CHALMERS
UNIVERSITY OF TECHNOLOGY

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

Citation for the original published paper (version of record):
Zhang, C., Berger, C. (2023). Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review. IEEE Transactions on Intelligent Transportation Systems, 24(10): 10279-10301. http://dx.doi.org/10.1109/TITS.2023.3281393

(article starts on next page)

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

Chi Zhang and Christian Berger

***Abstract*— The prediction of pedestrian behavior is essential for automated driving in urban traffic and has attracted increasing attention in the vehicle industry. This task is challenging because pedestrian behavior is driven by various factors, including their individual properties, the interactions with other road users, and the interactions with the environment. Deep learning approaches have become increasingly popular because of their superior performance in complex scenarios compared to traditional approaches such as the social force or constant velocity models. In this paper, we provide a comprehensive review of deep learning-based approaches for pedestrian behavior prediction. We review and categorize a large selection of scientific contributions covering both trajectory and intention prediction from the last five years. We categorize existing works by prediction tasks, input data, model features, and network structures. Besides, we provide an overview of existing datasets and the evaluation metrics. We analyze, compare, and discuss the performance of existing work. Finally, we point out the research gaps and outline possible directions for future research.**

***Index Terms*— Pedestrian behavior prediction, trajectory, intention, deep learning, neural networks, automated vehicles, survey.**

I. Introduction

ACCORDING to World Health Organization (WHO)'s report on road safety [1], about 1.35 million people are fatally injured by road crashes every year. Pedestrians constitute 23% of all road traffic deaths globally, which is unacceptably high. As the most vulnerable road users, pedestrians are important participants and need protection. Given that human errors are one of the main factors in most road traffic crashes [2], automated vehicles (AVs) may have the potential to reduce these figures and improve road safety. Hence, it is essential to predict the behavior of pedestrians for AVs to better understand the AV's surroundings for making better and safer driving decisions and preventing potential hazardous situations. In recent years, the interest in AVs has attracted increasing attention to research related to pedestrian behavior prediction.

Predicting pedestrians' behavior is a great challenge. In contrast to the vehicles, whose behavior prediction has been well studied and reviewed by Lefèvre et al. [3] and Mozaffari et al. [4] for instance, pedestrians are more agile and can change their speed and direction unexpectedly [5] with unknown or hardly predictable moving patterns [6]. Pedestrian behavior is driven by complicated influencing factors. These factors include not only the properties of the pedestrians themselves such as the motion states, destination, age, and gender [7], but also the interactions with other pedestrians [8] and vehicles [9], [10], [11]. Furthermore, the environment can also influence the intention of pedestrians both explicitly and implicitly. The non-linearity arising from pedestrian interactions and the complexity of multiple influencing factors hinder accurate prediction using conventional knowledge-based models such as social force [12] and constant velocity model [13]. Deep learning is a subset of machine learning based on artificial neural networks with multiple layers. Inspired by the biological neuron, artificial neural networks are composed of nodes with linear weights and bias, and non-linear activation functions. Deep learning methods are powerful tools that can be used to extract high-level features from data, and can deal with the non-linearity of the data. Therefore, researchers are exploring the potential of deep learning models to represent and extract pedestrians' behavior patterns in a data-driven manner. In this paper, we analyze and categorize existing research and discuss how current challenges have been addressed so far.

As deep learning methods are data-driven, datasets are important for developing models. The report on pedestrian safety by WHO [14] has shown that about 70% of pedestrian fatalities occur in urban areas in the European Union, and in the United States, this number is about 76%. Pedestrian-vehicle collisions occur more in urban areas than rural areas in these countries, and hence, most of the publicly available datasets for developing pedestrian behavior prediction models used by researchers are collected in urban areas. Therefore, we review prediction methods and datasets in urban scenarios.

The scope of this paper covers studies that predicted pedestrian behavior, including the future trajectory and crossing intention. We focus on deep learning-based models. When it comes to datasets and model inputs, we focus on urban scenarios, and cover various inputs such as camera images, light detection and ranging (LiDAR) point clouds, or the speed of the ego vehicle to name a few. Various factors that influence pedestrian behavior are covered, such as pedestrians' own past motion states, interactions with other pedestrians and vehicles, and influences of the environment.

There are several published papers that reviewed existing works on pedestrian behavior prediction. Hirakawa et al. [15]

Manuscript received 25 December 2021; revised 4 August 2022 and 12 December 2022; accepted 21 May 2023. Date of publication 12 June 2023; date of current version 4 October 2023. This work was supported in part by the European Research Project SHAPE-IT—Supporting the Interaction of Humans and Automated Vehicles: Preparing for the Environment of Tomorrow and in part by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant 860410. The Associate Editor for this article was C. Wu. *(Corresponding author: Chi Zhang.)*

The authors are with the Department of Computer Science and Engineering, University of Gothenburg, 41296 Gothenburg, Sweden (e-mail: chi.zhang@gu.se; christian.berger@gu.se).

Digital Object Identifier 10.1109/TITS.2023.3281393

surveyed vision-based methods for pedestrian path prediction, where deep learning-based methods were only covered by a small extent. Rudenko et al. [16] reviewed the work related to human motion trajectory prediction and categorized existing methods by the modeling approach and contextual cues. Korbmacher and Tordeux [17] reviewed pedestrian trajectory prediction methods, compared deep learning methods and knowledge-based methods. These papers only covered the trajectory prediction and omitted the important prediction of intention that can be used for pedestrian-vehicle collision avoidance. Shirazi and Morris [9] focused on pedestrian intention at intersections and analyzed how crossing behavior is influencing intersection participants. Ohn-Bar et al. [10] provided a survey on interactions between humans and autonomous vehicles. Rasouli and Tsotsos [11] reviewed pedestrian behavior studies of both classical pedestrian–driver interactions and more recent autonomous vehicles and pedestrian interactions, but mainly focused on analyzing human factors and interactions instead of deep learning-based behavior prediction. Ridel et al. [18] reviewed and classified existing pedestrian behavior prediction models, but they classified previous works from only a single criterion, and many recently suggested deep learning methods were not covered. Most of the previous review papers focused on a single task, either the analysis of trajectories [15], [16], [17] or intention [9], or interactions between pedestrians and vehicles [10], [11], which did not cover the aspects in this paper's scope. Moreover, most of these papers classified the existing literature by a single criterion [17], [18], and only include methods with some particular input data [15].

To overcome the drawbacks listed above, we review, categorize, and analyze the existing research on pedestrian behavior including both the trajectory and intention prediction in this paper. We propose four criteria for classification to consider existing works from different dimensions. The main contributions of this paper are:

- We present a detailed analysis of the existing literature on pedestrian behavior prediction, including trajectory prediction, intention prediction, and the joint prediction of both. We categorize existing approaches from four criteria including a) prediction tasks b) input data, c) the features that are considered in existing models, and d) network structures, and emphasize the advantages and drawbacks of existing approaches.
- We include the most recently proposed existing publicly available datasets and commonly used evaluation metrics. We compare the trajectory and intention prediction tasks on the most commonly used open datasets and present state-of-the-art algorithms.
- We point out research gaps and outline the potential directions for future works.

II. Methodology and Taxonomy

A. Methodology

Our methodology to find and collect existing papers is based on direct search and snowballing. We used IEEE Xplore digital library and Google Scholar for direct search to include both scientific databases and open-access pre-prints. We used "pedestrian behavior prediction" OR "pedestrian trajectory prediction" OR "pedestrian intention prediction" filtered by: "deep learning" OR "network" as initial search strings. We did not set the time range explicitly, but after searching, the results originated mainly from 2016 to 2021. Then we went through the results to select relevant papers meaning that the research targets are pedestrians instead of drivers or robots, the research goal is behavior prediction instead of detection, tracking, or vehicle/robot path planning, and the methods are deep learning. We selected 50 papers from direct searching. Then we did backward and forward snowballing (as proposed in [19]) with their citations and references to include relevant publications, and got 42 papers from snowballing. We review 92 papers in total, including 44 on trajectory prediction, 17 on intention prediction, 6 on joint prediction, 18 on datasets and benchmarks, and 7 on literature review. The number of papers over the years[1] and the distribution of the papers is shown in Fig. 1. The rising trend of the papers indicates the growing interest in this field.

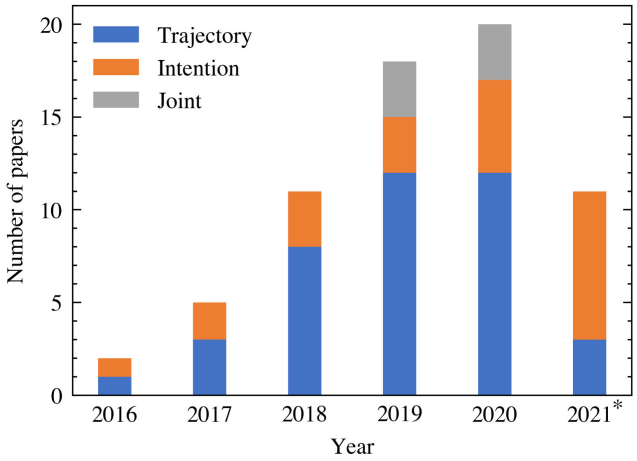


Fig. 1. The number of papers over the years and the distribution of the papers. The rising trend of the papers indicates the growing interest in deep learning-based pedestrian behavior prediction. *Note that in 2021 only the papers published in the **first half** of the year are included.

B. Taxonomy and Overview

We address our expansion of the taxonomy proposed by Hirakawa et al. [15] and Rudenko et al. [16], and categorize existing studies by the following four criteria. With the help of this taxonomy, one can easily get started with a model's desired input and output, decide the features that they want to consider in the model, and find a reasonable network structure.

1) Prediction tasks: The prediction tasks define the problem that a model is addressing, and a model's expected output. We classify previous models by three kinds of prediction tasks, including a) trajectory prediction, b) intention prediction, and c) joint prediction that predict both trajectory and intention.

[1]Note that in 2021 only the papers published in the first half of the year are included.

2) Input data: The input data show the information provided by sensors or annotations that are used as model inputs. We classify previous models by three kinds of input data that provide different types of information, including: a) the past trajectories of pedestrians from annotations, b) the information provided by sensors, and c) other supplementary information such as the map information, the information of the ego vehicle, etc.
3) Model features: There are many factors that influence the future behavior of pedestrians. It is hard to consider all factors, so previous studies tried to cover those factors that influence pedestrians most as model features. Model features are the observations and factors that were considered by previous studies in models as stimuli to the future behavior of pedestrians. We classify previous models by three types of model features, including a) the observed information of target pedestrians, b) the information of other agents that interact with target pedestrians, and c) the information of the environment.
4) Network structures: The network structures show how previous studies learned the moving pattern from observed information. There are several typical structures used in existing prediction models that can be classified into sequential networks and non-sequential networks.

We summarize the pedestrian behavior prediction framework in Fig.2 and show how these four criteria are related. We review and classify the existing works in detail from the proposed categories: prediction tasks as in Sec. III, input data as in Sec. IV, model features as in Sec. V, and network structures as in Sec. VI. Then, we outline the evaluation metrics and the datasets used in existing research in Sec. VII, and compare the performances on publicly available datasets to point out the research gaps and outlines potential research directions in Sec. VIII. Finally, we present our conclusions in Sec. IX.

III. Prediction Tasks

In this section, we classify previous studies based on prediction tasks, including trajectory prediction, intention prediction, and joint prediction that predicts both. We cover different output representations and training strategies for each type of task. Table I summarizes different types of prediction tasks, model features, and input data of existing studies.

A. Trajectory Prediction

1) Task Definition: The trajectory prediction methods provide low-level information of pedestrian behavior with detailed spatial and temporal information. This information can be used for collision avoidance or helping autonomous vehicles to plan their future path. We define the trajectory of a pedestrian as a sequence of x-y coordinate positions including their temporal order. A person's position in a scene is represented by the x-y-coordinate $X = (x, y)$. Given a set of n pedestrians with their observed positions over time steps t, $X_t^i = (x_t^i, y_t^i)$ where $i \in \{1, \ldots, n\}$, $1 \leq t \leq T_{obs}$, and other information I such as the information of the surrounding environment and objects, we aim to predict the likely trajectories of the target pedestrians $\hat{Y}_t^i = (\hat{x}_t^i, \hat{y}_t^i)$ in the future time steps $T_{obs} + 1 \leq t \leq T_{pred}$.

2) Output Representation: There are different kinds of output representations for trajectory prediction. Many researchers treated trajectory prediction as a regression problem. The output can be represented as: a) positions of (x, y) coordinates, b) uni-modal distributions, and c) multi-modal distributions. Representing output as positions is used by many studies, such as [21], [22], [23], [24], and [49]. Such models are simple compared to those models predicting distributions, and can get deterministic results, but they cannot include the randomness nature of the pedestrian movement. Uni-modal distributions are very popular for trajectory prediction and are used by studies such as [27], [28], [39], [41], [45], [52], [59], and [62]. Compared to multi-modal distribution models, the uni-modal prediction requires less computational cost, but the model may learn an "average behavior" that is not plausible. Multi-modal distributions can overcome the drawback of converging to average behaviors by outputting several plausible behaviors, and are used by studies such as [31], [33], [35], [38], [40], [44], [47], [50], [51], and [58]. But this representation requires higher computational resources with more complicated frameworks such as GANs, and are hard to converge.

Instead of treating the trajectory and positions as a continuous variable and directly regressing values, the trajectory prediction can also be represented as a discrete variable. The output can be represented as: a) discretizing the frame scene into grids, and b) discretizing the pedestrian velocity into bins. Grid-based representations are used by studies such as [48] and [56]. Using a grid-based representation to encode the location information enables a parameter-free approximation of distributions, but the discretization over the whole scene may require high dimensionality. Therefore, grids are more often used for representing local occupation information for interaction with neighbors or the environment as in [27] and [49]. The trajectory prediction can also be treated as a classification task by quantizing the input data into classes and represented by one-hot encoding. Giuliari et al. [24] used 1000 bins to represent the velocity of pedestrians and predicted the future velocity by classification. But the authors claimed that the classification generally gets worse results than regression models because of quantization errors. In addition to predicting only future trajectories, some work outputs both destination and trajectory prediction [86], or outputs the pedestrians' walking behavioral response in each footstep [57].

3) Training Strategies: For trajectory prediction, mean square error (MSE), also called L2 loss, is commonly used, especially for position representations, as in studies [21], [22], [23], [24], [32], [35], [49], [56]. For uni-modal distribution representations, the negative log-likelihood loss is used, as in studies [27], [28], [39], [41], [52], [59], [62]. For the multi-modal distributions representations such as GAN-based models, the adversarial loss is used, together with L2 loss to measure the distance between generated samples and the ground-truth, as in studies [31], [40], [44], [50], [51]. Amirian et al. [33] also used information loss in addition to discrimination loss and adversarial loss. Eiffert et al. [58] used

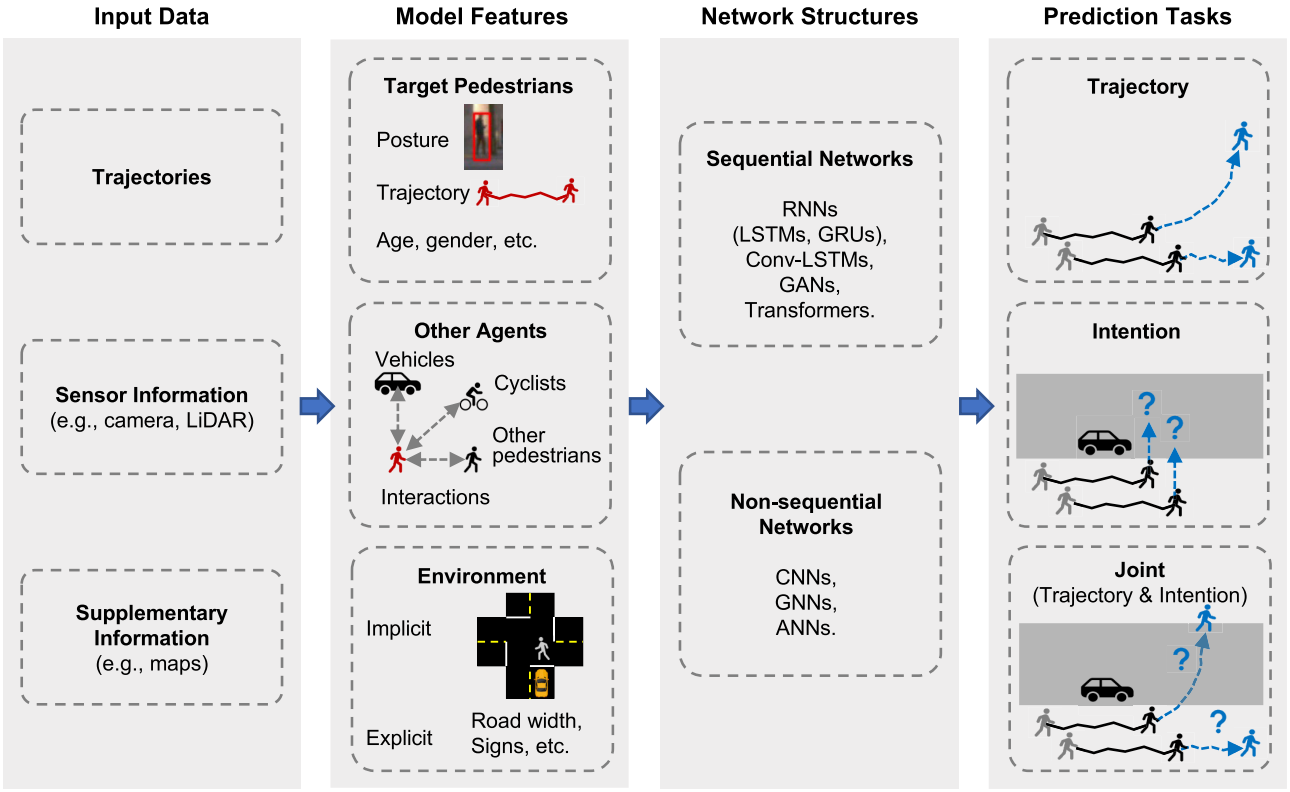


Fig. 2. Four criteria of the pedestrian behavior prediction system for categorizing existing studies. Input data are fed into the model. Model features are the stimuli of network structures. Different network structures are utilized to extract spatial and temporal information, and output different types of prediction tasks.

adversarial loss with the negative log-likelihood loss for the generator.

B. Intention Prediction

1) Task Definition: The intention prediction methods provide high-level information on pedestrian behavior. The intention or action can be predicted in different time horizons. Understanding and predicting the pedestrian intention, especially the crossing intention, is crucial for higher "Society of Automotive Engineers" (SAE) Levels aiming at automated driving. With the precise prediction of pedestrian intention in advance, automated vehicles can make better decisions and reduce the risk for potentially hazardous situations. Given the observed information of a pedestrian such as trajectories and postures, we aim to predict the intention of a pedestrian. The intention can be defined as discrete behavior types in the future. Many studies use "intention" interchangeably with "actual actions in the future", because labeling the "intention" of a pedestrian is usually a hard problem. Rasouli et al. [84] addressed and labeled intention by asking multiple annotation participants to observe the video of pedestrians and label the crossing intention, and then took the average. In this paper, we do not distinguish the intention and actual action.

2) Output Representation: The intention prediction is a classification problem. Many studies treated the problem as a binary classification with crossing or non-crossing (C/NC) action, such as in [64], [70], [75], [76], and [7]. Some other studies predicted multi-classification with several different action types. For instance, Fang et al. [69] predicted four types of behaviors including crossing, stopping, bending, and starting, using several binary classifications for multi-classification. Rasouli et al. [73] included four types of behaviors including walking, standing, looking towards the traffic, and not looking. Goldhammer et al. [81] classified pedestrians' motion states into waiting, starting, moving, and stopping. The multi-classification usually includes the whole process of crossing with a certain order, and contains more information.

3) Training Strategies: Rasouli et al. [73] used sigmoid cross entropy loss for classification. Besides, many studies used deep learning networks to extract features, and then use other machine learning classification methods. For example, the studies [69], [87] used SVM with hinge loss, and the studies [69], [70] used random forest (RF) for classification.

C. Joint Prediction

Pedestrian intention can be predicted jointly with trajectory prediction. There are mainly two kinds of joint prediction frameworks. One kind is that the trajectory and intention prediction tasks share the same feature extracting module. The extracted features are fed into two separate streams for different prediction tasks. For instance, Liang et al. [83] predicted both, the future positions (xy-coordinates) as well as estimating the possibilities of future activity labels simultaneously in one network. The trajectory generator and activity prediction modules share the features extracted from the images. In this framework, the trajectory and intention prediction share the same network, which can save computational resources.

Another kind is that the trajectory and intention are separately predicted, but the information is used to refine each other as suggested by Huang [80]. In works [81], [82], [84],

TABLE I
MODEL FEATURES AND INPUT DATA OF PEDESTRIAN BEHAVIOR PREDICTION

Prediction Tasks	Model Features			Input Data			Papers
	Target Pedestrians	Other Agents	Environment	Trajectory	Sensor Data	Supplementary information	
Trajectory (44 papers)	Trajectory	-	-	Yes	-	-	[20]–[26]
	Trajectory	Social interaction	-	Yes	-	-	[27]–[45]
	Trajectory, skeleton cue	Social interaction	-	Yes	Camera images	-	[46]
	Trajectory	-	Implicit	Yes	Camera images	-	[47], [48]
	Trajectory	Social interaction	Implicit	Yes	Camera images	-	[49]–[54]
	Trajectory	Person-ORU interaction	Implicit	Yes	Camera images	-	[55]
	Trajectory	Social interaction; Person-ORU interaction	Implicit	Yes	-	Scene image map	[40]
	Trajectory, motion states, destination	Social interaction	Implicit	Yes	-	Grid-based map, destination	[56]
	Trajectory, motion states, destination	Social interaction	Explicit	Yes	Camera images	Destination	[57]
	Trajectory, category	Social interaction; Person-ORU interaction	-	Yes	-	Agent Category	[58]–[60]
	Trajectory, category, direction	Person-ORU interaction	-	Yes	-	Agent Category	[61]
	Trajectory, velocity, agent shape	Person-ORU interaction	-	Yes	Camera images	Agents' states, traffic concentration	[62]
	Trajectory, appearance cue, VR information	Person-ORU interaction	Implicit	Yes	Camera images	Pedestrian VR information, vehicle's states	[63]
Intention (17 papers)	Motion states	Vehicle factors	Explicit	-	Lidar images	Static map	[64], [65]
	Trajectory	-	-	Yes	Lidar images	-	[66]
	Appearance cue	Vehicle factors	Explicit	-	Camera images	Vehicles' states	[67]
	Skeleton cue, motion states, individual information	Vehicle factors	Explicit	-	Camera images	-	[68]
	Skeleton and/or appearance cue	-	-	-	Camera images	-	[69]–[72]
	Appearance cue	-	Implicit	-	Camera images	-	[73]–[77]
	Appearance cue	Person-ORU interaction	Implicit	-	Camera images	-	[78]
	Speed, age, gender	Vehicle factors	Explicit	-	Lidar images, Camera images	Age, gender, environmental parameters	[7]
	Trajectory, skeleton and appearance cue	Vehicle states	Implicit	Yes	Camera images	Bounding boxes, the speed of ego-vehicle	[79]
Joint (6 papers)	Trajectory	-	-	Yes	-	-	[80]
	Trajectory, motion states	-	-	Yes	-	-	[81]
	Trajectory, skeleton cue, velocity	-	Explicit	Yes	Camera images	-	[82]
	Skeleton and appearance cue	Social interaction; Person-ORU interaction	Implicit	-	Camera images	-	[83]
	Trajectory, appearance cue	Vehicle factors	Implicit	Yes	Camera images	Bounding-boxes	[84], [85]

and [85], the researchers extracted features for the two tasks separately, and combined the two tasks based on the intention prediction results information to improve the trajectory prediction results. The combination of the two streams can then utilize more information to get better performance.

D. Summary of Prediction Tasks

The existing works for different prediction tasks are listed in Table I. We notice that there are more papers on trajectory prediction than the other two tasks. The application of different tasks is one of the reasons for this imbalance. The trajectory prediction can be used for many scenarios, not only for the automated vehicles in urban scenarios, but also for the development of social-aware robots in indoor scenarios, while the crossing intention prediction is mainly used for traffic scenarios. Therefore, there were more researchers from different research fields focused on trajectory prediction compared with intention prediction. There are other reasons related to the prediction methods and datasets that are used by these tasks. We discuss them in Sec. VI-C and Sec. VII-C.

IV. INPUT DATA

Previous models used various types of input data. The pre-processed data such as trajectories and raw sensor data

such as camera images can be used for training. Besides, other information such as the map and road parameters can be used to provide environment information. In this section, we classify previous studies based on the type of input data that provide different kinds of information.

Existing methods use one or multiple data sources as input to predict pedestrian trajectories:

1) Past trajectories, which can provide information of a pedestrian's motion state. It is used by most trajectory prediction methods.
2) Sensor data, such as the sequences of scene images recorded by the camera, and the point clouds recorded by LiDAR. The sensor data can provide more information of the pedestrian's posture and appearance, as well as provide the environmental context information.
3) Other supplementary information, including the pedestrian information (e.g., age and gender), the vehicle state (e.g., the speed and heading angle), and environment information (e.g., the road information and maps).

The input data of previous studies is shown in Table I, showing that different prediction tasks require different input data. The trajectory prediction requires trajectories as input. The trajectory can be labeled from either camera-recorded videos or LiDAR point cloud videos, or even generated from the simulation. In the studies that only require trajectories, raw sensor data is not required. For those trajectory prediction methods that require sensor data, the camera images and LiDAR point clouds can be used to provide visual behavior information. The intention prediction usually requires raw sensor data, that can provide visual or posture behavior cues for a pedestrian's intention. For joint prediction, both trajectory and raw sensor data can be utilized because this type of task requires both trajectory and visual behavior information. When the model needs the environment or other information, the supplementary information such as maps of the environment, the types of the object, and even virtual reality (VR) information can be required. With different types of input data, different features can be considered for modeling. More details about the model features are presented in Sec. V. As most of the existing studies used publicly available datasets for training and evaluation, we introduce more details about the sensors in Sec. VII-B for models that used raw sensor data.

V. Model Features

In this section, we categorize previous studies based on what features of pedestrian behavior have been considered in the model. Many factors can influence pedestrian behavior. Rasouli and Tsotsos [88] divided the factors that influence pedestrian behavior into pedestrian factors and environmental factors. Kotseruba et al. [85] analyzed the implicit and explicit factors that influence the pedestrians' crossing behavior, including the environment, communication with others, and their own states. Researchers consider one or several of these influencing factors as model features. In this paper, based on the internal and external stimuli of pedestrian behavior defined by Rudenko et al. [16] and the influencing factors mentioned in [85] and [88], we divide existing works by

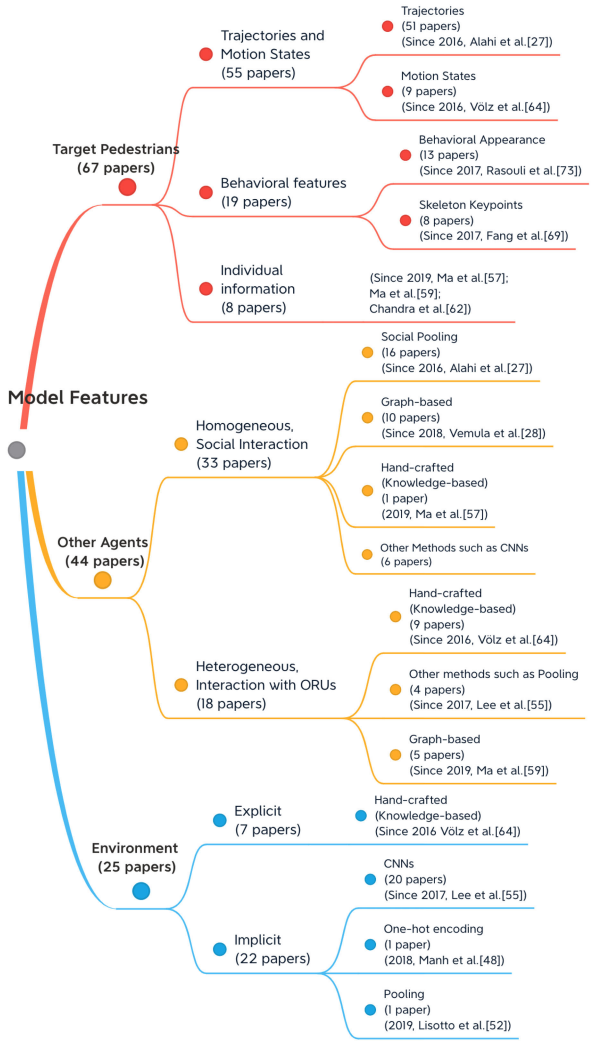


Fig. 3. The classification of the model features. The number of papers that use the corresponding features and the year that firstly used the factors/methods are listed. Please note that a paper can use **multiple** model features.

three types of model features. Fig. 3 shows the classification of model features, and the number of papers that used the corresponding features. Existing works use one or several combinations of these features:

1) The features related to *target pedestrians*, including trajectories and motion states, behavioral cues such as posture and appearance, as well as individual information such as the age and gender, etc.;
2) The features related to *other agents*, including homogeneous interaction, i.e., the social interactions between pedestrians; and heterogeneous interaction, i.e., the interaction between pedestrians and other road users (ORUs). Note that in this paper, we mean other types of road users except pedestrians when we say "ORUs";
3) The features related to the *environment*, including explicit factors and implicit interactions with context scenes.

A. Target Pedestrians

The states of target pedestrians are essential model features for predicting their future behaviors. A summary is listed in Table II.

TABLE II
INFORMATION OF TARGET PEDESTRIANS USED IN PREDICTION. PLEASE NOTE THAT A PAPER CAN USE **MULTIPLE** MODEL FEATURES

Target Pedestrian Information		Papers	Summary
Trajectories and Motion States (55 papers)	Trajectories (51 papers)	Trajectories only: Trajectory prediction: [20]–[24], [34] Intention prediction: [66] Joint prediction: [80]	**Advantages:** Contain the historical temporal information. The predicting models are usually simple and require less computing resources. **Drawbacks:** These models have not considered the interaction with other road agents and environment.
		Together with other factors: Trajectory prediction: [25]–[33], [35]–[63] Intention prediction: [79] Joint prediction: [81], [82], [84], [85]	**Advantages:** The models consider other features can get more accurate results. **Drawbacks:** The predicting models are complicated and require more computing resources.
	Motion States (9 papers)	Trajectory prediction: [56], [57], [61], [62] Intention prediction: [7], [64], [65], [68] Joint prediction [81]	**Advantages:** Provide simple but strong information. They are easy to get, and do not require complicated feature extraction and labelling. **Drawbacks:** These properties do not include other implicit information, and only related to current states. They are usually used together with other inputs.
Behavior features (19 papers)	Appearance-based (13 papers)	Trajectory prediction: [63] Intention prediction: [67], [72]–[79] Joint prediction: [83]–[85]	**Advantages:** The images can provide more information than just trajectories, the posture and appearance can reveal the future action. **Drawbacks:** These models require more powerful computing resources.
	Skeleton-based (8 papers)	Trajectory prediction: [46] Intention prediction: [68]–[71], [79] Joint prediction: [82], [83]	
Individual information (8 papers)	Category and type	Trajectory prediction: [58], [59], [61]	**Advantages:** These factors influence the pedestrian behavior and using them as features enables researchers to get accurate results. **Drawbacks:** Can be hard to get. If based on assumptions, it may not be precise.
	Others	Trajectory prediction: [63] Intention prediction: [7], [57], [62], [68]	

1) Trajectories and Motion States:

a) Trajectories: Most of the trajectory prediction models include the history of pedestrian trajectories, sometimes together with other model features. Trajectory prediction studies [20], [21], [22], [23], [24] considered only pedestrians' past trajectories for predicting their future trajectories. They extracted the features through embedding layers, and fed the features into deep learning structures for prediction. In addition to only trajectories, studies [25], [26], [44] encoded intermediate destinations from the trajectories and predict future trajectories conditioned on the destinations. For intention prediction, Zhao et al. [66] used trajectories extracted from roadside LiDAR sensors to predict the crossing intention. For joint prediction, Huang et al. [80] used trajectory to predict future intention and trajectory simultaneously, with the predicted results refining each other.

Although the context-based data are good indicators to include, the prediction can be faster by using only the trajectory as input. With recurrent networks, the temporal information of target pedestrians can be extracted from the trajectories, which usually provide rich historical information. There are several advantages of using only trajectories for prediction: It requires less annotation effort than annotating more semantic information on images, and the predicting framework is usually simple and requires less computing resources than those methods which consider the interaction with other road agents and with the environment. The drawbacks are that these methods have not considered the interaction with other road agents and the environment that could also affect the future behavior of pedestrians. The trajectories considered together with other model features usually take the trajectory as part of the input. These models extract the trajectory feature in an individual branch and utilize other compensation information resulting in higher accuracy.

b) Motion states: The motion states such as the velocity and position are also important features for human behavior prediction. For trajectory prediction, Ma et al. [57] focused on a microscopic level instead of estimating the positions at each time-step, and predicted the future trajectories by learning pedestrian's walking behavior at each footstep, considering the velocity and the step length as important inputs. Song et al. [56] also considered velocity as one of the features. Carrasco et al. [61] used orientation to build a graph representation for feature extraction. Chandra et al. [62] used position, velocity, and other factors as model features to define the state space of each road agent. For intention prediction methods, many studies [7], [64], [65], [68] included velocity to decide whether a pedestrian wants to cross the road or not. For the joint prediction, Goldhammer et al. [81] considered pedestrians' trajectory and velocity, as well as their ego-coordinate for prediction.

The motion states can provide simple but strong information about the moving behavior of pedestrians. The velocity and position information is easy to get, and does not require complicated feature extraction and labeling. However, these properties do not include other implicit information, and are only related to states at the current time step. Therefore, the motion states are usually used together with several other inputs as complements.

2) Behavior Features: As proposed by Schmidt and Färber [89], using only trajectory information for intention prediction is insufficient. The behavioral features, especially the appearance and posture, usually indicate a pedestrian's intention, and are used by many intention prediction and joint prediction works. The CNNs are usually used to extract visual cue information and/or get the key-point features of pedestrians. The behavioral information from the images can provide more behavior information of pedestrians than just trajectories, but requires more powerful computing resources.

a) Appearance-based: For intention prediction, the appearance behavioral feature can be extracted implicitly from images, usually using CNNs [72], [73], [78]. Three-dimensional CNNs (3D-CNNs) have been utilized to extract spatio-temporal features and recognize pedestrians' crossing intentions in [67], [75], and [76]. For the joint prediction, Rasouli et al. [84] and Kotseruba et al. [85] used separate streams for intention estimation, which extracts posture features from the local context and appearance with CNNs.

b) Skeleton-based: The postures of pedestrians are strong behavioral cues that can indicate their intentions. The postures can be represented and estimated by skeleton keypoints using pre-trained CNN-based networks, as in For intention prediction studies [46], [68], [69], [70], [71], [79] and joint prediction studies [82], [83]. The hourglass network [90] and OpenPose [91] are utilized to extract pose features.

3) Individual Information: Category, destination, and agent shape/size, age, gender, and the theory of mind information are considered in many existing papers. The trajectory prediction models that involve multi-agents [58], [59], [61] required the category of the target pedestrians in prediction. Ma et al. [59] and Carrasco et al. [61] used the category and coordinates of the agent as vertex information. Both methods build graph representations of the instances, and consider all types of agents in traffic, that can also be used as pedestrian predictors. In work [58] denotes the vehicle and pedestrian type, and used the information for vehicle-human interaction, which can be explained in detail in the following sections.

Individual information such as age and gender can provide supplementary information for pedestrian behavior prediction, and they are significant factors that influence pedestrian behavior [92], [93]. For intention prediction, age and gender are included as important model features in work [7] to provide necessary human factors-related information. Ma et al. [57] assumed the destination is a vertical line of the crossroad, and used the distance from the destination to the target pedestrian as input features. Chandra et al. [62] also considered the road agent's shape and size as implicit constraints in the trajectory prediction. Kim et al. [63] proposed the multiple stakeholder perspective structure (MSPM) that considered the information not only from the driver's view using sensors mounted on a vehicle, but also included the information from the pedestrian's view using VR devices. These individual factors can influence pedestrian behavior and enable the researchers to get more accurate results using them as model features. However, compared to the trajectories and images, many factors are much harder to get. The destination, age, and gender usually require questionnaires or additional annotation. Otherwise, they can be based on assumptions or output from previous perception modules but may not be precise enough.

B. Other Agents

In this section, we discuss the influence of the other agents on pedestrians' behaviors. The information and interaction with other agents are included by 65% of existing papers that we reviewed. A summary is presented in Table III.

1) Homogeneous - Social Interaction Between Pedestrians: According to Moussaid et al. [8], pedestrians' future behavior is not only dependent on their past states, but also driven by social interactions with other pedestrians nearby. Social interaction is an important factor for modeling pedestrians' future trajectories.

a) Hand-crafted features: For trajectory prediction, Ma et al. [57] used hand-crafted features to model the social relationship between pedestrians. They utilized relative positions and relative velocities between the pedestrian and the seven nearest neighbors in front of the target pedestrian as input features. While these hand-crafted features succeeded in this task, they are often hard to generalize to new scenarios. Therefore, deep learning methods are developed to be more powerful structures for extracting social interactions.

b) Social pooling and its variants: Social-LSTM [27] modeled social interactions in a learning-based approach for trajectory prediction. Instead of using knowledge-based methods as in social force [12], the authors proposed a social pooling layer over the hidden states of LSTMs to model the interactions between pedestrians. Several works including [29], [30], [32], [44], [46], [49], and [52] followed the social pooling trend and improved the interaction mechanism by attention pooling using various attention mechanisms. Fernando et al. [29] improved the social pooling module with a soft and hard-wired attention mechanism. Xu et al. [30] utilized a weighted spatial affinity function with calculated weights to determine the social interactions over the spatial features. Zhang et al. [32] proposed a state refinement module for future predictions. Sophie [51] assumed that people pay more attention to closer objects and sorted the attention by distance.

Later works [31], [33], [45] improved the interaction module with a more complicated pooling structure. Social-GAN [31] pointed out that local interaction information is not always sufficient, and hence, they use a multi-layer perceptron (MLP) followed by a max-pooling structure to capture the global social interaction information. Amirian et al. [33] improved the interaction module by using an attention pooling that relies on hand-crafted interaction features inspired by neuroscience and biomechanics. Zhang et al. [45] proposed the Social Interaction Extractor to learn interaction weights with a sub-network structure. Kothari et al. [39] categorized the existing interaction module into grid-based methods and non-grid-based methods, and proposed a grid-based directional pooling method and the DirectConcat method that achieved improvement. Bhujel et al. [53] calculated the social attention from the hidden state with designed physical and social attention functions. Col-GAN [36] proposed an attention module that used MLPs to learn the interaction and used a weighted sum to calculate the interaction feature.

The social pooling module enables the existing work to consider social interaction. The structure is simpler than the graph-based models with fewer parameters to learn.

c) Graph-based representation: The symmetric pooling (max or average pooling) operation assumes that the interactions between pedestrians are symmetric, which, however, is not always the case. To extract non-symmetric interactions, researchers use a graph to represent the relationship between pedestrians. In such graphs, the vertices represent the states of

TABLE III
INFORMATION OF OTHER ROAD USERS (ORUS) USED IN PEDESTRIAN BEHAVIOR PREDICTION

Other Agent Information		Papers	Summary
Homogeneous - Social Interaction Between Pedestrians (33 papers)	Hand-crafted features (1 paper)	Trajectory prediction: [57]	**Advantages:** Explainable. **Disadvantages:** Hard to generalize to new scenarios.
	Social pooling and its variant (16 papers)	Trajectory prediction: [27], [29]–[33], [36], [39], [44]–[46], [49], [51]–[53], [56]	**Advantages:** The social pooling module considers social interaction in the model. It is relatively simple compare with graph based models. **Disadvantages:** They mainly deal with symmetric interactions.
	Graph-based representation (10 papers)	Trajectory prediction: [28], [34], [35], [37], [41] [40], [50], [54], [58], [59]	**Advantages:** They can extract non-symmetric interactions. **Disadvantages:** The construction of the graph is computational- and time-consuming.
	Other methods (6 papers)	Trajectory prediction: [38], [42], [43], [56], [60] Joint prediction: [83]	**Comments:** The social norm is considered using sampling methods. Agent-aware attention and LSTMs are used to model social and time dimensions simultaneously. CNNs are applied on grid-based map.
Heterogeneous - Interaction with Other Road Users (18 papers)	Hand-crafted features (9 papers)	With single vehicle: Trajectory prediction: [63] Intention prediction: [7], [64], [65], [67], [79] Joint prediction: [84], [85] Vehicle volume: Intention prediction: [68]	**Advantages:** Explainable. **Disadvantages:** Can be hard to generalize to new scenarios.
	Graph-based representation (5 papers)	Trajectory prediction: [40], [58], [59], [61], [78]	**Comments:** The graph based module can extract non-symmetric interactions between pedestrians and other road users.
	Other methods (4 papers)	Trajectory prediction: [55], [60], [62] Joint prediction: [83]	**Comments:** Grid-based pooling, CNNs, One-hot coding, and reinforcement learning can be combined into the network.

the pedestrians and the edges represent the spatial or temporal relationships between pedestrians.

For trajectory prediction, Vemula et al. [28] represented the social attention by a spatio-temporal graph representation, using soft attention with calculated weights over hidden states of each node. Zhang et al. [34] used the graph representation and applied the social graph network directly on MLP embedded features from agents' locations and velocity status. STGAT [35] and Social-BiGAT [50] applied the graph attention networks (GAT) as proposed by Veličković et al. [94] to extract the social interactions between pedestrians over the hidden states of LSTMs. STGAT [35] calculates the relationship for each time step to get the state of pedestrians, while Social-BiGAT [50] calculates the interaction after extracting the hidden states from all observed time steps. Hu et al. [40] proposed an interaction branch with a graph structure, namely neural motion message passing (NMMP), which calculates k times the interacted actor embedding with graph neural network on the hidden states of each agent. Yu et al. [37] exploited a spatio-temporal graph transformer (STAR) to model the spatio-temporal interaction between pedestrians. Social-STGCNN [41] and STGT [54] used graph convolutional networks (GCNs) [95], which are defined as convolution operations over graphs to extract the spatio-temporal social interaction feature.

The graph-based module can extract non-symmetric interactions and get better results than the pooling structures, but the instruction of the graph takes more computational resources, and hence, can be more time-consuming.

d) Other methods: For the trajectory prediction, Social-NCE [38] considers unfavorable events like discomfort and collision situations when learning socially aware motion representations. The authors proposed a safety-driven sampling method, called the multi-agent contrastive sampling, to select negative samples from the neighborhood of other agents in the future. Yuan et al. [42] proposed AgentFormer that can simultaneously model the time and social dimensions using an agent-aware attention mechanism. Tra2Tra [43] proposed a spatial-temporal attention module, that embedded the spatial feature from the coordinates of all pedestrians, and used an LSTM network to extract the temporal dependency between spatial features. Song et al. [56] considered the target pedestrian's neighbors by considering their neighbors' speed. The speed is filled in cells of a grid-based map, and CNNs are used to extract the spatial relationship with the neighbors.

2) Heterogeneous - Interaction With Other Road Users (ORUs): The future behavior of pedestrians is influenced by the interaction with ORUs such as vehicles according to Shirazi et al. [9].

a) Hand-crafted features: In Schmidt and Färber's research [89], parameters such as the distance and velocity of the vehicles can influence the crossing intention. For the intention prediction, many researchers used hand-crafted features as inputs, such as in studies [7], [64], [67], [79], including vehicle's velocity or speed, relative velocity and distance between the pedestrian and vehicle, or time to collision (TTC). Zhang et al. [68] used vehicle-related information for crossing intention prediction, including vehicle volume, the green light time for vehicles, and the number of vehicles. For the joint prediction, Rasouli et al. [84] and Kotseruba et al. [85] utilized the ego-vehicle information including the speed and heading angle as complementary inputs.

b) Graph-based representation: As the interaction between different types of traffic agents is usually non-symmetric, graph-based methods can model heterogeneous interactions. For the trajectory prediction, Eiffert et al. [58] proposed a graph vehicle-pedestrian attention network (GVAT) to include both human-human interactions and human-vehicle interactions. Ma et al. [59] used a 4-dimensional graph that consists of the instance layer and the category layer to

represent the traffic sequence and to calculate their interaction. The instance layer represents the individual interaction, while the category layer ensures the motion pattern of different categories. Hu et al.'s framework [40] can jointly predict the trajectory of pedestrians and vehicles by the proposed NMMP module with a graph representation. Carrasco et al. [61] built the graph with coordinates, categories, headings as vertices, and exploited the graph attention layer to include the interaction. For the intention prediction, Liu et al. [78] captured the interaction between the pedestrians and other road users using graph convolution to include both spatial and temporal context.

c) Other methods: For trajectory prediction, Lee et al. [55] modeled the interaction for multi-agents with a spatial grid-based pooling layer, which is similar to the social-pooling layer. Chandra et al. [62] took sequences of images as input to predict the trajectories of heterogeneous traffic agents including pedestrians, using CNNs for extracting the appearance and behavioral information of different road agents. Li et al. [60] considered the existence of a vehicle, and combined reinforcement learning into the prediction. For joint prediction, Liang et al. [83] modeled the interaction between pedestrians and other road users in the scene by explicitly modeling the geometric relation with a knowledge-based function defined by the authors that considered the geometric distance and the box size, and modeled the object type using one-hot encoding.

C. Environment

The interaction with the environment also influences pedestrians' behaviors. The environmental information is included by 36% of the existing papers that we reviewed. To include the interactions with the environment scene as model features, studies either took explicitly defined environment features as inputs, or use sequences of camera images or a navigation map to learn the pedestrians' interaction with the surrounding environment implicitly. In this section, we present how the researchers address pedestrian-environment interactions. The summary is shown in Table IV.

1) Explicit Features: Explicit features are manually defined and usually explainable. Many researchers utilize information about zebra crosswalks and the curbs. For trajectory prediction, Ma et al. [57] used the distance to the left and right boundaries of the crossroad as input features for the prediction. For intention prediction, the distance between the vehicle and the crosswalk, the distance between the pedestrian and the crosswalk, and the distance between the pedestrian and the curb are important factors and were used by Völz et al. [64] and Zhang et al. [7] as model features. Yang et al. [67] considered the existence of stop signs, zebra markings, and traffic lights in local traffic scenes. They used the prior weight to represent different scenes. In addition to the geometry-related environment features, Zhang et al. [68] used temperature as an important factor to predict the crossing intention at red-light. For joint prediction, Wu et al. [82] used the crossable information at a crossroad to change the sampling weight when predicting the trajectory.

2) Implicit Features: The implicit features are not explicitly defined and are usually extracted from images or semantic maps.

a) CNN-based feature extractor: As CNNs are capable of extracting image features, pre-trained CNNs can be used to extract appearance features and implicit human-scene interaction features from a sequence of images. The trajectory prediction works [47], [49], [50], [51], [54], [55] followed this direction using pre-trained CNNs. Bhujel et al. [53] utilized CNN features and a physical attention function to learn the probability that a location is the right place to focus for predicting the next position. Instead of sensor image data, Hu et al. [40] used a 2D bird's-eye-view scene image map as input to provide prior knowledge about the traffic condition and rules, and extracted the environment information with the CNN structure [96] to extract scene embedding. Song et al. [56] used a grid-based map with occupied cells to indicate the fixed obstacles in the scenes using CNNs to extract the environmental features. For intention prediction, Rasouli et al. [73], Hoy et al. [74], and Kotseruba et al. [79] used CNNs to extract visual context features implicitly. Liu et al. [78] segmented the images into pedestrians and objects with binary masks using a segmentation model [97]. Then, they captured the context feature by encoding the segmented binary masks with the ResNet backbone. Works [75], [76] utilized 3D-convolutional networks for image feature extraction in the observed time period. For joint prediction, Liang et al. [83] used a pre-trained scene segmentation model [98] for environmental feature extraction. The integer scene semantic features are transformed into binary masks, then two convolutional layers are applied to the mask features to get CNN features. Rasouli et al. [84] and Kotseruba et al. [85] used CNNs to extract the local visual context around the pedestrian with a bounding box implicitly along with the appearance feature.

b) Other methods: For trajectory prediction, Scene-LSTM [48] takes the scene information into consideration by using grid cells to represent the input scene image. The calculated hidden states of each grid cell are used as input to a scene data filter to pass the scene constraints information to get better trajectory prediction results. Lisotto et al. [52] utilized a semantic map and the navigation map, and applied semantic and navigation pooling to extract the environmental interaction feature. The semantic map, which contains the scene context, is generated from the image using semantic segmentation, and the navigation map which embodies the most frequently crossed areas is generated from the observed data by counting the crossing frequency of squared patches.

D. Summary of Model Features

Model features play important roles in pedestrian behavior prediction. As we summarized in Fig. 3, a method can use multiple model features. For the target pedestrians, the trend is also to include more information. In 2016, the trajectories and motion states are included [27]. In 2017, the behavioral features are included [69], [73], and in 2019, the individual information are added [57], [59], [62]. For the interaction with

TABLE IV
INFORMATION OF ENVIRONMENT USED IN PEDESTRIAN BEHAVIOR PREDICTION

Environment	Descriptions	Papers	Summary
Explicit (7 papers)	Hand-crafted features	Trajectory prediction: [57] Intention prediction: [7], [64], [65], [67], [68] Joint prediction: [82]	**Comments:** They are manually defined, simple and usually explainable.
Implicit (22 papers)	CNN based feature extractor (20 papers)	Trajectory prediction: [40], [47], [49]–[51], [53]–[56] Intention prediction: [63], [73]–[79] Joint prediction: [83]–[85]	**Comments:** CNNs are capable of extracting image features, and can be used to extract the interaction between pedestrians and the environment implicitly.
	Others (2 papers)	Trajectory prediction: [48], [52]	**Comments:** One-hot encoding and pooling can be used to encode the location information. But when encoding location information with one-hot vectors, the dimensionality might become very high.

other agents, the social interactions are included mainly in trajectory prediction. The social pooling methods was proposed in 2016 [27], and the graph-based model was proposed in 2018 [28]. In 2019, Ma et al. [57] added knowledge-based information to model the interaction. The interaction with other road users such as vehicles is included mainly in the intention prediction works. In 2016, researchers started to use hand-crafted features to model the interaction [64]. in 2017, the learning-based feature extractor such as pooling method [55] and graph-based methods [59] are proposed. For the environment feature, researchers first model it with hand-crafted features explicitly in 2016 [64], then used a CNN-based model to learn it in 2017 [55]. In 2018 and 2019, other attempts on one-hot encoding [48] and pooling [52] are tried by researchers. The hand-crafted features used in existing works are explainable but hard to generalize, while the learning-based features have achieved more accurate results but are difficult to explain. Future works can focus on how to combine these features.

VI. NETWORK STRUCTURES

In this section, we list commonly used network structures, and classify them into sequential networks and non-sequential networks. These structures can be combined to form a prediction model. For instance, a model can use CNNs for extracting visual information, and use LSTMs for temporal prediction. Fig. 4 shows the classification of the network structures. Table V presents the summary of the network structures used by existing research.

A. Sequential Networks

The sequential networks typically deal with time-series information by assuming the moving state at one time step is conditionally dependent on previous states. Traditional models used for predicting the pedestrian's future action such as hidden Markov models (HMM) [99], [100], partially observable Markov decision processes (POMDP) [101], and Gaussian processes [5], [102], [103] require accurate and precise segmentation and tracking of pedestrians. However, this is challenging due to the difficulty of extracting reliable image features as outlined by Völz et al. [64]. With the help of deep learning, the models are able to extract features from images with CNNs and to extend the long-term memory with Recurrent neural networks (RNNs) including long short-term

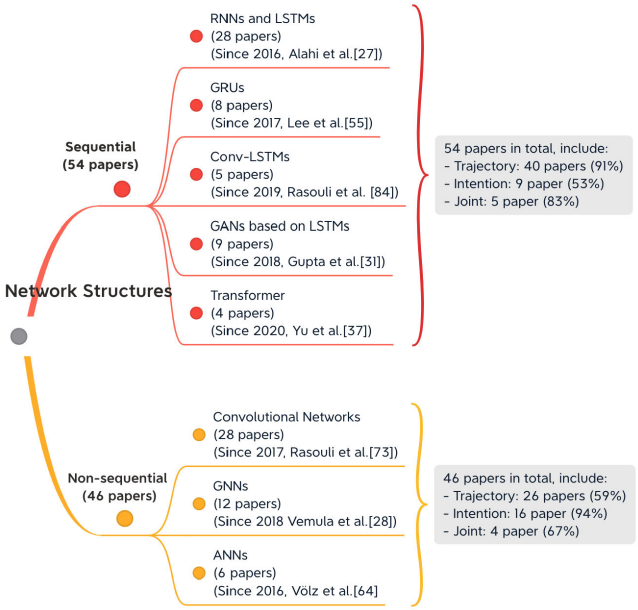


Fig. 4. The classification of the network structures. The number of papers that used corresponding methods and the year that firstly used the network structures are listed. Please note that a paper can use **multiple** network structures. For example, a model can use CNNs for extracting the visual information, and use LSTMs for the temporal prediction. The distribution of the papers is summarized in the boxes on the right side.

memory (LSTMs) and gate recurrent units (GRUs), convolutional LSTMs (Conv-LSTMs), and transformer networks (TFs) to overcome the limitation of traditional models.

1) Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs): RNNs and their improved version, LSTMs are preferred by many researchers because of their strong ability to handle the trajectory sequence information. For trajectory prediction, Vemula et al. [28] used spatio-temporal graph within the RNN structure. Alahi et al. [27] utilized LSTMs to learn the motion state of a pedestrian and proposed Social-LSTM model to predict a pedestrian's trajectory. Later trajectory prediction methods such as [20], [21], [22], [29], [30], [32], [34], [35], [39], [43], [48], [49], [52], [53], [59], and [62] followed this trend of using LSTM-based methods to cope with time-series information.

For intention prediction, Zhang et al. [7] used LSTMs with the attention mechanism for prediction that outperforms the SVM model. Pop et al. [77] proposed a multi-task network that combines the CNNs for extracting visual features and the LSTM network for estimating the time to cross the street. The FuSSI-Net proposed by Piccoli et al. [71] used a CNN-based

TABLE V
NETWORK STRUCTURES FOR PEDESTRIAN BEHAVIOR PREDICTION. PLEASE NOTE THAT A PAPER CAN USE **MULTIPLE** NETWORK STRUCTURES

Network Structures (Earliest Used Time)		Papers	Summary
Sequential Networks (54 papers)	RNNs and LSTMs (Since 2016, Alahi et al. [27])	Trajectory: RNNs: [28]; LSTMs: [20]–[22], [27], [29], [30], [32], [34], [35], [38], [39], [43], [48], [49], [52], [53], [59], [62], [63], [84] Intention: LSTMs: [7], [71], [72], [77] Joint: LSTMs: [80], [82], [83]	**Advantages:** RNNs (including LSTMs, GRUs) are more capable to handle long term prediction than the traditional models. **Disadvantages:** They cannot be parallelized, and cannot handle too long sequences.
	GRUs (Since 2017, Lee et al. [55])	Trajectory: [25], [26], [47], [55] Intention: [74], [78], [79] Joint: [85]	
	Conv-LSTMs (Since 2019, Rasouli et al. [84])	Trajectory: [56], [63] Intention: [75], [76] Joint: [84]	**Advantages:** Conv-LSTMs can extract spatial and temporal features simultaneously. **Disadvantages:** The computational cost is higher than for LSTMs.
	GANs (Since 2018, Gupta et al. [31])	Trajectory: [31], [33], [36], [40], [46], [50], [51], [58], [60]	**Advantages:** The GANs as generative models can predict multiple plausible trajectories. **Disadvantages:** Hard to train, and requires techniques for convergence.
	Transformers (Since 2020, Yu et al. [37])	Trajectory: [24], [37], [42], [54]	**Advantages:** They can handle long sequences and allow parallelization. **Disadvantages:** Implemented with a fixed-length, not flexible enough.
Non-sequential Networks (46 papers)	Convolutional Networks (Since 2017, Rasouli et al. [73])	Trajectory: [23], [40], [41], [45], [47], [49]–[51], [53]–[56], [61] Intention: [46], [67]–[73], [75], [76], [78], [79] Joint: [83]–[85]	**Comments:** CNNs can be used for both extracting spatial features and sequential features. For the sequential prediction, as there is not dependency of the previous time steps, the prediction error do not accumulate like the RNNs, and it allows parallel computation.
	GNNs (Since 2018, Vemula et al. [28]	Trajectory: [28], [34], [35], [37], [40], [41], [50], [54], [58], [59], [61] Intention: [78]	**Comments:** GNNs can be used for extracting non-symmetric interactions and capturing spatio-temporal features.
	Other ANNs (Since 2016, Völz et al. [64])	Trajectory: [44], [57] Intention: [64]–[66] Joint: [81]	**Advantages:** Structures are simple; can handle the non-linearity. **Disadvantages:** For 2D image input, ANNs will lose the spatial information, and require a huge amount of trainable parameters. For sequential input, ANNs cannot capture sequential information.

network for detection and skeleton keypoints extraction, and then used LSTMs to extract temporal information. For joint prediction, Huang et al. [80] proposed warp LSTM to deal with neighboring time steps in place of global positions and to allow for long-term trajectory prediction. They proposed the mutable intention filter to generate potential intentions, and then predicted the intention-aware trajectories. Lorenzo et al. [72] employed CNNs to extract pedestrians' behavioral features and applied various RNNs including LSTMs, GRUs, and the bidirectional variants of LSTMs and GRUs for crossing probability prediction. Kim et al. [63] proposed the MSPM model, that includes a driver perspective network and a pedestrian perspective network. The driver perspective network used LSTMs to encode the speed and trajectory information of the driver's perspective and other structures for image feature extraction, and used LSTMs to predict a pedestrian's behavior.

For joint prediction, Liang et al. [83] extracted the feature with CNNs, and then the extracted features are fed into a trajectory generator and activity predictor separately. In the trajectory generator, LSTMs are used for sequence prediction, while in the activity predictor, two separate convolution layers are used on a multi-scale Manhattan Grid for classification and regression to predict the label and location. Wu et al. [82] first extracted skeleton features with CNN-based methods and then used LSTMs to predict behavior classes (i.e. standing, walking, running), and used the dynamic Bayesian network to identify crossing intention. The predicted intention information is used for deciding the weights for trajectory sampling to improve the results. Rasouli et al. [84] used LSTMs in the pedestrian trajectory and vehicle speed prediction stream, and used LSTMs together with other structures in the intention estimation stream.

2) Gate Recurrent Units (GRUs): GRUs are another improved version of RNNs that are also popularly used for sequential prediction. For intention prediction, Hoy et al. [74] explored a variant of variational recurrent neural networks (VRNNs), namely the deep variational Bayes filters [104] for extracting tracking features, using GRU layers in VRNN cells with CNNs' extracted visual features as inputs. Liu et al. [78] used GRUs for behavior prediction after using a CNN-based segmentation model [97] for appearance feature encoding. Kotseruba et al.'s later work [79] used 3D-CNN for local visual context extraction and used GRUs for non-visual features encoding from bounding boxes, poses, and ego-vehicle speed. For joint prediction, Kotseruba et al. [85] employed GRUs for trajectory prediction, connected with the intention feature extracted from images using CNNs, and fed into a fully connected layer for future action classification.

GRUs can be combined with generative models for pedestrian trajectory prediction. These multi-modal models can provide multiple feasible results by incorporating prior knowledge into pedestrian behavior learning. Recently, conditional variational autoencoders (CVAEs) with sequential encoders and decoders have been adopted to predict multi-modal distributions. The BiTraP [25], SGNet [26], CGNS [47] and

DESIRE [55] used GRU encoder-decoders based on CVAE method for trajectory prediction with multi-modal goal estimation. Social-NCE [38] applied LSTM model based on the noise-contrastive estimation (NCE) methods [105] by introducing a social contrastive loss, namely the InfoNCE loss [106].

RNNs and their variants including LSTMs and GRUs use hidden states to represent the time-varying motion properties. They are more capable of dealing with long-term prediction than traditional models because of their capability of learning the dependencies between temporally correlated data. However, the sequential computation of RNN-based models inhibits parallelization. Besides, the networks cannot do well for long sequences because the "temporal distance" between two sample positions is linear, and the network tends to "forget" the information of the previous sample in the sequence. Furthermore, it is hard to explain the physical meaning of the hidden states that represent the moving states.

3) Convolutional LSTMs (Conv-LSTMs): Conv-LSTMs as proposed by Shi et al. [107] have been used to extract spatial and temporal information. For trajectory prediction, Kim et al. [63] used CNNs, Conv-LSTMs, and LSTMs for encoding image information in the driver perspective network. Song et al. [56] used a grid-based map with social and scene information filled in the cells, and used deep conv-LSTM to predict the future trajectories. For intention prediction, Gujjar et al. [76] and Chaabane et al. [75] used 3D-CNN layers as the encoder and conv-LSTM layers as the decoder in their encoder-decoder structure. For joint prediction, Rasouli et al. [84] proposed the PIE model that used LSTMs, CNNs, and conv-LSTMs for prediction. In the intention estimation stream, CNNs are used for appearance behavioral feature extraction with conv-LSTMs as the encoder, and LSTMs as the decoder.

4) Generative Adversarial Networks (GANs) Based on LSTMs: The previously mentioned models follow a uni-modal distribution. As there could be multiple socially acceptable trajectories, Gupta et al. [31] proposed Social-GAN, which assumed that the pedestrian trajectories follow a multi-modal distribution, which means that multiple future trajectories are potentially plausible. They utilized the GANs with an LSTM-based generator for trajectory prediction. Social-BiGAT [50] and studies [33], [40], [46], [51], [58] followed this trend and used LSTMs as generators of the GANs, with various structures of extracting the interactions with other objects. Li et al. [60] utilized Social-GAN and combined it with reinforcement learning in their prediction. The Col-GAN [36] used a GAN structure with LSTM encoder-decoder as the generator. But instead of using an LSTM-based discriminator like Social-GAN [31] and Sophie [51], they used CNNs as the discriminator and classify the segments of a trajectory are real or fake.

The GANs can predict multiple plausible and socially acceptable trajectories given a partial history instead of predicting only one "average behavior". The drawback of the GANs is that they are usually hard to train and require techniques to make the model converge.

5) Transformer Networks (TFs): The TFs [108] can alleviate the previously mentioned problems of RNN-based models. The TFs used the attention mechanism to help memorize the information in long sequences. The attention mechanism can create shortcuts between the context vector and the entire source input instead of only the last hidden state. TFs made ground-breaking progress recently in the Natural Language Processing domain and are becoming popular to be adopted for predicting pedestrian behaviors because of their capability of long-term prediction. Giuliari et al. [24] adopted both, the original TF and bidirection transformer (BERT) for trajectory prediction. The authors considered only the individual trajectory as model features yet still gained better performance than previous LSTM- and CNN-based methods. Yu et al. [37] further considered social interaction using graph-based representation to achieve more accurate results. The AgentFormer [42] applied the agent-aware transformer in a multi-agent trajectory prediction framework based on CVAE and modeled the future trajectory distribution conditioned on past trajectories and contextual information. Syed et al. [54] proposed the STGT model that used a CNN model (PSP-Net [109]) for segmentation and extracting the image environmental features, and the transformer is used for sequence prediction.

The TFs avoid recursion and allow parallel computation to reduce training time. With the attention mechanism, the TFs get more accurate results than RNNs. However, the transformers are implemented with a fixed length, and cannot model dependencies that are longer than the fixed length. Some other improvement versions of TFs such as the TransformerXL [110] and the compressive Transformer [111] could be used in the trajectory prediction or other sequence prediction tasks.

B. Non-Sequential Networks

The non-sequential networks are used to extract spatial and interaction features. Besides, they can also model the temporal information by directly modeling the final state or distribution over the entire history of observed states without the assumption of conditional dependency on previous states.

1) Convolutional Networks: CNNs are used in many models to extract implicit appearance features from images as discussed in Sec. V. Trajectory prediction studies used pre-trained CNNs to extract implicit features of the environment as in [40], [47], [49], [50], [51], [53], [54], [55], and [56]. Intention prediction studies used CNNs to extract appearance behavioral features ad in [72], [73], [78], and [79], and skeleton behavioral features as in [46], [68], [69], [70], [71], and [79]. 3D-CNNs are used to extract spatio-temporal features as in [67], [75], and [76]. For joint prediction, CNNs are used to extract posture features as in [84] and [85] and environment features as in [83].

In addition to extracting spatial features from images, CNNs can also be used to extract sequential features for pedestrian trajectory prediction. Many methods use hidden states of LSTMs to represent the pedestrian motion states. However, Nikhil and Morris [23] pointed out that trajectories are continuous in nature and do not have a complicated "state". The feature extraction of hidden states in previous models is indirect and the physical meaning of hidden states is difficult to interpret. Bai et al. [112] noticed that recurrent architectures

have limitations in inefficient parameters and the training can be inefficient. Therefore, instead of using LSTMs, Nikhil and Morris [23] proposed an algorithm using CNNs to predict the trajectories for computational efficiency, which yields competitive results with a faster speed. Lea et al. [113], [114] proposed temporal convolutional networks (TCNs) that dealt with time series and extracted features by convolutional layers on the temporal dimension. Mohamed et al. [41] proposed the Social-STGCNN model, which reached faster speed and better results on trajectory prediction, by using TCNs to extract spatio-temporal features from the spatial and social interaction features, and utilized CNNs as an extrapolator on the time dimension. Zhang et al. [45] proposed the Social-IWSTCNN, which followed the trend of using CNNs and TCNs for prediction. The convolutional-based methods enable parallelization and without the dependencies on the previous time step, the prediction can be faster and the prediction errors do not accumulate like with RNNs.

2) Graph Neural Networks (GNNs): GNNs are neural networks over graph-represented data. GNNs have achieved significant success in human action recognition [115], [116], [117]. GNNs can be used in pedestrian behavior prediction for extracting spatial and temporal interaction between pedestrians and other objects and are especially suitable for modeling non-symmetric interactions and spatio-temporal features as mentioned in Sec. V.

Graph convolutional networks (GCNs) proposed by Kipf and Welling [95] define the convolution operations over graphs. Social-STGCNN [41], STGT [54] use GCNs to extract the spatio-temporal social interaction features for trajectory prediction. Liu et al.'s [78] used GCNs captured the interaction between pedestrians and other road users using graph convolution to include both spatial and temporal context. In particular, the graph attention networks (GATs) as proposed by Veličković et al. [94] improved weighted message passing between nodes and are applied by STGAT [35], Social-BiGAT [50] and studies [58], [61]. Yu et al. [37] improved GAT by applying a transformer boosted attention mechanism and proposed spatio-temporal graph transformer (STAR) model. These methods model the interaction not only based on the current frame but also consider the influence of other time steps. Besides, commonly used network structures can be applied to graph representations. For trajectory prediction, Hu et al. [40] proposed a neural motion message passing (NMMP) structure, which used MLP embeddings to pass messages between nodes and edges. Zhang et al. [34] proposed the social graph network that applied a one-layer MLP on egde and nodes of a graph. Vemula et al. [28] applied structural RNN [118] on edges and nodes of spatio-temporal graphs to model the spatio-temporal interaction between pedestrians. Ma et al. [59] applied LSTM on the nodes of a 4-dimensional graph to model the interaction of different instances and categories.

3) Other Artificial Neural Networks (ANNs): For the trajectory prediction, Ma et al. [57] used an ANN with hidden layers to model the mechanism of decision-making that employed human experience to make the approach more realistic for the prediction of microscopic pedestrian walking behavior. For the intention prediction, Völz et al. [64] designed a dense neural network using 15 hand-crafted features over five time steps, and the dense network outperformed the LSTM and SVM methods. Zhao et al. [66] compared the intention prediction with Naive Bayes methods using trajectories as input, and claimed the results of ANN is worse than the Bayes methods. This may be because they only include the trajectories as inputs, which is too simple to demonstrate the power of neural networks, and other networks such as RNNs can be used for sequence prediction and CNNs can be used for image inputs. CVAEs can also be combined with ANNs. PCENet [44] considered the intermediate stochastic destinations of the pedestrians into prediction by using an endpoint CVAE, where the prediction is conditioned on the features extracted from the past encoder using MLPs. For joint prediction, Goldhammer et al. [81] proposed the PolyMLP model that uses an MLP network to predict polynomial approximation of time series.

The structures of ANNs are simple and can handle non-linearity. ANNs can be used for multiple tasks when the number of input features is small, especially for the intention prediction with hand-crafted features. However, for a 2D image that is a common kind of input in pedestrian behavior prediction, ANNs will lose the spatial information because of squeezing the image into a 1D vector, and can require a huge amount of trainable parameters, where CNNs could be the better choice because they share weights and can keep the spatial information. Besides, ANNs cannot capture sequential information in the input data, where RNNs could handle better.

C. Summary of Network Structures

From the distribution of the papers in Fig. 4, we see that sequential methods are mainly used for trajectory prediction. This is because trajectory prediction requires time series information. Trajectory prediction also employed GNNs for extracting interactions with other road users. The intention prediction usually used non-sequential networks, because they usually need the visual behavior features, which are extracted by CNNs. The joint prediction used both sequential networks and non-sequential networks, as they needed both spatial and temporal information.

The prediction methods also influenced the development of different prediction tasks. For the sequential methods that are commonly used by the trajectory prediction, research followed the trend from LSTMs in 2016 [27], GRUs in 2017 [55], to GANs in 2018 [31] and Conv-LSTMs in 2019 [84], and to the recently used Transformers in 2020 [37]. Each time the development of sequential methods stimulated the research on trajectory prediction. In contrast, for the intention prediction, most works used non-sequential. These models rely on the CNNs to process the images, which usually require more computing resources. This influences the development of the intention prediction. In future work, we need to investigate how much effort we should put into intention prediction. We need to trade off the additional gain from adding intention information for the application domain (e.g., for increased safety in an operational design domain for an autonomously

driving vehicle) and the cost of increased computing resources and accuracy and reliability for the perception system.

VII. Evaluation and Datasets

In this section, we firstly present the evaluation metrics that are commonly used for pedestrian behavior prediction. Then, we provide a review of the most commonly used datasets. There are some benchmarks for the trajectory prediction [39], [119] and intention prediction [79] that evaluated parts of the existing works.

A. Evaluation Metrics

1) Trajectory Prediction: The evaluation metrics for trajectory prediction are listed below.

- The average displacement error (ADE) (or the mean squared error (MSE)): The average distance between ground-truth and prediction trajectories over all predicted time steps, as defined below, where the predicted position for i^{th} pedestrian at time-step t is $\hat{Y}_t^i = (\hat{x}_t^i, \hat{y}_t^i)$, and the ground-truth is Y_t^i, $i \in \{1, \ldots, n\}$, $T_{obs} + 1 \leq t \leq T_{pred}$.

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

- The final displacement error (FDE): The average distance between ground-truth and prediction trajectories for the final predicted time-step, as defined below:

$$FDE = \frac{\sum_{i \in n} \| X_t^i - \hat{X}_t^i \|_2}{n}, \quad t = T_{pred} \tag{2}$$

Some other evaluation metrics such as the collision rate and negative log-likelihood are mentioned in the TrajNet++ benchmark [39]. The average non-linear displacement error is also used by some papers [27], [29], [30], [48], which is the MSE at the non-linear regions of a trajectory.

2) Intention Prediction: The evaluation metrics for intention prediction are listed below, with the number of positives P, negatives N, true positives TP, true negatives TN, false positives FP, and false negatives FN.

- Accuracy (ACC): $ACC = (TP + TN)/(P + N)$
- F1 score (F_1): $F_1 = 2TP/(2TP + FP + FN)$
- Precision: $Precision = TP/(TP + FP)$
- Recall (True Positive Rate): $Recall = TP/(TP + FN)$
- Average precision (AP): $AP = \sum_{k=1}^{n} (P(k)\Delta r(k))$. AP is defined as the area under the precision-recall curve, where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k-1$ to k.

3) Joint Prediction: For the joint prediction, the intention and trajectory results can be evaluated separately.

B. Datasets

High-quality and large-scale datasets are crucial for data-driven deep learning algorithms. Yin et al. [120] and Kang et al. [121] explored publicly available datasets to investigate their properties for developing autonomous driving features. In this part, we briefly introduce the publicly available datasets that that are commonly used for pedestrian behavior prediction. Table VI lists the publicly available datasets that are used by existing works and the summaries.

1) Trajectory Prediction: **ETH** [122] and **UCY** [123] datasets are widely used for evaluating pedestrian trajectories prediction. These two datasets contain five scenes of bird's-eye-view (BEV) videos collected in various scenarios, including crowded urban scenes. The ETH dataset contains two scenes with 750 annotated pedestrians, and UCY dataset contains three components with 786 annotated pedestrians. However, these two datasets are limited to pedestrians in crowds, and do not consider other road users.

KITTI [124] dataset contains driving scenarios collected by multi-sensors from the vehicle's view. The data is collected with a 64-layer LiDAR and two high-resolution stereo cameras (grayscale and color) with a resolution of 1392×512 pixels at 10 fps. It contains over 200,000 3D objects annotated in synchronized and calibrated LiDAR and stereo images. This dataset enables 3D detection and tracking estimation, and can also be used for pedestrian trajectory prediction.

Daimler [5] dataset consists of 68 sequences of images captured from the vehicle's view, of which 12,485 images contain pedestrians. The videos are recorded with a stereo camera with a resolution of 1176×640 pixels at 16 fps. The dataset contains four typical types of pedestrian behaviors, including crossing, stopping, starting, and bending in, and can be used to evaluate pedestrian trajectory prediction and intention classification.

New York Grand Central (GC) Dataset [125] contains more than 12,000 trajectories annotated in a one-hour-long BEV video. The video is recorded at 25 fps with a resolution of 1920×1080 pixels. This dataset includes crowd pedestrian scenes but is not collected in traffic scenarios.

Stanford Drone Dataset (SDD) [126] contains 20 scenes of BEV videos collected in a university campus. The videos are captured with a 4k camera on a quadcopter platform with a resolution of 1400×1904 pixels. It includes over 11,000 unique pedestrians and other road users, such as vehicles and bikers with their interactions captured.

Waymo Open Dataset [127] contains 1,150 scenes collected by multi-sensors from the vehicle's view in traffic scenarios. The sensors include five LiDAR sensors, and five high-resolution pinhole cameras. Three front cameras have a resolution of 1920×1280 pixels, two side cameras have a resolution of 1920×1040 pixels. The LiDAR on top has a scan range of 75m, the other four LiDAR have a scan range of 20m. Each scene is 20 seconds long, containing 2D and 3D objects labeled in LiDAR and camera images sampled at 10 Hz. The objects include pedestrians, cyclists, vehicles, and signs. This dataset has become increasingly popular for detection and tracking evaluation, and can also be used for evaluating trajectory prediction.

To evaluate existing pedestrian trajectory prediction algorithms, TrajNet benchmark [119] is proposed, based on selected trajectories from the ETH, UCY, and SDD datasets and uses the ADE and FDE evaluation metrics, and

TABLE VI
EVALUATION METRICS AND DATASETS FOR PEDESTRIAN BEHAVIOR PREDICTION

Dataset (Year)	Citation (Total/ Per year/Last year)	Prediction Tasks and Used in Papers	Summary
ETH (2009) [122]; UCY (2007) [123]	1188 / 99 / 364 710 / 51 / 265	Trajectory: [23], [25]–[28], [30]–[37], [40]–[44], [47]–[54], [58]	Collected in crowded urban scenes in bird's-eye-view (BEV). There are five scenes with more than 1500 people. Drawbacks: Do not include other traffic agents, and they are not collected in traffic scenarios.
KITTI (2012) [124]	7952 / 884 / 3520	Trajectory: [55]	Collected in traffic scenarios from vehicle's view. The data is collected with a 64-layer LiDAR and two high-resolution stereo cameras (grayscale and color) with a resolution of 1392×512 pixels at 10 fps. It contains over 200,000 3D objects annotated in synchronized LiDAR and stereo images.
Daimler (2013) [5]	214 / 27 / 75	Trajectory: [21], [22], [74]	Collected in traffic scenarios from the vehicle's view. The videos are recorded with a stereo camera with a resolution of 1176×640 pixels at 16 fps. It consists of 68 sequences of stereo images, with four types of pedestrian behaviors. It can be used to evaluate trajectory and intention prediction.
New York Grand Central (GC) (2015) [125]	209 / 35 / 59	Trajectory: [29], [30]	Collected in New York grand central in BEV. The video is recorded at 25 fps with a resolution of 1920×1080 pixels. It consists of more than 12,000 trajectories in a one-hour video. Drawbacks: Do not include other traffic agents, and they are not collected in traffic scenarios.
SDD (2016) [126]	485 / 97 / 284	Trajectory: [40], [44], [47], [51], [55]	Collected in a university campus in BEV. The videos have a resolution of 1400×1904 pixels at 30 fps. It contains 20 scenes with over 11,000 pedestrians, and other road users such as vehicles and bikers.
Waymo (2020) [127]	453 / 453 / 449	Trajectory: [45]	Collected in traffic scenarios from vehicle's view. It consists of 1,150 scenes collected by multi-sensors including five LiDAR sensors and five high-resolution pinhole cameras. Three front cameras have a resolution of 1920×1280 pixels, two side cameras have a resolution of 1920×1040 pixels. The dataset contains 2D and 3D objects (pedestrians, cyclists, vehicles, and signs) labeled in LiDAR and camera images sampled at 10 Hz. There are over 23k 3D-tracked pedestrians and 45k 2D-tracked pedestrians labeled.
JAAD (2017) [73]	128 / 32 / 93	Trajectory: [25], [26], [63] Intention: [67], [70]–[73], [75]–[79] Joint: [84]	Collected in traffic scenarios from the vehicle's view. There are over 300 video clips. The HD videos are recorded with on-board monocular camera at 30 fps. Most of the videos have a resolution of 1920×1080 pixels. The duration is between 5 to 15 seconds. The dataset contains approximately 82,000 frames and 2,000 unique pedestrian samples. The number of pedestrians with behavior annotations is 686.
PIE (2019) [84]	86 / 43 / 86	Trajectory: [25], [26], [63] Intention: [79] Joint: [84], [85]	Collected in traffic scenarios from the vehicle's view. There are six sets consisting of over 6 hours of driving videos. The HD videos with a resolution of 1920×1080 pixels are recorded with on-board monocular camera at 30 fps. The average duration is 10 min. The dataset contains approximately 290,000 annotated frames. The number of pedestrians with behavior annotations is 1842. The annotations include the bounding boxes with occlusion flags, crossing intention confidence, and text labels for pedestrians' actions.
ActEV/VIRAT (2018) [128]	97 / 32 / 42	Joint: [83]	Collected in traffic scenarios in BEV. Includes 455 videos from 12 traffic scenes, with more than 12 hours of recordings. Most of the videos have a high resolution of 1920×1080 pixels.

is expanded to TrajNet++ by Kothari et al. [39] with larger-scale data and more evaluation metrics.

For the trajectory prediction, there are datasets that only contains pedestrians, such as the Subway Station dataset [129] and the CUHK Crowd Dataset [130] used by Xu et al. [30]; and the Town Center Dataset [131] used by Xue et al. [49]. Besides, there are several datasets that contain urban traffic, such as ApolloScape [132] as used by Ma et al. [59], Interaction Dataset [133] as used by Li et al. [47], and nuScenes [134] as used by Yao et al. [25]. But these datasets are mainly designed for detection or for vehicle behavior prediction instead of pedestrian behavior prediction.

2) Intention Prediction: For the intention prediction, many previous works are based on data collected by the authors themselves [7], [64], because they can design what information to include in the data collection. We outline the publicly available datasets that are commonly used for pedestrian intention prediction.

Joint Attention for Autonomous Driving (JAAD) [73] dataset contains over 300 video scenes, and each scene ranges from 5 to 15 seconds in duration. The videos are recorded with three types of onboard cameras at 30 fps. 60 clips are collected in North America by a camera with a resolution of 1920×1080, 276 clips are collected in Europe by a camera with a resolution of 1920×1080, and 10 clips are collected in Europe by a camera with a resolution of 1280×720 pixels. This dataset contains approximately 82,000 frames and 2,000 unique pedestrian samples comprising a total number of 337,000 bounding boxes with behavioral and contextual tags. The number of pedestrians with behavior annotations is 686.

Pedestrian Intention Estimation (PIE) [84] dataset contains over 6 hours of driving footage captured from the vehicle's view, and the videos are split into approximately 10 minutes long pieces and grouped into 6 sets. The HD videos with a resolution of 1920×1080 pixels are recorded with an onboard camera at 30 fps. The dataset contains approximately 290,000 annotated frames. The number of pedestrians with behavior annotations is 1842. The dataset provides pedestrian behaviors and continuous sequences at the point of crossing. The pedestrians are annotated with the bounding boxes with

occlusion flags, and crossing intention confidence and text tags for their actions.

3) Joint Prediction: The **JAAD** and **PIE** datasets can be used for evaluating both trajectory and intention prediction, as well as joint prediction.

The **ActEV/VIRAT** [128] dataset includes 455 videos at 30 fps from 12 traffic scenes in BEV with more than 12 hours of recordings, and can be used for the evaluation of both trajectory and intention prediction. Most of the videos have a resolution of 1920 × 1080 pixels.

Other datasets such as the one proposed by Kooij et al. [135], which consists of sequences including single pedestrians with the intention to cross the street, can be used to evaluate the trajectories at crossing areas and intention prediction.

C. Summary and Discussion of Datasets

Table VI lists the publicly available datasets that are used by existing works and the summaries. We presented the number of citations of each dataset in the table to show the popularity of the dataset, including the number of total citations, the citation per year after released, and the citation in the last year. The KITTI dataset and Waymo Open dataset can also be used for other tasks such as detection and tracking, so there are more citations. ETH and UCY datasets are the most popular for trajectory prediction. SDD is also popular as it contains the annotation of pedestrians and other road users and can be used to study the interactions. JAAD and PIE datasets are the most popular for intention prediction. These two datasets can also be used for joint prediction.

The ETH and UCY datasets, the most commonly used datasets for trajectory prediction, were proposed in 2007 and 2009. While the JAAD and PIE datasets, the most commonly used datasets for intention prediction, were proposed in 2017 and 2019, which are ten years later than the datasets for trajectory prediction. This is because the information of pedestrian intention is more implicit compared to trajectories, and hence, the labeling of intention is more difficult compared to the labeling of trajectories. On the other hand, the dataset used for training and evaluation can influence the development of the prediction models. The earlier appearance of the commonly used dataset for trajectory prediction is another reason for more papers on this topic compared to intention prediction.

We also looked into the places where the data was captured and found they are mainly collected in North America, Europe, and Asia, including the USA, Canada, Germany, Switzerland, Bulgaria, Cyprus, and China. There are few datasets with urban scenarios captured in South America, Africa, and Oceania. Future research could focus on developing more datasets for these places. Furthermore, the comparability of findings across datasets is another issue that needs to be tackled to enable the transferability of results as well as applicability for certain geographic regions.

VIII. Comparison and Discussion

A. Performance of Existing Models

In this section, we compare the performance of some of the reviewed prediction methods. To align and compare the results, we select the works that used the most common publicly available datasets and metrics. The joint prediction is evaluated separately for trajectory and intention, so we compare them with the trajectory and intention prediction on corresponding datasets.

1) Trajectory Prediction: For the trajectory prediction, we compare the ADE and FDE values in meters, with 3.2s observation time and 4.8s prediction time on the ETH and UCY datasets. In Table VII, we list the evaluation results, model features, and summarize the methods used for feature extraction and modeling. From the first LSTM-based network for trajectory prediction, Social-LSTM [27], to the most recent model, AgentFormer [42], the ADE has improved from 0.72m to 0.18m, and the FDE improved from 1.54m to 0.29m.

The models intended to consider more model features to improve the accuracy, including the consideration of social interaction and the interaction within a scene. For the social interactions, the social pooling method improved to more complicated attention pooling networks, and afterwards, the graph-based spatio-temporal attention network took place. Recently, researchers have focused on the interactions with other road users, i.e., the heterogeneous interaction, to model real traffic scenarios. The graph-based representation is a powerful tool to model non-symmetric interactions. The environment and appearance features encoded by CNNs from the images help to improve the results. Besides, the instant destination is increasingly popular to be considered while predicting in goal-driven networks.

For prediction methods, instead of only using sequential or non-sequential methods, many models combine the CNNs and the sequential models to extract both the spatial and temporal features. The multi-modal GAN and CVAE models that can provide multiple plausible predictions are becoming increasingly popular compared to the uni-modal methods that predict a single distribution. The recurrent LSTM models are gradually replaced by the TCN models and TF models that have made a breakthrough in performance and can be paralleled to reduce training time. The current state-of-the-art algorithm AngentFormer [42] used the TF-based CVAE model and use agent-aware attention to model the spatio-temporal interaction at the same time.

2) Intention Prediction: For the intention prediction, we compare the AP and ACC for the C/NC classification on the most commonly used JAAD dataset. Table VIII lists the selected algorithms, their observation and prediction time horizon, the evaluation results, model features, and the summary. From the baseline method provided in the JAAD dataset [73] to the most recent intention prediction work [67], the AP is increased from 0.63 to 0.90.

Early works considered the appearance and skeleton of pedestrians and the environment context. Recent research included the vehicle states and the interaction with other road users to improve the precision. Off-the-shelf CNN-based segmentation and detection models are used for appearance and environmental feature extraction. 3D-CNNs can be used to extract both spatial and temporal information. A longer observation time improves the results [70], [78] showing that time series-related information contributes to the intention

TABLE VII
COMPARISON FOR TRAJECTORY PREDICTION

Paper, Author	Year	ADE / FDE	Model Features	Summary of Network Structures
Social-LSTM [27] (Alahi et al.)	2016	0.72 / 1.54	Trajectory, social interaction	LSTMs for sequence prediction; social pooling to model social interaction.
Social-GAN [31] (Gupta et al.)	2018	0.58 / 1.18	Trajectory, social interaction	LSTM-based GAN for multi-modal sequence prediction; social pooling network to model social interaction.
[23] (Nikhil et al.)	2018	0.59 / 1.22	Trajectory	CNNs instead of LSTMs for sequence prediction, enables parallelization.
SNS [52] (Lisotto et al.)	2019	0.36 / 1.81	Trajectory, social interaction, environment	LSTMs for sequence prediction; social, navigation and semantic pooling to model social interaction and environmental interaction.
Sophie [51] (Sadeghian et al.)	2019	0.54 / 1.15	Trajectory, social interaction, environment	LSTM-based GAN for multi-modal sequence prediction; CNNs for environmental feature extraction; soft-attention to model social interaction.
[34] (Zhang et al.)	2019	0.48 / 0.99	Trajectory, social interaction	LSTM encoder-decoder for sequence prediction; social graph network to model social interaction.
CGNS [47] (LI et al.)	2019	0.49 / 0.97	Trajectory, social interaction, environment	GRU-based CVAE for multi-modal sequence prediction; CNNs for environmental feature extraction; soft-attention to model social interaction.
Social-BiGAT [50] (Kosaraju et al.)	2019	0.48 / 1.00	Trajectory, social interaction, environment	LSTM-based GAN for multi-modal sequence prediction; CNNs for environmental feature extraction; GAT to model social interaction.
[83] (Liang et al.) (Joint Prediction)	2019	0.46 / 1.00	Person behavior, social interaction, Person-ORU interaction environment	LSTM for sequence prediction; CNNs for environmental and appearance feature extraction; geometric relation function for person-object interaction modeling.
SR-LSTM [32] (Zhang et al.)	2019	0.45 / 0.94	Trajectory, social interaction	LSTMs for sequence prediction; social-aware information selection and state refinement module to model social interaction.
Social-ways [33] (Amirian et al.)	2019	0.46 / 0.83	Trajectory, social interaction	LSTM-based Info-GAN for multi-modal sequence prediction; attention pooling to model social interaction.
STGAT [35] (Huang et al.)	2019	0.43 / 0.83	Trajectory, social interaction	LSTM encoder-decoder for sequence prediction; GAT for social interaction modeling.
Social-STGCNN [41] (Mohamed et al.)	2020	0.44 / 0.75	Trajectory, social interaction	TCNs and CNNs for sequence prediction, enables parallelization; spatio-temporal GCNs to model social interaction.
NMMP [40] (Hu et al.)	2020	0.41 / 0.82	Trajectory, social interaction, Person-ORU interaction	LSTM-based GAN for multi-modal sequence prediction; graph-based NMMP module to model the interaction with other road users.
[58] (Eiffert et al.)	2020	0.34 / 0.77	Trajectory, social interaction, Person-ORU interaction	LSTM-based GAN for multi-modal sequence prediction; Mixture Density Networks (MDN) and GVAT module to model the interaction with other road users.
Transformer (TF) [24] (Giuliari et al.)	2020	0.31 / 0.55	Trajectory	TF for sequence prediction, enables parallelization for encoder-phase.
STAR [37] (Yu et al.)	2020	0.26 / 0.53	Trajectory, social interaction	TF for sequence prediction; GCNs to model social interaction.
PECNet [44] (Mangalam et al.)	2020	0.29 / 0.48	Trajectory, social interaction, destinations	CVAE for multi-modal sequence prediction with an endpoint encoder for destinations; social pooling to model social interaction.
Tra2Tra [43] (Xu et al.)	2021	0.20 / 0.54	Trajectory, social interaction	LSTM for sequence prediction; LSTM-based spatio-temporal attention module to model social interaction.
SGNet [26] (Wang et al.)	2021	0.18 / 0.35	Trajectory, destinations	GRU-based CVAE for multi-modal sequence prediction; a stepwise goal estimator (SGE) for destination estimation.
Bitrap [25] (Yao et al.)	2021	0.18 / 0.35	Trajectory, destination	GRU-based CVAE for multi-modal sequence prediction with a GRU-based encoder and goal estimation, and a bi-directional decoder.
AgentFormer [42] (Yuan et al.)	2021	**0.18 / 0.29**	Trajectory, social interaction	TF-based CVAE for multi-modal sequence prediction; agent-aware TF to model social interaction on both time and social dimensions.

prediction. Recent work combined the CNN-based model with sequential models, including LSTMs and conv-LSTMs, to better extract the temporal information. The current state-of-the-art model [67] used a 3D-CNN to extract spatial and temporal behavioral feature, and encode the environmental and vehicle interaction feature with an additional distance encoding module.

B. Research Gaps and Future Opportunities

Next, we discuss the current research gaps in pedestrian behavior prediction that could be improved for future research.

1) Trajectory Prediction: Most existing trajectory prediction studies relied on past trajectories, and did not take full use of the appearance and skeleton behavioral features like intention prediction studies. Only a few of them (e.g., [46]) consider the pedestrians' visual behavioral features. In future works, the visual behavioral features can be considered even more. Another problem of existing trajectory prediction is that the prediction considers the "perfect" detection and tracking (i.e., the ground-truth of past trajectories). However, this is usually not feasible in practice. Future work should look at how to predict under conditions of imperfect detection and tracking and how to develop an end-to-end prediction from raw sensor

TABLE VIII
COMPARISON FOR INTENTION PREDICTION

Paper, Author	Year	Observation / Prediction Time (sec)	AP / ACC	Model Features	Summary of Network Structures
ATGC [73] (Rasouli et al.)	2017	0.33-0.5 / 0.03 (Next frame)	0.63 / –	Appearance cue, environment	Used CNNs to extract behavioral features for prediction.
[70] (Fang et al.)	2018	0.46 / 0.03 (Next frame)	– / **0.88**	Skeleton cue	Used CNNs and skeleton fitting to extract skeleton-based behavior features for prediction.
Res-EnDec [76] (Gujjar et al.)	2019	0.53 / 0.53	0.81 / –	Appearance cue, environment	Used 3D-CNNs as the encoder and conv-LSTMs as the decoder for generating future video, and appended a binary classifier to the generator for intention classification.
[75] (Chaabane et al.)	2020	0.53 / 0.53	0.87 / –	Appearance cue, environment	Used 3D-CNNs as the encoder and depth-wise separable conv-LSTMs as the decoder for generating future video, and appended a binary classifier to the generator for intention classification.
[72] (Lorenzo et al.)	2020	– / 1.0	0.83 / –	Appearance cue	Used CNNs to extract behavioral features, and applied LSTMs, GRUs, and the bidirectional variants of LSTMs and GRUs for crossing probability prediction.
[78] (Liu et al.)	2020	1.0 / 1.0	– / 0.79	Appearance cue, Person-ORU interaction, environment	Used a CNN-based model for image parsing, and encoding the appearance features. Used graph convolution for spatio-temporal interaction extraction. GRUs are used for capturing the temporal features and for behavior prediction.
PCPA [79] (Kotseruba et al.)	2021	0.53 / 0.03 (Next frame)	0.86 / 0.85	Skeleton and appearance cue, vehicle state, environment	Used 3D-CNNs for local visual context extracting, and used GRUs for non-visual features encoding. The temporal attention and modality attention modules are applied to learn the interaction.
[67] (Yang et al.)	2021	0.53 / 0.03 (Next frame)	**0.90** / –	Appearance cue, vehicle states, environment	Used 3D-CNNs to Extract spatial and temporal behavioral features, and used a distance encoding module to extract environmental contextual cues and vehicle features.
PIE [84] (Rasouli et al.) (Joint Prediction)	2019	0.5 / 0.03 (Next frame)	– / 0.79	Appearance cue, vehicle states, environment	Used LSTMs, CNNs and conv-LSTMs for joint prediction. For the trajectory prediction and vehicle speed prediction stream, the authors used LSTMs with temporal attention in encoder inputs, and self-attention in decoder inputs. For the intention estimation, CNNs are used for appearance behavioral feature extraction, and conv-LSTMs are used as encoders, and LSTMs are used as decoders.

data. Besides, existing works have used static graph-based models to extract spatio-temporal features. As dynamic graph-based models have shown a potential of better reflecting the spatio-temproal features compared with the static graph-based model in traffic flow prediction as used by Peng et al. [136], in future trajectory prediction works, researchers can also consider using dynamic graph-based models.

2) Intention Prediction: Only a few works (e.g., [7], [57], [67]) considered the traffic rules and signals while predicting the crossing behaviors. The existence of crosswalks and traffic signal lights are easy to get while strongly influencing the crossing behavior. Hence, such factors can be combined with other implicit environmental context features for intention prediction in future works. The interaction with vehicles and other road users can influence the pedestrian's decision. Unlike trajectory prediction, which considered various interactions between different traffic agents, most intention prediction studies used hand-crafted features to define the relationship with a single vehicle as shown in Table III. In future works, the graph-based or attention-pooling method can also be employed to extract the interaction relationships in crossing intention prediction.

As discussed, intention prediction usually requires large computational resources. More research could focus on investigating whether adding the intention prediction can bring noticeable improvements to an application domain.

3) Joint Prediction: The predicted results of trajectory and intention can be used to improve each other. Future works can focus on joint prediction, which could use past trajectories and interaction information that is usually used in trajectory prediction, and appearance behavioral cue that is typically used in intention prediction. The two prediction branches can share the extracted features to compensate for each other.

4) Hybrid Models: The behavior of pedestrians in urban traffic usually includes interactions between multiple road users. As we summarized in Table III, the interactions can either be learned implicitly by deep learning models that can include as much information as possible without requiring expert knowledge but that are hard to explain, or be represented by using knowledge-based hand-crafted features that are explainable but requires prior knowledge instead. In future works, we can develop hybrid models to take advantage of both approaches. For example, we can use conventional models with parameters learned from deep learning networks such as the Deep Social Force proposed by Kreiss [137], or implement the conventional knowledge-based model as a layer in the deep learning network.

5) Benchmark: As we reviewed and summarized in Sec VII, existing works use various datasets and metrics. The most popularly used datasets for trajectory prediction, ETH and UCY, are limited to crowds but not designed for traffic scenarios, and hence, they are not suitable to represent the performance for automated driving usage. The recently proposed popular benchmark, TrajNet, and TrajNet++ are not designed for automated driving scenarios and do not cover enough traffic scenes. For intention prediction, many researchers still use

self-collected datasets on a selected intersection, which makes it difficult for others to replicate and compare the work. For joint prediction, many existing works evaluate the trajectory and intention prediction separately with different datasets for comparison with previous works. Existing benchmarks either focus on trajectory or intention prediction. In future works, a benchmark can be defined and explored for the behavior prediction that includes both tasks and to thoroughly compare the performance for the joint prediction.

IX. Conclusion

In this paper, we have presented a thorough review of pedestrian behavior prediction models that use deep learning methods extracted from 92 papers. Compared with previous literature review papers, the original contributions of our review paper are as follows:

- Both trajectory and intention predictions are considered and analyzed, instead of only focusing on a single type of task;
- 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 introduced widely used datasets containing urban scenarios and we have evaluated and compared previous methods on such publicly available datasets.
- We included the most recent papers from 2016 to 2021.

We have discussed the model features used by existing models, and how they extracted these features. We have presented, categorized, and discussed the prediction methods used by existing works. The advantages and drawbacks of using different model features, and the properties of different prediction methods are discussed in detail. We have discussed why there is more research on trajectory prediction than intention prediction, how much effort we should put into intention prediction, which prediction methods we should use for which task, and the distribution of the datasets in the world. Finally, we outline the research gaps and possible research directions for improving the performance of prediction algorithms for urban scenarios.

Acknowledgment

The authors would like to thank Prof. Marco Dozza for his valuable comments and also would like to thank Zhongjun Ni for his suggestions on visualization.

References

[1] *Global Status Report on Road Safety 2018: Summary*, World Health Organization, Geneva, Switzerland, 2018.

[2] K. Rumar, "Transport safety visions, targets and strategies: Beyond 2000," in *Proc. 1st Eur. Transp. Saf. Lect.* Brussels, Belgium: European Transport Safety Council, 1999, pp. 6–8.

[3] S. Lefèvre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," *ROBOMECH J.*, vol. 1, no. 1, pp. 1–14, Dec. 2014.

[4] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 33–47, Jan. 2022.

[5] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2013, pp. 174–183.

[6] 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*. Berlin, Germany: Springer, 2015, pp. 161–177.

[7] 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, Mar. 2020.

[8] 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, Apr. 2010, Art. no. e10047.

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

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

[11] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.

[12] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 5, pp. 4282–4286, May 1995.

[13] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1696–1703, Apr. 2020.

[14] *Pedestrian Safety: A Road Safety Manual for Decision-Makers and Practitioners*, World Health Organization, Geneva, Switzerland, 2013.

[15] T. Hirakawa, T. Yamashita, T. Tamaki, and H. Fujiyoshi, "Survey on vision-based path prediction," in *Proc. Int. Conf. Distrib., Ambient, Pervasive Interact.* Cham, Switzerland: Springer, 2018, pp. 48–64.

[16] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *Int. J. Robot. Res.*, vol. 39, no. 8, pp. 895–935, Jul. 2020.

[17] R. Korbmacher and A. Tordeux, "Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches," 2021, *arXiv:2111.06740*.

[18] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A literature review on the prediction of pedestrian behavior in urban scenarios," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3105–3112.

[19] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng.*, May 2014, pp. 1–10.

[20] J. Bock, T. Beemelmanns, M. Klosges, and J. Kotte, "Self-learning trajectory prediction with recurrent neural networks at intelligent intersections," in *Proc. VEHITS*, 2017, pp. 346–351.

[21] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 327–332.

[22] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 4, pp. 414–424, Dec. 2018.

[23] N. Nikhil and B. Tran Morris, "Convolutional neural network for trajectory prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 186–196.

[24] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10335–10342.

[25] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1463–1470, Apr. 2021.

[26] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," 2021, *arXiv:2103.14107*.

[27] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[28] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4601–4607.

[29] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + hard-wired attention: An LSTM framework for human trajectory prediction and abnormal event detection," *Neural Netw.*, vol. 108, pp. 466–478, Dec. 2018.

[30] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5275–5284.

[31] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[32] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12077–12086.

[33] J. Amirian, J. Hayet, and J. Pettré, "Social ways: Learning multi-modal distributions of pedestrian trajectories with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2964–2972.

[34] L. Zhang, Q. She, and P. Guo, "Stochastic trajectory prediction with social graph network," 2019, *arXiv:1907.10233*.

[35] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial–temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6271–6280.

[36] S. Liu, H. Liu, H. Bi, and T. Mao, "CoL-GAN: Plausible and collision-less trajectory prediction by attention-based GAN," *IEEE Access*, vol. 8, pp. 101662–101671, 2020.

[37] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 507–523.

[38] Y. Liu, Q. Yan, and A. Alahi, "Social NCE: Contrastive learning of socially-aware motion representations," 2020, *arXiv:2012.11717*.

[39] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," 2020, *arXiv:2007.03639*.

[40] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6318–6327.

[41] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14412–14420.

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

[43] Y. Xu, D. Ren, M. Li, Y. Chen, M. Fan, and H. Xia, "Tra2Tra: Trajectory-to-trajectory prediction with a global social spatial–temporal attentive neural network," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1574–1581, Apr. 2021.

[44] K. Mangalam et al., "It is not the journey but the destination: Endpoint conditioned trajectory prediction," 2020, *arXiv:2004.02025*.

[45] 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 *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 1515–1522.

[46] J. Zhong, H. Sun, W. Cao, and Z. He, "Pedestrian motion trajectory prediction with stereo-based 3D deep pose estimation and trajectory learning," *IEEE Access*, vol. 8, pp. 23480–23486, 2020.

[47] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6150–6156.

[48] H. Manh and G. Alaghband, "Scene-LSTM: A model for human trajectory prediction," 2018, *arXiv:1808.04018*.

[49] H. Xue, D. Q. Huynh, and M. Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1186–1194.

[50] V. Kosaraju, A. Sadeghian, R. Martin-Martin, I. Reid, H. Rezatofighi, and S. Savarese, " Social-BiGAT: Multimodal trajectory forecasting using bicycle-GAN and graph attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 137–146.

[51] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.

[52] M. Lisotto, P. Coscia, and L. Ballan, "Social and scene-aware trajectory prediction in crowded spaces," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2567–2574.

[53] N. Bhujel, E. K. Teoh, and W. Yau, "Pedestrian trajectory prediction using RNN encoder–decoder with spatio-temporal attentions," in *Proc. IEEE 5th Int. Conf. Mechatronics Syst. Robots (ICMSR)*, May 2019, pp. 110–114.

[54] A. Syed and B. Morris, "STGT: Forecasting pedestrian motion using spatio-temporal graph transformer," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 1553–1558.

[55] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2165–2174.

[56] X. Song et al., "Pedestrian trajectory prediction based on deep convolutional LSTM network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3285–3302, Jun. 2021.

[57] 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 Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3964–3980, Oct. 2019.

[58] 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 Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5026–5033, Oct. 2020.

[59] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "TrafficPredict: Trajectory prediction for heterogeneous traffic-agents," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6120–6127.

[60] K. Li, M. Shan, K. Narula, S. Worrall, and E. Nebot, "Socially aware crowd navigation with multimodal pedestrian trajectory prediction for autonomous vehicles," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–8.

[61] 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 *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 1501–1508.

[62] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8475–8484.

[63] K. Kim, Y. K. Lee, H. Ahn, S. Hahn, and S. Oh, "Pedestrian intention prediction for autonomous driving using a multiple stakeholder perspective model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 7957–7962.

[64] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 2607–2612.

[65] B. Völz, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "Inferring pedestrian motions at urban crosswalks," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 2, pp. 544–555, Feb. 2019.

[66] J. Zhao, Y. Li, H. Xu, and H. Liu, "Probabilistic prediction of pedestrian crossing intention using roadside LiDAR data," *IEEE Access*, vol. 7, pp. 93781–93790, 2019.

[67] 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 Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5338–5349, Jun. 2022.

[68] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2331–2339, Mar. 2022.

[69] Z. Fang, D. Vázquez, and A. López, "On-board detection of pedestrian intentions," *Sensors*, vol. 17, no. 10, p. 2193, Sep. 2017.

[70] Z. Fang and A. M. López, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1271–1276.

[71] F. Piccoli et al., "FuSSI-Net: Fusion of spatio-temporal skeletons for intention prediction network," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2020, pp. 68–72.

[72] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo, "RNN-based pedestrian crossing prediction using activity and pose-related features," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1801–1806.

[73] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 206–213.

[74] M. Hoy, Z. Tu, K. Dang, and J. Dauwels, "Learning to predict pedestrian intention via variational tracking networks," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3132–3137.

[75] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2286–2295.

[76] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2097–2103.

[77] D. O. Pop, A. Rogozan, C. Chatelain, F. Nashashibi, and A. Bensrhair, "Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction," *IEEE Access*, vol. 7, pp. 149318–149327, 2019.

[78] B. Liu et al., "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3485–3492, Apr. 2020.

[79] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1257–1267.

[80] Z. Huang, A. Hasan, K. Shin, R. Li, and K. Driggs-Campbell, "Long-term pedestrian trajectory prediction using mutable intention filter and warp LSTM," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 542–549, Apr. 2021.

[81] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of vulnerable road users—Detection and forecasting by means of machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3035–3045, Jul. 2020.

[82] H. Wu, L. Wang, S. Zheng, Q. Xu, and J. Wang, "Crossing-road pedestrian trajectory prediction based on intention and behavior identification," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[83] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5718–5727.

[84] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6261–6270.

[85] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do they want to cross? Understanding pedestrian intention for behavior prediction," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1688–1693.

[86] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5903–5908.

[87] B. Völz, H. Mielenz, G. Agamennoni, and R. Siegwart, "Feature relevance estimation for learning pedestrian behavior at crosswalks," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 854–860.

[88] A. Rasouli and J. K. Tsotsos, "Joint attention in driver-pedestrian interaction: From theory to practice," 2018, *arXiv:1802.02522*.

[89] S. Schmidt and B. Färber, "Pedestrians at the kerb—Recognising the action intentions of humans," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 12, no. 4, pp. 300–310, Jul. 2009.

[90] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Cham, Switzerland: Springer, Jun. 2021, pp. 483–499.

[91] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[92] M. M. Hamed, "Analysis of pedestrians' behavior at pedestrian crossings," *Saf. Sci.*, vol. 38, no. 1, pp. 63–82, Jun. 2001.

[93] H. Guo, Z. Gao, X. Yang, and X. Jiang, "Modeling pedestrian violation behavior at signalized crosswalks in China: A hazards-based duration approach," *Traffic Injury Prevention*, vol. 12, no. 1, pp. 96–103, Jan. 2011.

[94] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–12.

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

[96] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[97] J. Li, A. Raventos, A. Bhargava, T. Tagawa, and A. Gaidon, "Learning to fuse things and stuff," 2018, *arXiv:1812.01192*.

[98] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[99] Q. Zhu, "Hidden Markov model for dynamic obstacle avoidance of mobile robot navigation," *IEEE Trans. Robot. Autom.*, vol. 7, no. 3, pp. 390–397, Jun. 1991.

[100] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, "Understanding human intentions via hidden Markov models in autonomous mobile robots," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interact.*, Mar. 2008, pp. 367–374.

[101] T. Bandyopadhyay, C. Z. Jie, D. Hsu, M. H. Ang, D. Rus, and E. Frazzoli, "Intention-aware pedestrian avoidance," in *Experimental Robotics*. Berlin, Germany: Springer, 2013, pp. 963–977.

[102] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[103] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2543–2549.

[104] M. Karl, M. Soelch, J. Bayer, and P. van der Smagt, "Deep variational Bayes filters: Unsupervised learning of state space models from raw data," 2016, *arXiv:1605.06432*.

[105] M. Gutmann and A. Hyvarinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.

[106] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[107] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[108] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.

[109] A. Syed and B. T. Morris, "CNN, segmentation or semantic embeddings: Evaluating scene context for trajectory prediction," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2020, pp. 706–717.

[110] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.

[111] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap, "Compressive transformers for long-range sequence modelling," 2019, *arXiv:1911.05507*.

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

[113] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 47–54.

[114] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.

[115] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[116] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.

[117] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3590–3598.

[118] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5308–5317.

[119] S. Becker, R. Hug, W. Hübner, and M. Arens, "An evaluation of trajectory prediction approaches and notes on the TrajNet benchmark," 2018, *arXiv:1805.07663*.

[120] H. Yin and C. Berger, "When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–8.

[121] Y. Kang, H. Yin, and C. Berger, "Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments," *IEEE Trans. Intell. Vehicles*, vol. 4, no. 2, pp. 171–185, Jun. 2019.

[122] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[123] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, Sep. 2007.

[124] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[125] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3488–3496.

[126] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 549–565.

[127] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2443–2451.

[128] G. Awad et al., "TRECVID 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search," in *Proc. TRECVID*, 2018, pp. 1–38.

[129] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proc. CVPR*, Jun. 2011, pp. 3441–3448.

[130] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2227–2234.

[131] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. CVPR*, Jun. 2011, pp. 3457–3464.

[132] X. Huang et al., "The ApolloScape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 954–960.

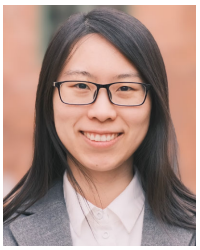
[133] W. Zhan et al., "INTERACTION dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," 2019, *arXiv:1910.03088*.

[134] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[135] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 618–633.

[136] H. Peng et al., "Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting," *Inf. Sci.*, vol. 521, pp. 277–290, Jun. 2020.

[137] S. Kreiss, "Deep social force," 2021, *arXiv:2109.12081*.



Chi Zhang received the B.E. and M.E. degrees in control science and engineering from Zhejiang University, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Gothenburg, Gothenburg, Sweden. From 2017 to 2020, she was a Research and Development Engineer in automated driving perception with the Intelligence Driving Group, Baidu, Beijing, China. She is a Marie Curie Early Stage Researcher. Her research interests include applying deep learning on pedestrian behavior prediction and learning interactions between vulnerable road users and vehicles.

Christian Berger received the Ph.D. degree from RWTH Aachen University, Germany, in 2010. He is currently a Full Professor with the Department of Computer Science and Engineering, University of Gothenburg, Sweden. He coordinated the project for the vehicle "Caroline," which participated in the world's first urban robot race 2007 DARPA Urban Challenge Final. He co-led the Chalmers Truck Team during the 2016 Grand Cooperative Driving Challenge (GCDC), and is one of the two leading architects behind Open Driverless Vehicle (OpenDLV). His research interests include architecting complex and distributed real-time software systems, micro-services for cyber-physical and the IoT-systems, continuous integration/deployment/experimentation, and data-driven software engineering.