Behavioral Intention Prediction in Driving Scenes: A Survey

Jianwu Fang, Member, IEEE, Fan Wang, Jianru Xue, Member, IEEE, and Tat-Seng Chua

Abstract—In driving scenes, road agents often engage in frequent interaction and strive to understand their surroundings. Ego-agent (each road agent itself) predicts what behavior will be engaged by other road users all the time and expects a shared and consistent understanding for safe movement. To achieve this, Behavioral Intention Prediction (BIP) simulates such a human consideration process to anticipate specific behaviors, and the rapid development of BIP inevitably leads to new issues and challenges. To catalyze future research, this work provides a comprehensive review of BIP from the available datasets, key factors, challenges, pedestrian-centric and vehicle-centric BIP approaches, and BIP-aware applications. The investigation reveals that data-driven deep learning approaches have become the primary pipelines, while the behavioral intention types are still limited in most current datasets and methods (e.g., Crossing (C) and Not Crossing (NC) for pedestrians and Lane Changing (LC) for vehicles in this field. In addition, current research on BIP in safe-critical scenarios (e.g., near-crashing situations) is limited. Through this investigation, we identify open issues in behavioral intention prediction and suggest possible insights for future research.

Index Terms—Behavioral intention prediction, challenges, promising approaches, road agents, benchmarks.

I. INTRODUCTION

The driving scene is highly socialized which necessitates the effective and precise understanding of the intentions of surrounding road users. For safe driving, the decision-making is influenced by the actions of pedestrians, vehicles, cyclists, and other agents on the road. Behavioral intention, in the context of driving scenes, links the anticipated actions or behaviors of the road agents, such as “crossing the street” for pedestrians/cyclists and “changing lanes” for vehicles. It reveals the deliberate inclination of road agents to take certain actions or approach specific goals, which is typically seen as the internal motivation behind their behaviors [1], [2].

This work defines the term “Behavioral Intention Prediction” (BIP) as the prediction of intended actions by pedestrians/cyclists or maneuvers by vehicles (as shown in Fig. 1) within the context of surrounding driving scenes. However, accurately understanding road structure [26], road user interaction [27], determining moving goals, and other prior knowledge, such as skills, gender, social and cultural factors [25], pose challenges to BIP. These clues permeate the social and causal relations, as depicted in Fig. 1. The behavioral intention understanding for each agent is crucial for facilitating the interactive function for autonomous systems [28]. Developing advanced techniques to predict the intentions of road agents can enhance the cognitive abilities of autonomous systems and ensure the safety of all road users. Nowadays, with the vigorous demand for self-driving systems at home and abroad, the corresponding scale of data also grows rapidly, which provides fertile soil for deep learning-based behavioral intention prediction [29], [30].

A. Distinction to Previous Surveys

Over the past decades, numerous works have concentrated on the detection, segmentation, and tracking of road agents on enhancing the safety and intelligence of self-driving systems. Some previous surveys [31], [32] have comprehensively summarized the pipelines within those fields.

Fig. 1. An intersection scenario, where the Ego Vehicle (A) wants to turn left and arrives at the goal region (F). It needs to estimate whether the Target Vehicle (B) will move straight or turn left, and the crossing intention of the pedestrian group (D). This estimation process implies a causal relation of $A \iff B$ and $A \iff D$ conditioned by $F \rightarrow A$. Certainly, the movement of other road agents also involves complex and causal relation reasoning.

Manuscript received 31 October 2022; revised 1 September 2023 and 20 December 2023; accepted 4 March 2024. Date of publication 22 March 2024; date of current version 1 August 2024. This work was supported in part by NSFC under Grant 62036008 and Grant 62273057 and in part by the Foundation Strengthening Program Technology Field Fund under Grant 2021-JCQ-JJ-0004. The Associate Editor for this article was Y. Zhang.

Jianwu Fang and Jianru Xue are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an 710049, China (e-mail: fangjianwu@mail.xjtu.edu.cn; jrxue@mail.xjtu.edu.cn).

Fan Wang is with the College of Transportation Engineering, Chang’an University, Xi’an 710064, China.

Tat-Seng Chua is with the Sea-NExT++ Joint Research Centre, School of Computing, National University of Singapore, Singapore 119077 (e-mail: dcscs@nus.edu.sg).

Digital Object Identifier 10.1109/TITS.2024.3374342
To provide a clearer distinction from previous surveys, we conducted a comprehensive search on Google Scholar for related surveys on “behavior prediction”, “intention prediction”, “trajectory prediction”, and “crash anticipation”. Fig. 2 presents a chronological overview of 22 high-quality surveys spanning nearly ten years. It is evident from the figure that trajectory prediction has garnered greater attention compared to other prediction tasks in recent years. Additionally, an increasing number of studies have acknowledged the significance of road agents’ intention or interaction in driving scenarios. Nevertheless, our work differs from the related surveys in the following aspects.

1) The majority of surveys concentrate on trajectory prediction and specifically target deep learning models [10], [13], [17]. However, there is a limited review of behavioral intention prediction, particularly regarding essential factors such as prediction uncertainty and the latest advancements.

2) The most related surveys to our work are [5], [6], and [7]. Among them, these surveys focus on a single type of road agent, such as vehicles [6], [7], pedestrians [5], [23], [24], or human drivers [4]. In addition, the intention types in these surveys are monotonous, such as crossing for pedestrians, and lane changing for vehicles.

Complementary to these surveys, we concentrate on the latest progress in Behavioral Intention Prediction (BIP) for both pedestrians and vehicles, where the trajectory prediction is only an application conditioned by BIP and takes a limited space. Our main objective is to draw inspiration for BIP research from key factors, challenges, and promising models such as causality, multimodality, and synthetic-real data collaboration. In addition, we present the most comprehensive review of the available datasets for BIP.

B. Taxonomy and Contributions

This paper reviews the latest works of BIP in driving scenes and presents a full portrait of the problem definition, available datasets, key factors and challenges, agent-centric BIP, and BIP-aware applications. Fig. 3 depicts the detailed taxonomy of this survey, where all parts are tightly coupled with clear relations. Based on this work, we want to showcase the progress of BIP through the research pipeline of problems and datasets→factors and challenges→approaches→applications→research insights. The contributions of this survey are as follows.

- Different from previous surveys, we clarify the definition of different prediction tasks, and provide a more targeted and comprehensive survey on BIP from the available datasets (17 ones), intention types, key factors, approaches, applications, and future insights.
- The latest progress in behavioral intention prediction is extensively investigated and the new research pipelines are chronologically reviewed and discussed.
- We provide more discussion for the promising formulations and insights, such as causality, parallel testing, prediction uncertainty modeling, BEV representation, etc.

The remainder of this work is organized as follows. Sec. II presents the background of the prediction task definition, available datasets, and intention types, which form the concept and denotation basis for the following descriptions. Sec. III briefly reviews the key factors and challenges in BIP. The method progress of agent-centric BIP including pedestrians and vehicles is described and discussed in Sec. IV. BIP-aware applications including trajectory prediction and behavior prediction are summarized in Sec. V. Sec. VI presents the discussion for current research and provides potential insights for future research. The conclusion is given in Sec. VII.

II. BACKGROUND

Based on the investigation, we find that there is a concept confusion for the tasks of trajectory prediction, behavior prediction, and behavioral intention prediction, and are interchangeably used in this field [33], [34]. These concepts vary with the output and have different targets of interest.

A. Prediction Task Definition in Driving Scenes

To begin with, we introduce two observation views in data collection in driving scenes: 1) Ego-View: capturing the video or data point in first-person view (Fig. 4 (a)(c)); 2) Bird’s Eye View (BEV): commonly observed in the world coordinate system (Fig. 4(b)). With this setting, we present the terminology for different prediction tasks.

1) Trajectory Prediction (TP) is well studied in this field [4], [21], which refers to the process of estimating the future trajectories of various entities (vehicles, pedestrians, cyclists, etc.) on the road. It analyzes the historical observation states of the entities, e.g., their positions, velocity, and heading, and encodes the information along with surrounding contextual data (e.g., the road map) to predict the future trajectories measured either in the Ego-View or the BEV observation.

2) Behavior Prediction (BP) refers to the process of estimating the future behaviors of various entities (vehicles, pedestrians, cyclists, etc.) on the road [37]. It can also analyze the historical observation states of the entities, encode them,

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The content taxonomy in this survey. The prediction task definition and available dataset investigation provide a concept basis for this survey. The key factors and challenges are analyzed comprehensively. Then, the agent-centric BIP works for pedestrians and vehicles are chronologically reviewed and discussed. Furthermore, we review the BIP-aware applications from trajectory prediction and behavior prediction, and one-to-one discussions to previous sections are presented for potential issues and possible insights.

Fig. 3. The content taxonomy in this survey. The prediction task definition and available dataset investigation provide a concept basis for this survey. The key factors and challenges are analyzed comprehensively. Then, the agent-centric BIP works for pedestrians and vehicles are chronologically reviewed and discussed. Furthermore, we review the BIP-aware applications from trajectory prediction and behavior prediction, and one-to-one discussions to previous sections are presented for potential issues and possible insights.

Fig. 4. Examples for the behavioral intention prediction, behavior prediction, and crash anticipation, where (a) and (c) captured in the Ego-View are sampled from JAAD [35] and DADA-2000 [27] datasets, respectively. (b) denotes the frames of Bird’s Eye View (BEV) sampled from the nuScenes dataset [36].

3) Behavioral Intention Prediction (BIP) refers to the process of estimating the intended actions or behaviors of various entities on the road. It similarly can analyze the historical observation states of the entities, and the contextual information, to infer the intended actions of agents in the near future, as shown in the first column of Fig. 4 (a-c).

4) Crash Anticipation (CA) refers to the ability to predict and foresee potentially dangerous situations or collisions before they occur on the road. It involves actively analyzing various factors such as the behavior of surrounding vehicles, pedestrians, and road conditions to identify potential hazards or risky situations that may lead to accidents, seeing Fig. 4(c).

Based on these definitions, it is evident that BP and BIP share a consistent input under the same observation view, apart from CA and TP. However, their outputs differ. It is important to note that the output of BP and BIP deviates from the concept of “intended” behavior, as BIP refers to conscious and deliberate actions or goals [53], which occurs earlier in the timeline compared to specific behaviors. In certain situations, CA can to some extent integrate behavioral intentions, behaviors, and trajectories. Each prediction task owns its timeline, and the success of BIP can provide an earlier prompt for safe decisions than other prediction tasks.

With the clarification of different prediction tasks, we overview the available BIP datasets and intention types.

B. Available Datasets and Intention Types

For a targeted review, we exhaustively investigate and elaborate on the publicly available datasets for the behavioral intention prediction task. Table. 1 presents the attributes of 17 datasets, and the samples are shown in Fig. 5. Almost all pedestrian-centric BIP datasets have pedestrian crossing or not crossing intention. In the following, we describe the

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE I
CHRONOLOGICAL OVERVIEW OF 17 DATASETS IN BEHAVIORAL INTENTION PREDICTION GENERATED BY REAL DATA (R) OR SYNTHETIC DATA (S) WITH THE INTENTION TYPES, ANNOTATIONS, AND THE SERVICEABLE PREDICTION TASKS (PRED. TASKS)

| Datasets   | Years/book/title | Seq. num. | Annotations | Intention Types | S/R | Pred. Tasks |
|------------|------------------|-----------|-------------|-----------------|-----|-------------|
| Daimler [38]| 2014/ECVC | 58 | I, T, BVV  | C, NC           | R   | BIP, TP     |
| NGSIM [39]  | 2016/IPPO         | 17,179p  | T, VT, MV   | LLC, RLC, LK    | R   | BIP, TP     |
| JAAD [35]   | 2017/ICCVW        | 346      | I, 2DB, W, O, Beh, A, BO | C, NC | R   | BIP, TP     |
| HighD [40]  | 2018/TTSC         | 110,500t | T, VT, MTV, | LLC, RLC, LK    | R   | BIP, TP     |
| VIENA2 [41] | 2018/ACCV         | 13000    | I, Beh      | S, TL, TR, LLC, RLC, C, NC, W | S   | BIP, CA     |
| INTERACTION [42] | 2019/aXtr  | -        | I, 2DB, SS, Beh | SRD, NC, LLC, RLC, M | R   | BIP, TP, BP |
| PIE [43]    | 2019/ICCV        | 53       | I, 2DB, Beh, EVV | W, STOP, C, NC   | R   | BIP, TP, TP |
| BLVD [44]   | 2019/ICRA         | 654      | I, 2DB, Beh, 3DB  | 22 types       | R   | BIP, TP, TP |
| PREVENTION [45] | 2019/TTSC    | 11       | I, 2DB, VT, Beh | LLC, RLC, CI, CO | R   | BIP, BP     |
| BPI [46]    | 2020/TTSC         | 120      | T, PO, T    | C, NC          | R   | BIP, TP     |
| TITAN [47]  | 2020/IEEE RAL     | 700      | I, 2DB, Beh | ST, C, NC      | R   | BIP, TP     |
| STIP [48]   | 2020/IEEE RAL     | -        | I, 2DB, Beh | C, NC          | R   | BIP, TP     |
| PePScenes [49]| 2020/NeuroIPS   | 850      | I, Beh, SS, BO, STD | C, NC | R   | BIP, TP, BP |
| PSI [50]    | 2021/aXtr         | 110      | I, 2DB, SS, Beh | C, NC          | R   | BIP, TP     |
| LOKI [51]   | 2021/ICCV        | 664      | I, 2DB, Beh, 3DB, SS, Age, D, GE, W | C, NC | R   | BIP, TP     |
| Virtual PedCross-4667 [52]| 2022/TTSC     | 4667     | I, W, 2DB, Beh | C, NC          | S   | BIP, TP     |
| DADA-2000 [27]| 2022/IEEE TITS  | 3000     | I, Beh, DA, W | LLC, RLC, VO, C, NC | R   | BIP, DAP, CA |

P*: Irrelevant. The vehicle trajectory data is collected on the highway with a sampling frequency of 10 Hz. T*: trajectories of vehicles.

Intention Types: Crossing (C); Non-Crossing (NC); Walking (W); Standing (ST); Straight Moving (SM); Turning Left (TL); Turning Right (TR); U-Turn (UT); Lane Keeping (LK); Cutting In (CI); Cutting Out (CO); Vehicle Overtaking (VO); Left Lane Changing (LLC); Right Lane Changing (RLC); Stopping (STOP); Pushing (P); Yielding (Y); Merging (M); Moving along the Roundabout (MRD); Accelerating (ACCE); Decelerating (DECE).

Annotation: Image (I); 2D Boxes (2DB); 3D boxes (3DB); Vehicle Type (VT); Ego Vehicle Velocity (EVV); Motion of Target Vehicle (MTV); Driver Attention (DA); Trajectory (T); Weather (W); Behavior (Beh); Pose (PO); Occasions (O); Age (Age); Gender (G); Depth image (D); Human Body Orientation (BO); Destination (DES); Semantic Segments (SS); Scene Text Description (STD).

main differences among these datasets from the aspects of observation views, annotation details, and intention types.

1) Observation Views: Observation views have a direct influence on behavioral intention annotation and prediction model designing. Daimler is the pioneering dataset for the prediction of pedestrian crossing or not-crossing, which only consists of four pedestrians captured in grayscale images. From Fig. 5, the top seven datasets (i.e., JAAD [35], PIE [43], TITAN [47], BPI [46], PePScenes [49], STIP [48], and PSI [50]) concentrate on the pedestrian-centric BIP with the Ego-View observation. These datasets mainly categorize intentions into crossing and not crossing. However, TITAN offers a wider range of intention types and provides interaction labels between pedestrians and road scenes.

We can see that NGSIM [39], HighD [40], and INTERACTION [42] are collected from the BEV view. Compared to the Ego-View, the BEV observation captures a larger spatial range of view and provides a complete movement observation, making it a valuable source for ground-truth verification in Behavior Intention Prediction (BIP) and other prediction tasks. However, the Ego-View offers clearer information about the pose and height of road agents and provides the opportunity for collision avoidance by controlling vehicles in real-time. The 3D point clouds, like the ones in the BLVD dataset [44], capture both the BEV and Ego-View perspectives. However, these raw 3D point cloud datasets lack semantic labels and fine-grained pose information of road agents. Some recent works have been focused on capturing panoramic views by multiple cameras, such as the Argoverse 3D dataset [54] (with seven cameras) or the nuScenes dataset [36] (with six cameras). However, these 3D point cloud datasets do not provide the behavioral intention label. Therefore, in the future, these panoramic view datasets can be extended with behavioral intention or behavior labeling.

2) Annotation Details: The datasets mentioned above offer intricate annotation details that provide valuable insights and different modeling approaches for predicting behavioral intentions. These annotations include object bounding boxes, trajectories, road entities, driver attention, long-term intended goal regions, risk levels, and even demographic information like age and gender, which are all crucial for safety evaluations in driving scenarios. PSI [50] provides the Scene Text Description (STD) for different situations, which provides an additional perspective for scene understanding. From Fig. 5 and Table I, DADA-2000 is the only one that considers crash scenarios, making it suitable for counterfactual analysis in behavioral intention prediction (BIP) related to crash...
anticipation or collision avoidance. With the development of text-to-video diffusion models [55], [56], there is potential for creating editable driving scenes and exploring their impact on BIP. Another interesting direction for BIP cross-validation or counterfactual analysis is crow-view annotation, which involves calibrating between the Ego-View and BEV perspectives.

3) Intention Types: Different datasets focus on distinct intention types. In Table I, we mainly summarize 21 common intention types. Most of the datasets only contain the Crossing (C) or Not Crossing (NC) of pedestrians. However, TITAN [47] provides more detailed pedestrian behavior labels, such as Pushing (P) or Standing (ST), based on the road structure. In TITAN, each road agent is assigned a sequential action label as its intention. When it comes to vehicle-centric intention types, Lane Changing (LC) is the primary intention provided in datasets like NGSIM [39] and HighD [40]. PREVENTION includes the intentions of “Cutting In” (CI) and “Cutting Out” (CO), along with the provision of “risk levels” for the Lane Changing (LC) intention. Additionally, the INTERACTION dataset [42] offers the intention of “Moving along the Roundabout” (MRD) due to the presence of round paths in its scenarios. Based on the comparison, it is evident that the behavioral intention types for various road agents are not sufficiently detailed in most current datasets. Many crucial and safety-critical behavioral intention types, such as “Vehicle running conversely” and “Braking,” are not captured. Moreover, there is a lack of intention types that involve interactions between road agents and road entities, such as the sidewalks, bus stations, and steep slopes. To address this limitation, the BLVD dataset provides a more fine-grained annotation of intention types, encompassing 22 specific types for vehicles, pedestrians, and riders (cyclists or motorbikes). These detailed intention types in BLVD enable a more nuanced analysis and modeling of behavioral intention in the driving scene [44].

4) Evaluation Metrics: Most of the BIP works in this field determine the future intention as a classification problem. Consequently, the Accuracy (Acc), Precision (Pre), Recall (Rec), and F1-measure (F1) are four common metrics for evaluation, where the computing methods are: Acc=$\frac{TP}{TP+FN+FP+TN}$, Pre=$\frac{TP}{TP+FP}$, Rec=$\frac{TP}{TP+FN}$, and F1=$2 \times \frac{Rec \times Pre}{Rec + Pre}$. TP and FP are respectively the predicted positive intention samples in the true positive and negative set, and TN and FN are respectively the predicted negative intention samples in the true negative and positive set.

With the definition of the prediction tasks and the introduction of datasets and intention types, we describe the key factors, challenges, promising models, and BIP-aware applications in the following, where the dataset name, intention types, and annotations are consistent throughout the whole paper.

III. KEY FACTORS AND CHALLENGES

The mixed traffic scene adds complexity to the prediction of behavioral intentions (BIP) [25], [57]. BIP of surrounding agents needs to consider the agent type, road structure, social relation, action tendency, and intention types, which encompasses factors from robust road structure representation, social interaction modeling, and prediction uncertainty estimation, posing various challenges for BIP research.

A. Road Structure Representation

The road scene is a highly structured environment, with consistent traffic rules that road users must adhere to [33] and [58]. Within this environment, there are static road entities such as road lanes and boundaries, as well as dynamic agents like pedestrians and vehicles. All of these elements together form the contextual information necessary for safe driving [59]. Hence, a crucial factor in predicting behavioral intentions is the representation of the road structure, which involves modeling both the static road entities and the dynamic agents.

Pedestrians, cyclists, motorbikes, buses, trucks, cars, trailers, etc., are common entities in a mixed road scene. Their intentions vary depending on factors such as movement patterns, road layout, and observation views (e.g., Ego-View and BEV). For pedestrians, agent-centric intentions such as “crossing”, “walking”, “running”, and “stopping” are of interest, while vehicles are typically concerned with “lane keeping”, “turning left or right”, “braking”, “accelerating”, and “lane changing”. These intentions often correspond to specific situations.

1) Road Lane Representation: The road lane is a primary clue for the road structure representation and is of great interest to the autonomous driving community [60], [61], [62], [63]. In this field, most researchers detect the road lanes in camera videos [64] or differentiate the reflectance of the scanned targets by 3D-LiDAR [65]. Many studies investigate the use of lane centerlines as anchors to constrain trajectory prediction [66], [67]. Lane graph representations [67], [68] are modeled from raw map data to explicitly explore the complex road topology and long-range dependencies, where there are three types of interaction between agents and the road map (i.e., lane-to-lane, lane-to-agent, agent-to-lane) in the lane graph representation. For an in-depth utilization of the information on road lanes, Hong et al. [33] unify the representation which encodes the road map in a spatial road grid, allowing the use of fusing complex scene context of entity-entity and entity-environment interactions. However, these lane graph representations only consider the road boundaries or the lane centerlines of the pre-built High-Definition (HD) maps. As for behavioral intention prediction, road lane-based representation only can be reflected on the Lane Changing (LC) intention or Lane Keeping (LK) intentions.

2) BEV Representation: HD maps are often expensive to build in advance, leading researchers to explore road topology representation using on-board camera data [70], [71]. This gives rise to the road structure representation by Bird’s Eye View (BEV) [72], [73], [74] (as shown by Fig. 6) transferred from the raw camera videos by deep feature transformation [69]. The BEV view offers a clearer depiction of the road layout compared to the Ego-View, allowing for the perception of object scale, movement attributes (e.g., velocity and orientation), and social interactions without geometric distortion. However, BEV representation relies on the inter-correlation between the camera videos and the road map grid. The inter-correlation is easily affected by the perception distortion in...
Social interaction modeling is the key factor for driving scene perception all the time. Previous studies [25], [79], [80], [81], [82], [83], [84] have focused on the interaction and relationship among road agents to accurately predict their future states. The intermediate feature representation of the occupancy map taken by agents [81], [83] is commonly used for interaction modeling in intention or trajectory prediction. This interaction [15] provides the prior for potential moving tendency, such as the reachability prior [26].

1) Agent-to-Agent Interaction: There are a vast number of models for agent-to-agent interaction in this field, and the social-LSTM [79] is the pioneering model for future state prediction of agents. Inspired by this, many variants of the social-LSTM-based interaction have emerged since 2016 [14], [85], [86]. In the social-LSTM, each agent is represented as an individual LSTM and shares social relations through a social pooling system, which has a promising ability for modeling the temporal dependence of the agent state but is limited when meeting crowd agents. To address this issue, Graph Neural Networks (GCN) offer a flexible solution for modeling social relations [87], [88], [89], [90], [91], [92]. Each node in the graph can represent diverse information of locations, velocity, agent types, etc. However, a drawback of GCN-based methods is their requirement for a temporally consistent number of agents, which is challenging to achieve in highly dynamic driving scenes.

2) Agent-to-Scene Interaction: For a long time, the road map or the occupancy map is encoded with a dense rasterized processing, which has been adopted in many popular trajectory prediction methods, such as DESIRE [93], IntentNet [84], CoverNet [94], Trajector++ [95], MultiPath [96], Target Driven Trajectory (TNT) [97], and so on. These methods typically encode the road map with Convolutional Neural Networks (CNN), while the structure of the road layout is not modeled well with the restricted perception field of CNN. MultiPath++ [98] extends MultiPath with an efficient polyline encoding for agent-to-scene relations, which exploits the region-to-region relation for a better prediction ability. However, the polyline representation requires accurate annotation, which can be challenging to obtain.

Agent-to-scene interaction shows a promising constraint for reducing the implausible trajectories [99]. The aforementioned agent-to-scene interactions need to pre-annotate or build the road map effectively, which limits flexibility in various situations when accurate road map information is not available. Recently, for agent-to-scene interaction, the scene graph has attracted more attention [100], [101]. Scene graphs capture social interaction knowledge from large-scale data, such as raw video or point cloud data, and can be effectively vectorized, as demonstrated by RoadScene2Vec [102] (see Fig. 7). Building on this insight, Song et al. [103] improved pedestrian crossing prediction performance by vectorizing the traffic scene graph in each frame.

3) Agent-to-Goal Interaction: Along with the driving scenes, BIP is commonly influenced by different intended goal areas. With this in mind, we can see that the importance of the road agents is different and changes depending on the specific driving scenes [104], [105], as illustrated in Fig 8(a). Within this context, the driver attention serves as a direct indicator of the important and preferred goals. Fig. 8(b) demonstrates that the intended fixation of drivers not only reveals their desired destinations but also helps identify hazardous objects [27].

For the agent-to-goal interaction, the “goal” is commonly represented as the intended destination coordinates [93], [107]. If the goals are known in advance, the goal-conditioned prediction can be inferred by the inverse optimal control [108] or inverse reinforcement learning [109]. However, in scenarios where the agent’s goals are not predefined, the agent-to-goal interaction becomes dynamic, as the agents’ behaviors and locations change over time, making it crucial to estimate the intended goals. In this case, the agent-to-scene interaction...
plays a significant role in constraining the potential goals of the agents [107]. For example, map-adaptive goal path [110] generates a set of possible goal-directed future path anchors based on the road lane constraint. The Goal Area Network (GANet) [111] models possible goal areas instead of exact goal coordinates for motion prediction. GANet estimates the possible goals by calculating the loss between the inferred goal locations and the endpoint of the ground-truth trajectories. Typically, these agent goal estimation methods require the pre-definition of multiple goal anchors (candidate goal coordinates) and conduct heuristic or rule-based goal selection. The quality of these goal anchors significantly impacts the prediction accuracy, as an incorrect estimation can directly affect the predicted future intentions. Target-driveN Trajectory prediction accuracy, as an incorrect estimation can directly coordinates) and conduct heuristic or rule-based goal selection.

Different types of interactions mentioned above are inter-connected. As for the BIP, the highly socialized driving scenes permeate various interactions but are challenging because of the dynamic behaviors of agents, frequent disappearance or the emergence of new objects, and complex road structures.

C. Prediction Uncertainty Estimation

“...It is far better to foresee even without certainty than not to foresee at all.” –Henri Poincare, Foundations of Science [113].

The inherent multi-modality, partial observability, short time scales, data limitation, intention type imbalance [114], domain gap [115], and deficiency can all cause uncertainty. Moreover, deep learning models, due to their generalizability, may introduce bias into the predicted distribution of behavioral intentions. There are two kinds of uncertainties: 1) the aleatoric uncertainty, also known as observation uncertainty, which stems from the inherent randomness or variability that presents in a system or process, and 2) the epistemic uncertainty, also referred to model uncertainty, which arises from limited knowledge or information about a system or process.

In particular, aleatoric uncertainty is inherent in physical systems (sensor ability) or environment (severe weather, low light condition, etc.). It cannot be eliminated even with complete knowledge or understanding of the underlying factors [116]. On the contrary, epistemic uncertainty, accounting for the model parameters, can be reduced through improved data collection, enhanced modeling techniques, or increased knowledge about the system.

1) Aleatoric Uncertainty in Prediction: In agent state prediction, handling aleatoric and epistemic uncertainties requires different approaches. To mitigate aleatoric uncertainty, additional information, such as High-Definition Map (HD Map), Birds’ Eye View (BEV), etc., are taken into account for future prediction. One recent approach, called StretchBEV [117], utilizes a full-range BEV representation to extend spatial scene understanding over longer time horizons compared to previous methods. Another method, MultiPath [96], leverages HD Map to generate multiple probabilistic anchor trajectory hypotheses and models future states using a Gaussian Mixture Model (GMM). This model incorporates an “intention uncertainty” to infer the coarse-scale intention or desired goals. Yalamanchi et al. [118] address the long-term future prediction with the uncertainty-aware trajectories with lane-based paths. To model the aleatoric uncertainty, various kinds of probability models are developed, such as the Gaussian model [95], GMM [120], [121]. However, Gaussian distributions often fail to adequately represent scene sensitivity due to their static and objective nature. The inherent multimodal nature of future road agent states further increases uncertainty. For example, a pedestrian in Fig 9 (a) may either continue along a sidewalk or cross a crosswalk.

2) Epistemic Uncertainty in Prediction: For epistemic uncertainty, various models incorporate multiple kinds of information or prior knowledge to reduce the prediction uncertainty. For example, in the context of interaction between different road agents, there is a concept called collaborative uncertainty (CU) [122] which arises from the dynamic nature of interactions. By considering CU, it becomes possible to evaluate the uncertainty associated with the predicted multi-modal states resulting from these interactions. Another kind of approach for handling model uncertainty is the incorporation of data knowledge from different datasets through cross-dataset domain adaptation. For example, Gesnouin et al. [123] have studied the cross-dataset generalization for pedestrian crossing intention prediction, and find that the dataset shift can degrade the quality of predictions, regardless of the model used, even when the training and testing distributions within each dataset are well-calibrated. To mitigate this issue, deep ensembles of multiple networks seem promising, as they have shown to be beneficial in improving the model performance under data shift [124].

a) What uncertainties do we need to consider for BIP?: Although there is little work for this question, we can seek the answer from the work on Bayesian deep learning [126], [127] in computer vision [116]. Aleatoric uncertainty becomes relevant when we have sufficient data or real-time demands, whereas epistemic uncertainty becomes crucial in safety-critical applications with limited data. As for BIP in the driving scene, the intention types of road agents are multitudinous. Consequently, it is impossible to collect enough data for practical use for each type of behavioral intention and may
involve many types with small-scale samples. In the meantime, each agent in the driving scene may have different intentions at each time step, which implies natural aleatoric uncertainty. Moreover, considering the influence of partial observation, addressing epistemic uncertainty can be approached through few-shot [128] or zero-shot [129] learning models with limited or important labels [125] (as shown in Fig. 9(b)). Human-machine hybrid intelligence will have a vital role in future prediction, as humans can help correct prediction errors in an active learning setting [130].

D. In Summary

Factors in BIP mainly have an essential impact on model design. Here is a summary of this section:

1) The Ego-View has a limited observation range, predicting crossing and not-crossing intentions are the primary tasks. However, with the development of BEV representation for autonomous driving systems, observation views are expected to become unified, providing a full range and clear observation of surrounding scenes. Currently, BEV observation is transformed from raw camera videos with deep feature learning. However, the performance of fine-grained road entities (static entities and dynamic agents) projection in BEV representation still needs to be enhanced.

2) The aforementioned interaction types coexist and depend on specific driving scenarios such as highways, urban roads, etc. Accurate agent attributes are essential for agent-to-agent interaction. To be concise and effective, agent-to-agent interaction needs to consider the neighborhood range and the importance of the agents because there is only a small proportion of the target agents that have influences on the ego agent. Agent-to-scene and agent-to-goal interaction require accurate road structure representations, which need to pre-construct the HD map in LIDAR sensors but are easily influenced by the weather and light conditions if the BEV representation transformed by cameras is used.

3) The main challenge for prediction uncertainty lies in the data shift and intention type imbalance. Recent approaches like Digital Twining (DT) [131] or Parallel Intelligence (PI) [132] show promise in addressing epistemic uncertainty by generating large-scale behavioral intention data in long-tailed and critical situations.

IV. AGENT-CENTRIC BIP

With the background definition and key factor descriptions, this section elaborates on the progress of pedestrian-centric and vehicle-centric BIP. These two kinds of agents are studied in different observation views, different formulations, and different scenarios in this field. In particular, we present the key novelties and the latest progress.

A. Pedestrian-Centric BIP

With the successful application of deep learning, pedestrian-centric BIP methods based on Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), Graph Neural Networks (GCNs), and Transformer networks have become popular in this field. The JAAD [35] and PIE [43] datasets have the dominant position for performance evaluation. Table. II provides a chronological overview of the pedestrian-centric BIP methods. These methods can be categorized into three stages of development.

1) Spatial-Temporal Modeling in Pedestrian-Centric BIP:

To our best knowledge, Volz et al. [133] for the first time introduce Deep Neural Networks (DNNs) into this field. They employ dense neural networks for pedestrian intention classification and utilize CNNs and LSTMs to predict pedestrian crossing intention. However, the approach solely considers image inputs without incorporating other features. Compared to traditional machine learning methods such as Support Vector Machines (SVMs), their DNN-based approach achieves a 10-20% increase in accuracy on their self-collected data. Another significant contribution is made by Saleh et al. [134], who transform the pedestrian-centric BIP problem into a sequential prediction task, employing stacked LSTM models to predict pedestrian crossings based on historical trajectories. Experimental results on the Daimler dataset indicate that this method exhibits lower displacement bias to the ground truth than traditional models.

Researchers have also started considering the impact of multiple sources of information on performance. Fang and López [135] build upon image inputs and introduce pedestrian Pose (PO) and 2D Boxes (2DB) information for predicting pedestrian crossing intention. Ablation studies show that including the PO and 2DB information significantly improves the accuracy of the predictions. Furthermore, Ghorii et al. [136] extend this work by incorporating LSTM structures to capture the temporal dynamics of human poses. Experimental results demonstrate that combining spatial and temporal information achieves a 0.72 F1 score for one-second predictions on the Daimler dataset. Varytimidis et al. [137] discover that the combination of SVM and CNN is also useful for leveraging the deep features in estimating pedestrian head direction and motion. This approach yields an accuracy of 0.89 for intention prediction in the JAAD dataset. Rasouli et al. [138] propose a stacked GRU structure, progressively fusing pedestrian locations, ego-vehicle motion, and pedestrian appearance. Their results show that the fusion of multiple sources of information (pedestrian context, surrounding context, full context, pose, displacement, bounding box, and speed) achieves an accuracy of 0.844 in the PIE dataset. Ablation experiments on feature fusion reveal that a multi-scale feature stacking in stacked GRU layers is promising in model performance. Concurrently, the work [139] demonstrates that the integration of appearance, environmental information, and pedestrian actions resulted in better results in JAAD and PIE datasets.

Previous studies have addressed the spatial and temporal aspects by separate modules in [140], [141], and [148], they only employ image inputs as cues but introduce 3DCNN and ConvLSTM to jointly model the spatial-temporal information for predicting pedestrian crossing intention. The experimental results have shown improved performance, with an average accuracy of 0.867 on the JAAD dataset compared to previous methods. Consequently, 3DCNN and ConvLSTM have
become popular in subsequent works [50], [149], [153], [157] in 2021. Nevertheless, due to the high computation cost associated with 3DCNN, the CNN, LSTM, and GRU remain common choices [29], [171], [173].

2) Interaction Modeling in Pedestrian-Centric BIP: With the emergence of Graph Convolutional Networks (GCNs), Cadena et al. [142] are among the early adopters of GCNs in this context. In their work, an adjacency matrix is computed to represent relationships between human pose key points, and these key points’ coordinates are taken as input for GCN, which generates an accuracy of 0.92 in the JAAD dataset. This work serves as a foundation for subsequent research in this area [48], [142], [163], [166]. Building upon these findings, GCN shows the powerful ability to model the key point
relation and agent location interaction in pedestrian crossing intention prediction. For instance, Liu et al. [48] construct the spatiotemporal graphs with each frame’s segments as nodes and model the spatiotemporal relationships between pedestrians and other objects. Similarly, Zhang et al. [163] propose a human skeleton data-based spatiotemporal GCN to simultaneously learn the spatial and temporal patterns, which achieves fast inference speed and promising performance.

Naik et al. [166], on the other hand, employ both images and Ego Vehicle Velocity (EVV) as inputs to construct a Spatial-Temporal Scene GCN (STS-GCN) for encoding the dynamic relationships between pedestrians and other objects. Experimental results highlight the positive influence of traffic lights, traffic signs, and zebra crossings on pedestrian crossing intention. To further enhance the capabilities of GCNs, Cadena et al. [162] extend their previous work [142] and propose Pedestrian-Graph+ [162]. In this approach, they modify the GCN structure from [142] to incorporate multimodal data information by modeling the node representation by 1D or 2D CNN for vehicle speed and image features, respectively. The ablation studies show that ego-vehicle speed affects accuracy the most. While GCN models show dominant superiority in human pose and object location interaction modeling, they struggle to explicitly reduce redundant relationships. Therefore, the affinity matrix often becomes high-dimensional when dealing with a large number of objects.

3) Attention Modeling in Pedestrian-Centric BIP: Since the invention of Transformer architecture [176], its capability to capture essential and attentive information has been widely explored. Pedestrian Crossing Prediction with Attention (PCPA) [158] is a milestone work for pedestrian crossing intention prediction with attentive modeling for different information, and has become a baseline in this field. IntFormer [152] is the first work that applies the Transformer structure to pedestrian crossing intention prediction, and better performance than PCPA [158] is obtained. Consistent with previous findings [162], experiments show that self-vehicle speed is the most crucial variable. Based on this, the works of [160], [165], [169], and [172] adopt the Transformer as the backbone model. Notably, Achaji et al. [165] achieve an accuracy of 0.91 on the PIE dataset using only bounding boxes. Recently, Rasouli et al. propose the Pedformer [172], which combines the historical Trajectory (T), Ego Vehicle Velocity (EVV), Image (I), and Image Segments (SS) through Multi-Head Attention (MHA) in Transformer. They also introduce a trajectory prediction task for multi-task learning alongside pedestrian crossing intention prediction. Pedformer achieves the highest accuracy (0.93) on the PIE dataset to date. Therefore, due to the inherent prediction uncertainty, multi-task learning and attention modeling have become a prominent pipeline in this field.

B. Vehicle-Centric BIP

Unlike the pedestrian-centric BIP task, vehicle-centric BIP primarily focuses on intention types such as Lane Changing (LC), Merging (M), Turning Left/Right (TL/TR), and Lane Keeping (LK). Table III presents a chronological overview of the vehicle-centric BIP methods.
1) Surrounding Vehicle (SV)-Centric BIP: The Surrounding Vehicle (SV)-centric BIP can provide an interactive understanding of scenes for the Ego Vehicle (EV) with a BEV observation, as depicted in Fig. 10, where the NGSIM and HighD are two common datasets in these situations. In particular, the Lane Changing (LC) intention and Lane Keeping intention are two primary types. Because of the trajectory data form in this situation, sequential networks, such as LSTM, are popular for modeling the temporal locations of vehicles [187]. For example, Dang et al. [177] treat LC prediction as a regression problem, which employs an LSTM network to predict the Time-To-Lane-Change (TTLC) by incorporating driver status, vehicle information, and environmental cues. Through the ablation experiments, this work obtains a 3.2% improvement in F1 score over the traditional SVM method in a self-collected dataset. Scheel et al. [180] introduce the Bidirectional LSTM (BiLSTM) to model the temporal correlation between vehicle trajectory points and LC intention. Compared with LSTM, BiLSTM obtains a 4% improvement, with an accuracy of 0.926 on the NGSIM dataset. Furthermore, the interaction between vehicles is considered in [182], which infers the Hybrid State System (HSS) between the Target Vehicle (TV) and SVs. Their experiments on the NGSIM dataset show accuracies of 0.953, 0.980, and 0.963 for LLC, RLS, and LK intentions, respectively. Similarly, the social interaction of SVs is also modeled in recent works [188], [189], [194], while differently, they take the Transformer to fulfill an attentive feature extraction for maneuver prediction [189], [194], while differently, they take the Transformer to fulfill an attentive feature extraction for maneuver prediction and maneuver-aware trajectory prediction of SVs. The ablation results on the NGSIM and HD maps demonstrate that the inclusion of the interaction module significantly improves the BIP performance.

SV-centric BIP commonly takes the trajectory under the BEV observation as input, allowing for direct measurement of LC, LK, CO, and CI intentions without geometrical distortion. However, the Surrounding-Vehicle (SV) intention is based on the exhibited behaviors, making the SV-centric BIP more like behavior prediction with a short prediction horizon. Contrarily, the Ego-Vehicle (EV)-centric BIP can leverage the maneuver status of the vehicle itself to fulfill more reasonable BIP.

2) Ego Vehicle (EV)-Centric BIP: For the Ego Vehicle (EV)-centric BIP, different from SV-centric BIP, the steering angle is a focused intention indicator. For example, an Adaptive Fuzzy Neural Network (AFFN) [179] fuses vehicle sensor data to predict Steering Angles (SA) of the Ego Vehicle (EV), thereby achieving LC intention prediction. In addition, compared with the monotonous LC or LK intention types, EV-centric BIP involves more diverse intention types. In this category, some formulations utilize voxelized 3D LiDAR data to obtain a rasterized road map. The Intent Network (IntentNet) [84] is a typical model that predicts seven kinds of intentions of EV by inputting BEV representation from the voxelized LiDAR data and the High-Definition (HD) map. Ablation experiment results demonstrate that the HD map can significantly enhance model performance, which verifies the crucial role of road structure representation in vehicle-centric BIP. Furthermore, EV-centric BIP can encode the camera videos, road maps, road entity segments, and social interaction in model inference. The scene context feature can be obtained to facilitate the accurate BIP. For example, Izquierdo et al. [183] introduce a CNN-LSTM model to encode RGB video frames, local and global scene context features, and temporal information of agent to achieve LC intention prediction, where the PREVENTION dataset provides a Cutting In (CI) and Cutting Out (CO) intention labels.

C. In Summary

To improve the differentiation between various intentions, most existing works exploit multiple clues. Pedestrian intention prediction typically relies on images (I), pose (P), and 2D Boxes (2DB), while vehicles often incorporate road structure information such as images (I), HD map (HD), Distance to Centerline (D2C), and vehicle velocity (v, EVV or MTV). From Table. II and Table. III, it is apparent that the intention types of pedestrians and vehicles differ, with Crossing (C) and Not Crossing (NC) being the main concerns for pedestrians, and Lane Changing (LC) being the primary focus for vehicles. This distinction is reasonable as “crossing warning” is a crucial function for assisted driving systems [197], and vehicle lane changing poses a frequent behavior that can potentially endanger other vehicles [8].

Most multi-clue intention prediction works do not evaluate the importance of different clues but simply fuse them in a “concat” strategy. Contrarily, “attentive fusion” provides a mechanism for selecting the important information, leading to better adaptation in different situations.

In retrospect, Deep Neural Networks (DNNs) have been widely applied in agent-centric BIP, with CNNs, LSTMs, and GRUs consistently delivering impressive performance. The integration of diverse input sources and feature fusion strategies has proven critical for achieving high model performance. Attentive fusion methods have shown notable enhancements and have attracted repeated concentration in recent works. The adoption of widely recognized benchmarks such as JAAD and PIE has shifted the evaluation of models from self-collected datasets to a more consistent and comparable framework.

V. BIP-AWARE APPLICATIONS

The accurate BIP is essential for understanding the future movement tendencies of road agents. It is useful for the
following trajectory prediction and behavior prediction tasks. How BIP promotes the other prediction tasks is described here.

A. BIP-Aware Agent Trajectory Prediction

The first prediction task is BIP-aware trajectory prediction, where the problem can be currently inferred by a multi-task learning prototype and a parameter conditioning prototype.

1) Multi-Task Learning Prototype: Formulating the joint prediction of behavioral intention and trajectory as a multi-task learning prototype can be easily considered and implemented. By incorporating an additional loss function with trajectory prediction loss, these two interconnected tasks can be inferred simultaneously [92], [144], [198], [199], [200].

In the Ego-View situations, the pedestrian crossing intention prediction recently leverages the future location or trajectory to make performance assistance. For example, Su et al. [199] treat the pedestrian crossing intention as an additional signal and incorporate an intention loss (cross-entropy of the intention labels) along with the L2 loss of the trajectory endpoint to facilitate trajectory prediction. The results show that crossing intention has promoted the trajectory prediction for the end-time step significantly. Rasouli et al. [150] formulate a multi-task prediction (BiPed) for pedestrian crossing intention, trajectories, and final grid location, as shown in Fig. 11(a). BiPed enhances the prediction performance of both pedestrian crossing intention prediction and trajectory prediction. The binary cross-entropy loss is used for the pedestrian crossing intention prediction. Sui et al. [200] introduce the Transformer to model the cross-attention of different information (locations and images) and also formulate the multi-task learning of pedestrian crossing intention and trajectory prediction. PedFormer [172] is a new work for pedestrian motion prediction with multi-task learning, where the learned feature is decoded by a Hybrid Gated Decoder constructed by stacked LSTM for crossing intention and future trajectory prediction.

Under the BEV observation, one kind of formulation for BIP-conditioned trajectory prediction is to leverage the future intended goal points to guide the prediction. The Retrospective-Memory-based Trajectory Prediction (RememNet) [196] combines the future intended goal points (named as future location intention) and trajectory prediction together, and infers the intention prediction with the MemoNet to reconstruct the compatible future trajectory and future intended goal points jointly, as shown in Fig. 11(b). The results of RememNet demonstrate that a suitable number selection of intended goal points is important for avoiding the intention modality missing and irrelevant instances. DROGON [92] fulfills a goal-oriented trajectory prediction network, which computes the probability of intended goal points based on the inferred interaction of vehicles, and estimates the label of future intention goal points by cross-entropy loss. Actually, the aforementioned goal-oriented trajectory prediction is also another kind of intention for conditioning the trajectory prediction. The goal estimation is also fulfilled by the cross-entropy loss [97], [110], [112], as investigated in Sec. III-B.3.

2) Parameter Conditioning Prototype: Parameter Conditioning Prototype for trajectory prediction usually models the intent as extra information to re-weight or re-constrain the trajectory distribution sampling function [201], [202], [203]. The Conditional VAE (CVAE) models [204] defined as follows are commonly adopted.

\[ p_\theta(y_i|X) = \int p_\theta(y_i|z_i, X) p_\theta(z_i|x_i) dz_i, \]

where \( p_\theta(z_i|x_i) \) denotes the conditional independence of the latent variables \( z_i \) under the agent observation \( x_i \in X \). Commonly, the intention is encoded in \( p_\theta(z_i|x_i) \), where the other conditions, such as interaction and road scene knowledge may also be encoded. Euro-PVI [205] models the interactive intention between the surrounding objects and ego-vehicles (e.g., yielding, decelerating, and crossing, etc.), and develops a Joint-β-CVAE to conduct the trajectory prediction, where the interaction intention is encoded as the latent variables in the CVAE formulation. The results verify that involving the interactive intention between pedestrians and vehicles could significantly reduce the ADE and FDE values. Sun et al. [206] also propose a CVAE model to jointly predict the intended goals and trajectories, which embeds the predicted goals and the interaction of agents with a Multiple layers Perception (MLP) at each time step. Recent work LOKI [51] treats the intended goals as a condition for scene graph construction, where the outputs of a Goal Proposal Network (GPN) and the agent intention prediction model are added to decode future trajectories, as shown in Fig. 11(c).

Recently, the trajectory prediction of vehicles in highway scenarios usually takes the parameter conditioning prototype. For example, Gao et al. [192] propose a dual Transformer to encode the past trajectories and interactive information, and decode the future Lane Changing (LC) intention. Then, the predicted LC intention probability is fed into the feature of past trajectories for the subsequent trajectory prediction. Compared
with the naïve Transformer, the Root-Mean-Squared Error (RMSE) value for future predicted trajectories is reduced by 7.52% and 27.3% on the NGSIM and HighD datasets, respectively [192]. Do et al. [195] take the LC intention as a prior for future trajectory prediction. Differently, they initialize the path generation step by a cubic spline curve in the Frenet Coordinate System (FCS), then predict the LC intention and future trajectories by a dynamic estimation of path probabilities.

Besides, other parameter-based intention prediction models, such as the Dynamic Bayesian Network (DBN) [46], [190], [207], [208], [209], are also explored. As for the deep learning era, the framework of DBN will be popular, where the feature extraction of the inference model may be fulfilled by deep learning modules. The work [210] firstly predicts the vehicle intention on the BEV sequence by a CNN model with the binary cross-entropy loss, and then fuses the predicted intent to the trajectory prediction with a multi-head attention decoder model. Ma et al. [34] propose a continual multi-agent behavior prediction work, which designs an episodic memory buffer and a conditionally generative memory to capture the historical interaction trajectories with the labeling of goal position and interaction intention. Wu et al. [46] fuse the pedestrians’ behavior, intention, and scene context to tackle the trajectory prediction problem. The pedestrian intention is inferred by DBN with the variables for the existence of the crossing area, waiting time, distance to curb, etc. The pedestrian crossing intention is treated as a bool variable to change the trajectory sampling function. In some works, the researchers fuse the intention and trajectory prediction as a sequential prediction problem, where the predicted trajectories are also useful for the intention prediction tasks. For example, Saleh et al. [134], [211] predict the long-term intention of pedestrians by a stacked LSTM over the trajectory points.

Discussion: Multi-task learning and parameter conditioning prototypes have gained popularity in recent years for BIP-aware trajectory prediction. When it comes to jointing the BIP task with trajectory prediction, the cross-entropy loss is frequently utilized. In the multi-task learning prototype, the parameters of modules on two tasks commonly share the weights. Contrarily, the parameter conditioning prototypes usually have separate weight sets. Explainability still is a core issue in the multi-task learning prototype. The parameter conditioning prototype seems to be appropriate because of the flexible dynamic networks and can provide a parameterized explainability for information importance modeling. Of course, the parameter conditioning prototype can also involve multi-task learning to optimize the models.

B. BIP-Aware Agent Behavior Prediction

In many related works, the terms behavior prediction and trajectory prediction have been interchanged and used by treating trajectory prediction as behavior prediction. We think these two tasks have intrinsic differences, where behavior prediction is a classification problem but trajectory prediction is commonly a regression problem.

Compared with trajectory prediction, behavior prediction shares a similar problem formulation with BIP, while behavior prediction determines the behavior label over a longer time window [212], [213]. In some cases, behavior prediction can be formulated as a sequential classification task for multiple future time steps. For example, when predicting pedestrian behavior, the prediction may change from the “will cross” to “crossing” for pedestrians, as shown in Fig. 12. Yao et al. [156] couple the crossing intention and crossing behavior of pedestrians, where the “standing”, “walking towards”, “crossing” and “crossed” actions are combined with the intention of “will cross or not cross”. The pedestrian behavior prediction is modeled as a sequential prediction problem solved by a multi-task inference and verifies that intention affects actions and future action is also useful for accurate intention prediction. Inspired by [156], Zhai et al. [214] propose a multi-task learning model for pedestrian crossing intention and behavior prediction. Differently, they propose a Spatial-Temporal Heterogeneous Graph (STHG) to model the relationships between pedestrians and surrounding dynamic and static road entities and improve the Average Precision (mAP) of [156] to 0.26 (+0.03) on the JAAD dataset for pedestrian behavior prediction within one second time. Banijamali et al. [215] develop an action-conditioned behavior prediction framework, where the prediction problem is formulated as latent probabilistic generative process \( p(o_{t+1} | o_{1:t}, a_t) \), where \( a_t \) represents the action at time \( t \) and \( o_{1:t} \) denotes the observation feature. They alternate between predicting the action \( a_t \) and the future state \( o_{t+1} \) to fulfill a “Prediction by Anticipation” framework. Li et al. [216] propose an interaction and behavior-aware driving behavior prediction framework based on joint predictions of intentions and motions of surrounding vehicles, which is fulfilled by a multi-modal hierarchical Inverse Reinforcement Learning (IRL) over the driving trajectory data. The driving behaviors are defined as aggressive, conservative, and moderate driving.

From the investigation, we find that compared with BIP-aware trajectory prediction, the research on BIP-aware behavior prediction is limited. Sometimes, this field treats behavioral intention and behavior prediction in the same concept, while they are rather different from the problem connotation [217]. Manifestly, the behavioral intention of road agents has a positive promotion role for long-term behavior prediction. In addition, similar to BIP-aware trajectory prediction, current BIP-aware behavior prediction works also utilize multi-task learning or parameter-conditioning prototypes. Differently, behavior prediction is a classification problem, where the explainability of models will be promising for trustworthy prediction but not explored in current works.
VI. EXTENSION AND DISCUSSION

Through the exhaustive investigation of behavioral intention prediction and its roles in other prediction tasks, we arrive at a full portrait of this topic. Here, we make an extension and discussion for BIP.

A. Benchmarks and Theories

1) Benchmarks: As discussed in Sec. II-B, most of the available benchmarks for BIP focus on the behavioral intention of Crossing (C), Not Crossing (NC) for pedestrians, Lane Changing (LC), and Lane Keeping (LK) for vehicles. In addition, the data observation views concentrate on the Ego-View, which cannot capture the full range of the road scene, and many types of behavioral intentions cannot be found, such as the “rear car following”, “overtaking from behind”, etc. A possible way is to add the behavioral intention label for the datasets with panoramic views, such as the Argoverse 3D dataset [54], nuScenes dataset [36], or KITTI-360 [218]. It is also interesting to introduce viewpoints from novel devices, e.g., drone and satellite, for comprehensive scene understanding [219]. Besides, the intention types in current datasets are not fine-grained enough, and the intention type imbalance issue is universal. In the future, it would be beneficial to consider more detailed and nuanced interactive intention types between road agents with other road entities. For example, pedestrians of different ages and genders often show different behavioral intentions on the road. Furthermore, the safe-critical scenarios with long-tailed distribution or harsh environments (e.g., rainy, foggy, snowy, windy, and low-light conditions) also need to be considered.

It is worth mentioning that the evaluation of performance in current works is based on different datasets, leading to evident performance gaps due to data shifts [158]. For instance, a recent study [123] highlights the poor performance of state-of-the-art models for pedestrian crossing intention prediction when evaluated on different datasets such as JAAD and PIE. To address this, the authors introduce the confidence calibration metrics, i.e., Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [220] as complementary evaluation measures, and find significant differences between ECE and MCE. Additionally, leveraging the pre-trained model from diverse source datasets can enhance the generalization ability of target datasets. For the BIP problem, besides the crossing intention types, multiple kinds of behavioral intentions and the uncertainty estimation of model calibration in multi-label classification problems need to be explored. Furthermore, with the development of deep learning models, measuring the calibration of data and models [221] is also a fundamental concern for trustworthy implementation.

2) Theories: Despite the numerous works on BIP that have exhibited significant progress in performance, most of the current works on BIP are all based on CNN, LSTM, ConvLSTM, Transformer, GCN, etc., as described in Sec. IV. These deep learning models are all deterministic neural networks for achieving a mapping from input space to output space, which is usually overconfident in the testing phase. Consequently, one self-calibrated deep learning approach on one benchmark faces the data shift issue in the evaluation and may suffer from either over-fitting or under-fitting issues when faced with datasets that contain simpler or more diverse samples, respectively.

For the deterministic neural networks, current research efforts employ domain adaptation to address this problem by using a well-pretrained model on large-scale datasets or leveraging more complex architectures. For example, vision-language pre-trained models, such as BEnt-3 [222] and VinVL [223], learn an informative representation with the help of dense semantics in language. However, although these pre-trained models can generate a good representation, the domain gap in BIP is still large and needs further valuable inference models. We think the possible ways for developing the new theories on BIP should consider the influencing factors (described in Sec. III) as aforementioned, such as the better adoption of the road structure representation, social interaction modeling, and robust estimation of the prediction uncertainty. Standing at the natural characteristics of multiple clues and preferring aims in the BIP problem, incorporating more explainable scene representation with scene knowledge, such as scene graphs [224], could be beneficial.

Essentially, fusing more clues could reduce the aleatoric uncertainty as aforementioned. More information provides more constraints for future intention prediction, while it gives rise to a fundamental problem on how to fuse this information in the best way, as certain clues may be contradictory in certain situations. Various Dynamic Neural Networks (DNNs) [225] have shown promise for adaptively selecting multi-modal information in different situations. Two examples of such models are Dynamic Multimodal Fusion (DynMM) [226] and Dynamic Routing Network (DRN) [227]. DynMM utilizes a Gating Network (GN) to dynamically fuse modalities, while DRN achieves this through a router network. The GN selects the best expert network for the final decision, and it can also be used in the feature embedding stage to enable selective multi-modal encoding.

Besides the deterministic neural networks, stochastic neural networks are capable of estimating prediction distributions. This is particularly useful for addressing prediction uncertainty caused by data shift, Out-of-Distribution (OOD) samples (i.e., unfamiliar behavioral intention), the inherent nature of behavioral intention, and the long-term prediction scenarios. Existing works [228] estimate the distribution uncertainty by the Bayesian neural networks, generative adversarial networks, CVAE, or deep ensembles [229] by introducing uncertainty consistency loss in the Bayesian latent variable models [230]. Therefore, a promising direction is to develop models that explicitly consider prediction uncertainty.

B. Parallel Testing

Parallel testing refers to the collaborative use of real and synthetic data for the formulation and evaluation of BIP models. As mentioned earlier in Sec. III, it is crucial to identify the natural relationship between road entities and gather a sufficient number of data samples. However, in practical scenarios, it is challenging to collect enough samples that cover all causal relations, diversity, and long-tailed behavioral
intention types in safety-critical driving scenes. Consequently, we are constantly faced with the issue of data imbalance. Therefore, more and more works have started utilizing various virtual simulation tools (e.g., CARLA [231], GTA-V [232], etc.) to generate diverse driving scenes. We call it Simulation Augmentation (SA) in this paper.

The core problems in SA are to transfer the scene consistency from real to synthetic data, and maintaining the diversity from synthetic to real scenarios, which generates the possible future state with a parallel evolution [233].

1) Real-to-Synthetic Generation: Within this domain, many kinds of virtual engines are adopted with high-fidelity driving scene rendering [234]. These advanced simulators hold great potential for significantly advancing research on Behavioral Intention Prediction (BIP) in the coming years. For example, Chen and Krahenbuhl [235] have developed a virtual multi-vehicle collaboration environment to study the BIP of Ego Vehicle (EV) by predicting the future intentions of Surrounding Vehicles (SVs). TrafficSim [236] can flexibly generate the behaviors of “U-Turn”, “Yielding”, and “Merging”, etc., for road vehicles.

SA is gradually becoming an essential technique for the reasoning of safe-critical driving scenarios. It has a direct relation with Digital Twining (DT) [131] or Parallel Intelligence (PI) [132] in the driving scene. As the Metaverse continues to gain popularity, the interaction between the virtual and real world will become a core basis for understanding our world.

2) Synthetic-to-Real Adaptation: Synthetic-to-real adaptation can absorb the superiority of various simulators for generating vast amounts of data in different weather, light, and road conditions. The data with long-tailed distribution or adverse weather conditions can be collected efficiently.

This field has made some attempts at pedestrian crossing intention prediction with the assistance of synthetic data. For example, the work [165] transfers the dynamics of the bounding box from synthetic data to real data. Another work [52] constructs 4667 sequences with “C” or “NC” intention and models a virtual-to-real deep distillation for the lightweight pedestrian crossing intention prediction. Different from the works that address the BIP with the Ego-View observation, Kim et al. [237] propose a pedestrian crossing intention prediction model with the pedestrians’ view with a Virtual Reality (VR) apparatus. Although synthetic data can boost the diversity of the scenarios, there is a large distribution gap between the synthetic data and real data. Therefore, models trained on synthetic data often show degraded generalization to real data [238]. Recently, Zhou et al. [239] presented a survey for the domain generalization problem and exhibit the core solutions for better synthetic-to-real adaptation.

C. Counterfactual Analysis

1) Causality Inference in BIP: Upon reviewing the agent-centric BIP works in Sec. IV, we can see that there are few attempts to consider the explainable models. The attention mechanism (e.g., self-attention) shows an initial beginning for important feature learning [240]. However, what clue is crucial for BIP? One potential worth exploring is causality inference.

Causal relationships [190] or factor relations [241] among road agents are involved in future state prediction, and construct the non-visible “Dark Matter” [242] for motivating the behavioral intentions. Chen et al. propose a scene-consistent, policy-based trajectory prediction method that starts by constructing a scene graph based on the agents’ distance. This graph is then divided into multiple cliques, forming the foundation for constructing a factor graph. This approach enables the analysis of conditioning and counterfactual analysis [243] in prediction. Taking Fig. 13 as an example, the intention of the vehicle A is influenced by the causal chain involving vehicle B conditioned by the accelerating vehicle Hu et al. [190] contribute a causal-based time series domain generalization model for predicting vehicle intentions. The causal knowledge is derived from the road topology, speed limits, and traffic rules.

With the causal or factor relation, counterfactual analysis can find the primary cause or the scene knowledge for the specific prediction results by imagining a change in the input state. For example, Li et al. [244] explore the causality in identifying risky objects by masking the front agents. This formulation is also adopted by the STEEK model [245] for the intention decision model (e.g., Stopping or Moving Forward), where “region-targeted counterfactual explanations” is introduced and could generate meaningful counterfactuals with a preserved scene layout and relevant traffic light changing. In addition, some recent works [246], [247] begin to investigate the robustness of future prediction by attacking the input observations. These approaches aim the explainability by changing the semantic or scene state and checking the influence on the outcome for finding the primary input state. Causal relation has been observed by the safety-critical driving scenario generation, such as the CausalAF [248] that aligns with the behavioral graphs. CausalAF integrates the Causal Order Masks (COM) to generate possible cause-effect relations for the road scene and the Causal Visible Mask (CVM) to filter the non-causal information. The causality has a natural relationship with the social interaction of road agents. Therefore, the causality does not just correlate with the static road entities, but also the dynamic action or pose of the agents.

D. Promising BIP-Aware Applications

From the investigation, we find that there are few research efforts on BIP-aware behavior prediction. Behavioral intention is the most direct promotion for certain behaviors and can enlarge the Time-to-Collision (TTC) for collision avoidance.
The BIP problem has a direct link with risk assessment in driving, as shown in Fig. 2. Recently, the collision risk prediction work [249] is modeled by inferring the hidden intention of surrounding objects. Similarly, Kim et al. [250] learn to identify dangerous vehicles using a simulator, which learns the crash patterns in the real accident video data and constructs a GTACrash dataset. The crash label is refined by predicting the future paths of other vehicles. VIENA² [41] is a promising benchmark with the synthetic data for the prediction of crashes, pedestrian intention (e.g., Crossing, Walking, Stopping), and front car's intention (e.g., Stopping, Turning Right/Left, and Left/Right Lane Changing).

In addition, Vehicle-to-Vehicle (V2V) or Vehicle-to-Anying (V2X) cooperation (internet of vehicles) [251] and road-vehicle collaboration [252] are promising applications with the help of other vehicles’ perception and large-scale cloud data. For instance, some attempts [188], [253] predict the pedestrian crossing intention from the cooperative vehicles' view. This kind of formulation can capture a larger range for road structure representation than a single vehicle’s view. Within these applications, consistent and shared behavioral intention understanding is an important problem. For example, the LC intention for a vehicle may be understood as a Vehicle Overtaking (VO) intention because of the location difference for the Ego Vehicle (EV). Therefore, group-wise consistent understanding [254] in collaboration is promising with a reasonable spatial and temporal perception window partitioning. DeepAccident [255] is a new dataset for accident understanding in virtual V2V scenarios. This dataset annotates the collision events and vehicle trajectories, which may be useful for BIP-aware crash anticipation in V2V situations.

VII. CONCLUSION

This paper presents a comprehensive review of Behavioral Intention Prediction (BIP), with an investigation of the datasets, intention types, key factors, challenges, agent-centric BIP, and BIP-aware prediction applications. With the definition of different prediction tasks and the introduction of available datasets, the key factors and challenges are summarized from the aspects of road structure representation, social interaction modeling, and prediction uncertainty. Based on this, we chronologically review the pedestrian- and vehicle-centric BIPs from different modeling approaches. The BIP-aware trajectory prediction and behavior prediction are described and highlight the potential for further development in BIP-aware behavior prediction. With a one-to-one response to potential challenges and possible insights, we discuss the theories and benchmarks, counterfactual analysis, parallel testing, and promising BIP-aware applications. We hope this survey can provide a good promotion for future BIP research.

REFERENCES

[1] J. D. Velleman, “Intention, plans, and practical reason,” Phil. Rev., vol. 100, no. 2, pp. 277–284, 1991.
[2] B. F. Malle and J. Knobe, “The folk concept of intentionality,” J. Experim. Soc. Psychol., vol. 33, no. 2, pp. 101–121, Mar. 1997.
[3] J.-R. Xue, J.-W. Fang, and P. Zhang, “A survey of scene understanding by event reasoning in autonomous driving,” Int. J. Autom. Comput., vol. 15, no. 3, pp. 249–266, Jun. 2018.
[4] I. Teeti, S. Khan, A. Shahbaz, A. Bradley, and F. Cuzzolin, “Vision-based intention and trajectory prediction in autonomous vehicles: A survey,” in Proc. 31st Int. Joint Conf. Artif. Intell., Jul. 2022, pp. 5630–5637.
[5] N. Sharma, C. Dhiman, and S. Indu, “Pedestrian intention prediction for autonomous vehicles: A comprehensive survey,” Neurocomputing, vol. 508, pp. 120–152, Oct. 2022.
[6] D. E. Benachouch, S. Glaser, M. Ellenhaw, and A. Rakotoniray, “Use of social interaction and intention to improve motion prediction within automated vehicle framework: A review,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 12, pp. 22807–22837, Dec. 2022.
[7] I. Gomes and D. Wolf, “A review on intention-aware and interaction-aware trajectory prediction for autonomous vehicles,” 2023, techrxiv:19337447.v1.
[8] Y. Xing et al., “Driver lane change intention for intelligent vehicles: Framework, survey, and challenges,” IEEE Trans. Veh. Technol., vol. 68, no. 5, pp. 4377–4390, May 2019.
[9] 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.
[10] T. Chen and R. Tian, “A survey on deep-learning methods for pedestrian behavior prediction from the egocentric view,” in Proc. IEEE Int. Transp. Syst. Conf. (ITSC), Sep. 2021, pp. 1898–1905.
[11] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzaktis, “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. 2020.
[12] 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.
[13] A. Bighashdel and G. Dubbelman, “A survey on path prediction techniques for vulnerable road users: From traditional to deep-learning approaches,” in Proc. IEEE Intell. Transp. Syst. Conf. (ITSC), Oct. 2019, pp. 1039–1046.
[14] 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.
[15] R. Korbmacher and A. Tordeux, “Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 12, pp. 24126–24144, Dec. 2022.
[16] F. Leon and M. Gavrilescu, “A review of tracking and trajectory prediction methods for autonomous driving,” Mathematics, vol. 9, no. 6, p. 660, Mar. 2021.
[17] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, “A survey on trajectory-forecasting methods for autonomous driving,” IEEE Trans. Intell. Vehicles, vol. 7, no. 6, pp. 652–674, Sep. 2022.
[18] J. Liu, X. Mao, Y. Fang, D. Zhu, and M. Q.-H. Meng, “A survey on deep-learning approaches for vehicle trajectory prediction in autonomous driving,” in Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO), Dec. 2021, pp. 978–985.
[19] M. Guizar, Y. Muhammad, and N. Muhammad, “A survey on motion prediction of pedestrians and vehicles for autonomous driving,” IEEE Access, vol. 9, pp. 137957–137969, 2021.
[20] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” Int. J. Comput. Vis., vol. 130, no. 5, pp. 1366–1401, May 2022.
[21] R. Huang, H. Xue, M. Pagnucco, F. Salim, and Y. Song, “Multimodal trajectory prediction: A survey,” 2023, arXiv:2302.10463.
[22] Z. Ding and H. Zhao, “Incorporating driving knowledge in deep learning based vehicle trajectory prediction: A survey,” IEEE Trans. Intell. Vehicles, vol. 8, no. 8, pp. 3996–4015, Jul. 2023.
[23] C. Zhang and C. Berger, “Pedestrian behavior prediction using deep learning methods for urban scenarios: A review,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 10, pp. 10279–10301, Nov. 2023.
[24] M. Golchoubian, M. Ghafourian, K. Dautenhahn, and N. L. Azad, “Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 10, pp. 11544–11567, Nov. 2023.
[25] A. Rasouli and K. J. 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.
[26] O. Makansi, Ö. Çiçek, K. Bucichcio, and T. Brox, “Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4353–4362.
Y. Zhao, Y. Zhang, Z. Gong, and H. Zhu, “Scene representation in bird’s-eye view from surrounding cameras with transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2022, pp. 4510–4518.

D. Cui, J. Xue, and N. Zheng, “Real-time global localization of robotic cars in lane level via lane marking detection and shape registration,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 4, pp. 1039–1050, Aug. 2016.

H. Li, C. Xue, F. Wen, H. Zhang, and W. Gao, “BSP-MonoLoc: Basic semantic primitives based monocular localization on roads,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2021, pp. 5470–5475.

Z. Li et al., “BEVFormer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” in Proc. IEEE Multisensory Perception, Proc. Conf. Robot Learn. (CoRL), vol. 155, Aug. 2020, pp. 8784–8781.

S. H. Park et al., “Diverse and admissible trajectory forecasting through multimodal context understanding,” in Proc. ECCV, vol. 12356, 2020, pp. 282–298.

J. Fang, C. Zhu, P. Zhang, H. Yu, and J. Xue, “Heterogeneous trajectory forecasting via risk and scene graph learning,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 11, pp. 12078–12091, Aug. 2023.

T. Li et al., “Graph-based topology reasoning for driving scenes,” 2023, arXiv:2304.05227.

A. V. Malawade, S.-Y. Yu, B. Hsu, H. Kaeyle, A. Karra, and M. A. A. Faruque, “roadscene2vec: A tool for extracting and embedding road scene-graphs,” Knowl.-Based Syst., vol. 242, Apr. 2022, Art. no. 108245.

X. Song, M. Kang, S. Zhou, J. Wang, Y. Mao, and N. Zheng, “Pedestrian intention prediction based on traffic-aware scene graph model,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2022, pp. 9851–9858.

E. Ohn-Bar and M. M. Trivedi, “Are all objects equal? Deep spatio-temporal importance prediction in driving videos,” Pattern Recognit., vol. 64, pp. 425–436, Apr. 2017.

A. Rasouli and J. K. Tsotsos, “Joint attention in driver-pedestrian interaction: From theory to practice,” 2018, arXiv:1802.02522.

R. Rhinehart, R. Mcallister, K. Kitani, and S. Levine, “PRECOG: Prediction conditioned on goals in visual multi-agent settings,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2821–2830.

L. F. Chiara, P. Coscia, S. Das, S. Calderara, R. Cucchiara, and L. Ballan, “Goal-driven self-attentive recurrent networks for trajectory prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2022, pp. 2517–2526.

K. M. Kitani, B. D. Ziebart, and M. Hebert, “Activity forecasting,” in Proc. ECCV, vol. 7575, 2012, pp. 201–214.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, “Maximum entropy inverse reinforcement learning,” in Proc. 23rd AAAI Conf. Artif. Intell., vol. 8, Jul. 2008, pp. 1433–1438.

L. Zhang et al., “Map-adaptive goal-based trajectory prediction,” in Proc. Conf. Robot Learn. (CoRL), 2020, pp. 1371–1383.

M. Wang et al., “GANet: Goal area network for motion forecasting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2022, pp. 947–956.

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. 12085–12094.

W. Wang, L. Wang, C. Zhang, C. Liu, and L. Sun, “Social interactions for autonomous driving: A review and perspectives,” Found. Trends Robot., vol. 10, nos. 3–4, pp. 198–376, 2022.

A. Mohamed, K. Qian, M. Elohenise, and C. Claudel, “Social-STGNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14398–14407.

H. Xue, D. Q. Huynh, and M. Reynolds, “Scene gated social graph: Pedestrian trajectory prediction based on dynamic social graphs and scene constraints,” 2020, arXiv:2010.05507.

X. Li, X. Ying, and M. C. Chuah, “GRID: Graph-based interaction-aware trajectory prediction,” in Proc. ITSC, 2019, pp. 3960–3966.

V. Kosaraju, A. Sadeghian, R. Martin-Martin, I. D. Reid, H. Rezatofighi, and S. Savarese, “Social-BiGAT: Multimodal trajectory forecasting using bicycle-gan and graph attention networks,” in Proc. NeurIPS, 2019, pp. 137–146.

C. Choi, S. Malla, A. Patil, and J. H. Choi, “DROGON: A trajectory prediction model based on intention-conditioned behavior reasoning,” in Proc. CoRL, vol. 155, 2020, pp. 49–63.

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.

T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, “CoverNet: Multimodal behavior prediction using trajectory sets,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 14074–14083.

T. Salemann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectory++: Dynamically-feasible trajectory forecasting with heterogeneous data,” in Proc. Eur. Conf. Comput. Vis., Aug. 2020, pp. 683–700.

Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, “Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction,” in Proc. Conf. Robot Learn. (PMLR), Oct. 2019, pp. 86–99.

H. Zhao et al., “TNT: Target-driven trajectory prediction,” in Proc. Conf. Robot Learn. (CoRL), vol. 155, Aug. 2020, pp. 985–904.

B. Varadarajan et al., “MultiPath++: Efficient information fusion and trajectory aggregation for behavior prediction,” in Proc. ICRA, 2022, pp. 1976–1981.
C. Zhang et al., “Cross or wait? Predicting pedestrian interaction at unmarked crossings,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2023, pp. 1–8.

A. Rasoni and I. Kotseruba, “PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and guided multitask learning,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2023, pp. 9844–9851.

D. Guo, T. Mordan, and A. Alahi, “Pedestrian stop and go forecasting with hybrid feature fusion,” in Proc. Int. Conf. Robot. Autom. (ICRA), May 2022, pp. 940–947.

Y. Zhou, G. Tan, R. Zhong, Y. Li, and C. Gou, “PIT: Progressive interaction transformer for pedestrian crossing intention prediction,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 12, pp. 14213–14225, Dec. 2023.

S. Ahmed, A. A. Bazi, C. Saha, S. Rajbhandari, and M. N. Huda, “Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation,” Expert Syst. Appl., vol. 225, Sep. 2023, Art. no. 120077.

A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inform. Process. Syst. (NIPS), 2017, pp. 5998–6008.

H. Q. Dang, J. Fürnkranz, A. Biedermann, and M. Hoepf, “Time-to-lane-change prediction with deep learning,” in Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC), Oct. 2017, pp. 1–7.

Y. Hu, W. Zhan, and M. Tomizuka, “Probabilistic prediction of vehicle semantic intention and motion,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2018, pp. 307–313.

J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, “Lane-changes prediction based on adaptive neural network,” Expert Syst. Appl., vol. 91, pp. 452–463, Jan. 2018.

O. Scheel, L. Schwarz, N. Navab, and F. Tombari, “Situation assessment for planning lane changes: Combining recurrent models and prediction,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2018, pp. 2082–2088.

J. Tang, S. Yu, F. Liu, X. Chen, and H. Huang, “A hierarchical prediction model for lane changes based on combination of fuzzy C-means and adaptive neural network,” Expert Syst. Appl., vol. 130, pp. 265–275, Sep. 2019.

T. Han, J. Jing, and Ü. Özgüner, “Driving intention recognition and lane change prediction on the highway,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2019, pp. 957–962.

R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, “Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM,” in Proc. IEEE Intell. Transp. Syst. Conf. (ITSC), Oct. 2019, pp. 3657–3662.

A. Zynor, S. Worrall, and E. Nebot, “Naturalistic driver intention and path prediction using recurrent neural networks,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 4, pp. 1584–1594, Apr. 2020.

V. Mahajan, C. Katrakazas, and C. Antoniou, “Prediction of lane-changing maneuvers with automatic labeling and deep learning,” Transp. Res. Rec., J. Transp. Res. Board, vol. 2674, no. 7, pp. 336–347, Jul. 2020.

A. Girma, S. Amsalu, A. Workineh, M. Khan, and A. Homaifar, “Deep learning with attention mechanism for predicting driver intention at intersection,” in Proc. IEEE Intell. Vehicles Symp. (IV), Oct. 2020, pp. 1183–1188.
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.

Z. Li, C. Lu, Y. Yi, and J. Gong, “A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 7, pp. 9102–9114, Jul. 2022.

Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, “Multi-modal probabilistic prediction of interactive behavior via an interpretable model,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2019, pp. 557–563.

X. Zhai, Z. Hu, D. Yang, L. Zhou, and J. Liu, “Social aware multi-modal pedestrian crossing behavior prediction,” in Proc. ACCV, 2022, pp. 4428–4443.

E. Banjanami, M. Rohani, E. Amirloo, J. Luo, and P. Poupart, “Prediction by anticipation: An action-conditioned prediction method based on hierarchical learning,” in Proc. IEEE/CVF Int. Conf. Intell. Vis. (ICCV), Oct. 2021, pp. 15621–15630.

D. Li, Y. Wu, B. Bai, and Q. Hao, “Behavior and interaction-aware motion planning for autonomous driving vehicles based on hierarchical intention and motion prediction,” in Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC), Sep. 2020, pp. 1–8.

I. Ajzen and M. Fishbein, “Factors influencing intentions and the intention-behavior relation,” Human Relations, vol. 27, no. 1, pp. 1–15, Jan. 1974.

Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3292–3310, Mar. 2023.

Z. Zheng, Y. Wei, and Y. Yang, “University-1652: A multi-view multi-source benchmark for drone-based geo-localization,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1395–1403.

M. P. Naeini, F. Gregory Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using Bayesian binning,” in Proc. 29th AAAI Conf. Artif. Intell., 2015, pp. 2901–2907.

J. Nixon, M. W. Dunsterberry, L. Zhang, G. Jerfel, and D. Tran, “Measuring calibration in deep learning,” in Proc. CVPR Workshops, vol. 2, 2019, pp. 38–41.

W. Wang et al., “Image as a foreign language: BEIT pretraining for vision and vision-language tasks,” in Proc. IEEE/CVF Conf. Intell. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 19175–19183.

P. Zhang et al., “ViViL: Revisiting visual representations in vision-language models,” in Proc. IEEE/CVF Conf. Intell. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5579–5588.

J. Devarajan, A. Kar, and S. Fidler, “Meta-Sim2: Unsupervised learning of scene structure for synthetic data generation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 715–733.

Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, “Dynamic neural networks: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

Z. Xue and R. Marculescu, “Dynamic multimodal fusion,” in Proc. IEEE/CVF Conf. Intell. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2023, pp. 2575–2584.

S. Cai, Y. Shu, and W. Wang, “Dynamic routing networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2021, pp. 3587–3596.

J. Zhang et al., “Dense uncertainty estimation,” 2021, arXiv:2110.06427.

B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Proc. Annu. Neural Inf. Process. Syst., 2017, pp. 6402–6413.

S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft, “Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 84–93.

A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in Proc. Conf. Robot Learn., 2017, pp. 1–16.
Jianwu Fang (Member, IEEE) received the Ph.D. degree in signal and information processing (SIP) from the University of Chinese Academy of Sciences, China, in 2015. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an, China. He was a Visiting Scholar with NExT++ Research Centre, School of Computing, National University of Singapore (NUS), Singapore. He has published many papers in top-ranked journals and conferences, such as IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, AAAI, ICRA, and ITSC. His research interests include computer vision, multimedia analysis, and their applications in intelligent transportation. He received the Best Paper Nominees Award from ITSC 2023. He is the Chair of the special sessions of “Parallel Vision for Intelligent Vehicles” (ITSC 2022) and the “Safe-Critical Scenario Understanding in Intelligent Transportation Systems” (ITSC 2023).

Fan Wang received the B.S. degree in transportation from Chang’an University in 2022, where he is currently pursuing the master’s degree in transportation. His research interests include intelligent transportation and behavior prediction.

Jianru Xue (Member, IEEE) received the B.S. degree from Xi’an University of Technology in 1994 and the M.S. and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 1999 and 2003, respectively. He joined the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, in 1999, where he is currently a Full Professor. He worked with Fuji Xerox, Tokyo, Japan, from 2002 to 2003, and visited the University of California at Los Angeles, Los Angeles, CA, USA, from 2008 to 2009. His research interests include computer vision, visual localization and navigation, and video coding based on analysis. He and his team were winners of the IEEE ITS Institute Lead Award in 2014. He and his students won the Best Application Paper Award from the 2012 Asian Conference on Computer Vision.

Tat-Seng Chua received the Ph.D. degree from the University of Leeds, U.K. He was the Acting and Founding Dean of the School from 1998 to 2000. He is currently the KITHTC Chair Professor of the School of Computing, National University of Singapore. He is the Co-Director of NExT, a joint Center between NUS and Tsinghua University to develop technologies for live social media search. He is the Co-Founder of two technology startup companies in Singapore. His main research interests include multimedia information retrieval and social media analytics. In particular, his research focuses on the extraction, retrieval, and question-answering (QA) of text and rich media arising from the Web and multiple social networks. He was the 2015 winner of the prestigious ACM SIGMM Award for Outstanding Technical Contributions to Multimedia Computing, Communications, and Applications. He is the Chair of the Steering Committee of the ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. He is also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves on the editorial boards for four international journals.