PTH: 1 s. **Evaluation**: TPR and FPR.	Intention prediction (crossing).	OWN (on-board) Inner-city dataset, zebra dataset and combination of both ICZ. Inner-city model: 31% TPR, 0.0 FPR, PTH 0.72 s for the zebra dataset. TPR 29%, PTH 0.67 s for the inner-city dataset. TPR 31%, PTH 0.72 s for the ICZ dataset. Zebra crossing model: 100% TPR, 3.23 s PTH for the zebra dataset. 86% TPR, 28% FPR and 1.73 s PTH for the inner-city dataset. CMT model: 62% TPR, 2.59 s PTH for the ICZ dataset.
Neogi et al. (2017)	Approach: Dynamic + Context. FLDCRF. **Features**: pedestrian position (distance to curb, and left or right side of the road), pedestrian–vehicle interaction, optical flow. **Evaluation**: average probability, time to stop and time to cross.	Intention prediction.	**NTUC** (OWN, on-board and actors) Average probability > 0.7 predicting 1.2 s before the action.
Minguez et al. (2018)	Approach: Dynamic. Balanced-GDPMs to reduce 3-D time relevant information into low dimensional information and to assume future latent positions. **Features**: Skeleton motion analysis. Four models to predict start, stop, walk and stand actions. HMM is used to select which model to use to predict future pedestrian path and poses. **Evaluation**: MED against TTE.	Predict pedestrian actions.	**CMU-UAH** Achieved MED of 41.24 mm for TTE of 1 s, for starting activity; and MED of 238.01 mm for TTE of 1 s for stopping activity.
Fang et al. (2017)	Approach: Data Driven. **Features**: Skeleton. CNN for pose estimation. Deep association for tracking. **Evaluation**: Intention probability vs TTE.	Intention prediction (crossing/not crossing).	**Daimler** 0.8 predictability with TTE=12 (750 ms).
CV Fang and López (2018)	Approach: Data Driven. **Features**: Skeleton. CNN for pose estimation. Deep association for tracking. **Evaluation**: Accuracy.	Intention prediction (crossing/not crossing).	See Table 8
PIE (int) Rasouli et al. (2019)	Approach: Data driven. **Features**: bbox coord, image context, and image bbox. RNN (LSTM). **Evaluation**: Accuracy, and F1-score.	Intention prediction (crossing).	See Table 8
Bouhsain et al. (2020)	Approach: Data Driven. **Features**: bboxes coordinates and velocities. PV-LSTM Multi-task sequence to sequence learning **Evaluation**: ADE, FDE, Accuracy.	Pedestrian intention and pedestrian bbox predictions (crossing).	See Table 8.
Liu et al. (2020)	Approach: Context, Temporal, and Data driven. **Features**: Graph Convolution and GRU to learn spatio-temporal relationship. **Evaluation**: Accuracy.	Intention prediction (crossing).	**Stanford-TIR** A: 79.10%. **JAAD** A: 79.28%.

(*continued on next page*)

Table 7 (*continued*).

Work	Methods	Problem	Dataset/Results
Abughalieh and Alawneh (2020)	Approach: Data driven. **Features**: pedestrian body landmarks considering depth information. CNN. **Evaluation**: Accuracy.	Intention prediction (walking and crossing).	**OWN** (on-board) A: 89%.
FUSSI-net Piccoli et al. (2020)	Approach: Data driven, target-agent context. **Features**: Skeleton and bbox. DenseNet. **Evaluation**: Accuracy.	Intention prediction (crossing).	See Table 8
SFR-GRU Rasouli et al. (2020)	Approach: Data driven. **Features**: pose, 2D bbox, appearance, global context, and ego speed. Stacked-RNN (GRU). **Evaluation**: Accuracy, Precision, recall, F1-score, and AUC.	Intention prediction (crossing).	See Table 8
C+B+S+Int Kotseruba et al. (2020)	Approach: Data driven. **Features**: surrounding, appearance, context, bbox, and ego vehicle speed. single GRU. PTH: 2 s. **Evaluation**: Accuracy, AUC, F1, Precision, and recall.	Intention prediction (crossing). Studied human performance.	See Table 8
Razali et al. (2021)	Approach: Data driven and key body landmarks. **Features**: PAF and PIF. Uses only one RGB image. Multitask learning. CNN (ResNet). **Evaluation**: Precision for different prediction horizon.	Recognition and Intention prediction (crossing) in real-time.	**JAAD** Recognition: −0 s: 81.7%; −1 s: 83.6%; −2 s: 83.5%; −3 s: 83%; −4 s: 82.7%. Prediction: −1 s: 42.6%; −2 s: 46.1%; −3 s: 46.3%; −4 s: 46.0%. FPS: 5.
Zhang, Abdel-Aty, et al. (2021)	Approach: Data Driven. **Features**:: pose-key-points. Compared SVM, RF, GBM, and XGBoost models. **Evaluation**: Accuracy.	Intention prediction (crossing at red light).	**CCTV** A: 92%: 1 s; 92%: 2 s; 88.9%: 3 s; 92.5%: 4 s.
PCIR Yang et al. (2021)	Approach: Data driven, context, and behavioural. **Features**: pedestrians, ego vehicle, and environment. 3D-CNN. **Evaluation**: AP.	Intention detection (crossing).	See Table 8
Chen et al. (2021)	Approach: Data driven. **Features**: bbox, body pose, road objects. Graph encoder, CNN, and LSTM. PTH: 1.5 s. **Evaluation**: Balanced Accuracy and F1 score.	Intention prediction (crossing).	See Table 8
I+A+F+R Yao et al. (2021b)	Approach: Data driven, and multi-task. ARN Attentive Relation Network. CNN, MLP, and GRU. PTH: 1–2 s. **Features**: bbox context and coordinates, relation, and visual. **Evaluation**: Accuracy, F1-score, ROC-AUC, precision.	Intention and action prediction (crossing).	See Table 8 Inference: < 6 ms.
PCPA Kotseruba et al. (2021)	Approach: Data driven. **Features**: bbox, pose, local context, and ego vehicle speed. 3D CNN + single-RNN (GRU) + attention mechanism. **Evaluation**: Accuracy, AUC, and F1.	Intention prediction (crossing).	See Table 8
Yang, Zhang, et al. (2022)	Approach: Data driven. **Features**: local and global context, bbox, pose-key-points. Attention mechanism, 2D CNN, and RNN. **Evaluation**: Accuracy, F1, and recall.	Intention prediction (crossing).	See Table 8
Graph+ Cadena et al. (2022)	Approach: Data driven. **Features**: context, ego vehicle velocity, and key body landmarks. Graph Convolutional Network. **Evaluation**: Accuracy.	Intention Prediction (crossing).	See Table 8 Inference: 6 ms.
ST-CrossingPose Zhang, Angeloudis, and Demiris (2022)	Approach: Data driven. **Features**: skeleton-based. Spatio-Temporal GCN. **Evaluation**: Accuracy, AUC, F1-score, Precision, and Recall.	Intention prediction (crossing).	**JAAD** Recognition: 63%. See Table 8
Achaji et al. (2022)	Approach: Data Driven. **Features**: bbox. Transformer Networks. PTH: 1 s and 2 s. Test human ability for pedestrian action prediction. **Evaluation**: Accuracy and F1-Score.	Intention recognition and prediction (crossing).	**PIE** A:91%. F1:0.83. CP2A A:91%. F1:0.91.

(*continued on next page*)

Table 7 (*continued*).

Work	Methods	Problem	Dataset/Results
Scene-STGCN Naik et al. (2022)	Approach: Data Driven. **Features**: Scene Spatio-Temporal GCN. **Evaluation**: Accuracy, F1-score, AP, and ROC-AUC.	Intention recognition (crossing).	See Table 8
Zeng (2022)	Approach: Data driven. **Features**: body land-marks. SqueezeNet and GRU. Hardware: AMD Ryzen 5 3600, G Force RTX 3070. **Evaluation**: Accuracy and ROC-AUC.	Intention prediction (crossing). Light-weight and inference speed.	See Table 8
CA-LSTM Lian et al. (2022)	Approach: Data driven. context and dynamic. **Features**: appearance, velocity, and walking angle. Attention LSTM. **Evaluation**: Accuracy, F1-score, recall metrics.	Intention Prediction (crossing).	See Table 8
Gazzeh and Douik (2022)	Approach: Data driven. **Features**: pedestrian localisation and environment contest (lane lines). ML and DL. **Evaluation**: Accuracy.	Intention recognition in real-time.	See Table 8
Ma and Rong (2022)	Approach: Data driven. **Features**: pedestrian pose (skeleton), pedestrian to vehicle distance, and ego vehicle information. Multi-feature fusion. Random forest classifier. PTH: 0.6 s. **Evaluation**: Accuracy and AUC.	Intention prediction (crossing).	See Table 8
Ahmed et al. (2023)	**Approach**: Data driven. **Features**: Past trajectories, velocity, and 3D joint estimation. **Model(s)**: Position and Velocity LSTM. **PTH**: 0.4 s. **Evaluation**: Accuracy.	Intention prediction (crossing).	**JAAD and PIE** Accuracy: 89%/91%.

Transformers: Even though attention mechanism have the ability to focus on the most relevant features, it was reported by Achaji et al. (2022) that its effectiveness might be reduced when coupled with LSTM networks. For this reason, Achaji et al. (2022) proposed a framework based on three types of transformer networks: encoder-only, encoder-pooling, and encoder–decoder architectures. The proposed framework used only the pedestrian bounding box information as its input. The authors argued that their model outperformed other methods that used multiple input features. Transformer networks offer the advantage of parallel input processing, which accelerates training stage. On the other hand, the ability to process the input data in parallel restricts the model to take advantage of the sequential nature of the input.

Multiple Methods: many studies have used more than one method to predict pedestrian intention. Liu et al. (2020) used GCN to generate a pedestrian-centring graph for each observation frame. These graphs connect the target pedestrian to its surrounding, allowing the algorithm to learn relation between the pedestrian and the scene. In addition, edges were introduced between the pedestrian nodes in each pedestrian-centring graph to allow the algorithm to learn temporal information. The resulting interconnected graphs were then fed into a GRU network to predict crossing intention. Chen et al. (2021) used a combination of methods, including a CNN to extract features from traffic objects and pedestrian appearance, a GCN to auto encode the extracted features, another framework to extract human skeleton, and an LSTM network to predict crossing intentions. CNN, ARN, MLP and GRU were used by Yao et al. (2021b) to predict crossing intentions. The CNN was used to extract global features, ARN was used to extract relational features from detected traffic objects, MLP was used for intention classification, and the LSTM was used for intention prediction. One major difference of this work is that the network also takes the predicted intention output as input. Kotseruba et al. (2021) used 3D-CNN, RNN, and attention mechanism. The 3D-CNN was used to encode local features from a sequence of cropped bounding boxes, the RNN was used to encode the bounding-box coordinates, pose landmarks and the ego-vehicle speed. Finally, an attention mechanism was used to combine the most relevant features. Yang, Zhang, et al. (2022) used 2D-CNN, stacked-RNN, and attention mechanism. Spatio-temporal GCN was used by Naik et al. (2022) to encode the input image, image class and location information tensors. Then the output of the spatio-temporal GCN was fed into an LSTM network to generate long-term predictions. Zeng (2022) used SqueezeNet to extract visual features and used GRU to extract temporal dependencies. They also used a multi-tasking approach to predict both pedestrians' intentions and poses. One primary advantage of using multiple models is that each model can compensate for the limitations of others. For example, CNN, GCN, and attention mechanism can aid the limitations of an LSTM network to extract spatial information, handle non-Euclidean data, and prioritise relevant features, respectively.

Full-Pipeline: Gazzeh and Douik (2022) presented a full pipeline model which includes detection, tracking, and crossing intention prediction. They used YOLOv4 for object detection, DeepSort for tracking, Canny Edge for lane line detection, and linear SVM for intention prediction. Another full pipeline system was implemented by Piccoli et al. (2020), where they used YOLOv3 for detection, DeepSort for tracking, and spatio-temporal Densenet for intention prediction. YOLOv5, DeepSort, and an LSTM network with an attention mechanism were used by Lian et al. (2022) to detect, track, and predict pedestrian intention, respectively. A multi-task network was implemented by Razali et al. (2021) to recognise pose state and predict pedestrian intentions. ResNet was used to extract features, Part-Intensity-Fields (PIF), and Part-Association-Fields (PAF) to produce channels and pose joints, and a head network to predict pedestrian intentions.

Table 8 presents the results achieved by the most relevant pedestrian intention prediction works in the literature. Unfortunately, direct comparisons between these studies are not possible due to variations in different problem formulations, OTH, TTE, datasets, and metrics. For example, the work that achieved the best accuracy was Zhang, Angeloudis, and Demiris (2022), however, the authors used their own dataset. The second best was Bouhsain et al. (2020) but they used an observation horizon and TTE of 0.6 s.

5. Heterogeneous road agents

All the previously mentioned works primarily focused on predicting the behaviour of either pedestrians or vehicles. However, in a real-world traffic scenario, complex interactions occur among various types of agents, each with different dimensions and dynamics. Consequently, it is crucial to consider the interaction between heterogeneous agents. Several works have addressed the detection and behaviour prediction of heterogeneous agents.

For example, authors (Ma et al., 2019) introduced the **TrafficPredict** algorithm, which was developed to learn motion patterns and predict the trajectories of different types of traffic agents, including pedestrians, bicycles and cars. They adopted the 4D Graph network in conjunction with an RCNN LSTM to learn the movements and interactions of traffic agents. The authors used an OTH of 2 s to predict a horizon of 3 s. They achieved a state-of-the-art average displacement error of 0.085 and a final displacement error of 0.141. **DeepTAgent** is another heterogeneous system presented by Chandra, Randhavane, et al. (2019) in which they used Mask R-CNN to detect objects, a CNN to extract tracking features, and a Heterogeneous Interaction Model (HTMI) that considered collision avoidance behaviour to predict the agents' position, velocity and subsequently their trajectory and interactions. The authors (Chandra, Bhattacharya, et al., 2019) presented a hybrid network for predicting the trajectory of road agents and modelling their interactions. They used a CNN to capture local information, such as the agent's shape and position, and an LSTM network for trajectory prediction. In dense, diverse traffic situations, the algorithm demonstrated a notable performance of 30% over state-of-the-art methods. However, it did not outperform the state-of-the-art algorithms in sparse and homogeneous traffic scenes. Li, Yang, et al. (2020) presented a framework called **EvolveGraph**. In this framework, they encoded an observation graph to infer an interaction graph, and subsequently, decoded both the observation and interaction graphs to predict future trajectories. Zhang, Zhao, et al. (2022) implemented the Attention-based Interaction-aware Trajectory Prediction (AI-TP) model. This model used Graph Attention Network (GAT) to represent interaction among heterogeneous traffic agents and used a Convolutional GRU (ConvGRU) to make predictions. A multi-agent trajectory prediction system was performed by Mo et al. (2022) where a three-channel framework was used to account for dynamics, interactions and road structure. Moreover, a novel Heterogeneous Edge-enhanced graph ATtention network (HEAT) was proposed to extract interaction features. Dynamic features were extracted from the agents' previous trajectories, interaction patterns were represented through a directed edge-feature heterogeneous graph and extracted with the HEAT network. The road structure information was shared among all agents using a gate mechanism. Finally, all the information acquired from the previous process was combined to predict trajectories.

All the previously cited works have predicted the trajectories and interactions among the agents. However, they have not taken into consideration their intentions, such as crossing/not-crossing, braking/non-braking. Also, they have not incorporated the information provided by road static objects like traffic lights and road signs. Static road traffic objects play a crucial role in directing, informing, and controlling road users' behaviour. Furthermore, there is limited research on how to use detection and prediction information to identify potential and developing hazards.

The authors (Chen et al., 2018) proposed a multi-task learning model that combines both object detection and distance prediction to identify dangerous traffic road objects. They used SSD CNN to detect cars, vans, and pedestrians. The input image was divided into a grid map with four vertical and three horizontal distances. Depending on the category of the target vehicle and its location, the network assigned a danger level using blue, green, yellow, and red bounding boxes, where blue and red represented the least and the most dangerous levels, respectively. However, predicting the target vehicle's velocity using a grid map limits the velocity resolution and might not give realistic measurements. Also, relying solely on the distance between the ego and the target vehicle is not enough. For example, an ego vehicle might maintain a safe distance from the target vehicle, but the target vehicle can suddenly brake and change its velocity. Therefore, it would be beneficial for the ego vehicle to predict and recognise instances when the target vehicle is braking or experiencing a sudden change in velocity.

Authors (Li, Wang, et al., 2020) considered themselves pioneers in combining object detection and intention recognition to assess the risks in a complex traffic scenarios. Their objective was to detect both non-static objects such as vehicles and pedestrians, and static objects such as traffic lights, and then use the gained information to evaluate potential hazards ahead. In order to detect the objects, they used the YOLOv4 and the BDD100K dataset and achieved an mAP of 52.7%. For recognising the pedestrian intention (crossing or not-crossing), they used VGG-19 CNN and Part Affinity fields, achieving an accuracy of 97.5%. To predict vehicle intentions, including braking and turning, they employed the EfficientNet CNN, achieving a recognition accuracy of 94%. Lastly, for recognising traffic light state (red, green, or amber), they used the MobileNet CNN, achieving an accuracy of 97.75%. Nevertheless, using only the brake and the turn signal lights information to predict vehicle behaviour and assess danger is not sufficient since braking behaviour can exhibit varying intensities. For example, normal braking, characterised by a gradual decrease in the vehicle's velocity, is typically regarded as a potential hazard. In contrast, harsh braking, involving a sudden and significant change in the vehicle's velocity, is seen as a developing hazard. Furthermore, there are situations where the target vehicles abruptly change their direction without using their turn signal, which also poses a developing hazard. Therefore, the ego vehicle must be capable of detecting sudden changes in the vehicle's direction and velocity. Similarly, depending only on pedestrian crossing/not crossing intentions limits the system to make a long prediction horizon, as pedestrians can cross at different velocities, and may suddenly change their goal destination.

6. Discussion

This paper has surveyed several works that investigate the behaviour prediction of pedestrians and vehicles. Based on the findings, this section presents a general framework diagram, outlines risk assessment, discusses challenges, examines techniques, outlines requirements, and suggests potential future directions for pedestrian and vehicle behaviour prediction systems.

6.1. *General framework for a behaviour prediction system*

A proposed general framework for a behaviour prediction system is depicted in Fig. 6. The camera sensor outputs RGB images which are used by the detection and image processing algorithms.

The detection algorithm is responsible for detecting both static and non-static road objects, including road lanes, vehicles, vulnerable road users, traffic lights, and road signs. The position information of the detected objects, represented by bounding boxes, is then used by a tracking algorithm to assign a unique ID to each object. This ID assignment enables the system to track past trajectories of each detected object, which serves as input for subsequent processing.

The image processing algorithm uses the RGB images from the camera sensor as well as the past trajectories of the detected objects to generate optical flow, depth, appearance, global and local context images. An example of how image processing uses past trajectories is the use of the bounding box information to crop the RGB image at the specific location of the detected object. This cropping operation provides local context information for further analysis and decision-making within the system.

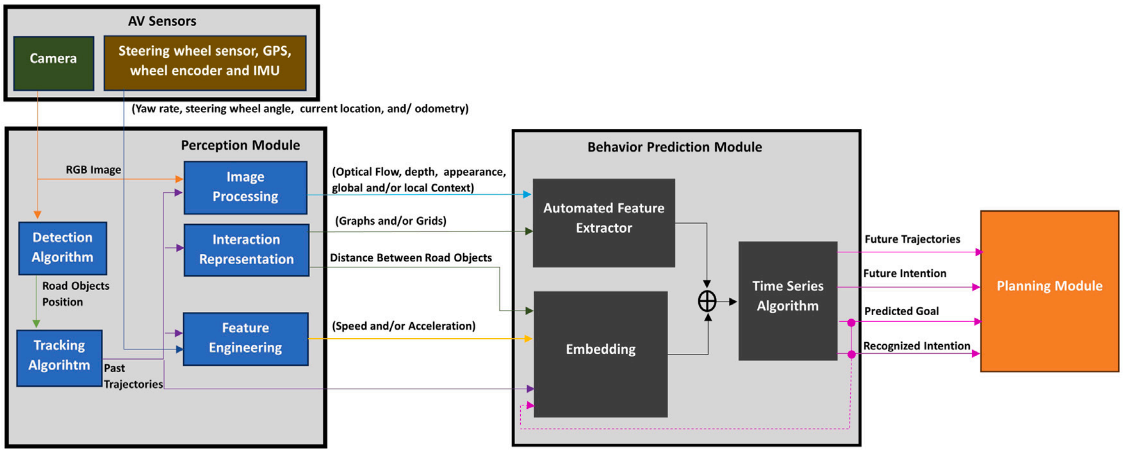
AV Sensors
Camera
Steering wheel sensor, GPS, wheel encoder and IMU
(Yaw rate, steering wheel angle, current location, and/ odometry)
Perception Module
RGB Image
Image Processing
Detection Algorithm
Interaction Representation
Road Objects Position
Feature Engineering
Tracking Algorihtm
Past Trajectories
(Optical Flow, depth, appearance, global and/or local Context)
(Graphs and/or Grids)
Distance Between Road Objects
(Speed and/or Acceleration)
Behavior Prediction Module
Automated Feature Extractor
Embedding
Time Series Algorithm
Future Trajectories
Future Intention
Predicted Goal
Recognized Intention
Planning Module

Fig. 6. General behaviour prediction framework. The behaviour prediction module **consists** of an automated feature extractor (CNN, 3D-CNN, GCN, FCN, CVAE, GAN, etc.), an embedding layer (FCN and ANN), and a time series algorithm (RNN, GRU, and LSTM). It is **dependent** on the perception module (Detection, tracking, image processing, interaction representation, and feature engineering) which is dependent on the ego vehicle sensors (camera, GPS, and wheel encoder). Additionally, the outputs of the behaviour prediction modules are **sent** to the planning module.

The interaction representation algorithm uses the past trajectories of the objects to calculate distances between the traffic agents, construct graph networks with vertices and edges, and generate grid maps that account for interactions between traffic agents.

The feature engineering algorithm uses the past trajectories of objects and internal sensors data from the AV (e.g., steering wheel angle, yaw rate, wheel encoder, etc.) to derive additional features. For example, to use the differences between the objects' positions between consecutive frames to calculate their velocities.

The outputs of the perception module are then fed into the automated feature extractor and the embedding algorithms within the behaviour prediction module. Automated feature extractors are deep learning algorithms designed to generate feature vectors representing spatial properties of the inputs. Embedding uses a linear transformation to transform the inputs into a desired output feature size. The time series algorithm uses the combined feature vectors generated by the automated feature extractor and the embedding layer to learn temporal information, enabling it to predict various aspects of object behaviour, including future trajectories, future intentions, goals, and current intentions. Note that the predicted goals and recognised intentions can be used by the embedding layer and the time series algorithm as extra information for predicting future trajectories.

Finally, the outputs of the behaviour prediction module are then used by the AV's Planning module, which in turn uses this information to plan the actions of the AV to achieve its final goal.

6.2. Risk assessment for behaviour prediction system

Authors (Bhavsar et al., 2017) proposed a risk assessment for a AV. They mentioned that AV failures can arise from various aspects, including vehicular components such as hardware, software, mechanical systems, communication infrastructure, and interactions between the passenger and the AV Human Machine Interface system. Based on their finding, this paper presents a risk assessment specifically for an AV behaviour prediction system. This assessment identifies, analyses, and provides recommendations for mitigating and controlling these identified risks.

6.2.1. Risk identification

Based on the general framework for a behaviour prediction system depicted in Fig. 6, the following risks have been identified:

- Camera sensor failure: this includes hardware malfunctions, blocked field of view, and noise (electricity, heat, and illumination).
- Computing components failure: computer or GPU failure.
- Sensor Failure: Failure in the steering wheel, wheel encoder, GPS, and IMU sensors.
- Detection algorithm failure: missed detections, poor intersection over union, false-positive and false-negative classification.
- Tracking algorithm failure: missed tracking and incorrect association of objects between frames. For instance, an object might not be tracked in the next frame or objects might swap their IDs due to overlap.
- Image processing failure: incorrect optical flow and depth estimation.
- Interaction representation failure: noisy and incorrect distance calculation, as well as incorrect graph or grid representation of the object interactions.
- Feature Engineering failure: redundant features, noisy estimates speed and acceleration due to poor detection and tracking performance.
- Cybersecurity failure: remote hacking, vehicle spoofing, insider threat, and tampering with sensor data.

6.2.2. Risk analysis

The authors (Bhavsar et al., 2017) discussed several methods for analysing risks in automotive contexts, including situation-based analysis, ontology-based analysis, failure modes and effects analysis (FMEA), and fault tree analysis (FTA). From their investigation, they concluded that FTA is the most suitable method for conducting a risk assessment on AV features. For this reason, this paper also adopts FTA to perform a risk analysis on the behaviour prediction system. FTA methods have the following advantages, being event-orientated, enabling the diagnosis of the root cause of failures, facilitating an understanding of how subsystems can impact each other, having a straightforward and graphical nature for ease of comprehension, and aiding in decision-making regarding the control of identified risks. The proposed FTA is depicted in Fig. 7. A qualitative analysis of the proposed FTA reveals that the system is highly vulnerable because any failure occurrence of the basic events (EVX) can lead to the failure of the behaviour prediction system. For instance, if the detection algorithm fails, it can cascade failures throughout the tracking algorithm, image processing, interaction representation, and feature engineering, ultimately in the failure of the behaviour predictions system.

In order to quantitatively analyse the behaviour prediction system, it is required to know the probability of failure for each event (EVX), which depends on the hardware, software, and cybersecurity in use. However, a general mathematical model to calculate the overall system

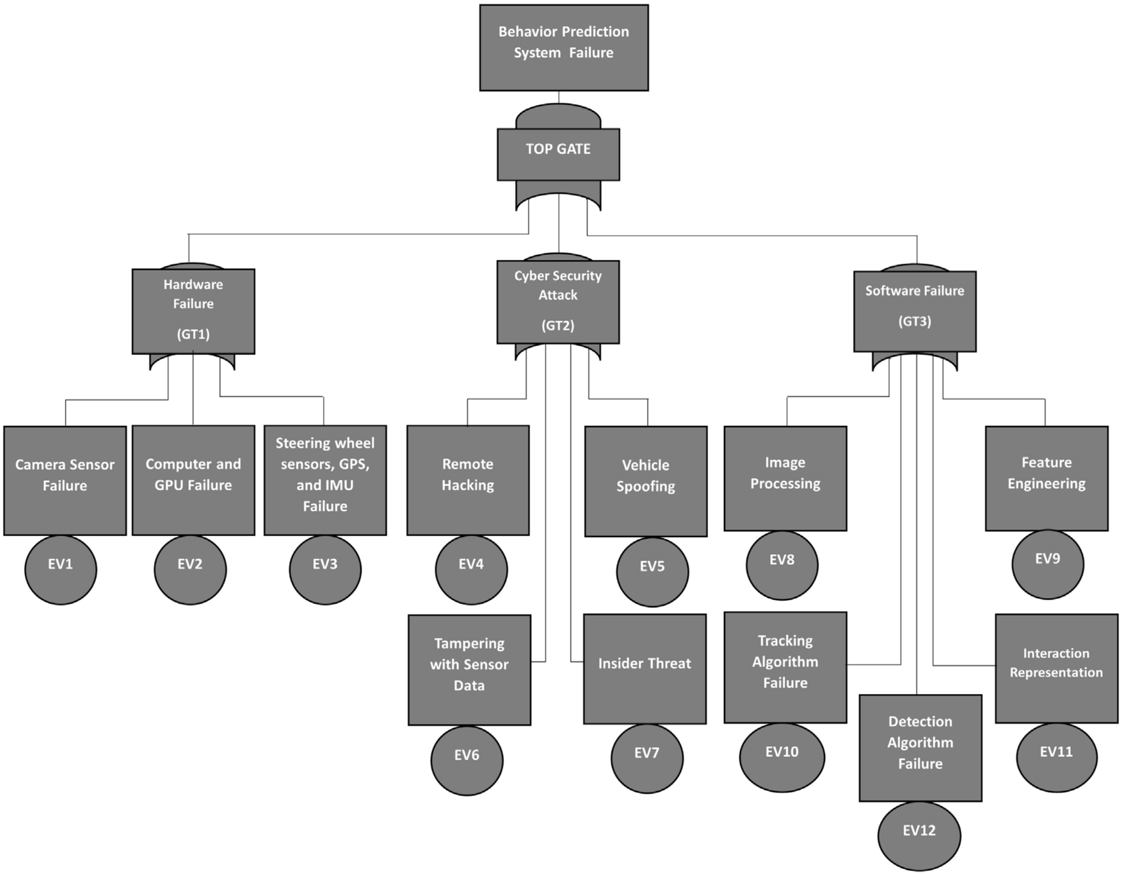

Fig. 7. Fault tree analysis for a Behaviour Prediction System. The circle shapes with the square shapes are the basic events that may lead to failures on the top events. The square shape after the TOP GATE is the top event which means the failure of the behaviour prediction system. The "OR" gates mean that if one of its input events occurs it will output an event as true.

failure from an FTA diagram depicted in Fig. 7 is given by the following equation (Ruijters & Stoelinga, 2015; Xing & Amari, 2008),

$$Q_0(t) \le (1 - \Pi_{j=1^k[1-\check{Q}_j(t)]}) \tag{3}$$

where $Q_0(t)$ is the top event (failure of the behaviour prediction system), $\check{Q_j}(t)$ is the failure probability of a minimal cut-set. For instance, the probability that the TOP GATE in the proposed FTA diagram happens is given by,

$$Q_0(t) \le (1 - [1 - P(GT1)] * [1 - P(GT2)] * [1 - P(GT3)]) \tag{4}$$

where

$$P(GT1) = (1 - [1 - P(EV1)] * [1 - P(EV2)] * [1 - P(EV3)]) \tag{5}$$

and,

$$P(GT2) = (1 - [1 - P(EV4)] * [1 - P(EV5)] * [1 - P(EV6)] * [1 - P(EV7)]) \tag{6}$$

and

$$P(GT3) = (1 - [1 - P(EV8)] * [1 - P(EV9)] * [1 - P(EV10)] * [1 - P(EV11)]). \tag{7}$$

6.2.3. Risk control

Based on the identification and analysis of risks, it has been concluded that a behaviour prediction system is vulnerable. Below are some recommendations to mitigate these risks:

- Given that any hardware failure can cause a top event, it is recommended to have backups for hardware components with a high probability of failure, for example, to have an extra camera sensor. The disadvantage of this approach is that it is expensive and requires more space in the vehicle.
- For the general prediction behaviour system in question, it is observed that it relies on three types of information (RGB image, engineering feature, and interaction) for predictions. Therefore it is recommended to enable the system to function in a degraded mode by using one or two pieces of information if one of them fails.
- The detection and tracking algorithms are important for the system, as their outputs are used by the other algorithms. Thus, it is recommended to make use of sensor fusion, since if one of the hardware or the algorithms responsible for detecting and tracking the object fails the system can work in a degraded mode.

6.3. Behaviour prediction system challenges

Table 9 summarises the main challenges in the research of behaviour prediction of traffic agents. These challenges are categorised into target agents, systems, resources, and uncertainties. Target agents refer to the unique characteristics of these agents that make their behaviour challenging to predict. System challenges are related to the inherent characteristics of the system, considering its design and evaluation. Resource challenges are associated with the hardware and data required for training and operating the system. Uncertainties include events such as hardware malfunctions, cybersecurity vulnerabilities, and software failures.

6.4. Behaviour prediction system requirements

An AV behaviour prediction system needs to meet several key requirements to ensure its effectiveness:

Table 8
Results for the most relevant **pedestrian** intention prediction works.

Work	Dataset	Obs. Hor.	TTE	Acc(%)	AUC(%)	F1(%)	Rec.(%)	Prec(%)	ROC-AUC(%)
Gazzeh and Douik (2022)	JAAD	–	Recog.	92.88	–	–	–	–	–
Fang and López (2018)	JAAD	0.5 s	Next-Frame	88	–	–	–	–	–
STRR-Graph Liu et al. (2020)	JAAD	0.5 s	Next-Frame	76.98	–	–	–	–	–
FUSSI-net Piccoli et al. (2020)	JAAD	0.5 s	Next-Frame	76.6	–	–	–	–	–
PIEint Rasouli et al. (2019)	PIE	0.5 s	Next-Frame	79	–	87	–	90	73
CA-LSTM Lian et al. (2022)	JAAD	0.5 s	Next-Frame	89.68	–	75.38	**85.96**	–	–
PV-LSTM Bouhsain et al. (2020)	JAAD	0.6 s	0.6 s	91.48	–	–	–	–	–
Ma and Rong (2022)	BPI	–	0.6 s	89.5	**99.2**	–	–	–	–
SFR-GRU Rasouli et al. (2020)	PIE	0.5 s	2 s	84.4	82.9	72.1	80	65.7	–
C+B+S+Int Kotseruba et al. (2020)	PIE	0.5 s	2 s	83	85	81	85	79	–
PCIR Yang et al. (2021)	JAAD	–	–	89.6	–	–	–	–	–
Chen et al. (2021)	PIE	0.5 s	1.5 s	79	–	78	–	–	–
I+A+F+R Yao et al. (2021b)	JAAD PIE	0.5 s –	1-2 s	87 84	92 88	70 **90**	– –	66 **96**	– –
Yang, Zhang, et al. (2022)	JAAD PIE	0.5 s	1-2 s	83 89	82 86	63 80	81 81	51 79	– –
GRAPH+ Cadena et al. (2022)	JAAD PIE	0.5 s	1–2 s	86 89	88 90	65 81	75 79	58 83	– –
Achaji et al. (2022)	PIE	0.5 s	1–2 s	91	91	83	–	–	–
Scene-STGCN Naik et al. (2022)	PIE	0.5 s	1–2 s	83	–	89	–	**96**	**85**
PCPA Kotseruba et al. (2021)	JAAD PIE	0.5 s –	0.5-1 s	85 87	86 86	68 77	– –	– –	– –
ST-CrossingPose Zhang, Angeloudis, and Demiris (2022)	OWN	0.5 s	1 s 2 s	**92** **92**	84.9 84.1	83.7 79.7	81.8 79.7	85.9 81.3	– –
Zeng (2022)	JAAD	-s	-s	84	–	–	–	–	**85**

- **Good Evaluation Metric Performance**: AV behaviour prediction system is a safety-critical system, therefore it must perform well in terms of evaluation metric performance to prevent traffic collisions. For example, if the system fails to predict that a pedestrian will cross the road, it could lead to a serious collision.
- **Long Prediction Horizon (PTH)**: A system with a long PTH can plan and react well in advance, reducing the chances of collisions and improving overall safety.
- **Fast Inference Time**: Given that an AV behaviour prediction system must operate in a real-time, it must have a low inference time and require a low hardware resource.
- **Low Cost**: To make AVs accessible to a wide range of people, the behaviour prediction system should be cost-effective, ensuring that AVs are affordable for all social classes
- **Low Hardware Resource Requirement**: Efficient utilisation of hardware resources is important, as it allows the system to run on hardware with limited capacity.
- **Robustness**: The system should be robust and able to handle various scenarios and conditions on the road, ensuring reliable performance in different situations.
- **Prediction of Various Non-Static Objects**: The system should be capable of predicting the behaviour of different types of non-static objects on the road, including pedestrians, vehicles, animals, and cyclists, to ensure comprehensive safety.

Evaluation metrics, long prediction horizons, and robustness are interrelated. For instance, as the prediction horizon increases the evaluation metric performance tends to decrease. In addition, as a system becomes more robust, its evaluation metric performance is expected to increase. The major challenges that limit behaviour prediction algorithms from meeting the previously mentioned requirements stem from the fact that an agent's behaviour depends on other agents in the scene, the local and global context, and their final goal. Various approaches have been proposed to address these challenges:

- Social pooling layers (Alahi et al., 2016; Deo & Trivedi, 2018a), Graph representation, GCN, self-attention based social pooling (Mangalam et al., 2020), message passing mechanism (Zhang et al., 2019), occupancy maps (Kasper et al., 2012; Park et al., 2018; Xue et al., 2018), view frustum social pooling (Hasan et al., 2018), and star-like networks to model interactions between agents (Zhu et al., 2019).
- CNNs to extract agents' appearance, body pose, local context, global context, and to classify intentions (Biparva et al., 2021; Chen et al., 2021; Fang et al., 2017; Fernández-Llorca et al., 2020; Izquierdo et al., 2021; Yang, Zhang, et al., 2022; Yao et al., 2021b; Zhao et al., 2019).
- Attention mechanisms and transformer networks to focus on the most relevant information (Achaji et al., 2022; Lian et al., 2022; Rasouli et al., 2019).

Table 9
Behaviour prediction research challenges.

Type of challenge	Class	Challenges
Target Agents	*Pedestrian*	Highly dynamic, can move in many directions and change them very quickly, be easily occluded, be distracted by their own objects or external environments, their motion can be affected by other traffic agents, might be under the influence of drugs or alcoholic drinks, and they are hard to see in poor visibility condition.
	Vehicle	Dependent on other vehicles' actions, traffic rules, road geometry, different driving environments, vehicles have multi-modal behaviour, different types of vehicles have different motion properties, drivers might be under the influence of drugs or alcoholic drinks, and target vehicles might be occluded.
*System	*Design*	To achieve a good evaluation metric performance, long PTH, real-time inference, low hardware resources, and robustness.
	Evaluation	Works have used different types of datasets, evaluation metrics, observation and prediction horizon, and hardware setup. Therefore, works cannot be directly compared and the actual progress of pedestrians and vehicle behaviour prediction research cannot be measured.
*Resources	*Hardware*	Smaller size GPUs that can process deep learning algorithms in real-time, sensors that enable the AV to perceive 360-degree road view, and affordable hardware to enable all social classes to afford AVs.
	Data	Several existing datasets are not publicly available and they are not standardised to enable cross-dataset evaluation and progressive training pipeline techniques.
Uncertainties *	*Hardware Failure*	Camera, GPS, IMU, steering wheel, and wheel encoder sensor failure.
	Cyber Attack	Remote hacking, vehicle spoofing, insider threat, and tampering with sensor data.
	Software Failure	Perception module (detection, tracking, image processing, interaction representation, and feature engineering) failure.

- 3D-CNNs and temporal-Densenet to learn short-term temporal information (Biparva et al., 2021; Kotseruba et al., 2021; Piccoli et al., 2020; Yang et al., 2021).
- LSTMs and GRUs to learn long-term temporal information (Bouhsain et al., 2020; Chung et al., 2014; Kotseruba et al., 2020; Rasouli et al., 2019, 2020).
- A modified version of the LSTM cell that accepts more than one input sequence set (Quan et al., 2021).
- CVAE was used to estimate the final goals of the agents to extend the prediction time horizon (Lee, Choi, et al., 2017; Mangalam et al., 2020; Wang et al., 2022; Yao et al., 2021a).
- Heterogeneous agent behaviour prediction works have been presented to enable the system to predict the behaviour of different non-static object behaviour (Chandra, Bhattacharya, et al., 2019; Chandra, Randhavane, et al., 2019; Chen et al., 2018; Li, Wang, et al., 2020; Li, Yang, et al., 2020; Ma et al., 2019; Mo et al., 2022). However, these works have primarily focused on pedestrians, cyclists, and vehicles, while there are other objects such as animals, disabled individuals, scooters, toys (balls), skate riders, etc.
- Combination of two or more methods to compensate their limitations (Chen et al., 2021; Kotseruba et al., 2021; Liu et al., 2020; Naik et al., 2022; Yang, Zhang, et al., 2022; Yao et al., 2021b; Zeng, 2022).
- Systems that can predict the behaviour of heterogeneous agents (Chandra, Bhattacharya, et al., 2019; Chandra, Randhavane, et al., 2019; Li, Wang, et al., 2020; Li, Yang, et al., 2020; Ma et al., 2019).

Inference time, low cost, and low hardware resource requirements are also interrelated. For example, if a system consumes less memory and computational power, it results in cheaper hardware requirements, making the overall system more cost-effective. Typically, when a system requires less memory, such as for processing image inputs, the system's overall inference time is expected to be shorter. However, there may be a trade-off between accuracy and inference time. For example, using multiple-feature information can increase the system's accuracy but may lead to longer inference times compared to a system using a single type of feature. The following methods have been proposed in order to achieve low inference time, low cost, and low hardware resource requirements:

- GCN, which represents interactions between agents effectively without relying on additional information like original images, cropped images, or contextual information (Li et al., 2019a, 2019b).
- Dual-LSTM, which allows the system to learn more information from past trajectories without requiring extra input features (Xin et al., 2018).
- Fusion of multiple input features (context, interaction, trajectories, and appearance) into an enriched image representation, rather than processing a sequence of images (Izquierdo et al., 2021).

6.5. Behaviour prediction system further work

Despite the techniques presented to meet the specified requirements, there is still work to be done from the authors' perspective. For example:

- Most of the works, both for pedestrians and vehicles, were implemented using either a top-view or BEV dataset, which may not be ideal for an AV system. Only in the past five years have researchers started implementing algorithms using on-board datasets such as PREVENTION, Appolo, JAAD, and PIE. Moreover, most of the works that used on-board datasets focused on implementing intention prediction algorithms, and most of proposed algorithms cannot be directly compared.

- While some works have used the same datasets, evaluation metrics, observation time horizon, and prediction time horizon, these works were implemented on top-view and BEV datasets. For example, many vehicle trajectory predictions have used the NGSIM dataset with an OTH of 3 s, a PTH of 5 s, and the MSE evaluation metric. Several pedestrian prediction trajectory algorithms adopted the ETH and UCY dataset, with an OTH of 3.2 s, a PTH of 4.8 s, and the ADE and FDE evaluation metric. If these datasets were ideal for AV systems, then the best vehicle trajectory prediction algorithms would be GRIP (Li et al., 2019b), GRIP++ (Li et al., 2019a), and AI-TP (Zhang, Zhao, et al., 2022), and the best pedestrian trajectory prediction algorithm would be the Bi-Trap algorithm (Yao et al., 2021a).
- There is a lack of research on unusual behaviour exhibited by pedestrians and vehicles. For example, pedestrians might exhibit unusual behaviour when under the influence of toxic substances, involved in fights, or disoriented. Similarly, vehicles may display unusual behaviour when the driver is under the influence of toxic substances, and is distracted with their personal belongings, or if the vehicle is an emergency vehicle, garbage truck, road sweeper, carrying an abnormal load, or experiencing mechanical malfunctioning.
- There is a limited research on decreasing inference time, and more emphasis should be placed on addressing this demand.
- Standardising datasets would enable cross-dataset evaluation and the development of progressive training pipeline techniques.
- Introducing universal metrics would allow for direct comparisons of algorithm performance.
- When considering a full pipeline system (detection, tracking and behaviour prediction), it is necessary to account for perception uncertainties due to sensor noise, fuzzy features, or unknown inputs (Liu et al., 2022). Since there are a limited number of works that have implemented a full pipeline system, more works considering the entire pipeline process are recommended to investigate the effect of possible noise.

Based on the literature review the following suggestions are given to further improve and accelerate the development of the Autonomous Vehicle Behaviour Prediction System:

- Encourage more research works to adopt on-board view datasets for predicting both pedestrian and vehicle behaviour, including intention and trajectories.
- Standardise existing dataset to enable cross-dataset evaluation and progressive training pipeline techniques.
- Choose or create a standard evaluation metric to enable direct comparison among algorithms.
- Develop datasets that have instances of abnormal pedestrian and vehicle behaviours to enable research on the recognition and prediction of abnormal pedestrian and vehicle behaviour.
- Implement behaviour prediction algorithms on resource-constrained hardware, such as Jetson Orin, and Jetson Xavier GPUs, which are low-cost, small in size, lightweight, and consume low power.
- Investigate more methods to select the target object and the objects that directly interact with the target object.

The general object detection problem serves as an example of the importance of having a large dataset and standard evaluation metrics. The field has achieved an acceptable level of maturity because researchers have access to publicly available large image benchmark datasets, such as the ImageNet (Russakovsky et al., 2015) and COCO (Lin et al., 2014). These datasets enabled the authors to directly compare their detection algorithm performance and to measure the advancement of object detection research.

7. Conclusion

AV systems must not only detect pedestrians and vehicles but also predict their behaviour to avoid or mitigate collisions. Therefore, the purpose of this literature review, was to survey the most relevant pedestrian and vehicle behaviour prediction algorithms to identify the requirements for a behaviour prediction algorithm, the challenges associated with predicting pedestrian and vehicle behaviour, whether current techniques have met these requirements, and what steps are needed to enable AVs to predict pedestrian and vehicle behaviours. In conclusion, the review shows that:

- An AV behaviour prediction system must have a good evaluation metric performance, long prediction horizon, fast inference time, must be cost-effective, robust, require minimal hardware resources, and predict various types of non-static objects on the road.
- The main challenges in predicting the behaviour of traffic agents involve modelling their interactions, establishing relationship between the agents and the scene, and achieving a balance between good evaluation metric performance and low inference times.
- Current techniques do not fully meet these requirements for several reasons:
  - when predicting for long-term horizons, evaluation metric performance significantly decreases;
  - while top-view and BEV datasets are commonly used in the literature, there are limited works that adopted on-board datasets, which are more suitable for AVs;
  - on-board datasets usually only use a single forward-facing camera, limiting the behaviour prediction system to consider only agents ahead, whereas considering agents around the ego vehicles using multiple cameras is essentials (Zhang, 2021);
  - more investigation is required to develop models that can predict intention and trajectory simultaneously; although some authors (Li et al., 2019a, 2019b) claimed that their system has achieved real-time inference times, they have used top-view cameras, whereas systems that use on-board sensors may require more processing time;
  - there are no works that consider abnormal behaviour exhibited by traffic agents.
- Most of the reviewed works have not considered the full pipeline behaviour prediction process, which consists of detection, classification, and tracking. More research should focus on the full pipeline process to assess the performance of each stage and its impact on the final prediction results.

Abbreviations

AV	Autonomous Vehicle.
ADAS	Advanced Driver Assistance System.
WHO	World Health Organisation.
DL	Deep Learning.
OTH	Observation Time Horizon.
PTH	Prediction Time Horizon.
EV	Ego Vehicle.
TTE	Time-To-Event.
KF	Kalman Filter.
EKF	Extended Kalman Filter.
HMM	Hidden Markov Model.
SVM	Support Vector Machine.
ANN	Artificial Neural Network.
OGM	Occupancy Grid Map.
CNN	Convolutional Neural Network.

FCN	Fully Connected Network.
RNN	Recurrent Neural Network.
GCNN	Graph Convolutional Neural Network.
LSTM	Long-short Term Memory.
RMSE	Root Mean Square Error.
GRIP	Graph-based Interaction-aware Trajectory Prediction.
LLC	Left Lane Change.
RLC	Right Lane Change.
NLC	No Lane Change.
DESIRE	Deep Stochastic Inverse Optimal Control RNN encoder–Decoder.
CVAE	Conditional Variational Auto Encoder.
IOC	Inverse Optimal Control.
BEV	Bird's Eye View.
ARIMA	Auto-Regressive Integrated Moving Average.
TSM	Target Selection Model.
TIM	Temporal Integration Method.
RoI	Region of Interest.
AUC	Area Under the Curve.
ROC-AUC	Receiver Operating Characteristic Curve - AUC.
ANDE	Average Non-Linear Displacement Error.
MAD	Mean Average Displacement.
FAD	Final Average Displacement.
GAN	Generative Adversarial Network.
MLP	Multi-layer Perceptron.
MDN	Mixture Density Network.
ST-GCN	Spatial–Temporal Graph Convolutional Network.
GRU	Gated Recurrent Unit.
PIF	Part-Intensity-Fields.
PAF	Part Association Fields.
AI-TP	Attention-Based Interaction-aware Trajectory Prediction.
HEAT	Heterogeneous Edge-enhanced Graph Attention Network.
MATF	Multi-Agent Tensor Fusion.
VGMM	Variational Gaussian Mixture Models.
WSADE	Weight Sum of Average Displacement Error.
WSFDE	Wight Sum of Final Displacement Error.
NGSIM-LP	NGSIM Lankershim and Peachtree.
RBF	Radial Basis Function.
MLPE	Mean Lateral Position Error.
ADE	Average Displacement Error.
FDE	Final Displacement Error.
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis

CRediT authorship contribution statement

Luiz G. Galvão: Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **M. Nazmul Huda:** Conceptualization, Writing – review & editing, Supervision, Resources, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Luiz G. Galvao reports financial support was provided by Engineering and Physical Sciences Research Council.

Data availability

No data was used for the research described in the article.

Acknowledgements

This research was funded by EPSRC DTP Ph.D. studentship at Brunel University London, United Kingdom.

References

Abbas, A. F., Sheikh, U. U., AL-Dhief, F. T., & Haji Mohd, M. N. (2021). A comprehensive review of vehicle detection using computer vision. *Telkomnika*, *19*(3).

Abdulrahim, K., & Salam, R. A. (2016). Traffic surveillance: A review of vision based vehicle detection, recognition and tracking. *International journal of applied engineering research*, *11*(1), 713–726.

Abughalieh, K. M., & Alawneh, S. G. (2020). Predicting pedestrian intention to cross the road. *IEEE Access*, *8*, 72558–72569, Publisher: IEEE.

Achaji, L., Moreau, J., Fouqueray, T., Aioun, F., & Charpillet, F. (2022). Is attention to bounding boxes all you need for pedestrian action prediction? In *2022 IEEE intelligent vehicles symposium* (pp. 895–902). IEEE.

Afrin, T., & Yodo, N. (2020). A survey of road traffic congestion measures towards a sustainable and resilient transportation system. *Sustainability*, *12*(11), 4660, Publisher: Multidisciplinary Digital Publishing Institute.

Ahmed, S., Al Bazi, A., Saha, C., Rajbhandari, S., & Huda, M. N. (2023). Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation. *Expert Systems with Applications*, *225*, Article 120077, Publisher: Elsevier.

Ahmed, S., Huda, M. N., Rajbhandari, S., Saha, C., Elshaw, M., & Kanarachos, S. (2019a). Pedestrian and cyclist detection and intent estimation for autonomous vehicles: A survey. *Applied Sciences*, *9*(11), 2335, Publisher: MDPI.

Ahmed, S., Huda, M. N., Rajbhandari, S., Saha, C., Elshaw, M., & Kanarachos, S. (2019b). Visual and thermal data for pedestrian and cyclist detection. In *Towards autonomous robotic systems: 20th annual conference, TAROS 2019, London, UK, July 3–5, 2019, proceedings, Part II 20* (pp. 223–234). Springer.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 961–971).

Altché, F., & de La Fortelle, A. (2017). An LSTM network for highway trajectory prediction. In *2017 IEEE 20th international conference on intelligent transportation systems* (pp. 353–359). IEEE.

Antonio, J. A., & Romero, M. (2018). Pedestrians' detection methods in video images: A literature review. In *2018 international conference on computational science and computational intelligence* (pp. 354–360). IEEE.

Benterki, A., Boukhnifer, M., Judalet, V., & Choubeila, M. (2019). Prediction of surrounding vehicles lane change intention using machine learning. In *2019 10th IEEE international conference on intelligent data acquisition and advanced computing systems: technology and applications, vol. 2* (pp. 839–843). IEEE.

Benterki, A., Boukhnifer, M., Judalet, V., & Maaoui, C. (2020). Artificial intelligence for vehicle behavior anticipation: Hybrid approach based on maneuver classification and trajectory prediction. *IEEE Access*, *8*, 56992–57002, Publisher: IEEE.

Berndt, H., & Dietmayer, K. (2009). Driver intention inference with vehicle onboard sensors. In *2009 IEEE international conference on vehicular electronics and safety* (pp. 102–107). IEEE.

Bhattacharyya, A., Fritz, M., & Schiele, B. (2018). Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4194–4202).

Bhavsar, P., Das, P., Paugh, M., Dey, K., & Chowdhury, M. (2017). Risk analysis of autonomous vehicles in mixed traffic streams. *Transportation Research Record*, *2625*(1), 51–61, Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Biparva, M., Fernández-Llorca, D., Izquierdo-Gonzalo, R., & Tsotsos, J. K. (2021). Video action recognition for lane-change classification and prediction of surrounding vehicles. arXiv preprint arXiv:2101.05043.

Bonnin, S., Weisswange, T. H., Kummert, F., & Schmüdderich, J. (2014). Pedestrian crossing prediction using multiple context-based models. In *17th international IEEE conference on intelligent transportation systems* (pp. 378–385). IEEE.

Bouhsain, S. A., Saadatnejad, S., & Alahi, A. (2020). Pedestrian intention prediction: A multi-task perspective. arXiv preprint arXiv:2010.10270.

Cadena, P. R. G., Qian, Y., Wang, C., & Yang, M. (2022). Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, Publisher: IEEE.

Chandra, R., Bhattacharya, U., Randhavane, T., Bera, A., & Manocha, D. (2019). RoadTrack: Realtime tracking of road agents in dense and heterogeneous environments. arXiv, arXiv–1906.

Chandra, R., Randhavane, T., Bhattacharya, U., Bera, A., & Manocha, D. (2019). *Deeptagent: Realtime tracking of dense traffic agents using heterogeneous interaction*: *Technical report*, 2018. [Online]. Available: http://gamma.cs.unc.edu/HTI.

Chen, L., Ding, Q., Zou, Q., Chen, Z., & Li, L. (2020). DenseLightNet: A light-weight vehicle detection network for autonomous driving. *IEEE Transactions on Industrial Electronics*, *67*(12), 10600–10609, Publisher: IEEE.

Chen, L., Ma, N., Wang, P., Li, J., Wang, P., Pang, G., & Shi, X. (2020). Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Science and Technology*, *25*(4), 458–470, Publisher: TUP.

Chen, T., Tian, R., & Ding, Z. (2021). Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3103–3109).

Chen, Y., Zhao, D., Lv, L., & Zhang, Q. (2018). Multi-task learning for dangerous object detection in autonomous driving. *Information Sciences*, *432*, 559–571, Publisher: Elsevier.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

COLONNA, M. (2018). Urbanisation worldwide. Knowledge for policy - European Commission, URL: https://ec.europa.eu/knowledge4policy/foresight/topic/continuing-urbanisation/urbanisation-worldwide_en.

Czech, P., Braun, M., Kreßel, U., & Yang, B. (2022). On-board pedestrian trajectory prediction using behavioral features. arXiv preprint arXiv:2210.11999.

Dai, S., Li, L., & Li, Z. (2019). Modeling vehicle interactions via modified LSTM models for trajectory prediction. *IEEE Access*, *7*, 38287–38296, Publisher: IEEE.

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, *129*, 845–881, Publisher: Springer.

Deo, N., Rangesh, A., & Trivedi, M. M. (2018). How would surround vehicles move? A unified framework for maneuver classification and motion prediction. *IEEE Transactions on Intelligent Vehicles*, *3*(2), 129–140, Publisher: IEEE.

Deo, N., & Trivedi, M. M. (2018a). Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1468–1476).

Deo, N., & Trivedi, M. M. (2018b). Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMS. In *2018 IEEE intelligent vehicles symposium* (pp. 1179–1184). IEEE.

Dueholm, J. V., Kristoffersen, M. S., Satzoda, R. K., Moeslund, T. B., & Trivedi, M. M. (2016). Trajectories and maneuvers of surrounding vehicles with panoramic camera arrays. *IEEE Transactions on Intelligent Vehicles*, *1*(2), 203–214, Publisher: IEEE.

Durrant-Whyte, H. (2001). A critical review of the state-of-the-art in autonomous land vehicle systems and technology. *Albuquerque (NM) andLivermore (CA), USA: SandiaNationalLaboratories*, *41*, 242.

Fang, Z., & López, A. M. (2018). Is the pedestrian going to cross? answering by 2D pose estimation. In *2018 IEEE intelligent vehicles symposium* (pp. 1271–1276). IEEE.

Fang, Z., Vázquez, D., & López, A. M. (2017). On-board detection of pedestrian intentions. *Sensors*, *17*(10), 2193, Publisher: MDPI.

Fernández-Llorca, D., Biparva, M., Izquierdo-Gonzalo, R., & Tsotsos, J. K. (2020). Two-stream networks for lane-change prediction of surrounding vehicles. In *2020 IEEE 23rd international conference on intelligent transportation systems* (pp. 1–6). IEEE.

Flohr, F. F., Kooij, J. F. K., Pool, E. A. P., & Gavrila, D. M. G. (2018). Context-based path prediction for targets with switching dynamics.

Galvao, L. G., Abbod, M., Kalganova, T., Palade, V., & Huda, M. N. (2021). Pedestrian and vehicle detection in autonomous vehicle perception systems—A review. *Sensors*, *21*(21), 7267, Publisher: MDPI.

Gazzeh, S., & Douik, A. (2022). Deep learning for pedestrian behavior understanding. In *2022 6th international conference on advanced technologies for signal and image processing* (pp. 1–5). IEEE.

Girma, A., Amsalu, S., Workineh, A., Khan, M., & Homaifar, A. (2020). Deep learning with attention mechanism for predicting driver intention at intersection. In *2020 IEEE intelligent vehicles symposium* (pp. 1183–1188). IEEE.

GOVUK, G. (2020). Reported road casualties Great Britain, annual report: 2020. GOV.UK, URL: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2020/reported-road-casualties-great-britain-annual-report-2020.

GOVUK, G. (2021). Reported road casualties in Great Britain, provisional estimates: year ending June 2021. GOV.UK, URL: https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-provisional-estimates-year-ending-june-2021/reported-road-casualties-in-great-britain-provisional-estimates-year-ending-june-2021.

Gulzar, M., Muhammad, Y., & Muhammad, N. (2021). A survey on motion prediction of pedestrians and vehicles for autonomous driving. *IEEE Access*, *9*, 137957–137969, Publisher: IEEE.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2255–2264).

Hasan, I., Setti, F., Tsesmelis, T., Del Bue, A., Galasso, F., & Cristani, M. (2018). Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6067–6076).

He, J.-H., Chen, Y.-L., Chen, X.-Z., & Chiang, H.-H. (2021). Vehicle turning intention prediction based on data-driven method with roadside radar and vision sensor. In *2021 IEEE international conference on consumer electronics-Taiwan* (pp. 1–2). IEEE.

Hermes, C., Wohler, C., Schenk, K., & Kummert, F. (2009). Long-term vehicle motion prediction. In *2009 IEEE intelligent vehicles symposium* (pp. 652–657). IEEE.

Huang, H., Zeng, Z., Yao, D., Pei, X., & Zhang, Y. (2021). Spatial–temporal ConvLSTM for vehicle driving intention prediction. *Tsinghua Science and Technology*, *27*, 599–609.

Izquierdo, R., Quintanar, A., Lorenzo, J., García-Daza, I., Parra, I., Fernández-Llorca, D., & Sotelo, M. A. (2021). Vehicle lane change prediction on highways using efficient environment representation and deep learning. *IEEE Access*, *9*, 119454–119465, Publisher: IEEE.

Izquierdo, R., Quintanar, A., Parra, I., Fernández-Llorca, D., & Sotelo, M. A. (2019). The prevention dataset: A novel benchmark for prediction of vehicles intentions. In *2019 IEEE intelligent transportation systems conference* (pp. 3114–3121). IEEE.

Karasev, V., Ayvaci, A., Heisele, B., & Soatto, S. (2016). Intent-aware long-term prediction of pedestrian motion. In *2016 IEEE international conference on robotics and automation* (pp. 2543–2549). IEEE.

Kasper, D., Weidl, G., Dang, T., Breuel, G., Tamke, A., Wedel, A., & Rosenstiel, W. (2012). Object-oriented Bayesian networks for detection of lane change maneuvers. *IEEE Intelligent Transportation Systems Magazine*, *4*(3), 19–31, Publisher: IEEE.

Keller, C. G., & Gavrila, D. M. (2013). Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, *15*(2), 494–506, Publisher: IEEE.

Khosroshahi, A., Ohn-Bar, E., & Trivedi, M. M. (2016). Surround vehicles trajectory analysis with recurrent neural networks. In *2016 IEEE 19th international conference on intelligent transportation systems* (pp. 2267–2272). IEEE.

Kim, B., Kang, C. M., Kim, J., Lee, S. H., Chung, C. C., & Choi, J. W. (2017). Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *2017 IEEE 20th international conference on intelligent transportation systems* (pp. 399–404). IEEE.

Kong, Y., & Fu, Y. (2018). Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230.

Kooij, J. F. P., Schneider, N., Flohr, F., & Gavrila, D. M. (2014). Context-based pedestrian path prediction. In *European conference on computer vision* (pp. 618–633). Springer.

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2020). Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE intelligent vehicles symposium* (pp. 1688–1693). IEEE.

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2021). Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1258–1268).

Kuefler, A., Morton, J., Wheeler, T., & Kochenderfer, M. (2017). Imitating driver behavior with generative adversarial networks. In *2017 IEEE intelligent vehicles symposium* (pp. 204–211). IEEE.

Kumar, P., Perrollaz, M., Lefevre, S., & Laugier, C. (2013). Learning-based approach for online lane change intention prediction. In *2013 IEEE intelligent vehicles symposium* (pp. 797–802). IEEE.

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H., & Chandraker, M. (2017). Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 336–345).

Lee, D., Kwon, Y. P., McMains, S., & Hedrick, J. K. (2017). Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC. In *2017 IEEE 20th international conference on intelligent transportation systems* (pp. 1–6). IEEE.

Lefèvre, S., Vasquez, D., & Laugier, C. (2014). A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, *1*(1), 1–14, Publisher: SpringerOpen.

Leon, F., & Gavrilescu, M. (2019). A review of tracking, prediction and decision making methods for autonomous driving. arXiv preprint arXiv:1909.07707.

Levy, J. I., Buonocore, J. J., & Von Stackelberg, K. (2010). Evaluation of the public health impacts of traffic congestion: A health risk assessment. *Environmental Health*, *9*(1), 1–12, Publisher: Springer.

Li, Y., Wang, H., Dang, L. M., Nguyen, T. N., Han, D., Lee, A., Jang, I., & Moon, H. (2020). A deep learning-based hybrid framework for object detection and recognition in autonomous driving. *IEEE Access*, *8*, 194228–194239, Publisher: IEEE.

Li, J., Yang, F., Tomizuka, M., & Choi, C. (2020). Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. In *Proceedings of the neural information processing systems*.

Li, X., Ying, X., & Chuah, M. C. (2019a). Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. arXiv preprint arXiv:1907.07792.

Li, X., Ying, X., & Chuah, M. C. (2019b). Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE intelligent transportation systems conference* (pp. 3960–3966). IEEE.

Lian, J., Yu, F., Li, L., & Zhou, Y. (2022). Early intention prediction of pedestrians using contextual attention-based LSTM. *Multimedia Tools and Applications*, 1–17, Publisher: Springer.

Lim, Y.-C., Lee, M., Lee, C.-H., Kwon, S., & Lee, J.-h. (2010). Improvement of stereo vision-based position and velocity estimation and tracking using a stripe-based disparity estimation and inverse perspective map-based extended Kalman filter. *Optics and Lasers in Engineering*, *48*(9), 859–868, Publisher: Elsevier.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, proceedings, Part V 13* (pp. 740–755). Springer.

Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Shenoi, A., Gaidon, A., & Niebles, J. C. (2020). Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, *5*(2), 3485–3492, Publisher: IEEE.

Liu, J., Wang, H., Peng, L., Cao, Z., Yang, D., & Li, J. (2022). PNNUAD: Perception neural networks uncertainty aware decision-making for autonomous vehicle. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 24355–24368, Publisher: IEEE.

Luan, Z., Huang, Y., Zhao, W., Zou, S., & Xu, C. (2022). A comprehensive lateral motion prediction method of surrounding vehicles integrating driver intention prediction and vehicle behavior recognition. *Proceedings of the Institution of Mechanical Engineers, Part D (Journal of Automobile Engineering)*, Article 09544070221078636, Publisher: SAGE Publications Sage UK: London, England.

Ma, J., & Rong, W. (2022). Pedestrian crossing intention prediction method based on multi-feature fusion. *World Electric Vehicle Journal*, *13*(8), 158, Publisher: MDPI.

Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., & Manocha, D. (2019). Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI conference on artificial intelligence, vol. 33* (pp. 6120–6127). Issue: 01.

Mangalam, K., Girase, H., Agarwal, S., Lee, K.-H., Adeli, E., Malik, J., & Gaidon, A. (2020). It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European conference on computer vision* (pp. 759–776). Springer.

Manh, H., & Alaghband, G. (2018). Scene-LSTM: A model for human trajectory prediction. arXiv preprint arXiv:1808.04018.

Messaoud, K., Yahiaoui, I., Verroust-Blondet, A., & Nashashibi, F. (2019). Non-local social pooling for vehicle trajectory prediction. In *2019 IEEE intelligent vehicles symposium* (pp. 975–980). IEEE.

Minguez, R. Q., Alonso, I. P., Fernandez-Llorca, D., & Sotelo, M. A. (2018). Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, *20*(5), 1803–1814, Publisher: IEEE.

Mo, X., Huang, Z., Xing, Y., & Lv, C. (2022). Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. *IEEE Transactions on Intelligent Transportation Systems*, Publisher: IEEE.

Mohamed, A., Qian, K., Elhoseiny, M., & Claudel, C. (2020). Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14424–14432).

Mozaffari, S., Al-Jarrah, O. Y., Dianati, M., Jennings, P., & Mouzakitis, A. (2020). Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, Publisher: IEEE.

Naik, A. Y., Bighashdel, A., Jancura, P., & Dubbelman, G. (2022). Scene spatio-temporal graph convolutional network for pedestrian intention estimation. In *2022 IEEE intelligent vehicles symposium* (pp. 874–881). IEEE.

Neogi, S., Hoy, M., Chaoqun, W., & Dauwels, J. (2017). Context based pedestrian intention prediction using factored latent dynamic conditional random fields. In *2017 IEEE symposium series on computational intelligence* (pp. 1–8). IEEE.

Park, S. H., Kim, B., Kang, C. M., Chung, C. C., & Choi, J. W. (2018). Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. In *2018 IEEE intelligent vehicles symposium* (pp. 1672–1678). IEEE.

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., Rus, D., & Ang, M. H. (2017). Perception, planning, control, and coordination for autonomous vehicles. *Machines*, *5*(1), 6, Publisher: Multidisciplinary Digital Publishing Institute.

Petrović, D., Mijailović, R., & Pešić, D. (2020). Traffic accidents with autonomous vehicles: Type of collisions, manoeuvres and errors of conventional vehicles' drivers. *Transportation Research Procedia*, *45*, 161–168, Publisher: Elsevier.

Piccoli, F., Balakrishnan, R., Perez, M. J., Sachdeo, M., Nunez, C., Tang, M., Andreasson, K., Bjurek, K., Raj, R. D., & Davidsson, E. (2020). Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network. In *2020 54th asilomar conference on signals, systems, and computers* (pp. 68–72). IEEE.

Quan, R., Zhu, L., Wu, Y., & Yang, Y. (2021). Holistic LSTM for pedestrian trajectory prediction. *IEEE Transactions on Image Processing*, *30*, 3229–3239, Publisher: IEEE.

Ragesh, N. K., & Rajesh, R. (2019). Pedestrian detection in automotive safety: understanding state-of-the-art. *IEEE Access*, *7*, 47864–47890, Publisher: IEEE.

Raimundo, V., & Favio, M. (2021). Driver intention prediction at roundabouts. In *2021 XIX workshop on information processing and control* (pp. 1–5). IEEE.

Rasouli, A., Kotseruba, I., Kunic, T., & Tsotsos, J. K. (2019). Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6262–6271).

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2020). Pedestrian action anticipation using contextual feature fusion in stacked rnns. arXiv preprint arXiv:2005.06582.

Razali, H., Mordan, T., & Alahi, A. (2021). Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation Research Part C: Emerging Technologies*, *130*, Article 103259, Publisher: Elsevier.

Rehder, E., Wirth, F., Lauer, M., & Stiller, C. (2018). Pedestrian prediction by planning using deep neural networks. In *2018 IEEE international conference on robotics and automation* (pp. 1–5). IEEE.

Ridel, D., Rehder, E., Lauer, M., Stiller, C., & Wolf, D. (2018). A literature review on the prediction of pedestrian behavior in urban scenarios. In *2018 21st international conference on intelligent transportation systems* (pp. 3105–3112). IEEE.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). Human motion trajectory prediction: A survey. *International Journal of Robotics Research*, *39*(8), 895–935, Publisher: Sage Publications Sage UK: London, England.

Ruijters, E., & Stoelinga, M. (2015). Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. *Computer Science Review*, *15*, 29–62, Publisher: Elsevier.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252, Publisher: Springer.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., & Savarese, S. (2019). Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1349–1358).

Schneider, N., & Gavrila, D. M. (2013). Pedestrian path prediction with recursive Bayesian filters: A comparative study. In *German conference on pattern recognition* (pp. 174–183). Springer.

Schwall, M., Daniel, T., Victor, T., Favaro, F., & Hohnhold, H. (2020). Waymo public road safety performance data. arXiv preprint arXiv:2011.00038.

Sharma, N., Dhiman, C., & Indu, S. (2022). Pedestrian intention prediction for autonomous vehicles: A comprehensive survey. *Neurocomputing*, Publisher: Elsevier.

Shirazi, M. S., & Morris, B. T. (2016). Looking at intersections: A survey of intersection monitoring, behavior and safety analysis of recent studies. *IEEE Transactions on Intelligent Transportation Systems*, *18*(1), 4–24, Publisher: IEEE.

Shobha, B. S., & Deepu, R. (2018). A review on video based vehicle detection, recognition and tracking. In *2018 3rd international conference on computational systems and information technology for sustainable solutions* (pp. 183–186). IEEE.

Siegwart, R., Nourbakhsh, I. R., & Scaramuzza, D. (2011). *Introduction to autonomous mobile robots*. MIT Press.

SIMulation, G. (2007). US highway 101 dataset.

Sivaraman, S., & Trivedi, M. M. (2013). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, *14*(4), 1773–1795, Publisher: IEEE.

Su, S., Muelling, K., Dolan, J., Palanisamy, P., & Mudalige, P. (2018). Learning vehicle surrounding-aware lane-changing behavior from observed trajectories. In *2018 IEEE intelligent vehicles symposium* (pp. 1412–1417). IEEE.

Sun, J., Jiang, Q., & Lu, C. (2020). Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 660–669).

Vemula, A., Muelling, K., & Oh, J. (2018). Social attention: Modeling attention in human crowds. In *2018 IEEE international conference on robotics and automation* (pp. 4601–4607). IEEE.

Vitas, D., Tomic, M., & Burul, M. (2020). Traffic light detection in autonomous driving systems. *IEEE Consumer Electronics Magazine*, *9*(4), 90–96, Publisher: IEEE.

Wang, C., Wang, Y., Xu, M., & Crandall, D. J. (2022). Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, *7*(2), 2716–2723, Publisher: IEEE.

Waymo, W. (2020). *Waymo safety report*. Waymo, URL: https://waymo.com/safety/.

WHO, W. H. O. (2018). *Global status report on road safety 2018: Summary*: *Technical report*, World Health Organization.

Xin, L., Wang, P., Chan, C.-Y., Chen, J., Li, S. E., & Cheng, B. (2018). Intention-aware long horizon trajectory prediction of surrounding vehicles using dual LSTM networks. In *2018 21st international conference on intelligent transportation systems* (pp. 1441–1446). IEEE.

Xing, L., & Amari, S. V. (2008). Fault tree analysis. In *Handbook of performability engineering* (pp. 595–620). Publisher: Springer.

Xing, Y., Lv, C., Huaji, W., Wang, H., & Cao, D. (2017). *Recognizing driver braking intention with vehicle data using unsupervised learning methods*: *Technical report*, SAE Technical Paper.

Xu, Y., Piao, Z., & Gao, S. (2018). Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5275–5284).

Xue, H., Huynh, D. Q., & Reynolds, M. (2018). SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In *2018 IEEE winter conference on applications of computer vision* (pp. 1186–1194). IEEE.

Xue, H., Huynh, D. Q., & Reynolds, M. (2020). A location-velocity-temporal attention LSTM model for pedestrian trajectory prediction. *IEEE Access*, *8*, 44576–44589, Publisher: IEEE.

Yang, J., Sun, X., Wang, R. G., & Xue, L. X. (2022). PTPGC: Pedestrian trajectory prediction by graph attention network with ConvLSTM. *Robotics and Autonomous Systems*, *148*, Article 103931, Publisher: Elsevier.

Yang, B., Zhan, W., Wang, P., Chan, C., Cai, Y., & Wang, N. (2021). Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 5338–5349, Publisher: IEEE.

Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., & Ozguner, U. (2022). Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, *7*(2), 221–230, Publisher: IEEE.

Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., & Du, X. (2021a). Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters*, *6*(2), 1463–1470, Publisher: IEEE.

Yao, Y., Atkins, E., Roberson, M. J., Vasudevan, R., & Du, X. (2021b). Coupling intent and action for pedestrian crossing behavior prediction. arXiv preprint arXiv:2105.04133.

Yoon, S., & Kum, D. (2016). The multilayer perceptron approach to lateral motion prediction of surrounding vehicles for autonomous vehicles. In *2016 IEEE intelligent vehicles symposium* (pp. 1307–1312). IEEE.

Zeng, Z. (2022). High efficiency pedestrian crossing prediction. arXiv preprint arXiv:2204.01862.

Zhang, J. (2021). Deep understanding Tesla FSD Part 1: HydraNet. Medium, URL: https://saneryee-studio.medium.com/deep-understanding-tesla-fsd-part-1-hydranet-1b46106d57.

Zhang, S., Abdel-Aty, M., Wu, Y., & Zheng, O. (2021). Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 2331–2339, Publisher: IEEE.

Zhang, X., Angeloudis, P., & Demiris, Y. (2022). ST CrossingPose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, Publisher: IEEE.

Zhang, X., Cheng, L., Li, B., & Hu, H.-M. (2018). Too far to see? Not really!—Pedestrian detection with scale-aware localization policy. *IEEE Transactions on Image Processing*, *27*(8), 3703–3715, Publisher: IEEE.

Zhang, H., & Fu, R. (2020). A hybrid approach for turning intention prediction based on time series forecasting and deep learning. *Sensors*, *20*(17), 4887, Publisher: Multidisciplinary Digital Publishing Institute.

Zhang, P., Ouyang, W., Zhang, P., Xue, J., & Zheng, N. (2019). Sr-LSTM: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12085–12094).

Zhang, T., Song, W., Fu, M., Yang, Y., & Wang, M. (2021). Vehicle motion prediction at intersections based on the turning intention and prior trajectories model. *IEEE/CAA Journal of Automatica Sinica*, *8*(10), 1657–1666, Publisher: IEEE.

Zhang, K., Zhao, L., Dong, C., Wu, L., & Zheng, L. (2022). AI-TP: Attention-based interaction-aware trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, Publisher: IEEE.

Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wang, Y., & Wu, Y. N. (2019). Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12126–12134).

Zhou, W., Berrio, J. S., De Alvis, C., Shan, M., Worrall, S., Ward, J., & Nebot, E. (2020). Developing and testing robust autonomy: The university of sydney campus data set. *IEEE Intelligent Transportation Systems Magazine*, *12*(4), 23–40, Publisher: IEEE.

Zhu, Y., Qian, D., Ren, D., & Xia, H. (2019). Starnet: Pedestrian trajectory prediction using deep neural network in star topology. In *2019 IEEE/RSJ international conference on intelligent robots and systems* (pp. 8075–8080). IEEE.

Zyner, A., Worrall, S., & Nebot, E. M. (2019). ACFR five roundabouts dataset: Naturalistic driving at unsignalized intersections. *IEEE Intelligent Transportation Systems Magazine*, *11*(4), 8–18, Publisher: IEEE.