Thesis for The Degree of Licentiate of Engineering

Predicting Pedestrian Behavior in Urban Traffic Scenarios Using Deep Learning Methods

Chi Zhang

UNIVERSITY OF GOTHENBURG

Division of Interaction Design and Software Engineering
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden, 2022

Predicting Pedestrian Behavior in Urban Traffic Scenarios Using Deep Learning Methods

Chi Zhang

Division of Interaction Design and Software Engineering
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden

This thesis has been prepared using LaTeX.
Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2022.

Abstract

Background: The statistics on global road safety show a great demand for reducing the fatalities caused by pedestrian-vehicle collisions. By utilizing artificial intelligence such as deep learning, human drivers can be supported by better driver assistance systems, and thereby the fatalities caused by human errors can be reduced. Therefore, accurately predicting pedestrian behavior is crucial for drivers and automated vehicles to better understand pedestrians in complex scenarios to avoid pedestrian-vehicle collisions.
Objectives: This thesis aims to use deep learning to predict pedestrian behavior in urban traffic more accurately. The research goals are: 1) reviewing, categorizing, and analyzing existing research to identify research gaps in pedestrian behavior prediction, 2) developing a model that can more accurately predict pedestrian trajectories in urban traffic by using deep learning to model social interactions, and 3) considering pedestrian-vehicle interactions using deep learning methods when predicting pedestrian trajectories.
Methods: In Paper A, the methodology to find and collect existing papers is based on direct search and snowballing. The IEEE Xplore digital library and Google Scholar is used for direct search. Paper B and C have considered social interactions and pedestrian-vehicle interactions using deep learning methods when predicting pedestrian trajectories. A real-world, large-scale open dataset released by Waymo is used for training and evaluation. The average displacement error (ADE) and final displacement error (FDE) are used to quantitatively evaluate the prediction accuracy.
Results: Paper A has reviewed 92 papers, 50 from direct searching and 42 from snowballing, and analyzed the models that considered different factors influencing the pedestrian behavior. The advantages and drawbacks of using different prediction methods have been outlined. Research gaps and possible research directions have been pointed out. In Paper B, while the performance on ADE and FDE has been slightly improved by 1.50% and 1.82% compared to the state-of-the-art model, the inference speed has been significantly improved by 4.7 times faster on total inference speed and 54.8 times faster on data pre-processing speed. In Paper C, our proposed pedestrian-vehicle interaction extractor is applied to both sequential and non-sequential models. For sequential models, our model improved the ADE and FDE by 7.46% and 5.24% compared to the state-of-the-art models, and for non-sequential models, our model improved the ADE and FDE by 2.10% and 1.27%.
Conclusions: Paper A has shown that including more influencing factors in trajectory prediction has the potential to improve accuracy. Paper B and C have shown that including social interactions and pedestrian-vehicle interactions can improve the accuracy of pedestrian trajectory prediction. By reducing the predicting error and reducing the inference time, our research findings contribute to making approaches for the perception in automated vehicles and driver assistant systems safer than the current state-of-the-art.

Keywords

Automated driving, pedestrian behavior, pedestrian trajectory prediction, deep learning, social interaction, pedestrian-vehicle interaction, urban traffic

Acknowledgments

This research is funded by the European research project "SHAPE-IT – Supporting the Interaction of Humans and Automated Vehicles: Preparing for the Environment of Tomorrow". This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement 860410.

First of all, I would like to thank my supervisor Christian Berger for his continuous guidance and encouragement. His support and patience helped me throughout the research. He allowed me to explore my research topics freely and continuously supported me by sharing his technical knowledge and experience.

I would also like to thank my co-supervisor Marco Dozza for his insightful advice and feedback that has pushed me to sharpen my thinking. He always provided constructive advice to my research and brought my work to a higher level.

I would like to express my gratitude to my colleagues in the SHAPE-IT project. I would like to thank Jonas Bärgman, Jacqueline Plette, and other supervisors for organizing the training workshops and seminars where researchers can present, share, and discuss research ideas. I would like to thank other PhD students for their support and collaboration. The strong connections within SHAPE-IT make me feel that I am not alone during this journey.

I am very grateful to my colleagues at the Division of Interaction Design and Software Engineering. They have provided a great working environment and nice social activities. I also want to thank the Wallenberg AI, Autonomous Systems and Software Program (WASP) for providing excellent courses, summer schools, and events.

Finally, I want to express my gratitude to my family and friends. I want to thank my parents Guoliang Zhang and Dewen Liu for their unconditional love and support. I want to thank my husband Zhongjun Ni for his unlimited love and encouragement. I also want to thank my friends for their support whenever and wherever.

List of Publications

Appended publications

This thesis is based on the following publications:

[A] C. Zhang and C. Berger "Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review"
In submission to IEEE Transactions on Intelligent Transportation Systems.

[B] C. Zhang, C. Berger, and M. Dozza "Social-IWSTCNN: A Social Interaction-Weighted Spatio-Temporal Convolutional Neural Network for Pedestrian Trajectory Prediction in Urban Traffic Scenarios"
In proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2021.

[C] C. Zhang and C. Berger "Learning the Pedestrian-Vehicle Interaction for Pedestrian Trajectory Prediction"
In proceedings of 2022 the 8th International Conference on Control, Automation and Robotics (ICCAR). IEEE, 2022.

Other publications

The following publications are published or are currently in submission, but not appended to this thesis.

[a] C. Zhang and C. Berger "Analyzing Factors Influencing Pedestrian Behavior in Urban Traffic Scenarios Using Deep Learning"
Abstract accepted at Transport Research Arena (TRA) conference, 2022.

[b] C. Zhang, C. Berger, and M. Dozza "Towards Understanding Pedestrian Behavior Patterns from LiDAR Data"
Abstract and poster at the 32nd annual workshop of the Swedish Artificial Intelligence Society (SAIS), 2020.

Research Contributions

My contributions in Paper A are that: I collected and reviewed existing research studies on pedestrian behavior prediction that used deep learning methods. Compared to existing review papers, I extended the taxonomy into three criteria for classifying existing studies to provide perspectives from multiple dimensions instead of a single criterion, and I analyzed and discussed both trajectory and intention prediction over the past five years instead of only analyzing a single prediction task. I established the overall framework of the pedestrian behavior prediction, and addressed the progress and development of state-of-the-art algorithms on pedestrian behavior prediction. I analyzed and discussed the advantages and drawbacks of different methods. I reviewed and introduced commonly used datasets, and compared existing methods by evaluation metrics that are well accepted in the field. I identified challenges and research gaps of existing works that we should focus on in future work. When it comes to paper writing, I structured and wrote the majority of the paper.

My contributions in Paper B are that: I proposed a deep learning method to extract social interaction features for pedestrian trajectory prediction. I designed the network structure, built the model, and conducted the model training and evaluation. Compared with previous state-of-the-art models that use hand-crafted social interaction weights, I designed a deep learning sub-network namely the Social Interaction Extractor to learn the social interaction weights between pedestrians. I selected the urban traffic scenarios from the Waymo Open Dataset, and prepared and pre-processed the data for training and evaluation. I conducted the performance evaluation, compared the proposed model with the state-of-the-art baselines, and analyzed the results quantitatively and qualitatively. When it comes to paper writing, I structured and wrote the majority of the paper.

My contribution in Paper C are that: I proposed a deep learning method to extract the pedestrian-vehicle interaction features when predicting the pedestrian trajectory in urban traffic scenarios. I designed the network structure, built the model, and conducted the model training and evaluation. Compared with previous state-of-the-art models that did not consider pedestrian-vehicle interactions in prediction, I designed a deep learning sub-network namely the Pedestrian-Vehicle Interaction Extractor to learn the pedestrian-vehicle interaction features. I prepared and pre-processed the data for training and evaluation. I conducted the performance evaluation, compared the proposed model with the state-of-the-art baselines for both sequential models and non-sequential models. I analyzed the results quantitatively and qualitatively. When it comes to paper writing, I structured and wrote the majority of the paper.

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 The Evolution of Road Safety over the Years

According to the global status report on road safety by World Health Organization (WHO) in 2018 [1], fatalities of road traffic crashes have increased to 1.35 million annually, which is unacceptably high. Besides, road traffic injuries are the main cause of death for the young aged between 5 and 29 years [1]. As shown in Figure 1.1, although the fatality rate per 100,000 population has declined slightly from 18.8 to 18.2 from the year 2000 to 2016, the absolute death number has increased from 1.15 million to 1.35 million.

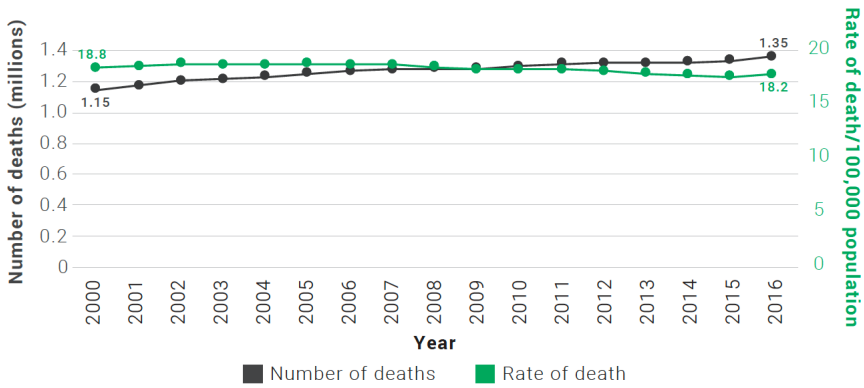


Figure 1.1: Number and rate of road traffic death per 100,000 population: 2000–2016 (cf. WHO's report 2018 [1]).

From 2000 to 2016, the number of motor vehicles has rapidly increased from 0.85 billion to 2.1 billion, which means we need to put more effort into reducing death rates to compensate for the proliferation of motor vehicles.

These numbers suggest that we need to put more effort to reduce the road traffic death number as proposed in the sustainable development goals (SDG) target 3.6 in the 2030 agenda for sustainable development [2]. Therefore, there

is an increasing need for safer vehicles to prevent hazardous situations and reduce fatalities.

Rumar [3] has stated that human errors are one of the main factors in most road accidents. As human error is a major source for road accidents being the top risk factor according to WHO [1], reducing the role human drivers take on public roads for operating vehicles is potentially addressing this risk by turning to vehicular automation.

Vehicular automation is using technologies to assist vehicle operation and make the vehicle intelligent. By involving artificial intelligence (AI), dynamic control, and other technologies, we can develop automated driving (AD) system and advanced driver assistant system (ADAS). An AD system is a system that is capable of sensing the car's driving environment and moving safely with little or no human input by incorporating automatics. Another type of using automated vehicle features to aid driving is ADAS, where such features are, for instance, used to assist drivers in driving and parking.

While it is still debated whether automated vehicles can provide a safer road [4], we can get us prepared for the more densely populated road in the future by developing the AD system and ADAS technologies.

1.1.2 Pedestrian Safety in Different Regions

The WHO's report in 2018 [1] shows that every year over 310,000 pedestrians lose their lives because of road crashes. This number has increased by 13.6% from 273,000 in 2010 [5], and constitutes 23% of all road deaths globally.

In Europe, the number is even higher, that 27% of all road deaths are pedestrians as shown in Figure 1.2. In International Transport Forum's (ITF's) report [6] on urban road safety, it is shown that although the death rate of road crashes has been reducing since 2010, the reductions for pedestrians were slower.

In the WHO's report on pedestrian safety [7], it is shown that pedestrian fatalities constitute a larger proportion of all road traffic deaths in low-income countries (LIC) with 36% compared with high-income countries (HIC) with 18% in 2010, as shown in Table 1.1.

World	Car occupants	Motorized 2-3 wheelers	Cyclists	Pedestrians	Other
LIC	31	15	6	**36**	12
MIC	27	25	4	**22**	22
HIC	56	16	5	**18**	5
All	31	23	5	**22**	19

Table 1.1: Road user fatalities as a **proportion (%)** of global road traffic deaths, 2010. The gross income per capita in 2010 used to categorize countries into: LIC (low-income countries) = US $1005 or less; MIC (middle-income countries) = US $1006 to 12,275; and HIC (high-income countries) = US $12,276 or more (cf. WHO's report on pedestrian safety [7]).

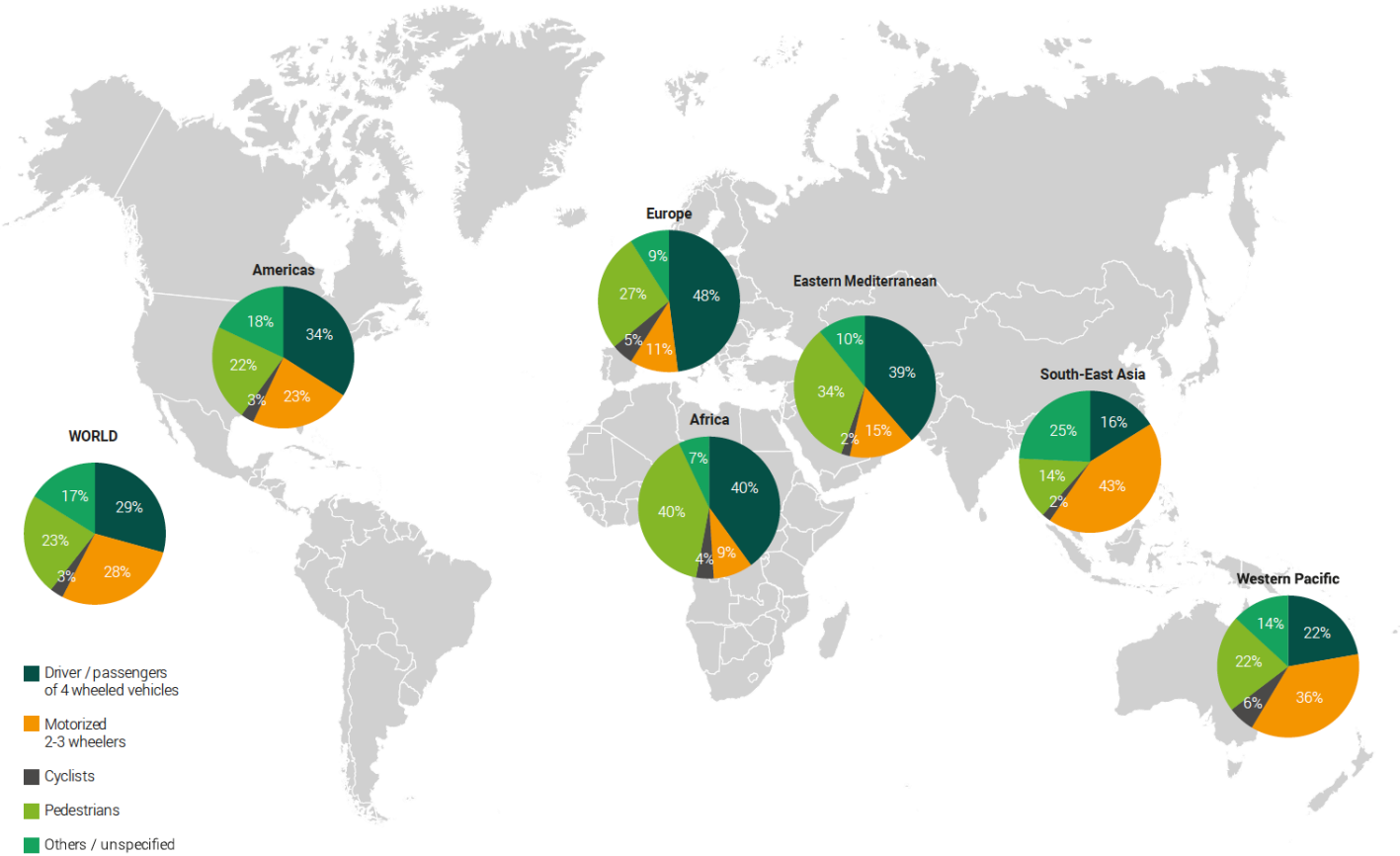


Figure 1.2: Distribution of deaths by road user type by WHO Region (cf. WHO's report [1]).

Sweden has a world-leading performance in road safety with 2.8 death per 100,000 inhabitants, and the rate is constantly decreasing [1]. Between 2006 to 2019, the absolute road traffic fatalities has decreased from 445 to 221 [8], and the absolute non-fatal road traffic injuries reduced from 26,636 to 17,719, as shown in Figure 1.3. However, the number of injuries is still high. The injuries caused by pedestrian-vehicle collisions have psychological, health, and economic costs to both individuals and society [7]. The traffic crashes in Sweden have caused significant costs to society at approximately 13.4 billion EUR, constituting 2.6% of national GDP in 2017 [8], as shown in Table 1.2. This number involves all road users, including pedestrians. In Sweden, pedestrians constitute 12% of all road traffic deaths, which is a large part that we need to consider.

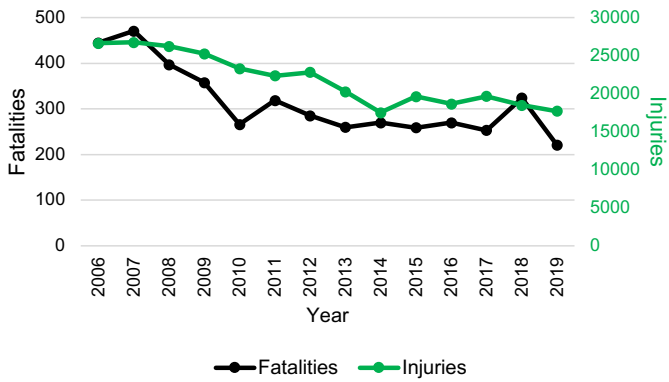


Figure 1.3: Trends in road fatalities and injuries in Sweden, 2006-2019.

	Costs (EUR)
Fatalities	1.32 billion
Other reported injuries	12.07 billion
Total (as % of GDP)	13.39 billion (2.6%)

Table 1.2: Costs of road crashes in Sweden, 2017 (cf. ITF's report [8]).

According to Peden et al. (cf. [9]), pedestrian-vehicle collisions are predictable and preventable. Therefore, the number of fatal and non-fatal injuries could be reduced if pedestrian-vehicle collisions could be better predicted to provide a chance for prevention. As reported by WHO [1], vehicles can be designed and built to better protect pedestrians.

Improving the pedestrian trajectory and intention prediction may have the potential to equip safer vehicles to protect the pedestrians. A precise and robust prediction of pedestrian behavior can reduce the misunderstanding of a pedestrian's intention and provide more reaction time to a pedestrian's unexpected movement and thereby preventing hazardous situations.

The information on pedestrian behavior can be used for the AD systems and ADAS to make better and safer decisions. For example, the prediction of pedestrian behavior can provide complementary information to the driver and

reduce the risk for pedestrian-vehicle collision during night-time travel, which is one of the key risk factors for pedestrians [7].

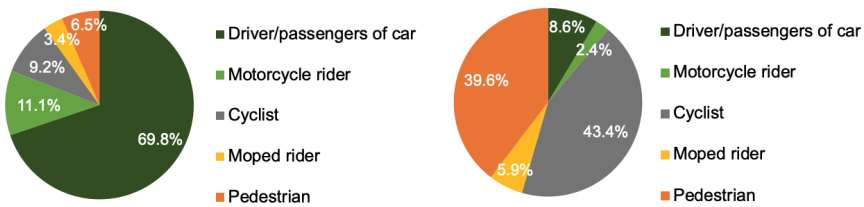
However, the prediction of pedestrian behavior is very challenging. The agility of pedestrians shows hardly predictable moving patterns [10], as pedestrians can change their speed and direction abruptly [11]. Furthermore, complicated factors such as the destination, age, and gender of a pedestrian [12] influence pedestrian behavior. In addition to the agility, pedestrians also tend to interact with other pedestrians [13] and vehicles [14–16] all the time, which makes it harder to precisely predict their behaviors.

It is crucial to study and predict pedestrian behavior to help prevent pedestrian-vehicle collisions on the road. Recently, the need for driving safety and automated driving has stimulated an increasing number of research studies on pedestrian behavior prediction in both industry and academia as presented in Chapter **??**.

1.1.3 Locations of Pedestrian-Vehicle Collisions

When we consider the locations where pedestrian-vehicle collisions occur, the results vary in different countries and regions. As stated by Do et. al [17], most pedestrian-vehicle collisions are likely to occur when pedestrians are crossing the road. According to WHO's report on pedestrian safety [7], in high-income countries, pedestrian-vehicle collisions occur more in urban areas than rural areas, while in some low- and middle-income countries the opposite is true. In the European Union, about 70% of pedestrian fatalities occur in urban areas [7]. The situation is similar in the United States, where 76% of total pedestrian deaths are in urban areas [7].

Research by Värnild et al. [18] in Sweden between 2003 to 2014 showed that the distribution of road user types who were seriously injured in road accidents varied between rural and urban areas. In rural areas, the pedestrians constituted 6.5%, while in urban areas, the pedestrians constituted 39.6%, which accounted for a large proportion of all serious injuries, as shown in Figure 1.4.



(a) Rural area, n=262 (b) Urban area, n=371

Figure 1.4: Serious injuries in Sweden (Region Västmanland) in rural and urban areas 2003-2014, N=633, with 262 in rural areas and 371 in urban areas (data from Värnild et al.'s report [18]).

In urban scenarios, the ITF collected traffic safety data in 48 cities from different continents to monitor the progress in urban road safety. In their

report [6], the distribution of road user types that were fatally injured varied in different densities of the city. As shown in Figure 1.5, in a more densely populated city, the proportion of pedestrians in road fatalities was higher. In high-density cities where there are more than 10,000 inhabitants per square kilometer, there are 51% of pedestrians were fatally injured in all types of road users. It was shown that the more densely a city is populated, the more dangerous it is for pedestrians.

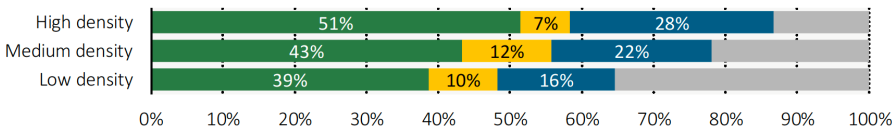


Figure 1.5: Distributions of road fatalities of road user types in different densities of the city, using the average values of figures available between 2014 and 2018. The low population density is less than 5,000 inhabitants per square kilometer. The medium density is less than 10,000, and the high density is 10,000 and above (cf. ITF's report [6]).

1.2 Motivation and Problem Domain

As highlighted in Section 1.1.3, pedestrians are more vulnerable compared to other road users. The pedestrian fatalities proportion in low-income countries is especially high, with 36% out of all fatalities compared to 18% in high-income countries. The statistics on global road safety show a great demand for reducing the death rate of road traffic crashes, which reveals the significance of developing safer vehicles to prevent crashes.

The demands for substantial reductions in the number of fatal and serious injuries for road traffic require us to develop AD technologies to ensure driving safety. The prediction of pedestrian behavior is essential for AD systems. By involving AI technologies, the human driver's operational load is reduced, and thereby the fatalities caused by human errors can be reduced. Accurately predicting pedestrian behavior can help automated vehicles better understand the pedestrians when interacting with vehicles in complex scenarios and to make safer decisions.

As outlined in Section 1.1.2, in Sweden and the European Union countries, serious injuries caused by pedestrian-vehicle collisions are more likely to happen in urban areas, and the proportion of road fatalities for pedestrians is higher in more densely populated cities. Therefore, in our research, we mainly focus on the data in urban scenarios to prevent hazardous situations.

The behavior of a pedestrian is influenced by the interactions with the other road users as outlined in Section 1.1.2. In this thesis, social interactions and pedestrian-vehicle interactions are considered and studied while predicting pedestrian behavior. The social interaction is the process of reciprocal influence of two or more individuals who modify their actions and reactions during social encounters [19]. In the context of pedestrian behavior prediction, social interaction can be interpreted as the influence on pedestrian behavior when

they interact with other pedestrians. The pedestrian-vehicle interaction refers to the impact of the interactions between pedestrians and vehicles on pedestrian behavior.

Due to the complexity and intricacy of pedestrian behavior, knowledge-based methods such as the rule-based models and statistics-based models can hardly predict pedestrian behavior precisely and reliably [20]. The non-linear behavior arising from interactions of the pedestrians are hard for the rule-based models to learn. Deep learning methods are strong tools that can handle complex scenarios. The non-linear activation functions of the deep learning network can learn the non-linearity of the model. The large number of learned parameters of the network can avoid data saturation, and benefit from large-scale datasets. Therefore, we explore deep learning methods to learn the patterns of pedestrian behavior in a data-driven manner.

We focus on pedestrian trajectory prediction specifically for urban scenarios because these areas are the ones where pedestrians are mostly affected by traffic accidents as shown in Section 1.1.3.

1.3 Research Goal and Questions

The overall goal of this PhD project is to use AI to better understand and predict how pedestrians behave when interacting with vehicles and automated vehicles in urban traffic. To contribute to this research goal, we need to develop deep learning methods for predicting the behaviors of pedestrians in interactions with other road users. AI tools, especially deep learning methods are applied to the large-scale real-world datasets in our research. The developed deep learning models are used in complex interactions and are compared with results from literature and other models.

In this licentiate thesis, we focus on predicting pedestrian trajectories. To achieve this goal, the following sub-goals are addressed:

G1: Reviewing, categorizing, analyzing, and discussing currently existing research to point out research gaps for the problem area of pedestrian behavior prediction.

G2: Developing the approach that can better predict pedestrian trajectories in urban traffic scenarios, and using deep learning to model social interactions between pedestrians.

G3: Considering pedestrian-vehicle interactions, and using deep learning to model the interactions between pedestrians and vehicles when predicting pedestrian trajectories.

We derive the following research questions from the corresponding goals, expressed as follows:

G1: Literature review on pedestrian behavior prediction:

RQ1-1: What are the state-of-the-art deep learning algorithms for predicting pedestrian behavior and how they performed in urban scenarios?

RQ1-2: What are the challenges and research gaps of existing works to improve the prediction performance?

G2: Pedestrian trajectory prediction considering social interactions using deep learning methods:

RQ2-1: How to use deep learning methods to model social interactions when predicting pedestrian trajectories in urban traffic scenarios?

RQ2-2: What are the improvements of using deep learning methods compared to existing methods?

G3: Pedestrian trajectory prediction considering pedestrian-vehicle interactions using deep learning methods:

RQ3-1: How to use deep learning methods to model pedestrian-vehicle interactions when predicting pedestrian trajectories in urban traffic scenarios?

RQ3-2: What are the improvements of considering pedestrian-vehicle interactions compared to existing methods?

1.4 Methodology

The methodology section is organized as: firstly going through the overall research philosophy and explaining how the appended papers are related, and then introducing the definition of the problem and how the methods are evaluated. After that, the methods that are used in this thesis are presented, including the deep learning methods for prediction, and the interactions of pedestrians. Finally, the dataset and data pre-processing method used in this thesis are introduced.

1.4.1 Research Methodology

1.4.1.1 Research Types

The corresponding research goals, research questions, and research types used in this study are highlighted in Figure 1.6. The appended Paper A is a literature review paper, in which we answered research questions RQ1-1 and RQ1-2. Paper B and C are approaches for pedestrian trajectory prediction, that answered research questions RQ2-1, RQ2-2, and RQ3-1, RQ3-2, respectively.

Kothari [21] described basic types of research as follows:

- Descriptive vs. Analytical. The purpose of descriptive research is to describe the existing state of affairs. For example, surveys are descriptive studies. Analytical research, in contrast, is used to analyze existing information to make a critical evaluation of the material. The appended Paper A (cf. Chapter **??**) is a review paper, which is a descriptive study, while Paper B (cf. Chapter **??**) and Paper C (cf. Chapter **??**) are analytical studies with analysis and evaluation.

- Applied vs. Fundamental. The purpose of applied research is to find a solution for an industrial or business problem, while fundamental research mainly deals with the formulation of a theory. All three appended papers are focusing on applying deep learning methods to the pedestrian prediction task, so they are applied studies.

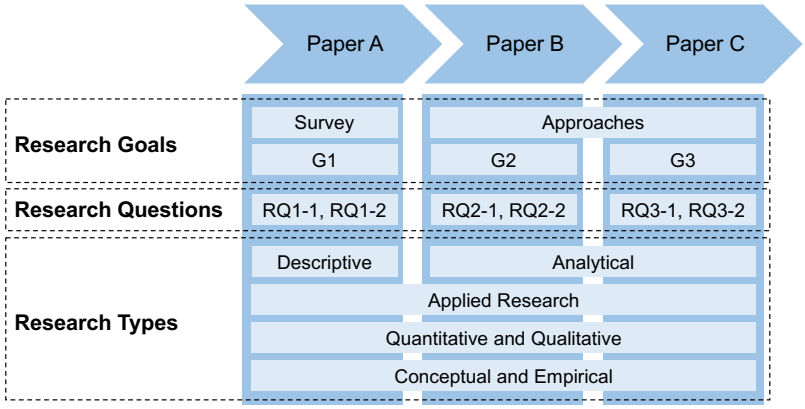


Figure 1.6: Research goals, research questions, and research types [21] of the appended papers in this thesis.

- Quantitative vs. Qualitative. Quantitative research is based on quantity measurement, while qualitative research is concerned with qualitative phenomena. In all three appended papers, we compare existing works (Paper A) or evaluate and analyze our proposed methods (Paper B and C) both quantitatively and qualitatively.
- Conceptual vs. Empirical. Conceptual research is related to abstract theory and ideas, while empirical research primarily relies on experience or observation. Our work is based on deep learning theory, which has a conceptual background. When it comes to data-driven methods, many hyper-parameters for network training are set empirically and the model is learned from data, so our work is also empiric. Therefore, all three appended papers are both conceptual and empirical studies.

The research methodology is a way to systematically solve a research problem [21]. Taking the licentiate thesis as a whole, the research process is as follows, and corresponding appended papers can be summarized as shown in Figure 1.7.

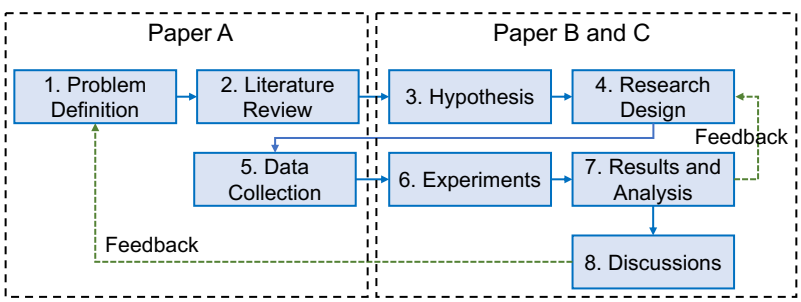


Figure 1.7: Research process [21] used in this thesis.

1. Define the research problem.
2. Review the literature, including the concepts and theories, and previous research findings.
3. Formulate hypotheses. After reviewing the literature, the hypotheses are developed.
4. Design the research. This thesis is based on experimentation, therefore the research design needs to be carefully prepared.
5. Collect data. There are many existing publicly available datasets for trajectory prediction. The dataset used for training and evaluation plays an important role because this thesis uses deep learning methods that learn the pattern from large amounts of data. Therefore, finding appropriate datasets that contain large-scale real-world urban scenes is essential for facilitating intelligent vehicles and automated driving in urban traffic scenarios.
6. Conduct experiments.
7. Analyze the result. This thesis includes both quantitative and qualitative analysis. This is used as feedback to modify the design of the research in step 4.
8. Interpret and discuss the results. The findings are used as feedback to address the research problem in greater depth in subsequent studies in step 1.

In Paper A, we define the problem and review existing works. Besides, we explore the properties of existing datasets and select appropriate datasets for training and testing. In Paper B and C, we go through from step 3 to step 8.

1.4.1.2 Research Methodology for Machine Learning

When it comes to the field of machine learning, research methods can significantly influence the accuracy and reliability of the results. Kamiri and Mariga [22] analyzed 100 papers published since 2019 in IEEE journals in machine learning and revealed that *quantitative research approaches* with *experimental research design* are mostly used. They stated that the research on machine learning is mainly quantitative because it requires the modeling of data that should make sense of the data. To understand the motivation of pedestrian behavior, this thesis also includes qualitative analysis by analyzing the moving pattern in different scenarios.

Quantitative research approaches rely on mathematical, numerical analysis, or other computational techniques applied to collected data [23]. In Paper B and C, we apply deep learning methods based on computational functions on publicly available data and quantitatively evaluate the results.

The general *experimental research design* for machine learning and deep learning is summarized by Kamiri and Mariga [22] and Sarker [24] and shown in Figure 1.8, and Paper B and C follow this design process:

- Data collection,

- Data understanding and pre-processing,
- Model building and training,
- Model testing and validation,
- Model evaluation and interpretation.

Figure 1.8: The general experimental research design process used in this thesis.

1.4.2 Problem Definition and Evaluation Metrics

1.4.2.1 Problem Definition

The prediction of pedestrian behavior includes the trajectory prediction and the intention prediction. This thesis mainly focuses on the low-level information, i.e., the trajectory prediction.

As shown in Figure 1.9, the trajectory of a pedestrian or a vehicle is defined as a sequence of x-y coordinate positions including their temporal order. The positions of pedestrians and vehicles in each frame are first pre-processed to x and y coordinates on a 2D map representation in bird's-eye-view. In a frame at time-step t with the number of pedestrians n_p and the number of vehicles n_v, the i^{th} person at time-step t is represented by x-y-coordinate $X_t^i = (x_t^i, y_t^i)$, where $i \in \{1, \ldots, n_p\}$. The j^{th} vehicle at time-step t is represented by x-y-coordinate $V_t^j = (x_t^j, y_t^j)$, where $j \in \{1, \ldots, n_v\}$. The observed pedestrians and vehicles can be denoted as $X_t = [X_t^1, X_t^2, \ldots, X_t^{n_p}]$, $V_t = [V_t^1, V_t^2, \ldots, V_t^{n_v}]$, with all observed time-steps $1 \leq t \leq T_{obs}$.

Given the observed pedestrians and vehicles, we aim to predict the most likely trajectories of pedestrians $\hat{Y}_t = [\hat{Y}_t^1, \hat{Y}_t^2, \ldots, \hat{Y}_t^{n_p}]$ in the future time-steps $T_{obs} + 1 \leq t \leq T_{pred}$. The ground truth of the future trajectories is denoted as $Y_t = [Y_t^1, Y_t^2, \ldots, Y_t^{n_p}]$, where $T_{obs} + 1 \leq t \leq T_{pred}$.

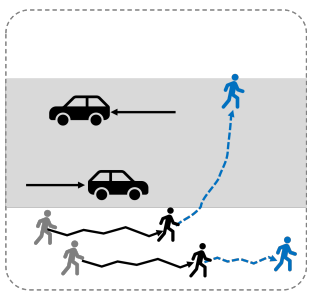


Figure 1.9: A birds-eye-view perspective illustration of pedestrian trajectory prediction. The solid lines are observed trajectories, and the dotted lines are predicted trajectories.

The predicted positions of pedestrians $\hat{Y}_t^i = (\hat{x}_t^i, \hat{y}_t^i)$ are treated as random variables, where $i \in \{1, ..., n_p\}$, $T_{obs} + 1 \leq t \leq T_{pred}$, by assuming that the i^{th} pedestrian's position at time t follows bi-variate Gaussian distribution $\hat{Y}_t^i \sim \mathcal{N}(\mu_t^i, \sigma_t^i, \rho_t^i)$. At time-step t, the mean value of the position is $\mu_t^i = (\mu_x, \mu_y)_t^i$. The standard deviation is $\sigma_t^i = (\sigma_x, \sigma_y)_t^i$, and the correlation coefficient is ρ_t^i. Our network predicts the Gaussian distribution parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)_t^i$, and samples from the distribution to get predicted trajectories.

1.4.2.2 Evaluation Metrics

The following two metrics are commonly used by researchers to report the prediction error and to evaluate the performance of the algorithms:

- The Average Displacement Error (ADE): the average distance between ground truth and prediction trajectories over all predicted time-steps, as defined in Eq. 1.1:

$$ADE = \frac{\sum_{i \in n_p} \sum_{t=T_{obs}+1}^{T_{pred}} \|Y_t^i - \hat{Y}_t^i\|_2}{n_p \times (T_{pred} - T_{obs})} \tag{1.1}$$

- The Final Displacement Error (FDE): the average distance between ground truth and prediction trajectories for the final predicted time-step, as defined in Eq. 1.2:

$$FDE = \frac{\sum_{i \in n_p} \|Y_t^i - \hat{Y}_t^i\|_2}{n_p}, t = T_{pred} \tag{1.2}$$

In Paper A (cf. Chapter **??**), existing algorithms are reviewed and compared, and the state-of-the-art algorithms are listed. In Paper B (cf. Chapter **??**) and Paper C (cf. Chapter **??**), we use the aforementioned two evaluation metrics and compare our proposed models with the state-of-the-art algorithms. In addition to ADE and FDE that evaluate the accuracy, the inference speed of different models is evaluated to compare the computational performance.

1.4.3 Deep Learning Methods for Prediction

Deep learning methods are powerful tools that enable a system to behave intelligently and are sometimes interchangeably used with machine learning, and AI. As illustrated by Sarker [24], deep learning is a branch of machine learning and AI, as shown in Figure 1.10.

Compared with other machine learning methods, deep learning models can benefit more from large-scale datasets, as shown in Figure 1.11 proposed by Sarker [24].

Much work has focused on leveraging deep learning and neural networks for behavior prediction in recent years. Many behavior prediction models use sequential methods such as the recurrent neural networks (RNNs), including their variant long short-term memory (LSTM) networks and gate recurrent units (GRUs), and LSTM-based generative adversarial networks (GANs). Recently, transformer (TF) networks have been used for pedestrian trajectory prediction.

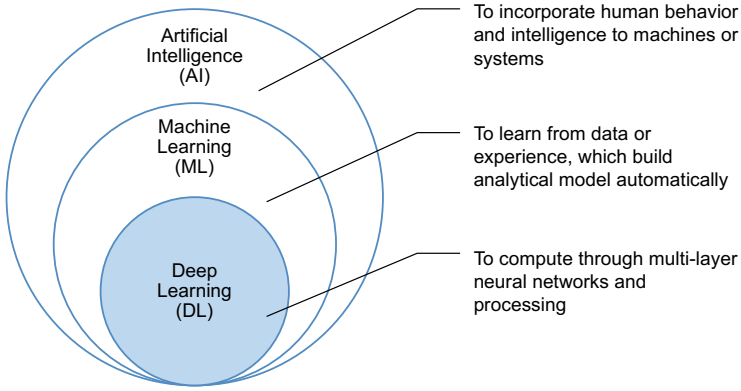


Figure 1.10: The relationship of deep learning, machine learning, and artificial intelligence (cf. Sarker [24]).

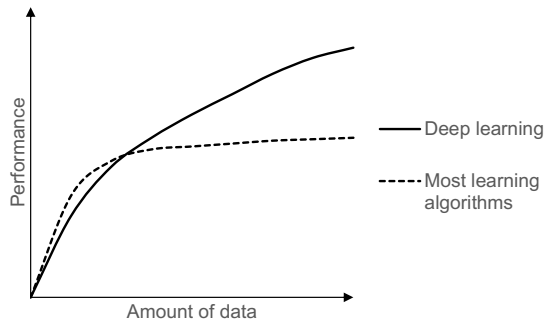


Figure 1.11: The comparison between deep learning and other machine learning algorithms. The performance of the deep learning model can increase with a larger amount of data (cf. Sarker [24]).

In addition to sequential methods, non-sequential methods such as convolutional neural networks (CNNs) are also employed to extract spatial and behavioral features. In the following, the most commonly used methods are described.

LSTM-based methods LSTM networks are the improved version of RNNs. LSTMs have both feedforward and feedback connections to capture long and short-term information, and are especially suitable for predictions based on sequential data. LSTM networks have been successfully used for sequential prediction tasks such as handwriting recognition [25] and speech recognition [26]. Due to the ability of LSTMs on sequential prediction tasks, they have been adopted by researchers to predict pedestrian trajectories (e.g., [27–30]). Alahi et al. [27] proposed Social-LSTM, which assumed the trajectories of pedestrians follow the bi-variate Gaussian distribution, and many researchers followed this uni-modal assumption. The drawback of the LSTM-based methods is that they cannot be parallized because the prediction on each time-step is dependent on the prediction of preceding time-steps.

GAN-based methods GANs are proposed by Goodfellow et al. [31], that use two neural networks called generator and discriminator to contest with each other. The generative network generates multiple possible candidates, and the discriminative network evaluates them. For predicting pedestrian trajectories, Gupta et al. [32] stated that by using the uni-modal distribution assumption, the researchers may learn the "average" trajectories instead of multiple "good behaviors". They assumed that the trajectories of pedestrians follow multi-modal distribution, and used GANs to predict the trajectories. Although GANs can predict multiple plausible results, they have a non-neglectable disadvantage: as the GANs need to train two deep networks in one structure, they are hard to train and require techniques for convergence.

CNN-based methods CNNs are widely and successfully used in image processing tasks such as image classification, image segmentation, because their capability of extracting spatial features. Many researchers utilized CNNs on pedestrian intention prediction, to extract the appearance and behavioral features implicitly.

In addition to the spatial feature, the convolutional networks on temporal space, also known as temporal convolutional networks (TCNs), can also be used for extracting the temporal feature and predicting trajectories. Bai et al. [33] claimed that the inefficient parameters used in RNNs can make the training expensive. Besides, the predictions for later time-steps depend on the predictions from preceding time-steps in LSTMs can cause the error accumulation in the prediction. Nikhil and Morris [34] utilized CNNs for predicting trajectories, that reached competitive results with a faster inference speed. Mohamed et al. [35] proposed the Social-STGCNN method that combined spatial and temporal (ST) features using TCNs and CNNs with graph structures, which is able to reach 20% improvement on FDE and is 48 times faster compared to sequential models. The appended Paper B (cf. Chapter **??**) proposed the "Social-IWSTCNN" model as one core contribution of this thesis, followed this trend of using spatio-temporal (ST) features and applying CNNs to predict the future sequences, and improved the accuracy and inference speed by 4.7 times by learning the social interaction weights (IW) with a learning-based sub-network. The CNN-based methods allow parallel computation and avoid accumulate errors, but as they predict the future trajectories using a fixed time horizon, they are not as flexible as LSTMs.

1.4.4 Interactions between Pedestrians and Other Road Users

Pedestrians in urban traffic always interact with others, including other pedestrians and vehicles. To study research questions RQ2-1, RQ2-2, RQ3-1 and RQ3-2, we use deep learning networks to learn the interaction between pedestrians and other road users, and study how the interaction influences the prediction, and what features can be used as inputs for extracting the interaction. In appended Paper B (cf. Chapter **??**), we learn social interactions with other pedestrians. In appended Paper C, (cf. Chapter **??**) we learn vehicle-pedestrian interactions.

Social interactions with other pedestrians Moussaid et al. [13] stated that pedestrian behavior is not only determined by themselves, but also influenced by social interactions with other pedestrians nearby. The modeling of the social interaction is one of the most concerning topics for pedestrian trajectory prediction lately. Alahi et al. [27] firstly introduced deep learning networks into trajectory prediction, that used LSTMs for prediction. They proposed the social pooling layer over the hidden states of each pedestrian to represent the social interaction instead of using hand-crafted knowledge-based function as in Social Force model [36]. Pooling layers are usually used in CNNs, that can combine small clusters and reduce dimensions of data. Pooling layers usually calculate the average or maximum within the pooling operation area. "Social" pooling [27] used the pooling operation to allow sharing of information with neighbors. Many researchers followed the trend of using pooling layers and improved the social pooling module with more complicated structures [32, 37].

Some other researchers proposed that social interactions are not symmetric as in pooling methods, therefore, they learn social interactions with graph-based networks [35, 38]. Graph neural networks (GNNs) construct a graph $< V, E >$ that uses vertices to represent the states of each road user, and use the edges to represent the interaction relationship between road users. STGAT [39] and Social-BiGAT [38] used the graph attention networks (GAT) proposed by Veličković et al. [40]. Mohamed et al. [35] used graph convolutional networks (GCN) [41] that implicitly assign the interaction weights of the target pedestrians' surrounding neighbors to model social interactions.

However, the construct of GNNs and the non-linear edge value computation are time consuming [42]. Multi-layer perceptrons (MLPs) have the capability to learn the interaction relationship with linear computation and activation function with a faster speed. In appended Paper B, we use MLPs to learn the interaction weights, and use weighted sum as aggregation function to calculate the influence of neighbouring pedestrians to avoid graph convolutional operation to accelerate the computing. In this way, we reduce the prediction error by 1.8% and speed up the inference by 4.7 times.

Interactions between pedestrians and vehicles In addition to the social interaction within crowds, another important factor that influences pedestrian behavior is the interaction between pedestrians and vehicles. Researchers tried to include the vehicle information while predicting pedestrian behavior. Many researchers used explicitly hand-crafted features such as speed, orientation, and distance to the pedestrians or time to collision (TTC) as input to feed into the networks to learn their influence on the pedestrians [12, 43–47]. However, the pedestrian-vehicle interaction can be complicated when there are more than one pedestrian and one vehicle, and the designed features are hard to be generalized to new scenarios. Recently there is an increasing number of works that use deep learning sub-networks to learn the interactions.

As pedestrians and vehicles usually have different motion patterns, the interactions between pedestrians and vehicles are asymmetric. Therefore, GNNs are used for learning the asymmetric interaction relationships between pedestrians and vehicles or other road users, as used by Ma et al. [48], Liu et al. [49], Eiffert et al. [50], Hu et al. [51], and Carrasco et al. [52]. The models

proposed by Chandra et al. [53–55] simultaneously predicted different types of road users, but they focused mainly on vehicles rather than pedestrians.

Considering the drawbacks of GNNs as mentioned previously, the MLPs can be used to learn the pedestrian-vehicle interaction relationship. The inputs of this sub-network can be the relative positions and relative velocities between pedestrians and vehicles. In the appended Paper C, we use a seperate MLP network to learn the pedestrian-vehicle interaction in addition to social interaction, and investigate which input performs better for learning, and how the pedestrian-vehicle interaction influences the accuracy.

1.4.5 Dataset and Data Pre-processing

Most existing research on pedestrian trajectory prediction used ETH [56] and UCY [57] datasets for training and evaluation. These two datasets contain five scenes collected in crowded urban scenarios from bird's-eye-view. There are two scenes in the ETH dataset with 750 unique pedestrians annotated, and three scenes in the UCY dataset with 786 unique pedestrians annotated. A snapshot of a scenario in ETH dataset is shown as in Figure 1.12.

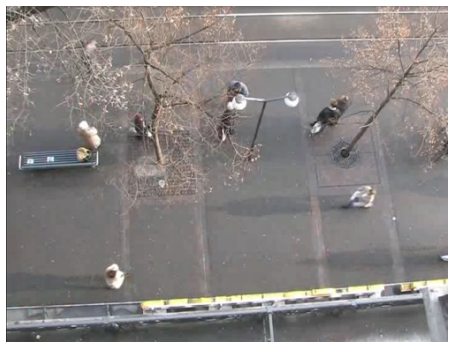


Figure 1.12: A snapshot of the hotel scenario in ETH [56] dataset.

However, these two datasets are not collected for traffic scenarios, and do not include the interaction with other road users such as vehicles. As we mentioned in Section 1.1, there are more serious injuries caused by pedestrian-vehicle collisions in urban areas than in rural areas in European countries and in the US. To study research questions RQ2-1 and RQ3-1 that particularly highlighted the prediction of pedestrian trajectories in urban traffic scenarios, we use the Waymo Open Dataset [58] that includes pedestrians and other road users collected in real traffic in appended Paper B and C. The Waymo Open Dataset is collected in urban traffic scenarios in the US from the vehicle's view, including 374 records used as training scenarios and 76 records used as test and evaluation scenarios. A snapshot of the urban traffic scenario in the Waymo Open Dataset is shown as in Figure 1.13.

The frequency of the record sequences in Waymo Open Dataset [58] is 10 Hz. To compare our algorithm with existing state-of-the-art models that are evaluated on ETH and UCY datasets sampled at 2.5 Hz, we keep the same settings and downsample the Waymo Open Dataset to 2.5 Hz. We use the pedestrians and vehicles labeled on LiDAR data with their real-world center

Figure 1.13: A snapshot of an urban traffic scenario in Waymo Open Dataset [58].

position (x, y, z), and we pre-process it into 2D position (x, y) sequences from a bird's-eye-view. We use all labeled pedestrians and vehicles in the LiDAR scan range which is 75 m. Each sequence of objects has its unique track id. The pedestrians and vehicles are taken as points without size information in this thesis.

The coordinates The previously commonly used datasets ETH [56] and UCY [57] are recorded from the bird's-eye-view, and use the global coordinate system. The Waymo Open Dataset [58] is recorded from the vehicle's view, and uses the local coordinate system with the ego-vehicle center as the origin in each frame. However, using the local coordinates will introduce the movement of ego-vehicle into the pedestrians' movement, which will affect the accuracy of prediction. To avoid the influence of the ego-vehicle, we have pre-processed the coordinates and transform them to global coordinates with the ego-vehicle's position of the first time-step of in the recording as the origin.

The sequences During training, validation, and evaluation, the sequences have been cut into pieces with a fixed sequence length, which equals to the sum of the observation length and prediction length. To aggregate the data amount, there are overlaps between each sequence piece, and the skip length is set to one time-step. Figure 1.14 shows how the training and evaluation sequences are cut.

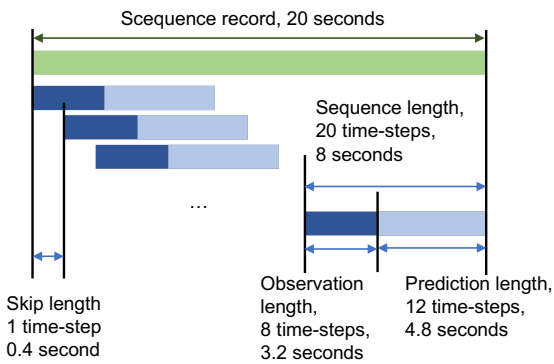


Figure 1.14: How the training and evaluation sequences are cut.

In Paper B and C, the sequence length is set to 20 time-steps with 8 observation time-steps which correspond to 3.2 seconds, and 12 prediction time-steps which correspond to 4.8 seconds. The total number of pedestrian sequences in the Waymo Open Dataset after pre-processing is 284,622, which is larger and more sufficient for training and evaluating compared to the ETH and UCY datasets with 1536 pedestrian sequences. The number of sequences used for training, validation and evaluation is shown as in Table 1.3.

	Training	Validation	Evaluation	Total
Number of scenarios	337	37	76	450
Number of pedestrians	7337	991	1978	10,306
Number of sequences (after cutting)	195,192	36,946	52,484	284,622

Table 1.3: The number of scenarios and sequences of the Waymo Open Dataset used for training, validation and evaluation.

1.5 Summaries of Studies

In this section, we summarize all the included studies, as shown in Figure 1.15. We briefly describe the research goal, scope, methodology, key results, and main contributions of each paper.

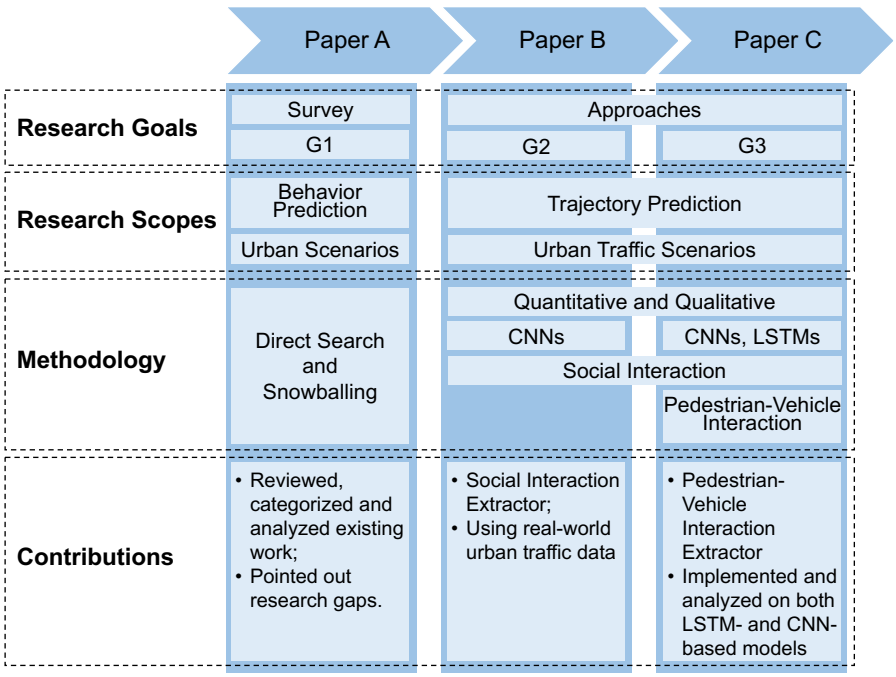


Figure 1.15: The summary of the appended papers.

1.5.1 Paper A: Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review

Research Goal The goals of this research are to address the progress and development of the state-of-the-art algorithms on pedestrian behavior prediction, and to identify the research gaps of existing works that we should focus on in future work.

Research Scope In this paper, we have reviewed the papers on behavior prediction including both trajectory and intention prediction. We have focused on the works using deep learning methods. We have also included datasets and benchmarks that are used in pedestrian behavior prediction. We have focused on urban scenarios, because the pedestrians are more vulnerable and need protection in urban areas as highlighted in Section 1.1.3.

Methodology This paper is a literature review paper. Our methodology to find and collect existing papers was based on direct search and snowballing (as proposed in [59]). We used IEEE Xplore digital library and Google Scholar for direct search to include both scientific databases and open access preprints. We have collected 50 papers from direct searching and 42 papers from snowballing. We did not set the time range explicitly, but after searching, the results originated mainly from 2016 to 2021. After paper collection, we have analyzed the existing works qualitatively and compared the methods quantitatively.

Results and Contributions In this paper, we have presented a thorough review of pedestrian behavior prediction models that used deep learning methods extracted from 92 papers. We have extended the taxonomy proposed by Hirakawa et al. [60] and Rudenko et al. [61], and have categorized existing studies by the following three criteria:

- Output types: The output types of the models include a) the trajectory prediction that provides the low-level information, b) the intention prediction that provides the high-level information, and c) the joint prediction for both trajectory and intention.
- Influencing factors: The factors that influence pedestrian behavior include a) target pedestrians whose behavior we aim to predict, b) other agents that interact with target pedestrians, and c) the environment.
- Prediction methods: The prediction methods can be classified into a) sequential methods, b) non-sequential methods, and c) the combination of the two.

We have presented advantages and drawbacks of using different influencing factors, and the properties of different prediction methods. The analysis of existing methods shows that including more influencing factors in trajectory prediction has the potential to improve accuracy, especially utilizing the features of social interaction and pedestrian-vehicle interaction. Finally, we have outlined

the research gaps and possible research directions for improving the performance of prediction algorithms for urban scenarios.

The original contributions of our review paper are as follows:

- We have presented a detailed analysis of the existing literature on pedestrian behavior prediction. Both trajectory and intention predictions are considered and analyzed, instead of only focusing on a single type of task. We included the most recent papers from 2016 to 2021.
- We have categorized existing works by three different criteria to provide a perspective from different dimensions, instead of reviewing the papers from a single criterion.
- We have introduced widely used datasets containing urban scenarios and commonly used evaluation metrics. We evaluated and compared previous methods on such publicly available datasets.
- We have pointed out research gaps and outlined the potential directions for future works.

1.5.2 Paper B: Social-IWSTCNN: A Social Interaction-Weighted Spatio-Temporal Convolutional Neural Network for Pedestrian Trajectory Prediction in Urban Traffic Scenarios

Research Goal The goal of this research is to propose a deep learning approach that can better understand the social interaction between pedestrians and more accurately predict pedestrian trajectories in urban traffic scenarios.

Research Scope In this research, we have focused on pedestrian trajectory prediction. The method is mainly developed for urban traffic scenarios. The dataset we used for training and evaluation is based on the urban traffic scenarios of the Waymo Open Dataset, which is collected in the US. We have considered the social interaction between pedestrians in the prediction, and used deep learning networks for modeling.

Methodology In order to reach our research goal, we have proposed the Social Interaction-Weighted Spatio-Temporal Convolutional Neural Network (Social-IWSTCNN) to predict pedestrian trajectories in urban traffic scenarios. We have developed the Social Interaction Extractor architecture to learn the interaction weights between pedestrians, and to improve the accuracy with a faster inference speed in contrast to the state-of-the-art model Social-STGCNN [35]. We use 3.2 s observation to predict 4.8 s trajectory in the future.

The overall Social-IWSTCNN model mainly includes three parts:

- The Social Interaction Extractor, which is proposed to learn the interaction weights and spatial features. Given observed frame sequences, we use the positions in each frame as input to learn the social interaction weights, and extract spatial and social interaction features using Social

Interaction Extractor. In our model, we do not build a graph representation of pedestrian trajectories. Instead, we directly use the observed locations relative to the last frame at each time-step as input for feature capturing. The spatial features of pedestrian i at time-step t are captured by embedding the input x-y-coordinate positions.

- The temporal convolutional networks (TCNs), which are used for temporal feature extracting. We apply TCNs on extracted spatial and social features to create spatio-temporal features for each pedestrian.
- The time-extrapolator convolutional networks, which are used for prediction. We apply time-exgrapolator CNNs to predict future trajectory distributions. Finally, we sample the learned Gaussian distributions to get the predicted trajectories.

Results and Contributions We have compared our proposed Social-IWSTCNN against the five baseline methods, including: linear regression (LR), Naïve LSTM without the influence of other individuals, Social-LSTM [27], Social-GAN [32], and Social-STGCNN [35].

The main contributions of this paper are as follows:

- We have proposed a novel structure, the Social Interaction Extractor, to better and faster capture interactions between pedestrians. Instead of using fixed attention weights with time-consuming non-linear computations as the state-of-the-art algorithm Social-STGCNN [35], our model learns the interaction attention weights in a data-driven manner. Compared to previous state-of-the-art model Social-STGCNN, the total inference speed of our proposed network is 4.7 times faster, and the data pre-processing speed is 54.8 times faster, while the prediction results are competitive, that improved the ADE and FDE by 1.50% and 1.82%, respectively.
- As we aim to solve the real-world task of predicting the pedestrian trajectories in urban traffic scenarios, the Waymo Open Dataset have been used for training and evaluation because this dataset contains more urban traffic scenarios and more sequences of pedestrians than the previously commonly used ETH [56] and UCY [57] datasets. Three state-of-the-art methods including Social-LSTM [27], Social-GAN [32], and Social-STGCNN [35] have been compared with our algorithm.

Our proposed model performs better than the other methods with lower error on both ADE and FDE on Waymo Open Dataset. Besides, compared to the state-of-the-art method Social-STGCNN, we can reach faster inference speed by removing the graph construction and avoiding non-liner interaction weights computation.

1.5.3 Paper C: Learning the Pedestrian-Vehicle Interaction for Pedestrian Trajectory Prediction

Research Goal The goal of this research is to propose a deep learning approach that can better understand the interaction between pedestrians and vehicles, and more accurately predict the trajectory of pedestrians in urban traffic scenarios.

Research Scope In this research, we have focused on pedestrian trajectory prediction, especially for urban traffic scenarios. Here, we also used the Waymo Open Dataset. In addition to the social interaction between pedestrians in the prediction, we have considered the interaction between pedestrians and vehicles, and used deep learning networks for modeling.

Methodology In this research, our network inputs include the trajectories of both pedestrians and vehicles. Three kinds of features are aggregated together and followed by the prediction backbones:

- The input spatial embedding features (e_t^i).
- The social interaction (SI) features between pedestrians (s_t^i).
- The pedestrian-vehicle interaction (PVI) features (v_t^i), In addition to the individual trajectory and social interactions between pedestrians, we introduce pedestrian-vehicle interactions in this paper.

We have applied the proposed PVI extractor to two different prediction backbones including an LSTM-based model as an example for sequential models, and a convolutional-based (Conv-based) model as an example for non-sequential models. We use 3.2 s observation to predict 4.8 s trajectory in the future.

Results and Contributions We have compared the performance of our proposed models against the following baseline methods:

- Sequential models: LSTM, Social-LSTM [27], and Social-GAN [32] are compared with our proposed SI-PVI-LSTM that considers both social interaction (SI) and pedestrian-vehicle interaction (PVI) using the LSTM algorithm.
- Non-sequential models: Linear Regression, Social-STGCNN [35], and Social-IWSTCNN [42] are compared with our proposed SI-PVI-Conv that considers both social interaction (SI) and pedestrian-vehicle interaction (PVI) using the convolutional algorithms including CNNs and TCNs.

The main contributions of this paper are as follows:

- We have proposed the Pedestrian-Vehicle Interaction (PVI) extractor to predict pedestrian trajectories. The features of interactions between pedestrians and vehicles are encoded by the vehicle feature embedding layers and pedestrian-vehicle interaction module.
- We have implemented, evaluated, and analyzed the proposed PVI extractor on both sequential (LSTM-based) and non-sequential (Conv-based) models. The LSTM-based model using our proposed PVI extractor is compared against Social-LSTM [27] and Social-GAN [32], and reduces ADE and FDE by 7.46% and 5.24%, respectively, compared to Social-LSTM. The Conv-based model using our proposed PVI extractor is compared against Social-STGCNN [35] and Social-IWSTCNN [42], and outperforms Social-STGCNN on ADE and FDE by 2.10% and 1.27%, respectively. The results show the efficiency of the proposed PVI extractor.

- The proposed algorithms have been trained and validated on real-world urban traffic data using the Waymo Open Dataset [58] as we aim to solve the real-world task of forecasting the trajectories in an urban traffic scenario.

The results have shown that the pedestrian-vehicle interaction influences pedestrian behavior, and models using the proposed PVI extractor can capture the interactions between pedestrians and vehicles. Therefore, our proposed models outperform the compared methods that only consider inter-personal interaction information.

1.6 Discussions

1.6.1 Existing Research on Pedestrian Behavior Prediction

1.6.1.1 Prediction Methods for Trajectory Prediction

Paper A reviews and analyzes existing research on pedestrian behavior prediction. It is shown that there are more papers on trajectory prediction (65.7%) than other behavior prediction (intention 25.4%, and joint prediction 8.9%). One possible reason could be that compared to the intention and joint prediction, trajectory prediction is used for more scenarios and attracted more researchers from different research fields. The trajectory prediction of pedestrians can be used not only for the automated vehicles in urban scenarios, but also for the development of social-aware robots in indoor scenarios. Another reason could be that the commonly used datasets for trajectory prediction appeared much earlier than the datasets for intention prediction. The most commonly used datasets for trajectory prediction - The ETH [56] and UCY [57] datasets were proposed in 2007 and 2009. The most commonly used datasets for intention prediction the JAAD [62] and PIE [63] datasets, were proposed much later, in 2017 and 2019, because the information of pedestrian intention is more implicit compared to trajectories, and more difficult to label.

As there are more applications and more available datasets, we mainly focus on trajectory prediction in Paper B and C. The deep learning methods that are used for trajectory prediction include sequential methods and non-sequential methods. For trajectory prediction, sequential methods including LSTMs and GANs have been mainly used since 2016 because this task requires time-series information. The LSTMs have advantages on long-term time series prediction compared with non-learning-based traditional models. But as each prediction time-step is dependent on preceding time-steps, LSTMs cannot be parallelized [34], and may accumulate errors when predicting long sequences [35]. Another problem with using LSTMs is that the model may learn the "average" of all possible trajectories instead of multiple feasible trajectories [32], as it is based on uni-modal assumption. The GAN-based models that are based on multi-modal assumptions can alleviate this problem and learn several feasible solutions. However, the GAN-based models are difficult to train and may not converge, because both the generator and discriminator in one framework need to be trained, and vanishing gradients may happen if there is an imbalance between the two deep networks [64].

Since 2018, non-sequential methods such as convolutional networks have been increasingly employed. Compared with LSTMs and other RNNs, the Conv-based networks including CNNs and TCNs can easily be parallelized to speed up the inference process [34], and can alleviate the problem of error accumulating [35]. The drawback of the Conv-based methods is that they predict the future trajectories in a fixed time horizon, and are less flexible than LSTMs.

We have investigated which type of deep learning prediction methods perform better in the pedestrian trajectory prediction task. In Paper B, the Conv-based models (our proposed Social-IWSTCNN and the previous state-of-the-art method Social-STGCNN) get more accurate results than the LSTM-based methods (Social-LSTM and Social-GAN). This is because that RNNs such as LSTMs and LSTM-based GANs accumulate the error, while the Conv-based models do not have this drawback as we discussed previously.

Furthermore, there is also evidence that the Conv-based methods can better represent the motion states embedding the features directly from spatial and temporal information, compared with the LSTMs that use hidden states. Moreover, we have noticed that for the Social-LSTM and Social-GAN methods, ADE and FDE of the most crowded scenarios are worse than the results of LSTMs. This can be interpreted as that the pooling structure on the hidden states of LSTMs cannot extract the interaction feature properly in dense urban traffic scenarios. Paper C has compared the results of applying the Pedestrian-Vehicle Interaction extractor on sequential and non-sequential methods, and the results provide further evidence that Conv-based models outperform LSTM-based models on pedestrian trajectory prediction.

Since 2020, a breakthrough appeared for sequential methods. Transformers [65] have been used for pedestrian trajectory prediction as in [66–68]. The transformers can overcome the aforementioned drawbacks of RNNs. They allow parallelization, and can avoid error accumulating. But similar to CNNs, they are implemented in a fixed length and are less flexible compared with LSTMs.

1.6.1.2 Influencing Factors of Pedestrian Behavior

The factors that influence pedestrian behavior play important roles in pedestrian behavior prediction. Paper A shows that many researchers use multiple influencing factors for prediction. We have summarized the influencing factors into three types, the information of target pedestrians, the influence of other road agents, and the influence of the environment. Of all the papers we reviewed, 20.9% used only one type of factor, the information of target pedestrians, 47.8% used two types of factors, and the remaining 31.3% used three types of factors. The performance of existing methods provides evidence that with more information provided, the algorithm is more likely to accurately approximate the future motion of pedestrians. Therefore, the current trend is to use as many influencing factors as possible to include more information.

Paper A reveals that the existing algorithms tend to include more information and use more complicated algorithms. For the target pedestrians, the trajectories and motion states are used since 2016 as in Social-LSTM [27]. In 2017, the behavioral features are included as in Rasouli et al.'s work [62], and in 2019, the individual information is added, as in [48, 53, 69]. For the influence

of other agents, the researchers model the social interaction with social pooling in 2016 [27], and then use more complicated algorithms such as the graph-based model in 2018 [70], and added knowledge-based information into the model in 2019 [69]. However, the computational costs and inference time are increasing accordingly. Therefore, we need to consider that while improving the model accuracy.

In this thesis, we consider the interaction between pedestrians and other road users. Paper A shows that many existing models considered the social interaction as symmetric so they used the pooling method to model it. Several asymmetric models that utilized graph-based algorithms to learn the influence of interactions but they used hand-crafted non-linear functions to represent the interaction relationship. Much work could be done to use deep learning networks to improve the model on accuracy performance and inference speed, as we have done in Paper B and C.

In Paper B and C, to improve the prediction accuracy, we have included more information, social interactions and pedestrian-vehicle interactions. To accelerate the prediction, we avoided graph constructing and non-linear calculation, thereby achieving a faster and more competitive inference speed. More discussion is shown in Section 1.6.2 and Section 1.6.3.

1.6.2 Social Interaction in Trajectory Prediction

In Paper B, we have proposed the "Social Interaction-Weighted Spatio-Temporal Convolutional Neural Network (Social-IWSTCNN)" model that learns the social interaction between pedestrians using a deep learning sub-network. The sub-network uses the velocity of pedestrians and the relative positions between pedestrians as inputs to learn the interaction relationship weights.

Our proposed model has been compared with the state-of-the-art methods Social-LSTM [27] and Social-GAN [32] that use a pooling module, and Social-STGCNN [35] that uses hand-crafted interaction weights between pedestrians. The quantitative evaluation results on ADE and FDE have shown that our proposed model outperforms the others. This shows that compared to using hand-crafted attention weights or pooling layers, using deep learning methods to learn the interaction weights can improve the accuracy.

In Paper B, the algorithms have been compared in scenarios of different traffic densities. We have divided the Waymo Open Dataset [58] into three groups with different densities by the number of pedestrians. We have noticed that for densely populated scenarios, the results of our model are only slightly better than the compared Social-STGCNN, while for less crowded scenarios, ADE is substantially improved by 17.3% and FDE is improved by 16.8%. A possible reason is that the hand-crafted function used by the Social-STGCNN model is designed for the ETH [56] and UCY [57] datasets, which are densely populated crowds scenarios. So it can well represent the movement of pedestrians in crowded scenarios and have competitive results with our model. While less crowded scenarios also occur in real traffic, our algorithm that learns from the data has a better performance. There is evidence that a model that performs well on one scenario may not perform well on the other, and the datasets are important for deep learning methods. This also shows that compared to the manually designed function of interaction weights, the deep learning method

has the potential to be used for scenarios of different traffic densities, both crowded and empty scenarios.

As we have discussed in Section 1.6.1.2, when we include more information and use more complicated algorithms for prediction to get more accurate results, we also need to consider whether the inference speed becomes too slow to use. In Paper B, we have compared the inference speed between the two methods that get competitive accuracy: our model and Social-STGCNN. Compared with the Social-STGCNN, our proposed model is 54.8 times faster on pre-processing speed, and 4.7 times faster on the total inference speed. Compared with Social-STGCNN, there are two changes: a) we have removed the non-linear calculation for attention weights computing, and b) we have avoided constructing the adjacent matrix of the graph as in Social-STGCNN. To evaluate the influence of the two changes, we have tested the inference time of the Social-IWSTCNN method with graph construction to see how much our method can speed up by only removing the non-linear calculation. The results show that both removing non-linear calculations and avoiding graph constructing improve the inference speed. Our algorithm is computationally more efficient and has a faster speed while reaching competitive results.

1.6.3 Pedestrian-vehicle Interaction in Trajectory Prediction

1.6.3.1 Extracting Pedestrian-vehicle Interaction

In Paper C, we have proposed the pedestrian-vehicle interaction (PVI) extractor, and applied it to both the sequential and the non-sequential models. Our proposed models have used both social interaction (SI) and pedestrian-vehicle interaction (PVI) features with separate sub-networks. For PVI features, the sub-network uses the velocities of vehicles, and the relative positions between pedestrians and vehicles as inputs to learn the pedestrian-vehicle interaction relationship weights. The results demonstrate that using pedestrian-vehicle interaction information improves the accuracy of pedestrian trajectory prediction.

For the sequential models, we have used LSTMs as the prediction backbone, so we refer to it as the SI-PVI-LSTM model. The proposed SI-PVI-LSTM model outperforms LSTM, Social-LSTM [27], and Social-GAN [32] that have not included the pedestrian-vehicle interaction feature while predicting. This shows that the pedestrian-vehicle interaction plays an important role in influencing pedestrian behavior. Compared with the Social-GAN, the SI-PVI-LSTM gets better results without using the complicated and hard-to-train GAN structure. This indicates that instead of improving the prediction backbone with complicated methods, the influencing factors should also be considered, as they may bring improvements at a small computational cost.

For the non-sequential models, our proposed model has used convolutional networks including CNNs and TCNs as the backbone, so we refer to it as the SI-PVI-Conv model. Compared with other non-sequential models including LR, Social-STGCNN, and Social-IWSTCNN, the proposed SI-PVI-Conv model achieves the best ADE result by using the pedestrian-vehicle interaction information in addition. However, it has not improved the performance of

FDE compared with Social-IWSTCNN. There are two possible reasons for this. Firstly, the model only uses the vehicle information of the observation period, so the information may be insufficient for a long-term prediction. As vehicles move much faster than pedestrians, in 12 time-steps covering 4.8 s the vehicles may travel a long distance (e.g., a vehicle with 30 km/h speed travels 40 m within 4.8 s), so some other vehicles may approach pedestrians and influence their behavior during predicting time horizon. Secondly, in SI-PVI-Conv, we calculate the interaction with all vehicles within the sensor scan range and do not consider the orientation or directions of pedestrians. However, the vehicles behind the pedestrians may not influence pedestrians as much as the other vehicles, so this may bring in extra noise. Therefore, the FDE result of SI-PVI-Conv has not been improved but still is comparative with the other two Conv-based models. Updating the vehicle information during the prediction time horizon, or simultaneously predicting both pedestrians and vehicles may alleviate the first problem. Adding the orientation and direction information of pedestrians may alleviate the second problem and improve the performance.

1.6.3.2 Analysis on Influencing Factors of Pedestrian Trajectory

In Paper C, we have investigated the influence of the social interaction (SI), the pedestrian-vehicle interaction (PVI), and different inputs including the relative positions and the relative velocities.

Comparing the LSTM model that only uses individual past trajectory information and the Social-LSTM that models the social interaction by pooling over hidden states from LSTMs, we surprisingly find that the Social-LSTM [27] reaches worse results than LSTM model on the Waymo Open Dataset [58]. This is consistent with the results on ETH [56] and UCY [57] datasets in Gupta et al.'s work [32].

There are two possible reasons for this result: a) the social interaction could not improve the accuracy, or b) the way the model extracts the social interaction features is not suitable. To evaluate the influence of these two components, we modify the way we extract the social interaction features. The Social-LSTM used the hidden states of LSTMs to represent the moving states of pedestrians, and applied pooling on the hidden states to learn the interaction. In our experiments, instead of using the hidden states of LSTMs to represent the moving states and calculate interactions on the hidden states, we directly extract the spatial features from the pedestrian positions, and use the spatial feature to calculate the interaction features of each frame. This is followed by LSTMs to compute the hidden states of the interactions. By using the improved social feature extractor, our social interaction model SI-LSTM reduces the ADE and FDE significantly compared to the LSTM model, which shows that social interaction can influence pedestrian behavior when it is well represented and learned. This also provides evidence that the hidden states cannot well represent the moving states and are not suitable for calculating the social interaction.

Paper C shows that with only SI or only PVI information, the results do not achieve the best accuracy. For both LSTM-based and Conv-based models, we get the best performance by including both social interaction and pedestrian-

vehicle information. This shows that both social interactions and pedestrian-vehicle interactions contribute to the improvement of the performance.

To investigate the inputs of extracting the interaction features, in addition to using the relative positions between objects as input, we compare the models with and without relative velocities between pedestrians and the other objects. The results are not improved by adding velocity information. This is because pedestrians are very agile, and the velocities of pedestrians can change all the time and may introduce noises into the network. Therefore, to improve the performance and learn the interaction weights with deep learning networks, the inputs of the networks should be carefully decided. More research could be done in the future to study the appropriate inputs.

1.6.4 Contribution to the Vehicular Automation

Our research outcomes can contribute to automated vehicles (AVs), driving safety, and driver assistant systems.

By reviewing existing works in Paper A, we have analyzed the strengths and weaknesses of different deep learning methods and have pointed out appropriate methods for various predicting tasks. We have proposed two novel approaches that consider social interactions (in Paper B) and pedestrian-vehicle interactions (in Paper C) in prediction, and have improved the prediction accuracy. By using our algorithms, automated vehicles and driver assistant systems can better avoid pedestrian-vehicle collisions. We have reached a faster inference speed, which means our algorithms have the potential to be used on the vehicle to get an earlier warning for the hazardous situation and leave more time for reaction and control.

Besides, the human behavior prediction methods we proposed have the potential to be transferred to other human-robot interaction scenarios and help to build socially aware robots. The output of our research can also provide the information for developing the human-AV and human-robot interfaces. For instance, the predicted trajectories can be provided to AVs/robots and guide AVs/robots to interact with humans accordingly.

1.6.5 Limitations

Paper A has summarized the places where the data was captured. As we mentioned in Section 1.1, the proportion of pedestrian fatalities in low-income countries is 36%, which is twice as in high-income countries. However, most of the existing datasets are captured in the high-income and middle-income countries, and there are few datasets that covered the low-income countries. There is a limitation that most of the existing research is not focusing on those low-income countries that especially need the concern. Future research could focus on developing more datasets and studies for these places.

Besides, as our methods are data-driven, the quality of data collection and annotation is important. Low-resolution images make it hard to learn useful features [43]. Without properly labeled data we cannot train our algorithms. However, high-quality data collection and annotation are costly, and existing datasets may not perfectly fit our needs. There is a limitation in the dataset we can use because of the requirement on quality.

To best simulate the real traffic, we use real-world datasets for training and evaluation. In Paper B and C, our algorithms are based on the real-world urban traffic scenarios of the Waymo Open Dataset which is collected in the US. There are limitations of using this dataset. For instance, the geographic features of the roads in the US can be different from other places such as in European countries. The crowd densities of the traffic vary in different regions. Culture and legal factors can also influence the behavior pattern of the pedestrians on the road. More research needs to be done to investigate the transferability and scalability of our models.

1.7 Conclusions and Future work

1.7.1 Conclusions

This thesis contributes to making AVs and ADAS safer by predicting the behavior of pedestrians and preventing pedestrian-vehicle collisions. Specifically, this thesis has advanced the prediction of pedestrian trajectories in urban traffic scenarios using deep learning methods. The goal of this thesis is to develop deep learning methods to predict the behavior of pedestrians in interactions with other road users. We have realized this goal by covering the following sub-goals. The novel and original contributions of each paper are highlighted below.

G1: Reviewing, categorizing, analyzing, and discussing currently existing research to point out research gaps for the problem area of pedestrian behavior prediction.

Paper A has reviewed the existing research on pedestrian behavior prediction that uses deep learning methods. Compared with existing literature review papers, our paper classifies existing studies by three criteria and provides a perspective from multiple dimensions. By including both trajectory and intention prediction instead of a single task, we are inspired to take the advantage of prediction methods on both tasks in future work. The most recent research studies over the past five years have been included. We have addressed the progress and development of state-of-the-art algorithms on pedestrian behavior prediction, and introduced commonly used datasets and evaluation metrics, and then compared existing methods. The advantages and drawbacks of different methods have been presented and discussed. In this research, we have established the overall framework of the pedestrian behavior prediction, and addressed the challenges and the research gaps of existing works that we should focus on in future work.

G2: Developing the approach that can better predict pedestrian trajectories in urban traffic scenarios, and using deep learning to model social interactions between pedestrians.

Paper B has proposed a deep learning method namely the Social-IWSTCNN model for pedestrian trajectory prediction. Compared with previous state-of-the-art algorithms, our proposed model uses a deep learning sub-network, namely the Social Interaction Extractor to learn the social interaction weights

between pedestrians. We use the velocities of pedestrians and the relative positions between pedestrians as inputs for learning the interaction relationship, and get more accurate prediction results with a faster inference speed. Besides, our model uses a large-scale real-world dataset, the Waymo Open Dataset in urban traffic scenarios for training and evaluation. Results in Paper B show that the social interaction information contributes to more accurate prediction, and using interaction attention weights learned from deep learning networks can improve performance and reduce inference time. This indicates that our proposed social interaction extractor can well learn the interaction feature.

G3: Considering pedestrian-vehicle interactions, and using deep learning to model the interactions between pedestrians and vehicles when predicting pedestrian trajectories.

Paper C has proposed a deep learning method that includes the pedestrian-vehicle interaction features when predicting the pedestrian trajectories in urban traffic scenarios. Compared with previous state-of-the-art algorithms, our proposed model learns the pedestrian-vehicle interaction features with a deep learning sub-network, namely the Pedestrian-Vehicle Interaction Extractor. By including both social interaction and pedestrian-vehicle interaction features, the model has achieved the best performance and outperforms previous state-of-the-art algorithms. Results in Paper C show that the Conv-based models get less prediction error than LSTM-based models. On both Conv- and LSTM-based methods, using our proposed pedestrian-vehicle interaction extractor can improve the prediction results, which indicates that our proposed extractor can well represent the interaction between pedestrians and vehicles.

1.7.2 Future Work

The findings in Paper A show that most existing trajectory predictions relied on trajectory information but did not leverage the appearance and skeleton behavioral features like in intention predictions. Future works can focus on joint predictions to predict trajectories and intentions simultaneously. The two prediction branches can share the extracted features and predicted results to compensate and improve each other. The appearance and skeleton behavioral cue that is typically used in intention prediction, and the predicted intention could be included to improve trajectory prediction. However, as we previously discussed, the computational costs could be high if we include the visual and skeleton feature of pedestrians. More work needs to be done to investigate how much improvement could be brought here if we include the intention information and appearance and skeleton behavioral features.

Paper C shows that the inputs for learning the interaction weights influence the prediction results. It is important to investigate and find the most suitable inputs and factors that influence the interaction. As mentioned in Paper A, the interactions can either be learned implicitly by deep learning models that can include as much information as possible without requiring expert knowledge and which are as a consequence hard to explain, or be represented by using knowledge-based hand-crafted features that are explainable but that require prior knowledge instead. In future works, we can develop hybrid models to take advantage of both approaches.

Therefore, based on the research gaps, the continued work in this PhD project can further investigate how:

G4: To improve the behavior prediction. We predict the crossing intention simultaneously in addition to the trajectory prediction, while considering the pedestrian-vehicle interaction.

G5: To make the prediction model more explainable and easier to be generalized to other scenarios. We develop hybrid methods by combining the knowledge-based methods with deep learning methods.

The continued future work will try to answer the following research questions:

RQ4: What improvements to trajectory prediction can we get from adding the intention information? How to accurately predict the crossing intention of pedestrians in urban traffic scenarios using deep learning models?

RQ5: How to combine the knowledge-based methods with deep learning methods? How to interpret the influence of the pedestrian-vehicle interaction on pedestrian behavior?

Bibliography

[1] WHO, "Global status report on road safety 2018: Summary," World Health Organization, Report, 2018.

[2] U. N. G. Assembly, "Transforming our world: The 2030 agenda for sustainable development," 2015.

[3] K. Rumar, "Transport safety visions, targets and strategies: beyond 2000," *1st European Transport Safety Lecture. European Transport Safety Council, Brussels, Tech. Rep*, 1999.

[4] P. Crist and T. Voege, "Safer roads with automated vehicles? (corporate partnership board report). international transport forum (itf)," 2018.

[5] WHO, "Global status report on road safety 2013: Supporting a decade of action," World Health Organization, Report, 2013.

[6] A. Santacreu, "Monitoring progress in urban road safety," *International Transport Forum Policy Papers, No. 79, OECD Publishing*, 2020.

[7] W. H. Organization *et al.*, "Pedestrian safety: a road safety manual for decision-makers and practitioners," 2013.

[8] I. R. Traffic and A. D. (IRTAD), "Road safety annual report 2020: Sweden," 2020.

[9] M. Peden, R. Scurfield, D. Sleet, C. Mathers, E. Jarawan, A. Hyder, D. Mohan, A. Hyder, E. Jarawan *et al.*, *World report on road traffic injury prevention*. World Health Organization, 2004.

[10] S. Ferguson, B. Luders, R. C. Grande, and J. P. How, "Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions," in *Algorithmic Foundations of Robotics XI*. Springer, 2015, pp. 161–177.

[11] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive bayesian filters: A comparative study," in *German Conference on Pattern Recognition*. Springer, 2013, pp. 174–183.

[12] H. Zhang, Y. Liu, C. Wang, R. Fu, Q. Sun, and Z. Li, "Research on a pedestrian crossing intention recognition model based on natural observation data," *Sensors*, vol. 20, no. 6, p. 1776, 2020.

[13] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS one*, vol. 5, no. 4, p. e10047, 2010.

[14] M. S. Shirazi and B. Morris, "Observing behaviors at intersections: A review of recent studies & developments," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 1258–1263.

[15] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 90–104, 2016.

[16] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[17] A. H. Do, S. A. Balk, and J. W. Shurbutt, "Why did the pedestrian cross the road?" *Public Roads*, vol. 77, no. 6, 2014.

[18] A. Värnild, P. Larm, and P. Tillgren, "Incidence of seriously injured road users in a swedish region, 2003–2014, from the perspective of a national road safety policy," *BMC public health*, vol. 19, no. 1, pp. 1–10, 2019.

[19] W. Little, R. McGivern, and N. Kerins, *Introduction to sociology-2nd Canadian edition*. BC Campus, 2016.

[20] B. Völz, H. Mielenz, G. Agamennoni, and R. Siegwart, "Feature relevance estimation for learning pedestrian behavior at crosswalks," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 854–860.

[21] C. R. Kothari, *Research methodology: Methods and techniques*. New Age International, 2004.

[22] J. Kamiri and G. Mariga, "Research methods in machine learning: A content analysis," *International Journal of Computer and Information Technology (2279-0764)*, vol. 10, no. 2, 2021.

[23] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[24] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021.

[25] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[26] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[27] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F.-F. Li, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.

[28] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[29] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1186–1194.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[32] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2255–2264.

[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[34] N. Nikhil and B. Tran Morris, "Convolutional neural network for trajectory prediction," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[35] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "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 (CVPR)*, 2020, pp. 14 424–14 432.

[36] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[37] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "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 (CVPR)*, 2019, pp. 1349–1358.

[38] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, S. H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting

using bicycle-gan and graph attention networks," *arXiv preprint arXiv:1907.03395*, 2019.

[39] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, Conference Proceedings, pp. 6272–6281.

[40] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017.

[41] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[42] C. Zhang, C. Berger, and M. Dozza, "Social-iwstcnn: A social interaction-weighted spatio-temporal convolutional neural network for pedestrian trajectory prediction in urban traffic scenarios," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 1515–1522.

[43] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 2607–2612.

[44] B. Völz, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "Inferring pedestrian motions at urban crosswalks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 544–555, 2018.

[45] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[46] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1258–1268.

[47] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[48] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, Conference Proceedings, pp. 6120–6127.

[49] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[50] S. Eiffert, K. Li, M. Shan, S. Worrall, S. Sukkarieh, and E. Nebot, "Probabilistic crowd gan: Multimodal pedestrian trajectory prediction using a graph vehicle-pedestrian attention network," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5026–5033, 2020.

[51] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, Conference Proceedings, pp. 6319–6328.

[52] S. Carrasco, D. F. Llorca, and M. A. Sotelo, "Scout: Socially-consistent and understandable graph attention network for trajectory prediction of vehicles and vrus," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, Conference Proceedings.

[53] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8483–8492.

[54] R. Chandra, U. Bhattacharya, C. Roncal, A. Bera, and D. Manocha, "Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs," in *ACM Computer Science in Cars Symposium*, 2019, Conference Proceedings, pp. 1–9.

[55] R. Chandra, T. Guan, S. Panuganti, T. Mittal, U. Bhattacharya, A. Bera, and D. Manocha, "Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4882–4890, 2020.

[56] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 261–268.

[57] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26. Wiley Online Library, 2007, pp. 655–664.

[58] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2446–2454.

[59] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, 2014, pp. 1–10.

[60] T. Hirakawa, T. Yamashita, T. Tamaki, and H. Fujiyoshi, "Survey on vision-based path prediction," in *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 2018, pp. 48–64.

[61] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[62] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[63] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "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 (ICCV)*, 2019, Conference Proceedings, pp. 6262–6271.

[64] J. Amirian, J.-B. Hayet, and J. Pettré, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[66] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *European Conference on Computer Vision*. Springer, 2020, Conference Proceedings, pp. 507–523.

[67] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 335–10 342.

[68] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," *arXiv preprint arXiv:2103.14023*, 2021.

[69] Y. Ma, E. W. Lee, Z. Hu, M. Shi, and R. K. Yuen, "An intelligence-based approach for prediction of microscopic pedestrian walking behavior," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 10, pp. 3964–3980, 2019.

[70] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2018, Conference Proceedings, pp. 4601–4607.