Behavioral Intention Prediction in Driving Scenes: A Survey

Jianwu Fang, Fan Wang, Peining Shen, Zhedong Zheng, Jianru Xue, and Tat-seng Chua

Abstract—In the driving scene, the road participants usually show frequent interaction and intention understanding with the surrounding. Ego-agent (each road participant itself) conducts the prediction of what behavior will be done by other road users all the time and expects a shared and consistent understanding. For instance, we need to predict the next movement of other road users and expect a consistent joint action to avoid unexpected accident. Behavioral Intention Prediction (BIP) is to simulate such a human consideration process and fulfill the beginning time prediction of specific behaviors. It provides an earlier signal promptly than the specific behaviors for whether the surrounding road participants will present specific behavior (crossing, overtaking, and turning, etc.) in near future or not. More and more works in BIP are based on deep learning models to take advantage of big data, and focus on developing effective inference approaches (e.g., explainable inference, cross-modality fusion, and simulation augmentation). Therefore, in this work, we focus on BIP-conditioned prediction tasks, including trajectory prediction, behavior prediction, and accident prediction and explore the differences among various works in this field. Based on this investigation and the findings, we discuss the open problems in behavioral intention prediction and propose future research directions.

Index Terms—Behavioral intention prediction, prediction uncertainty, explainable AI, cross-modality fusion, simulation augmentation

1 Introduction

Driving scene is a highly socialized environment. Any movement of road participants involves the adequate and accurate intention understanding of the surroundings. For example, the instance of whether the pedestrian or vehicle will cross the road or overtake the ego-car has a direct influence on the decision making of safe driving. Most often, we require the intention understanding to have an earlier timeline than the emergence of a specific behavior. However, signals for early prediction are often difficult to be observed though they can be commonly inferred by the road structure [1], road user attention [2], and other prior knowledge, such as the gender, social and culture factors, etc. [3].

Behavioral intention reveals the subjective tendency of road participants to take specific actions or achieve a specific target, which is usually understood as the internal reason for presenting specific behaviors [4], [5]. Therefore, the observation of behavioral intention is generally judged by the presented specific behavior. In the September 30, 2020 issue of Science Robotics, it is clarified that the neural and psychological mechanisms that increase human understanding of the surrounding environment for machine systems can provide ideas for designing the interactive autonomous systems [6]. Loading the behavioral intention understanding of traffic participants can make various perception systems in road vehicles to be integrated into the society better and improve the service level. Hence, Behavioral Intention Prediction (BIP) of road participants could improve the cognitive level of autonomous systems and help to guarantee the safety of all road users. Nowadays, with the demand for autonomous driving growing vigorously at home and abroad, the corresponding scale of data also grows rapidly, which provides a fertile soil for deep learning based behavioral intention prediction. This survey explores the issues, challenges and possible trends for the BIP in the deep learning era. It is worth noting that the intention prediction in this work is different from the psychological intention which is a common thinking in mind [7] without any behavior. Behavioral Intention Prediction (BIP) implies the forecasting of the beginning time of the specific behavior, e.g., the status of "*will cross*", "*will overtake*", *etc.*

With the beautiful vision for reducing the road fatalities by replacing the human drivers by self-controlled autonomous systems, numerous works have concentrated on the detection, segmentation, tracking, re-identification, and prediction of the behaviors of road participants over the past decades, and some previous surveys [8], [9] have summarized the pipelines and problems within this field. Complementary to our survey, Rasouli and Tsotsos [3] explore the interaction between vehicles and pedestrians with the theory and practice view and summarize the influencing factors for pedestrian behavior, especially for crossing. Recently, the same team releases a new survey for the algorithm and datasets for the driver attention modeling in assisted and autonomous driving systems [10], which focuses on the drivers' perspective for safe decision making. Xue *et al.* [11] propose the event reasoning survey in autonomous driving, where the long-term and short term intention prediction is reviewed while concentrated on the

This work is supported in part by the National Natural Science Foundation of China under Grant 62036008 and 62273057; in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2022JM-309.

J. Fang, F. Wang, and P. Shen are with the College of Transportation Engineering, Chang'an University, Xi'an, China, and J. Fang is also with the NExT++ Research Centre of School of Computing, National University of Singapore, Singapore (fangjianwu@chd.edu.cn).

Z. Zheng and T-S. Chua are with the Sea-NExT Joint Research Centre, School of Computing, National University of Singapore, Singapore (zdzheng,dcscts@nus.edu.sg).

J. Xue are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China (jrxue@mail.xjtu.edu.cn).

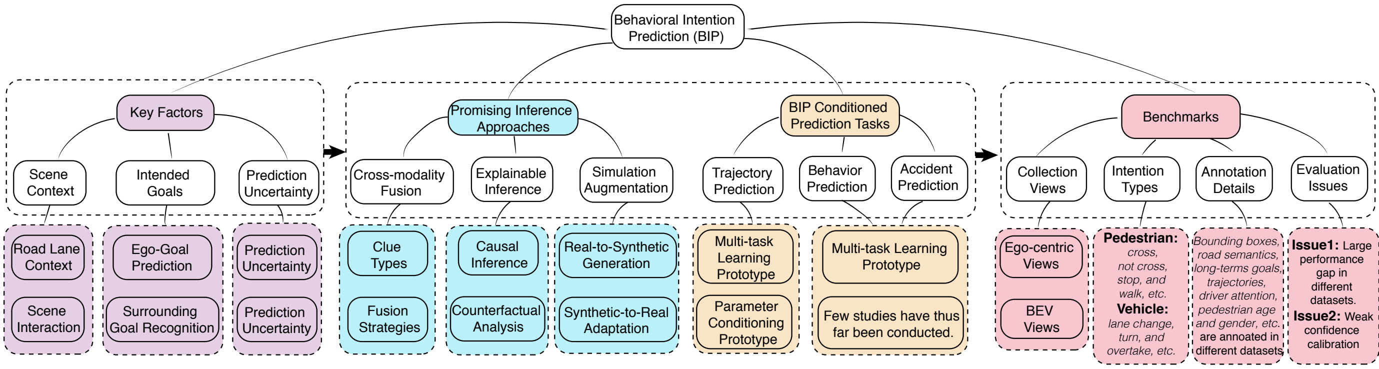


Fig. 1. The content taxonomy in this survey, where the Behavioral Intention Prediction (BIP) in driving scenes is reviewed with the presented items, where ego-goal means the intended destination of the road participants themselves, and the "Ego-centric" view and the "BEV" view denote the observation view from the forward view of the ego-agent and the Birds'-Eye-View, respectively.

traditional methods. Human or vehicle trajectory prediction is also investigated by the surveys of [12], [13], [14], which reviewed the last two decades of probabilistic and deep learning based trajectory prediction works. Sajjad *et al.* [15] study the deep learning based vehicle behavior prediction in autonomous driving applications, and investigate the relation between trajectory and manoeuvres. Recently, the explainability survey [16] of the deep vision based autonomous driving systems presents a detailed description for the counterfactual analysis in driving systems, which is promising for understanding the behavioral intention in the driving scene. A new survey summarized by Sharma *et al.* [17] presents the progress of pedestrian intention prediction for autonomous driving, while the content still focused on the trajectory prediction.

Through the review of the related surveys, we find that most of them concentrate on the trajectory prediction and has a concept confusion for the tasks of trajectory prediction, behavior prediction and intention prediction. Therefore, in this work, we focus on the clarification of the concept of these tasks and present a new comprehensive review for behavioral intention prediction and its role in various prediction tasks in the driving scene.

1.1 Motivation and Taxonomy

To fulfill a promising BIP, we must have an adequate understanding for the semantic context of the road scene. For example, the traffic rules and the participant density will influence the decision making of the vehicles directly. Hence, this survey will investigate the influencing factors in BIP, which will provide the possible consideration for the traffic scene representation. Besides, different road participants commonly have differing behavioral intention. Based on this prior, the role of BIP in different and popular prediction tasks, such as trajectory prediction and accident prediction, would have different focuses. In fact, BIP is a high-level perception task with many cognitive and psychological factors. It will certainly will change the research pipeline and pave the way for the promising inference approaches. Consequently, the large-scale benchmarks and methods of evaluation will emerge and hasten the development of new and advanced applications.

Fig. 1 shows the content taxonomy in this survey. To provide a comprehensive survey for BIP in driving scenes, we firstly start from the exploration of influencing factors in BIP and target the primary road participants of the behavior understanding of pedestrians and vehicles. Secondly, we identify the promising inference approaches in BIP, and focus mainly on the explainable AI models, multi-modal fusion models and simulation augmentation models. We then explore the behavioral intention role in the popular prediction tasks of trajectory prediction, behavior prediction and accident prediction. Finally, we summarize the available benchmarks and future research trends.

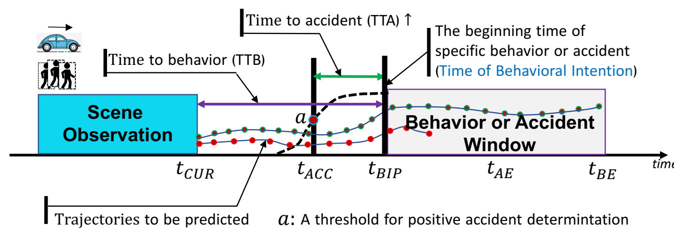


Fig. 2. The illustration for the timeline of behavioral intention prediction, trajectory prediction, behavior prediction, and accident prediction.

1.2 Problem Description and Organization

To be clear, this work makes a definition for the behavioral intention prediction, which is different from the trajectory prediction or behavior prediction. Fig. 2 shows an illustration of the timeline for behavioral intention prediction, trajectory prediction, behavior prediction, and accident prediction.

To be specific, behavioral intention prediction infers the behavioral intention label at time t_{BIP} with an interval of Time-to-Behavior (TTB>0)$[t_{CUR}, t_{BIP}]$.

Accident prediction implies a preference of larger Time-to-Accident (TTA>0) $[t_{ACC}, t_{BIP}]$, where there is a prediction probability a (set as 0.5 commonly) for positive accident prediction determination.

Behavior prediction is defined as the classification of behavior label in the $[t_{BIP}, t_{BE}]$ under the scene observation.

Trajectory prediction has no time interval with t_{CUR}, and is inferred as the future coordinate (x, y) chain prediction.

It is worth noting that the behavior prediction may involve a sequentially behavioral intention prediction, in which when TTB=0, BIP can be treated as the *behavioral intention detection*, and when TTB>0, some multi-task learning perspective may be useful for the BIP and behavior or trajectory prediction tasks. For example, Bouhsain and Saadatnejad [18] predict the crossing intention and locations of pedestrian in the future simultaneously.

With the aforementioned taxonomy in this survey, the reminder of this work is organized as follows. Sec. 2 briefly reviews the influencing factors in BIP. The promising inference approaches are described in Sec. 3. BIP-conditioned traffic scene prediction, including trajectory prediction, behavior prediction, and accident prediction, is summarized in Sec. 4. Sec. 5 elaborates the benchmarks. Sec. 6 presents the trend discussion, and the conclusion is given in Sec. 7.

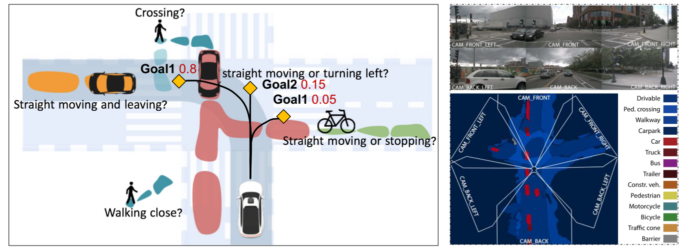
2 Key Factors of Behavioral Intention Prediction

The surveys [3], [21] have summarized different factors that influence the pedestrian-vehicle interaction, and mainly concentrated on the social factors (e.g., group size, social norms, imitation, gender, age, speed, and culture) and environmental factors (e.g., physical context, signals, road structure, and traffic density, etc.). Certainly, the mixed traffic scene makes the factors complex for behavioral intention prediction in the real world scenario. Different from existing surveys, we investigate the key factors for behavioral intention prediction in deep learning era. In this survey, we present the factors from scene context, moving tasks or goals, and prediction uncertainty.

2.1 Scene Context

Road scene is a highly-structured environment, and the road structure contains the consistent traffic rules. In the meantime, road users (pedestrians, vehicles, and cyclists, etc.) should obey the road etiquette [22], [23]. All these static and dynamic semantics in the road scene constitute the scene context for safe driving [24]. Therefore, the first kind of influencing factor is the scene context, which is often jointly modeled by the road static semantics and dynamic participants.

The most universal road participants in the mixed road scene are pedestrians, cyclists, motorbikes, buses, trucks, cars, and trailers, etc. Based on the movement patterns and the road layout, the behavioral intention is different. Compared with the participant category oriented intention, such as *"cross"*, *"walk"*, *"run"*, *"stop"* for pedestrians, or *"straight move"*, *"turn left or right"*, *"brake"*, *"accelerate"*, *"change lanes"* for vehicles, the behavioral intention usually correlates with different road semantics, as shown in Fig. 3(a), e.g., "*moving* close to the *bus station*", "*leaving* the *bus station*", or "*walking* close to a *car*", etc. In other words, the interaction or the relation link between different road participants or static semantics should be considered adequately, as shown by Fig. 3 (a). Actually, the interaction knowledge implies a fact where the road participant should emerge normally. For example, Makansi *et al.* predict the emergence of pedestrians with a reachability prior [1].



(a) (b)

Fig. 3. The scene context factor for behavioral intention prediction with: (a) the uncertainty interactions [19]; and (b) different road layout which can be obtained by the Birds-Eye-View (BEV) estimation [20].

2.1.1 Road Lane Context

The road lane is a manifest clue for the determination and is of great interest to the autonomous driving community [25], [26], [27], [28]. Using lane centerlines as the anchors for constraining the trajectory prediction is widely investigated [29], [30]. Lane graph representations are modeled from raw map data to explicitly explore the complex topology and long range dependencies, where four types of interaction between actors and road map (i.e., lane-to-lane, lane-to-actor, actor-to-lane, and actor-to-actor) are fused in the lane graph representation. For an in-depth utilization of the information on road lanes, Hong *et al.* [22] unify the representation which encodes high-level semantic information in a spatial road grid, allowing the use of fusing complex scene context of entity-entity and entity-environment interactions. Because of the critical role of the road map, some attempts concentrate on the road topology estimation with on-board camera data [31], [32]. This give rise to the hot topic of Bird's-Eye-View (BEV) estimation [33], [34], [35] (as shown by Fig. 3 (b)) with raw camera data in this community.

2.1.2 Scene Interaction Context

The research team led by Dr. Raquel Urtasun releases several works [38], [19], [39], [40], [41] on the interaction or relation of the road participants to fulfill the behavior or trajectory forecasting. The intermediate representation of semantic occupancy map [19], [40] is modeled for following intention or trajectory prediction. For a long time, the road map or the occupancy map is encoded with a dense *rasterized* processing, which are adopted in many popular trajectory prediction works, such as DESIRE [42], IntentNet [41], CoverNet [43], Trajectron++ [44], MultiPath [45], Target Driven Trajectory (TNT) [46], and so on. These methods typically encode the road map with a Convolution Neural Networks (CNN), while the semantic structure of road layout is not modeled well with the restricted perception field of CNN. MultiPath++[47] extends MultiPath with an efficient polyline encoding of road semantics, which exploits the region-to-region semantic relation with a better prediction ability. Actually, road layout has its intrinsic structure. For example, cars commonly drive in the road with different lanes marked by the road line. Pedestrians appear in the sidewalk and exhibit crossing behavior in the zebra region. Therefore, Park *et al.* [48] synthesize the environment's scene context and interactions between multiple surrounding agents to model the distribution of diverse and

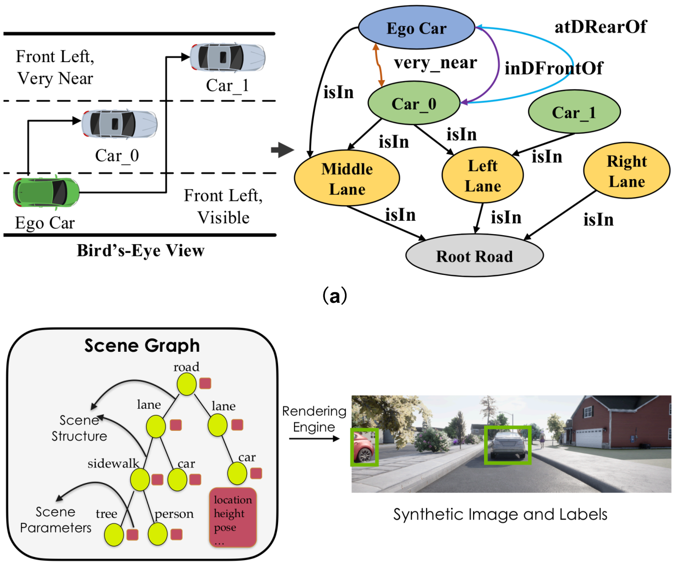


Fig. 4. The scene graph generation and utilization, where (a) is the scene graph generation by RoadScene2Vec [36], and (b) denotes the synthetic data generation by scene graph (credits to [37]).

admissible trajectory and give a high penalty for the *implausible* trajectories with the impossible relation in road scene context. Actually, for the scene context modeling, scene graph is a promising tool. Consequently, scene generation from raw data is important and some works have begun on the encoding learning of scene graph for scene context representation, such as the RoadScene2Vec [36], as shown Fig. 4(a). In addition, scene graph is found to be effective for synthetic data generation and very useful for the digital traffic scene modeling, as shown by the MetaSim works [37], [49] in Fig. 4 (b).

2.2 Intended Goals

Along with the scene context, the behavioral intention prediction is commonly influenced by different intended goals, and should be inferred with interpretablity. With this in mind, we can see that the importance of the road participants is different and changes with the varying driving scene [50], [51], as shown in Fig 5(a). In this community, for driving behavioral intention, the *driver attention* is a direct clue to reflect the important and preferring goals. As shown in Fig. 5(b), the intended fixation of drivers is not only feasible for inferring driving tasks, but can also help to discover the critical or dangerous objects [2].

For the goal-conditioned future prediction, the "*goal*" is modeled as the future state (defined as destination coordinates [52] or moving types [53]) that an agent desires. The Imitative Models (IM) is used to estimate the predictive likelihood of the trajectory by satisfying a goal determined by human experts. Through model training, implausible multiple predictions are omitted. Generally, the goals are similar to the destination [42], while the moving goal or destination is assumed as a priori. With the pre-known goals, the goal-conditioned prediction is interred by the inverse optimal control [54] or inverse reinforcement learning [55]. Manifestly, for future prediction, the coordinate based goals

Fig. 5. The goal-centric scene representation and prediction, where (a) denotes an attention change (credits to [51]), (b) is a sample for driver attention evolution in accident scenario (credits to [2]), and (c) specifies a goal-centric intention prediction (credits to [53]).

may be unknown and hard to be predicted beforehand. Therefore, the coordinate goal condition is relaxed in the following two kinds of goal inference situations.

2.2.1 *Ego-Goal Prediction*

Ego-goal prediction aims to forecast the intended destination or area (static goals) for the ego-vehicle. This formulation can provide a possible future path proposal or locations that the ego-vehicle movement will be satisfied [52]. For example, map-adaptive goal path [56] generates a set of possible goal-directed future path anchors by the road lane constraint. A recent work [57] proposes a Goal Area Network (GANet) for motion prediction, which modeled the goal areas rather than the exact goal coordinates as the preconditions for motion prediction. Within GANet, the possible goals are predicted by calculating the loss between the predicted goal with the endpoint of the ground-truth trajectory. Commonly, these ego-goal prediction works need to pre-define many goal anchors and conduct heuristic or rule-based goal selection. Apparently, the quality of the goal anchors has a heavy impact on the prediction accuracy. In addition, Target-driveN Trajectory (TNT) [46] and DenseTNT [58] are two popular vehicle trajectory prediction models with ego-goal prediction, where DenseTNT fulfilled the ego-goal prediction by estimating the probabilities of dense goal candidates without relying on the heuristic anchors, instead of the sparse goal sampling strategy.

2.2.2 *Surrounding Goal Recognition*

Different from the ego-goal prediction that focuses on the intended goal prediction for the ego-agents, surrounding goal recognition focuses on the goal recognition of other agents around the ego-agent. This aims to achieve a shared temporal-spatial safe space for future movement. In this domain, different from the ego-goal prediction works, the "*goal*" is defined as many kinds of behavioral intentions (dynamic goals), such as "straight-on", "turn left", "u-turn", and "stop", etc. The *interpretability* of the goal recognition is concentrated because of the safe-guaranteed demand for collision-free, real-time, accurate, and verifiable recognition. For example, Brewitt *et al.* [59] propose a Goal Recognition method by Interpretable Trees (GRIT), i.e., Decision Trees (DTs). Since GRIT can only work on fixed frame scenarios, Interpretable Trees under Occlusion (OGRIT) is further proposed recently in the work of [60], where the DTs are also used for training the past observed trajectories to infer the likelihood of the dynamic goals of other vehicles. Albrecht *et al.* [61] recognize surrounding goals by generating a set of possible static locations (road end) and dynamic intention of other vehicles heuristically based on the locations and traffic context, and recognize the surrounding goals by inverse planning model (A* search [62]).

For the behavioral intention prediction, all the agents in the road scene will be considered, and the ego-goal prediction (static goals) and surrounding-goal recognition (dynamic goals) are all important for the safe decision making, which is promising for the collision avoidance. Compared with the static goal prediction, dynamic goal recognition concentrates on the interpretable decision trees. We advocate some deep decision trees, such as Neural-Backed Decision Trees (NBDT) [63], and Neural Prototype Trees (NPT) [64], as promising approaches to leverage of the advantages of deep learning.

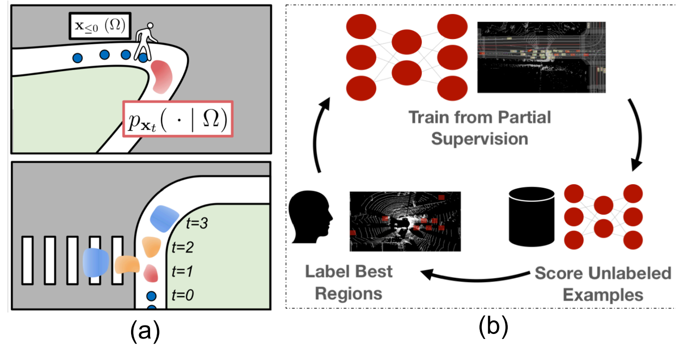


Fig. 6. The prediction uncertainty with: (a) Gaussian distribution for future paths [68] and (b) the observation uncertainty with partially labeled data [69].

2.3 Prediction Uncertainty

Prediction uncertainty is a natural factor in the behavioral intention prediction because of the objective and frequent change of the surrounding driving scene context. The inherent multi-modality, partial observability, short time scales, data limitation and imbalance [65], domain gap [66], and deficiency can all cause uncertainty. In addition, because of the generalizability of deep learning models, the predicted distribution may involve bias. Consequently, in the future state prediction, there are two kinds of uncertainty: the *aleatoric uncertainty* that mainly concerns with the latent decision variables of intended goals, and the *epistemic uncertainty*, the uncertainty that evaluates the a model generalizability owning to the lack od knowledge. These two types of uncertainties construct the basis for sensitivity analysis in the prediction [67].

In particular, aleatoric uncertainty refers to the irreducible, objective, or stochastic uncertainty of a physical system (sensor ability) or environment (severe weather, low light condition), and is modeled by the random variables or stochastic processes in probability theory. The inherent observation noise is captured in aleatoric uncertainty [70]. On the contrary, epistemic uncertainty accounts for uncertainties in the model parameters, and can be called *model uncertainty* with the assessment of the model generalizability.

2.3.1 Aleatoric Uncertainty in Prediction

In future prediction, the ways for these two types of uncertainties are different. For example, in order to weaken the aleatoric uncertainty, extra clues, such as High-definition Map (HD Map), Birds' Eye View (BEV), and many other priori, are taken account for the future prediction. Full-range BEV representation is adopted in a recent work StretchBEV [71] for the future instance prediction, and stretched the spatial scene for longer time horizons than previous works. MultiPath [45] proposes multiple probabilistic anchor trajectory hypotheses with the aid of HD Map, and models the future state as a Gaussian Mixture Model (GMM), where an *intention uncertainty* is defined for inferring the latent coarse-scale intention or desired goals. Yalamanchi *et al.* [72] address the long-term future prediction with the uncertainty-aware trajectories with lane-based paths. In order to model the aleatoric uncertainty, various kinds of probability models are developed, such as Gaussian model [73], [44], GMM [68], [74]. Actually, because of the dynamic and objective intention, Gaussian distribution usually expresses the scene-sensitivity poorly and the inherent multimodality of the road participants increases the uncertainty of movement. For example, the pedestrian may continue along a sidewalk or cross a crosswalk, as shown in Fig 6 (a).

2.3.2 Epistemic Uncertainty in Prediction

For the epistemic uncertainty, various models with the consideration of long-term dependency and large-range spatial interaction are proposed. Discrete Residual Flow (DRF) [68] is proposed for the long-term prediction with a residual update for the marginal distributions over the time span conditioned by the scene context. There are numerous works that focused on the interaction models in future prediction, and the interaction of different road agents can also raise a *collaborative uncertainty* (CU) [75], [76] because of the dynamics of the interaction. The consideration of CU enables to evaluate the interaction uncertainty in the multi-modal prediction. In addition to the model uncertainty on a single dataset, the uncertainty for cross-dataset generation in future prediction is recently proposed [77], which estimates the model uncertainty by the heatmap distribution of the predicted points, and computed by $\sum_p H(p)||p-E||^2$ with $E = \sum_p H(p)p$, where p is the point position, and $H(p)$ denotes the probability value for the given position. Actually, data shift has a large influence on the accuracy of prediction even with good calibration between the model and data. Deep ensembles of multiple networks seem to be beneficial for boosting the model performance under data shift [78].

2.3.3 Potential Trends

With the demand accuracy of long-term prediction, the aleatoric uncertainty and epistemic uncertainty are considered simultaneously in this field all the time. What uncertainties do we need for behavioral intention prediction? Although there is few work for this question, we can seek the answer from the work on *Bayesian* deep learning in computer vision [70]. Aleatoric uncertainty can be focused when we have sufficient data or with real-time demand, and epistemic uncertainty is important when we encounter safety-critical applications with small data. Recently, Bayesian deep learning models have become a favorite [80] in future

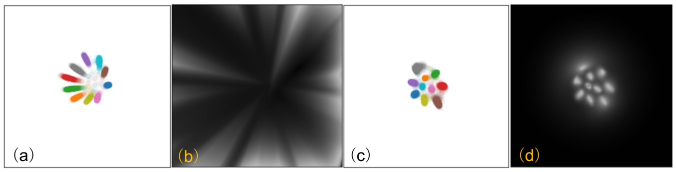


Fig. 7. The uncertainty estimation in the latent space by PriorNet (b) and PostNet (d) with the learned latent position of labels (a) and (c), respectively. Credits to [79].

prediction. Similarly, Itkina and Kochenderfer [81] propose an evidential deep learning model to estimate the epistemic uncertainty for an interpretable trajectory prediction, where the past agent behavior, social context, and road structure are exploited. They are based on the assumption that the unfamiliar road structure, input behavior, and social context would give rise to high epistemic uncertainty. The evidential deep learning in [81] is inspired by a Posterior Network (PostNet) [79] which has a promising ability for distinguishing the latent space of the Out-of-Distribution (OOD) samples, as shown in Fig. 7.

As for BIP in the driving scene, the intention types of the road participants are multitudinous. Consequently, it is impossible to collect enough data in practical use for each type of behavioral intention, and may involve many types with only few samples. In the meantime, each participant in the driving scene may have different intentions at each time step, which implies natural aleatoric uncertainty. Therefore, aleatoric and epistemic uncertainty will coexist for a long time. Recently, Digital Twining (DT) [135] or Parallel Intelligence (PI) [136] may be promising for epistemic uncertainty by generating large-scale behavioral intention data with the consideration of long-term, abundant and diverse samples. The real-synthetic data interaction and collaborative simulation may be useful for the long-tailed [137] or few-shot samples in the prediction (will be described in Sec.3.3 in detail). Furthermore, with the influence of partial observation, *few-shot* [138] or *zero-shot* [139] learning models with limited labels or important labels [69] (as shown in Fig. 6(b)) can also be taken as epistemic uncertainty. Human-machine hybrid intelligence will have an important role in future prediction with the humans help to correct prediction error [140] in an active learning setting.

3 Promising Inference Approaches in BIP

In this section, we elaborate the promising inference approaches in BIP. We mainly focus on cross-modality fusion, explainable AI, and simulation augmentation models in BIP. It is worth noting that, we do not include the works of driver intention (can be reviewed in other surveys [141], [142]) but elaborate the existing works that concentrate on intention prediction of pedestrians and surrounding vehicles.

3.1 Cross Modality Fusion in BIP

In order to enhance the category margin of different intentions, most existing works exploited multiple clues. As for pedestrians and vehicles, the types of clues are different, where pedestrians prone to use harness images (I), pose (P), estimated velocity (V), locations (L), while vehicles often fuse many road structure information, such as images (I), HD map (HD), Distance to Centerline (D2C), and velocity (V). Table. 1 presents the related works for pedestrian BIP and vehicle BIP, from the view of years, inference models, clue types, fusion strategies, and intention types. It is manifested that the intention types of pedestrians and vehicles are different, where crossing (C) and non-crossing (NC) are the main concerns of the pedestrians, while lane change (LC) is the key focus for vehicles. This observation is reasonable that "*crossing warning*" often takes the most important demand for assisted driving systems [143], and lane change of vehicles is the most frequent behavior with the potential threat to the other vehicles[144], [145].

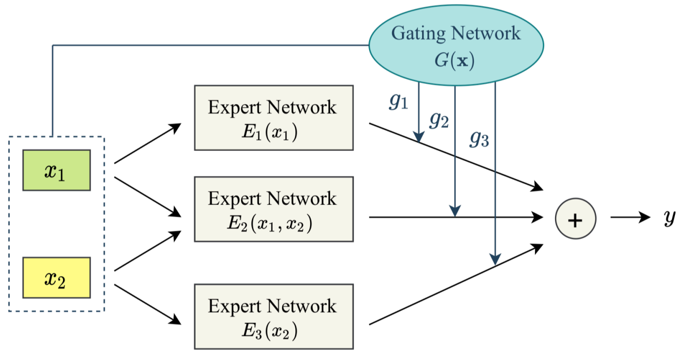


Fig. 8. The Dynamic Multi-modal Fusion flowchart. Credits to [146].

With the development of deep learning, the inference models for different clues still present a stable appreciation for Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) and their related variants. Some recent works [114], [115], [119], [106] adopt Graph Convolution Networks (GCN) to infer the embedding of different clues for pedestrian crossing intention prediction. Actually, most multi-clue intention prediction works do not evaluate the importance of different clues, but simply fuse them together with the feature embedding models by combining the embeddings in "*concat*" or "*attentive fusion*". Certainly, "*attentive fusion*" provides a mechanism for selecting the important information, while it does not give an explicit modeling and is unexplainable. Recently, Xu *et al.* [111] imitate the mechanism of retrospective memory in neuropsychology and propose a MemoNet to store the representative instance set and predicted the intention of road agent by looking for similar scenarios in the training data. They infer the future destination clues for the future intention feature encoding.

Essentially, fusing more clues could reduce the aleatoric uncertainty as aforementioned before. More information provide more constraints for future intention prediction, while it gives rise to a fundamental problem on how to fuse these information in the best way. That is because, in some situations, some information may be counteractive. Various Dynamic Neural Networks (DNNs) [147] may be promising for adaptively selecting the multi-modal information in different situations. Dynamic Multimodal Fusion (DynMM) [146] and Dynamic Routing Network (DRN) [148] are two kinds of models, where the "dynamics" in modality fusion is fulfilled by a Gating Network (GN) in DynMM and

TABLE 1
The comparison of the cross-modality fusion models for behavioral intention prediction of pedestrians and vehicles, *w.r.t.*, **Years**, **Inference Models**, **Fusion Strategies**, **Clue Types** (I: image; P:pose; L: location, E: ego-motion; V: velocity; M: other agent motion; T: trajectory; D: depth; S: semantics; F: future state; H: head orientation; W: waiting time; D2C: distance to centerline; D2V: distance to vehicles; 3DP: 3D point cloud; DM: dynamic map; R: raw-rate; SA: steering angle; HD: road map; RL: road lines; LC: longitudinal coordinate; VT: vehicle type), and **Intention Types** (C: cross; NC: non-cross; L:look; NL: not look; W: walk; ST: stand; SM: straight move; TL: turn left; TR: turn right; UT: u-turn; KL: keep lane; LC: lane change; S: stop; P: park; Pa: Pass; Y: yield; M: merge).

Ref.	Years/booktitle	Inference Models	Clue Types	Fusion Strategies	Intention Types
[82]	2016/ITSC	CNN, LSTM, SVM	I	concat	C, NC
[83]	2018/IEEE TIV	LSTM	T	concat	C, NC
[84]	2018/IV	CNN	I, P, L	concat	C, NC
[85]	2018/IV	CNN, LSTM	I, P, L	concat	C, NC
[86]	2018/SITIS	SVM,ANN,kNN,Decision Trees	I, L, M, H	concat	C, NC
[87]	2019/BMVC	GRU	I, P, L, E	concat	C, NC
[88]	2019/ICCV	DenseNet-121	I, P, L, E	concat	C, NC
[89]	2019/ICCV	LSTM	I, L	concat	C, NC, W, L, NL
[90]	2019/ICRA	Residual Encoder-Decoder,3DCNN	I	concat	C, NC
[91]	2019/ICRA	Spatial-Temporal (ST) DenseNet	I	w/o fusion	C, NC
[92]	2019/ITSC	GCN	P	concat	C, NC
[93]	2020/ACSSC	STDenseNet	I, P, L	Early-Middle-Late concat	C, NC
[18]	2020/hEART	LSTM	L, V	concat	C, NC
[94]	2020/IEEE TITS	CNN	I, P, L	concat	C, NC, TL, TR, S
[95]	2020/IEEE RAL	GCN	I, L	concat	C, NC
[96]	2020/ITSC	LSTM, Dynamic Bayesian Network	I, P	concat	C, NC
[97]	2020/IV	GRU	I, L, P, E	concat	C, NC
[98]	2020/IV	Logistic Regression Classifier	I, L, M	concat	C, NC
[99]	2020/WACV	C3D, Conv-LSTM	I	w/o fusion	C, NC
[100]	2020/Journal of Physics	GCN, Conv-LSTM	I, L, P, E	concat	C, NC, W, ST, L, NL
[101]	2021/ICCV	LSTM	T, I, L, E, S	concat	C, NC
[102]	2021/ICCVW	LSTM, GCN	I, L, P	concat	C, NC
[103]	2021/arxiv	LSTM, Conv-LSTM	I, L, P	Attentive fusion	C, NC
[104]	2021/arxiv	Transformer	I, L, P, V	concat	C, NC
[105]	2021/ICCVW	3DCNN	P, L	concat	C, NC
[106]	2021/ICRA	STGCN, LSTM	M, E	concat	C, NC
[107]	2021/IEEE TITS	Factor-CRF	I, S, M, D, E	concat	C, NC
[108]	2021/IJCAI	CNN, MLP	I, L, S	concat	C, NC
[109]	2021/TRC	3DResNet50	I, P	Attentive fusion	C, NC
[110]	2022/arXiv	CNN, GRU	I, P, L, E	concat	C, NC
[111]	2022/CVPR	Self-Attention, Memory network	T, F	Attentive fusion	C, NC
[112]	2022/IEEE SPL	Vison Transformer	I, L, P	concat	C, NC
[113]	2022/IEEE TITS	SVM	P, V, H, W	concat	C, NC
[114]	2022/IEEE TITS	GCN, CNN	I, P, L, E, S	concat	C, NC
[115]	2022/IEEE TITS	GCN	P	concat	C, NC
[116]	2022/IEEE TIV	CNN, GRU	I, L, P, E	Attentive fusion	C, NC
[117]	2022/IV	Transformer	L	concat	C, NC
[118]	2022/IV	CNN, LSTM	I, S, L, E	Attentive fusion	C, NC
[119]	2022/IV	GCN	I, E	concat	C, NC
[120]	2017/ITSC	LSTM	E, HD	concat	LC
[121]	2018/IV	CNN, GMM	V, D2C	concat	SM; TL,TR, LC, S
[41]	2018/CoRL	CNN	3DP, DM	concat	KL, TR, TL, LC, S, P
[122]	2018/Expert Syst. Appl.	Adaptive Fuzzy Neural Network	V, D2V, SA	concat	LC
[123]	2018/ICRA	Bi-LSTM	V,LC	One-hot vector	LC
[124]	2019/Expert Syst. Appl.	MLP+Fuzzy C-Means	V, D2V, SA	concat	LC
[125]	2019/IV	LSTM	LC, V	concat	LC
[126]	2019/ITSC	CNN, LSTM	I	Muti-Channel Stacking	LC
[127]	2020/IEEE TITS	RNN	L,V, SA	concat	SM, TL, TR, UT
[128]	2020/Transport. RR	LSTM	L,V, D2C	concat	SM, TL, TR, UT
[129]	2020/IV	LSTM +Attention	V, R	Muti-Channel Stacking	SM, TL, TR, S
[130]	2021/IEEE TVT	RNN	RL, V, SA	concat	LC
[131]	2022/IEEE TITS	Echo State Network, LSTM	SA	concat	LC
[132]	2022/IEEE TITS	LSTM+Attention	L, V, VT	concat	LC
[133]	2022/ICRA	Variational RNN	HD, V, E	causal	Pa, Y, M
[134]	2022/TIE	LSTM	L,V	concat	LC

achieved by the router network in DRN. Fig. 8 illustrates the modality level fusion in DynMM, where the input modalities, x_1 and x_2, are inferred by three expert networks and each network has a link with the output y. The gating network will select the best expert network in the final decision. Certainly, the gating network can also be added in the feature embedding part to fulfill a selective multi-modal encoding.

3.2 Explainable AI in BIP

For explainable AI models in the driving scene, the attention mechanism initially designed in Natural Language Processing (NLP) applications [149] are fundamental and have been exploited in many different applications. It learns the scores of different regions for the final decision process. Certainly, attention mechanism has also been adopted for

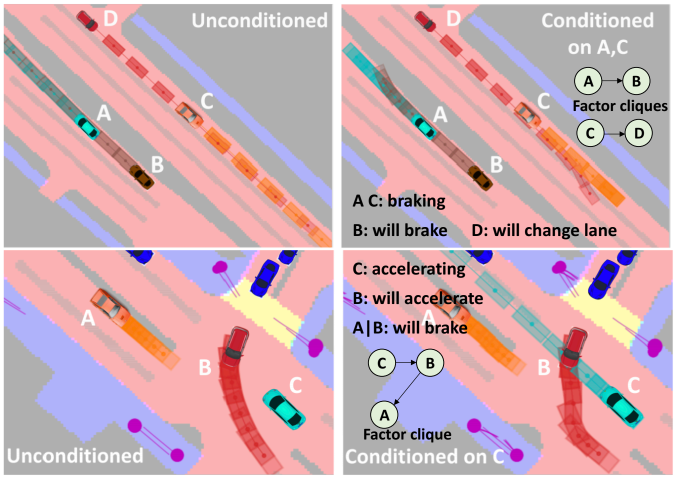


Fig. 9. The conditioning analysis for future prediction with a causal intention linking to different vehicles. Credits to [152].

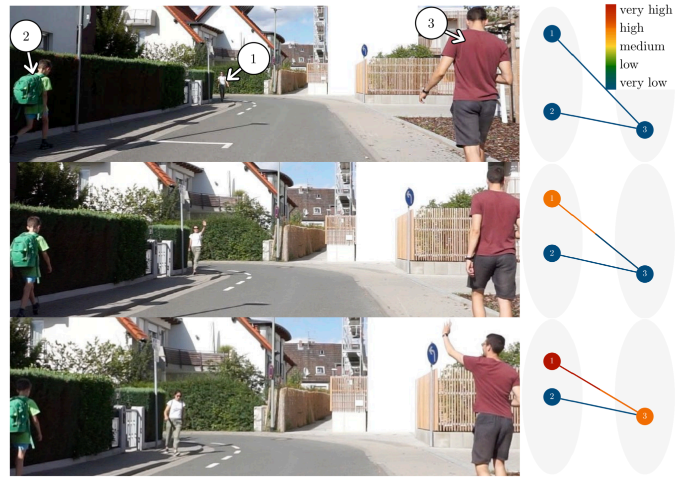


Fig. 10. An example for the pedestrian interaction "say hello", which provides a strong crossing street intention. Credits to [159].

the explainable AI models in the driving scene. The survey [16] has investigated the works correlated with the autonomous driving system, and will not be further described in this work. Here, we mainly concentrate on the causality inference in explainable AI in the driving scene.

Causal [133] or factor relation [150] amongst the road agents is involved in future prediction, and constructs the non-visible "Dark Matter" [151] for motivating the behavioral intentions. Chen *et al.* propose a scene-consistent, policy-based trajectory predictions, which firstly build a scene graph by the agents' distance, and partition the graph into several cliques. The factor graph is constructed on these cliques, which paved the way for the *conditioning* and *counterfactual* analysis [152] for the prediction. Taking Fig. 9 as an example, the intention of vehicle **A** will be influenced by the causal chain of vehicle **B** conditioned by the accelerating vehicle **C**. Hu *et al.* [133] contribute a causal-based time series domain generalization model for vehicle intention prediction, which construct a Structural Causal Model (SCM) based on the domain knowledge of the generation pattern of interactive trajectories within each vehicle pair. The domain knowledge originates from the road topology, speed limit, and traffic rules, and the interactive pattern corresponds to the driver's aggressiveness level and the relation with other road agents. These information form the causal relation of the features for the specific intention.

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 of the input state. For example, Li *et al.* [153] explore the Causal Inference (CI) on the identification of risk objects by masking the front participants. This formulation is also adopted by the STEEK model [154] for the intention decision model (e.g., *stop* or *move forward*), where "*region-targeted counterfactual explanations*" is introduced and could generate meaningful counterfactual with a preserved scene layout and relevant semantic changes. Based on CI, some recent works [155], [156], [157] begin to investigate the robustness of future prediction by attacking the input observations. These approaches aim toward explainability by changing the semantic or scene state and check the influence on the outcome for finding the primary input state. Actually, causal relation has been observed by the safety-critical driving scenario generation, such as tje CausalAF [158] that aligns with the behavioral graphs; it integrates the Causal Order Masks (COM) to generate possible cause-effect relations for the road scene and Causal Visible Mask (CVM) to filter the non-causal information. The causality has the natural relation with the interaction of road participants. For example, if two persons stand on different sides of the road and say hello, one of them is likely to cross the road to meet each other, as shown in Fig. 10. Therefore, the causality does not just correlate with the static road semantics, but also the dynamic action or pose of the participants.

3.3 Simulation Augmentation for BIP

From the aforementioned review, it is apparently that we need to find the natural relation of the road semantics and collect sufficient data samples. However, in practical use, it is difficult to gather adequate samples that cover all of the causal relation, diversity, and long-tailed future state distribution in safety-critical driving scenes. Therefore, more and more works begin to borrow multitudinous virtual simulation tools (e.g., CARLA [163], AirSim [164], GTA-V [165], and MetaDrive [166], etc.) to generate diverse driving scenes in this field. We call it as Simulation Augmentation (SA) in the following content.

The core problems in SA are to transfer the scene consistency from real to synthetic data, and transfer the diversity from synthetic to real scenarios, which generates the possible future state with a parallel or transcendental evolution [167], [168].

3.3.1 Real-to-Synthetic Generation

Within this domain, many kinds of virtual engines are adopted with high fidelity rendering. A recent survey [162] presents a methodological perspective review for the safety-critical driving scene generation. It categorizes the generation models as data-driven generation (e.g., TrafficSim [160]), adversarial generation (e.g., AdvSim [161]), and knowledge-based generation (e.g., MetaSim2 [37]), which

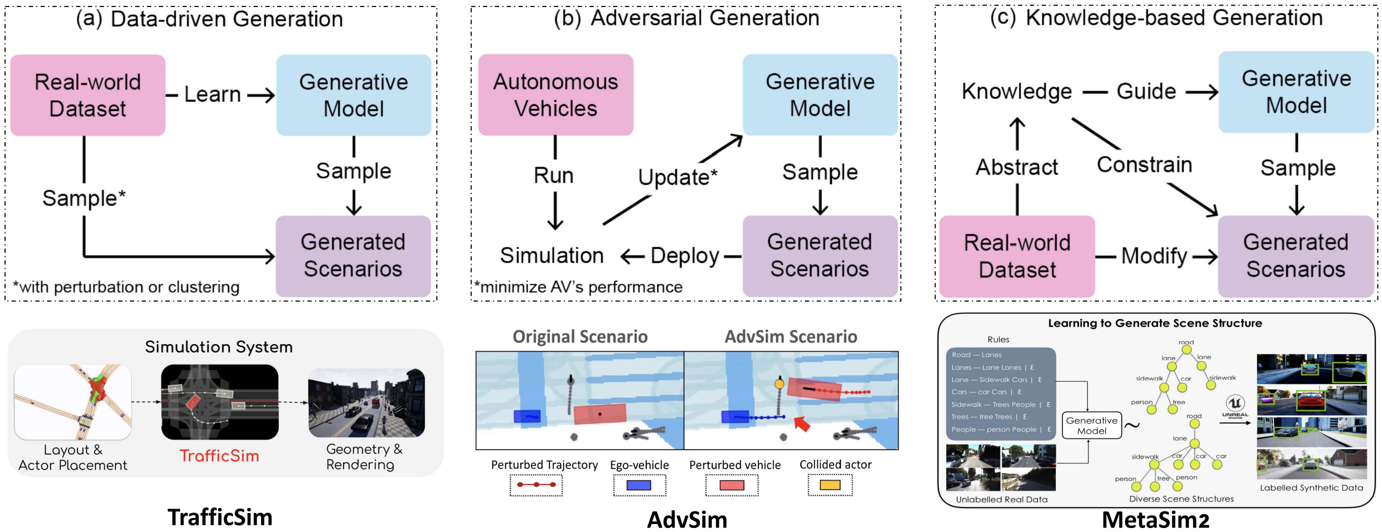


Fig. 11. Different generation pipelines, where the TrafficSim [160], AdvSim [161], and MetaSim2 [37] are the representative examples in data-driven generation, adversarial generation, and knowledge-based generation, respectively. The summary of different generation pipelines credits to [162].

are illustrated in Fig. 11. With these excellent simulators in the driving scene, the research on the behavioral intention prediction booms in recent years. For example, the AdvSim formulates the perturbed vehicles with an adversarial future behavior, which provides the robust training for the behavior prediction of the ego-vehicle. Chen and Krahenbuhl [169] create a virtual multi-vehicle collaboration environment for the behavioral intention prediction of ego-vehicle with the learning of future intention of surrounding vehicles. The behavioral intention in the vehicles are defined as "*turn-left*, *turn-right*, *go-straight*, *follow-lane*, *change-lane-to-left*, and *change-lane-to-right*." TrafficSim [160] leverages the real-world data to learn the rules of humans' experience for the behavior demonstration, and generates multi-agent behaviors with socially-consistent plans for all actors in the scene. The behaviors of *"U-turn"*, *"yielding"*, *"merging"*, *and "passing"*, *etc.*, are generated for road vehicles.

3.3.2 Synthetic-to-Real Adaptation

Synthetic-to-real adaptation absorbs the superiority of various simulators for generating vast amount of data in different weather, light, and road conditions. The data with long-tailed distribution or adverse weather condition can be collected efficiently.

In order to cover the dynamics of pedestrian crossing intention, some works leverage the diversity in synthetic data to the real scenarios. For example, the work [117] transfers the dynamics of bounding box from synthetic data to real data. Another work [170] constructs 4667 sequences with "crossing" or "non-crossing" intention and models a virtual-to-real deep distillation for the lightweight pedestrian crossing intention prediction. Different from the works that addresses the intention prediction from the ego-vehicle perspective, Kim *et al.* [171] propose a pedestrian crossing intention prediction model with the pedestrians' view with a Virtual Reality (VR) apparatus, which aims to learn a shared understanding for the pedestrian crossing intention. Although synthetic data can boost the diversity of the scenarios, there is large distribution gap between the synthetic data and real data. Therefore, models trained on synthetic data often show degraded generalization to real data [172], [173]. Recently, Zhou *et al.* [174] present a survey for the domain generalization problem and exhibit the core solutions for better synthetic-to-real adaptation.

3.3.3 Risk and Crash Inference

Within BIP problem, it has direct link with the risk assessment in driving. Recently, the collision risk prediction work [175] is modeled by inferring the hidden intention of surrounding objects. Similarly, Kim *et al.* [176] 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. VIENA^2 [177] is a promising benchmark with the synthetic data for the prediction of accidents, pedestrian intention (e.g., *cross*, *walk alongside*, *stop*), and front car intention (*forward*, *stop*, *turn right/left*, and *change lane left/right*).

Finally, SA is gradually becoming an indispensable technique for the reasoning of safe-critical driving scenarios; it has the direct relation with the Digital Twining (DT) [135] or Parallel Intelligence (PI) [136] in the driving scene. With the booming of Metaverse, the interaction between the virtual and real world will become a fundamental and core basis for understanding the world.

4 BIP-Conditioned Scene Prediction

The accurate BIP provides the future movement tendency of road agents. it is apparently useful for the following tasks of trajectory prediction, behavior prediction, and potential accident prediction. Specifically, how does BIP promote the other prediction tasks will be described in this section.

4.1 BIP-Conditioned Trajectory Prediction

The first prediction task is BIP conditioned trajectory prediction, where the conditioning function can be divided into multi-task learning prototype and parameter conditioning prototype.

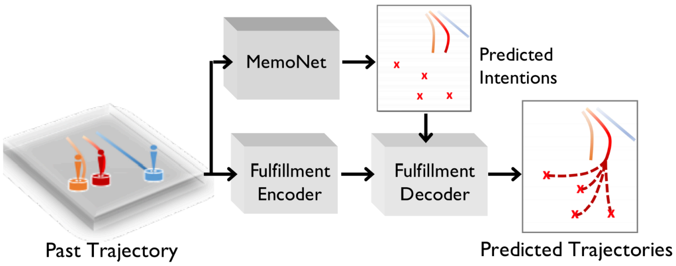


Fig. 12. The Retrospective-Memory-based Trajectory Prediction framework in [111].

4.1.1 *Multi-task Learning Prototype*

Formulating the intention and trajectory prediction as a multi-task learning prototype can be easily considered and implemented. By adding an extra loss function with trajectory prediction loss, these two coupling tasks can be inferred simultaneously [178], [179], [180], [181], [18]. For example, Su *et al.* [179] treat the pedestrian crossing intention as an extra signal and fulfill the trajectory prediction by adding an intention loss (cross entropy of the intention labels) to a terminal time focusing on L2 loss of trajectory. The results show that crossing intention had promoted the trajectory prediction for the end time step significantly. Rasouli *et al.* [101] formulate a multi-task prediction for pedestrian crossing intention, trajectories, and final grid location. The binary cross-entropy loss is used for the pedestrian crossing intention prediction. The Retrospective-Memory-based Trajectory Prediction [111] combines the pedestrian crossing intention and trajectory prediction together, and infers the intention prediction with the MemoNet to jointly reconstruct the compatible past trajectory and future intention features. As shown in Fig. 12, the MemoNet and the encoder-decoder for trajectory prediction is optimized alternatively. DROGON [180] fulfills a goal-oriented trajectory prediction network, which computes the probability of intended goals based on the inferred interaction of vehicles, and estimated the label of intention by cross-entropy loss. Sui *et al.* [181] introduce the Transformer to model the cross-attention of different information (locations and images) and formulate the multi-task learning of intention and trajectory prediction.

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 crossing-entropy loss [59], [56], [58], [46], [61], which is investigated in Sec. 2.2. Recent work LOKI [182] combines the intended goals and behavioral intention together in a new behavioral intention prediction benchmark with 15 types of intentions. The goal proposal network (GPN) and the intention prediction model are added with the trajectory prediction in a multi-task inference model.

4.1.2 *Parameter Conditioning Prototype*

Parameter Conditioning Prototype for trajectory prediction usually models the intent as an extra information to re-weight or re-constrain the trajectory distribution sampling function [183], [184], [185]. The Conditional VAE (CVAE) models [186] defined as follows are commonly adopted.

$$p_\theta(y_i|\mathbf{X}) = \int p_\theta(y_i|\mathbf{z}_i, \mathbf{X}) p_\theta(\mathbf{z}_i|\mathbf{x}_i) d\mathbf{z}_i, \qquad (1)$$

where $p_\theta(\mathbf{z}_i|\mathbf{x}_i)$ denotes the conditional independence of the latent variables $\mathbf{z}_i$ under the agent observation $\mathbf{x}_i \in \mathbf{X}$. Commonly, the intention is encoded in $p_\theta(\mathbf{z}_i|\mathbf{x}_i)$, where the other conditions, such as interaction and road scene knowledge may also be encoded. Euro-PVI [187] models the interactive intention between the surrounding objects and ego-vehicles (e.g., *yield, slow down, and cross, 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.* [188] 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.

Besides, other parameter based intention prediction models, such as the Dynamic Bayesian Network (DBN) [189], [190], [133], [191], [192], are also be explored. As for the deep learning era, the framework of DBN will be popular, where the feature extraction of inference model may be fulfilled by deep learning modules. The work [193] 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. Wu *et al.* [190] fuse the pedestrians' behavior, intention and the scene context together to tackle the trajectory prediction problem. The pedestrian intention is inferred by DBN with the variables for the existence of anxiety and longitudinal danger, crossing area, waiting time, and 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.* [83], [194] predict the long-term intention of pedestrians sequentially by a deep stacked LSTM over the trajectory points.

4.2 BIP-Conditioned Behavior Prediction

In many related works, the behavior prediction and the trajectory prediction has confusion by treating the trajectory prediction as the behavior prediction. We think these two tasks have intrinsic difference, where the behavior prediction prefers the semantics of certain movement, while the trajectory prediction only forecasts the future locations without the specific semantics.

Compared with trajectory prediction, behavior prediction has the most similar problem formulation with BIP, where they are usually formulated as a classification task for future state [202], [203]. Besides, since the behavior may last for a while with a time window, the behavior prediction can also be formulated as a sequential classification task for the future states, e.g., that the prediction changes from the "*will cross*" to "*crossing*" for pedestrians. For example, Yao *et al.* [108] couple the intention and action for pedestrian crossing behavior prediction, where the "*standing*", "*walking towards*", "*crossing*" and "*crossed*" actions are combined with the intention of "will cross or not cross". The pedestrian

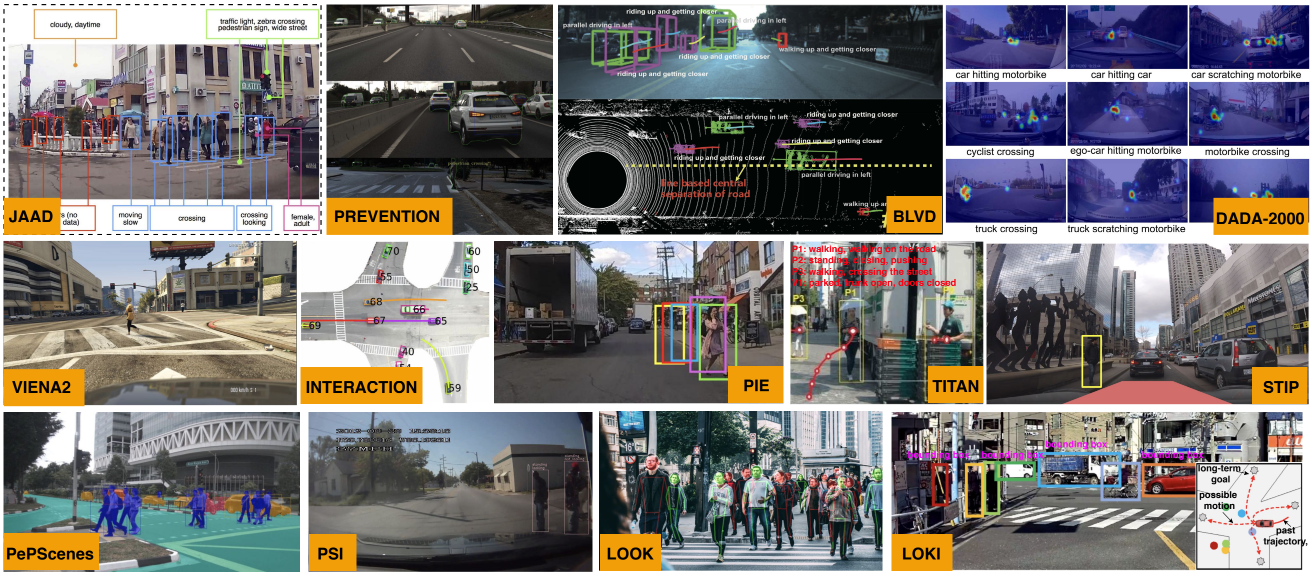


Fig. 13. The benchmarks of JAAD [195], PREVENTION [196], BLVD [197], DADA-2000 [2], VIENA2 [177], INTERACTION [198], PIE [199], TITAN [200], STIP [95], PePScenes [201], PSI [103], LOOK[23], and LOKI[182].

crossing behavior is modeled as a sequential prediction problem solved by a multi-task inference. Similarly, Rasouli *et al.* [101] fulfill the pedestrian behavior prediction by assigning the action state on the prediction trajectories. Ma *et al.* [204] 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, and adopts the CVAE to infer the prediction, respectively. Banijamali *et al.* [205] construct an action-conditioned behavior prediction, where the prediction problem is modeled as learning $p(\mathbf{o}_{t+1}|\mathbf{o}_{1:t}, \mathbf{a}_t)$ with the action $\mathbf{a}_t$ and observation feature $\mathbf{o}_{1:t}$. The action $\mathbf{a}_t$ at time t and the future state $\mathbf{o}_{t+1}$ are alternatively predicted to fulfill a "Prediction by Anticipation" framework. Li *et al.* [206] 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 there are many works on pure behavior prediction, while the intention conditioned behavior prediction of pedestrians and vehicles is a relatively undeveloped field. Actually, the behavioral intention of road agents has a manifested promotion role for long-term behavior prediction.

4.3 BIP-Conditioned Accident Prediction

Accidents are the special events in the driving scene and many works concentrated on how to reduce the occurrence probability of them. The accident prediction implies the collision avoidance problem. For example, these works [207], [208] have conducted the relation modeling between pedestrian intention and trajectory with the collision avoidance. Some previous works exploited the role of the behavioral intention for the collision avoidance by changing the intention or the moving speed [209], [210], [211], [212]. However, these works are all based on Gaussian Process (GP) [212], social force [211], or Markov Decision Process (MDP) [207] by learning the transition model for each vehicle given the driving intention. As for deep learning era, few studies have thus far been conducted on the intention-aware accident prediction or collision avoidance.

In this field, there are deep learning based works which predict the accident in dashcam video data [213], [214], [175], [215], [216], [217], [218], [176], where the interaction or relation of road participants are considered similar to the models in other prediction tasks. Nevertheless, the behavioral intention is also seldom involved. For the explainable AI model for behavioral intention prediction as aforementioned before, the causal inference or the attack based models are promising for this field. We think that it is urgent to open this problem because the behavioral intention change will have the largest influence on the avoidance or accident prediction. If we change the behavioral intention earlier, we can get the larger Time-to-Accident (TTA) for collision avoidance or decision making. As aforementioned, some works contribute the counterfactual analysis or attack operation on the factual scene by changing the status of road semantic or masking the objects [155], [154], [153]. We think that compared with these scenarios, the exploration for the behavioral intention changing mechanism will be more useful for accident prediction.

5 Available Benchmarks

In this section, we elaborate the publicly available benchmarks for the behavioral intention prediction task. Here, we only present the benchmarks for the behavioral intention prediction, while the datasets for trajectory prediction and accident prediction are not contained. Table. 2 presents the attributes of fourteen datasets, and the samples are shown in Fig. 13. Almost all the datasets have the pedestrian crossing

TABLE 2
Behavioral intention prediction benchmarks generated by real data (R) or simulated data (S) with the **Intention Types** (C: cross; NC: non-cross; L:look; NL: not look; W: walk; ST: stand; SM: straight move; TL: turn left; TR: turn right; UT: u-turn; KL: keep lane; CI: cut in; CO: cut out; VO: vehicle overtake; LK: looking; NLK: not look; LC: lane change; S: stop; Y: yield; M: merge; NC: near collision; RD: move along the roundabout) and the **Annotations** (L: locations (boxes); 3DB: 3D boxes; VT: vehicle type; V: velocity; DA: driver attention; T: trajectory; W: weather, B: behavior; O: occasions; A: age; G: gender; BO: human body orientation; DES: destination; SM: semantic map; SD: scene description). The serviceable prediction tasks (**Pred. Tasks**) by these benchmarks are denoted as behavioral intention prediction (BIP), trajectory prediction (TP), behavior prediction (BP), future location prediction (FLP), accident prediction (AP), and driver attention prediction (DAP).

Datasets	Years/booktitle	Seq. num	Annotations	Intention Types	S/R	Pred. Tasks
JAAD [195]	2017/ICCVW	346	I, L, W, O, B, G, A, BO	C, NC	R	BIP, TP, FLP
VIENA2 [177]	2018/ACCV	15000	I, B	S, TL, TR, LC, C,NC,W	S	BIP, AP
INTERACTION [198]	2019/arXiv	-	I, L, SM, B	RD, NC, LC, M	R	BIP, TP, BP
PIE [199]	2019/ICCV	53	I, L, B, GPS, V	W, S, LK, NLK, C, NC	R	BIP, BP, TP, FLP
BLVD [197]	2019/ICRA	654	I,L,T, B,3DB	22 types	R	BIP, BP, TP, FIP
PREVENTION [196]	2019/ITSC	11	I,L,VT, B	LC, CI, CO	R	BIP, BP, FLP
TITAN [200]	2020/CVPR	700	VC, I, L, B	ST, C, NC	R	BIP, BP, TP, FLP
STIP [95]	2020/IEEE RAL	-	I, L, B	C,NC	R	BIP, TP, FIP
PePScenes [201]	2020/NeurIPS	850	I, B, SM, BO, SD	C,NC	R	BIP, TP, FLP, BP
PSI [103]	2021/arXiv	110	I, L, S, B	C,NC	R	BIP, FIP, TP
LOOK [23]	2021/arXiv	-	I, DA, L	C,NC	R	BIP, DAP, FLP, TP
LOKI [182]	2021/ICCV	664	I, L, 3DB, SM, A, G, DES, W	C,NC	R	BIP, FLP, TP
Virtual-PedCross-4667 [170]	2022/ITSC	4667	I, W, L, B	C, NC	S	BIP, BP, TP, FIP
DADA-2000 [2]	2022/IEEE TITS	2000	I, B, DA	LC, VO, C, NC	R	BIP, DAP, AP

or not crossing intention. In the following, we describe the main differences among these benchmarks from the aspects of collection views, intention types, annotation details, and evaluation issues.

5.1 Collection Views

From Fig. 13, we can see that only INTERACTION [198] is collected from BEV view. Compared with the ego-centric view, BEV view can capture a large spatial range of view and provide a complete movement process. BEV view can provide a good ground-truth verification for the BIP or other prediction tasks. Besides INTERACTION, the ego-centric view has a limited range of perception, while the pose and the height of the road agents are clearer. In addition, the ego-centric view perception provides the opportunity for collision avoidance. Apart from the BEV and ego-centric views, 3D point cloud can also capture the BEV and ego-centric view jointly, such as BLVD [197]. However, the raw 3D point cloud has no semantic label or fine-grained pose information of the road agents. Some recent works capture are done on capturing the pedestrian point clouds [219], and the panoramic view by multiple cameras, such as Argoverse 3D dataset [220] (with seven cameras) or NuScenes dataset [221] (with six cameras). However, these 3D point cloud datasets do not provide the behavioral intention label. Therefore, in future, these panoramic view datasets should be extended with behavioral intention or behavior labeling.

5.2 Intention Types

From Table. 2, the intention types on most of the benchmarks only contain the pedestrian crossing or not crossing. Specially, BLVD [197] annotates 22 types of intention, including 12 types, 7 types, and 8 types of behavioral intention for vehicles, pedestrians, and riders (cyclists or motorbikes), respectively. It is promising for fine-grained BIP in the driving scene. In addition, PREVENTION [196] has the *"cut in"* (CI) and *"cut out"* (CO) intention, in which a "hazardous status" is provided for the vehicle lane change intention. Based on the road structure, INTERACTION [198] offers the behavioral intention of *"move along the roundabout"* (RD), and TITAN [200] provides the fine-grained behavior labels, such as *pedestrian crossing*, *pedestrian pushing*, or *pedestrian standing*. The label in TITAN is actually a sequential action label for each road agent. Based on the comparison, we can see that the behavioral intention types for different road participants in the current benchmarks are far from being meticulous. Some safe-critical behavioral intention types, such as *"vehicle runs conversely"* and *"brake"*, etc., and many kinds intention types involving the interactions between different road agents with road semantics (e.g., sidewalk, bus station, and steep slope, etc.) are not exploited.

5.3 Annotation Details

The annotation details in these benchmarks implies an issue that there are many problems in the traffic scene prediction, and the factors are intricate. The object bounding boxes, trajectories, road semantics, driver attention, long-term goals, eye-contact status, hazardous status, pedestrian age and gender are all important for the safe-evaluation in the driving scene. PSI [103] provides the scene description (DES) for different situations, which presents another perspective for scene understanding. From the demonstration in Fig. 13, we can see that only DADA-2000 considers the accident scenarios. If we want to check the counterfactual analysis for the BIP on accident prediction or collision avoidance, current benchmarks are all not feasible. Hence, some attempts for creating editable driving scene may be promising. In addition, crow-view annotation of the behavioral intention is another direction for the cross-validation or counterfactual analysis of BIP.

5.4 Core Issues in Evaluation

For the evaluation on the single dataset of BIP, the precision, recall, F1-measure, and accuracy metrics are commonly adopted in most of the works. For the trajectory prediction, the Average Distance Error (ADE) and Final Distance Error

(FDE), Miss Rate are taken for the evaluation, and mean Time to Accident (mTTA) over all samples are used for the accident prediction task. It is noted that, all current works evaluate the performance on different datasets, and the evaluations have obvious performance gaps. For example, the investigation of [222] finds that different methods on pedestrian crossing intention prediction generate different accuracy and generalization capability for the crossing or not crossing localization in prediction.

For cross-dataset evaluation on pedestrian crossing intention prediction, a recent work [223] shows that current state-of-the-art pedestrian crossing prediction models generated poor performance in cross-dataset evaluation (JAAD and PIE are used). They introduce the confidence calibration metrics, i.e., Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [224], to provide a complement evaluation, and find that ECE and MCE differ drastically. In the meantime, the pre-trained model on density and diverse source datasets can boost the generalization ability on other target datasets. For the BIP problem, beside the two classes of pedestrian crossing or not crossing intention prediction, multiple kinds of behavioral intentions and the uncertainty estimation of model calibration in multi-label classification problem will be exploited [225]. Furthermore, with the development of deep learning models, the calibration measurement [226] is also a core issue in trustworthy implementation.

6 Future Trends and Discussion

Through exhaustive investigation on the behavioral intention prediction and its roles for other prediction tasks, we arrive at a full portrait of this topic. Here, we discuss the future trends for BIP in the deep learning era.

6.1 Theories and Benchmarks

6.1.1 Theories

Despite the numerous works on BIP that have exhibited a significant progress in the performance, most of current works on BIP are all based on the CNN, LSTM, Conv-LSTM, Transformer, and GCN, etc. 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 cross-dataset evaluation, and may cause overfitting or under-fitting problem when encounters the dataset with simpler and more diverse samples, respectively.

For the deterministic neural networks, current research efforts employ the domain adaptation to address this problem by using a well pre-trained model on large scale datasets or leveraging more complex architectures. For example, vision-language pre-trained models, such as BEit-3 [227] and VinVL [228], 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 the behavioral intention prediction 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 as aforementioned before, such as better adoption of the scene context, clearer modeling for the intended goals, and robust estimation of the prediction uncertainty. Standing at the natural characteristics of multiple clues and preferring goals in BIP problem, more explainable scene representation with the aid of scene knowledge should be involved, such as the scene graph [37], [49]. Actually, recent BEV estimation works [33], [34] provide a road scene layout for the scene graph generation. For cross-modal information fusion, various dynamic neural networks can be considered, such as the dynamic multi-modal fusion [146].

Beside the deterministic neural networks, stochastic neural networks aim to estimate the prediction distribution, which is possible for providing the prediction uncertainty estimation, caused by data shift, Out-of-Distribution (OOD) samples (i.e., unfamiliar behavioral intention), the objective property of behavioral intention, and the long-term prediction situations. Existing works [229] estimate the distribution prediction uncertainty for the Bayesian neural networks, generative adversarial networks, CVAE, or deep ensembles [230] by adding the uncertainty consistency loss in Bayesian latent variable model [231]. Therefore, a possible direction for new theories in BIP is to develop the models with more consideration of prediction uncertainty.

6.1.2 Benchmarks

Most of the available benchmarks for BIP focus on the behavioral intention of pedestrian crossing, not crossing, vehicle lane change, and turning intention. In addition, the collection views concentrate on the ego-centric views, which cannot capture the full-range of the road scene, and many types of behavioral intentions cannot be found, such as the “rear car follow”, and “overtake from behind”, etc. A possible way is to add the behavioral intention label for the datasets with panoramic view, such as the Argoverse 3D dataset [220], NuScenes dataset [221], or KITTI-360 [232]. It is also interesting to introduce viewpoints from novel devices, *e.g.*, drone and satellite, for comprehensive scene understanding [233]. Besides, the intention types in current datasets are not fine-grained, and the type imbalance issue is universal. In future, more fine-grained interaction between the road participants with other road semantics involving the specific application and the category of road participants should be considered. For example, pedestrians with different age and gender often show different behavioral intention on the road. Furthermore, the safe-critical scenarios with long-tailed distribution or harsh environment (e.g., rainy, fogy, snowy, windy, and low-light conditions) are also the issue and should be considered.

6.2 Counterfactual Analysis and Parallel Testing

As aforementioned, the ultimate goal of BIP is to avoid road accident or collisions. In practice, we cannot collect enough dataset in the safe-critical scenarios for data-hungry models. Pre-trainning is one straight-forward solution. Zheng *et al.* collect the public real data to pre-train the model and then conduct the domain adaptation fine-tuning to the specific scenario [234]. Meanwhile, the occurred scenarios usually cannot be changed and edited. Therefore, if we want to further check the role of BIP for following accident prediction or collision avoidance, another way is to create the parallel scenarios which can be edited and changed [235]. With this setting, the causal reasoning with cause setting for the behavioral intention, accident question answering, counterfactual analysis is the indispensible problem to be solved. Manifestly, the parallel testing framework [236] should be considered for the testing in synthetic data, joint testing in synthetic and real data, testing in real data conditioned with counterfactual behavioral intention. A recent work [237] generates realistic and diverse trajectories with feasible agent behaviors for automated vehicles. It learns the behavior patterns from the aforementioned INTERACTION [198]. For more complex urban scene, the interaction and attributes are rather complex, and more promising generation pipelines are needed. Meanwhile, the intention pairs of crashing-intention (with collision tendency) and free-crash-intention (without the collision tendency) are the issue to be solved. Usually, a crashing intention may be changed to many free-crash-intentions. We think the changing should be minimal, which implies a problem for finding the most essential intention reason for a certain collision.

6.3 End-to-End BIP

Currently, the prediction tasks usually are based on the pre-detected or collected object locations; this implies that some pre-detection or tracking works have already been conducted. However, we all know that each pre-detection or tracking work will generate the generalization error which is inherited to the following prediction tasks [238]. In order to address the error accumulation problem, it is urgent to develop end-to-end behavioral intention prediction models with the raw video frames as input, so as to fulfill a joint optimization of sequentially tandem tasks [239]. Some works have began to make some attempts on trajectory prediction with the uncertainty motion encoding [240] or sparse agent queries [241] in multi-task learning perspective.

6.4 Promising Applications

From the investigation, we find that there are far research efforts on BIP conditioned behavior prediction and accident prediction. Many intention-aware collision avoidance works are based on traditional methods, such as Gaussian Process (GP) [212] and Markov Decision Process (MDP)[207]. Manifestly, in the big data era, the learning ability of these methods are limited. Meanwhile, the behavioral intention is the most direct promotion for certain behavior and can obtain the largest TTA for collision avoidance. Therefore, BIP-conditioned behavior or accident prediction need to be urgently developed.

In addition, multi-vehicle cooperation (internet of vehicles) [242] and road-vehicle collaboration [243] are the promising applications with the help of other vehicles' perception and large-scale cloud data. For instance, some works [244], [132], [245], [246] predict the pedestrian crossing intention from the cooperative vehicles' view. This kind of formulation can capture a larger range scene context in inference than ever before. Within these applications, the consistent and shared behavioral intention understanding is an important issue. For example, the lane change intention for a vehicle may be understood as an overtaking intention because of the location difference. Therefore, group-wise consistent understanding [247], [248] in collaboration is promising with a reasonable spatial and temporal perception window partitioning.

7 Conclusion

This paper presented a literature review for the behavioral intention prediction, with a comprehensive investigation for the influencing factors, promising inference approaches, and the intention conditioned prediction tasks. Related researches around the influencing factors are summarized from the aspects of scene context, intended tasks or goals, and the prediction uncertainty. Cross-modality fusion model, explainable AI models, and the simulation augmentation models are investigated for the promising inference approaches. The behavioral intention conditioned trajectory prediction, behavior prediction, and accident prediction are described and exhibited that the intention-aware behavior prediction and accident prediction have a large space to be developed. Through the investigation, we also found that significant challenges still exist for the behavioral intention prediction topic from the aspects of explainable prediction models, fine-grained benchmarks, and the testing prototype. Finally, we also discuss future trends and possible solutions.

References

[1] O. Makansi, Ö. Çiçek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *CVPR*, 2020, pp. 4353–4362.

[2] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, 2022.

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

[4] J. D. Velleman, "Intention, plans, and practical reason," *The Philosophical Review*, vol. 100, no. 2, pp. 277–284, 1991.

[5] B. F. Malle and J. Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology*, vol. 33, no. 2, pp. 101–121, 1997.

[6] T. Ziemke, "Understanding robots," *Sci. Robotics*, vol. 5, no. 46, p. 2987, 2020.

[7] H. C. Lau, R. D. Rogers, P. Haggard, and R. E. Passingham, "Attention to intention," *science*, vol. 303, no. 5661, pp. 1208–1210, 2004.

[8] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, 2020.

[9] L. Liu, W. Ouyang, X. Wang, P. W. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[10] I. Kotseruba and J. K. Tsotsos, "Attention for vision-based assistive and automated driving: A review of algorithms and datasets," *IEEE Trans. Intell. Transp. Syst.*, 2022.

[11] 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, 2018.

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

[13] I. Gomes and D. Wolf, "A review on intention-aware and interaction-aware trajectory prediction for autonomous vehicles," 2022.

[14] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Trans. Intell. Veh.*, 2022.

[15] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 33–47, 2020.

[16] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, pp. 1–28, 2022.

[17] N. Sharma, C. Dhiman, and S. Indu, "Pedestrian intention prediction for autonomous vehicles: A comprehensive survey," *Neurocomputing*, vol. 508, pp. 120–152, 2022.

[18] S. Bouhsain, S. Saadatnejad, and A. Alahi, "Pedestrian intention prediction: A multi-task perspective," in *European Association for Research in Transportation (hEART)*, 2020.

[19] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *ECCV*, vol. 12368, 2020, pp. 414–430.

[20] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *CVPR*, 2020, pp. 11 135–11 144.

[21] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3135–3151, 2019.

[22] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *CVPR*, 2019, pp. 8454–8462.

[23] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, "Do pedestrians pay attention? eye contact detection in the wild," *arXiv preprint arXiv:2112.04212*, 2021.

[24] S. Garg and so on, "Semantics for robotic mapping, perception and interaction: A survey," *Found. Trends Robotics*, vol. 8, no. 1-2, pp. 1–224, 2020.

[25] T. Sato and Q. A. Chen, "Towards driving-oriented metric for lane detection models," in *CVPR*, 2022, pp. 17 153–17 162.

[26] Z. Feng, S. Guo, X. Tan, K. Xu, M. Wang, and L. Ma, "Rethinking efficient lane detection via curve modeling," in *CVPR*, 2022, pp. 17 062–17 070.

[27] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, "K-lane: Lidar lane dataset and benchmark for urban roads and highways," in *CVPR Workshops*, 2022, pp. 4450–4459.

[28] C. Li, B. Zhang, J. Shi, and G. Cheng, "Multi-level domain adaptation for lane detection," in *CVPR*, 2022, pp. 4380–4389.

[29] S. Narayanan, R. Moslemi, F. Pittaluga, B. Liu, and M. Chandraker, "Divide-and-conquer for lane-aware diverse trajectory prediction," in *CVPR*, 2021, pp. 15 799–15 808.

[30] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *ECCV*, vol. 12347, 2020, pp. 541–556.

[31] Y. B. Can, A. Liniger, D. P. Paudel, and L. V. Gool, "Topology preserving local road network estimation from single onboard camera image," *CoRR*, vol. abs/2112.10155, 2021.

[32] N. Homayounfar, J. Liang, W. Ma, J. Fan, X. Wu, and R. Urtasun, "Dagmapper: Learning to map by discovering lane topology," in *ICCV*, 2019, pp. 2911–2920.

[33] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *CVPR*, 2022, pp. 13 760–13 769.

[34] A. Saha, O. Mendez, C. Russell, and R. Bowden, "" the pedestrian next to the lamppost" adaptive object graphs for better instantaneous mapping," in *CVPR*, 2022, pp. 19 528–19 537.

[35] Y. Zhao, Y. Zhang, Z. Gong, and H. Zhu, "Scene representation in bird's-eye view from surrounding cameras with transformers," in *CVPR*, 2022, pp. 4511–4519.

[36] A. V. Malawade, S.-Y. Yu, B. Hsu, H. Kaeley, A. Karra, and M. A. Al Faruque, "Roadscene2vec: A tool for extracting and embedding road scene-graphs," *Knowl. Based Syst.*, vol. 242, p. 108245, 2022.

[37] J. Devaranjan, A. Kar, and S. Fidler, "Meta-Sim2: Unsupervised learning of scene structure for synthetic data generation," in *ECCV*, 2020, pp. 715–733.

[38] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data," in *ICRA*, 2020, pp. 9491–9497.

[39] K. Luo, S. Casas, R. Liao, X. Yan, Y. Xiong, W. Zeng, and R. Urtasun, "Safety-oriented pedestrian motion and scene occupancy forecasting," *CoRR*, vol. abs/2101.02385, 2021.

[40] S. Casas, A. Sadat, and R. Urtasun, "MP3: A unified model to map, perceive, predict and plan," in *CVPR*, 2021, pp. 14 403–14 412.

[41] S. Casas, W. Luo, and R. Urtasun, "Intentnet: Learning to predict intention from raw sensor data," in *CoRL*, vol. 87, 2018, pp. 947–956.

[42] 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 *CVPR*, 2017, pp. 2165–2174.

[43] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "Covernet: Multimodal behavior prediction using trajectory sets," in *CVPR*, 2020, pp. 14 074–14 083.

[44] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*, 2020, pp. 683–700.

[45] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *CoRL*, vol. 100, 2019, pp. 86–99.

[46] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, C. Li, and D. Anguelov, "TNT: target-driven trajectory prediction," in *CoRL*, vol. 155, 2020, pp. 895–904.

[47] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. Lam, D. Anguelov, and B. Sapp, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *ICRA*, 2022, pp. 7814–7821.

[48] S. H. Park, G. Lee, J. Seo, M. Bhat, M. Kang, J. Francis, A. R. Jadhav, P. P. Liang, and L. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," in *ECCV*, vol. 12356, 2020, pp. 282–298.

[49] A. Kar, A. Prakash, M. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, "Meta-Sim: Learning to generate synthetic datasets," in *ICCV*, 2019, pp. 4550–4559.

[50] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? deep spatio-temporal importance prediction in driving videos," *Pattern Recognition*, vol. 64, pp. 425–436, 2017.

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

[52] L. F. Chiara, P. Coscia, S. Das, S. Calderara, R. Cucchiara, and L. Ballan, "Goal-driven self-attentive recurrent networks for trajectory prediction," in *CVPR Workshops*, 2022, pp. 2517–2526.

[53] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *ICCV*, 2019, pp. 2821–2830.

[54] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, vol. 7575, 2012, pp. 201–214.

[55] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI*, D. Fox and C. P. Gomes, Eds., 2008, pp. 1433–1438.

[56] L. Zhang, P. Su, J. Hoang, G. C. Haynes, and M. Marchetti-Bowick, "Map-adaptive goal-based trajectory prediction," in *CoRL*, vol. 155, 2020, pp. 1371–1383.

[57] M. Wang, X. Zhu, C. Yu, W. Li, Y. Ma, R. Jin, X. Ren, D. Ren, M. Wang, and W. Yang, "Ganet: Goal area network for motion forecasting," *arXiv preprint arXiv:2209.09723*, 2022.

[58] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *ICCV*, 2021, pp. 15 283–15 292.

[59] C. Brewitt, B. Gyevnar, S. Garcin, and S. V. Albrecht, "GRIT: fast, interpretable, and verifiable goal recognition with learned decision trees for autonomous driving," in *IROS*, 2021, pp. 1023–1030.

[60] C. Brewitt, M. Tamborski, and S. V. Albrecht, "Verifiable goal recognition for autonomous driving with occlusions," *CoRR*, vol. abs/2206.14163, 2022.

[61] S. V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, and S. Ramamoorthy, "Interpretable goal-based prediction and planning for autonomous driving," in *ICRA*, 2021, pp. 1043–1049.

[62] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 2, pp. 100–107, 1968.

[63] A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, H. Jin, S. Petryk, S. A. Bargal, and J. E. Gonzalez, "NBDT: neural-backed decision trees," *CoRR*, vol. abs/2004.00221, 2020.

[64] M. Nauta, R. van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *CVPR*, 2021, pp. 14 933–14 943.

[65] T. Westny, E. Frisk, and B. Olofsson, "Vehicle behavior prediction and generalization using imbalanced learning techniques," in *ITSC*, 2021, pp. 2003–2010.

[66] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1106–1120, 2021.

[67] J. Guo and X. Du, "Sensitivity analysis with mixture of epistemic and aleatory uncertainties," *AIAA journal*, vol. 45, no. 9, pp. 2337–2349, 2007.

[68] A. Jain, S. Casas, R. Liao, Y. Xiong, S. Feng, S. Segal, and R. Urtasun, "Discrete residual flow for probabilistic pedestrian behavior prediction," in *CoRL*, vol. 100, 2019, pp. 407–419.

[69] S. Segal, N. Kumar, S. Casas, W. Zeng, M. Ren, J. Wang, and R. Urtasun, "Just label what you need: Fine-grained active selection for p&p through partially labeled scenes," in *CoRL*, 2022, pp. 816–826.

[70] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NIPS*, 2017, pp. 5574–5584.

[71] A. K. Akan and F. Güney, "Stretchbev: Stretching future instance prediction spatially and temporally," *arXiv preprint arXiv:2203.13641*, 2022.

[72] S. Yalamanchi, T. Huang, G. C. Haynes, and N. Djuric, "Long-term prediction of vehicle behavior using short-term uncertainty-aware trajectories and high-definition maps," in *ITSC*, 2020, pp. 1–6.

[73] B. Ivanovic, Y. Lin, S. Shrivastava, P. Chakravarty, and M. Pavone, "Propagating state uncertainty through trajectory forecasting," in *ICRA*, 2022, pp. 2351–2358.

[74] F. Zheng, L. Wang, S. Zhou, W. Tang, Z. Niu, N. Zheng, and G. Hua, "Unlimited neighborhood interaction for heterogeneous trajectory prediction," in *ICCV*, 2021, pp. 13 148–13 157.

[75] B. Tang, Y. Zhong, U. Neumann, G. Wang, S. Chen, and Y. Zhang, "Collaborative uncertainty in multi-agent trajectory forecasting," in *NeurIPS*, 2021, pp. 6328–6340.

[76] B. Tang, Y. Zhong, C. Xu, W. Wu, U. Neumann, Y. Wang, Y. Zhang, and S. Chen, "Collaborative uncertainty benefits multi-agent multi-modal trajectory forecasting," *CoRR*, vol. abs/2207.05195, 2022.

[77] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "Uncertainty estimation for cross-dataset performance in trajectory prediction," *CoRR*, vol. abs/2205.07310, 2022.

[78] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *NeurIPS*, vol. 32, 2019.

[79] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts," in *NeurIPS*, 2020.

[80] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[81] M. Itkina and M. Kochenderfer, "Interpretable self-aware neural networks for robust trajectory prediction," in *CoRL*, 2022.

[82] B. Volz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. I. Nieto, "A data-driven approach for pedestrian intention estimation," in *ITSC*, 2016, pp. 2607–2612.

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

[84] Z. Fang and A. M. López, "Is the pedestrian going to cross? answering by 2d pose estimation," in *IV*, 2018, pp. 1271–1276.

[85] O. Ghori, R. Mackowiak, M. Á. Bautista, N. Beuter, L. Drumond, F. Diego, and B. Ommer, "Learning to forecast pedestrian intention from pose dynamics," in *IV*, 2018, pp. 1277–1284.

[86] D. Varytimidis, F. Alonso-Fernandez, B. Durán, and C. Englund, "Action and intention recognition of pedestrians in urban traffic," in *SITIS*, 2018, pp. 676–682.

[87] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," in *BMVC*, 2019, p. 171.

[88] K. Kim, W. Choi, Y. J. Koh, S. Jeong, and C. Kim, "Instance-level future motion estimation in a single image based on ordinal regression," in *ICCV*, 2019, pp. 273–282.

[89] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019, pp. 6261–6270.

[90] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *ICRA*, 2019, pp. 2097–2103.

[91] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet," in *ICRA*, 2019, pp. 9704–9710.

[92] P. R. G. Cadena, M. Yang, Y. Qian, and C. Wang, "Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks," in *ITSC*, 2019, pp. 2000–2005.

[93] F. Piccoli, R. Balakrishnan, M. J. Perez, M. Sachdeo, C. Nuñez, M. Tang, K. A. andf Kalle Bjurek, R. D. Raj, E. Davidsson, C. Eriksson, V. Hagman, J. Sjöberg, Y. Li, L. S. Muppirisetty, and S. Roychowdhury, "Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network," in *ACSCC*, 2020, pp. 68–72.

[94] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2d pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4773–4783, 2020.

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

[96] H. Wu, L. Wang, S. Zheng, Q. Xu, and J. Wang, "Crossing-road pedestrian trajectory prediction based on intention and behavior identification," in *ITSC*, 2020, pp. 1–6.

[97] W. Morales-Alvarez, F. M. Moreno, O. Sipele, N. Smirnov, and C. Olaverri-Monreal, "Autonomous driving: Framework for pedestrian intention estimation in a real world scenario," in *IV*, 2020, pp. 39–44.

[98] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do they want to cross? understanding pedestrian intention for behavior prediction," in *IV*, 2020, pp. 1688–1693.

[99] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *WACV*, 2020, pp. 2297–2306.

[100] D. Cao and Y. Fu, "Using graph convolutional networks skeleton-based pedestrian intention estimation models for trajectory prediction," *Journal of Physics*, vol. 1621, no. 1, 2020.

[101] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *ICCV*, 2021, pp. 15 580–15 590.

[102] T. Chen, R. Tian, and Z. Ding, "Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention," in *ICCV Workshops*, 2021, pp. 3096–3102.

[103] T. Chen, R. Tian, Y. Chen, J. Domeyer, H. Toyoda, R. Sherony, T. Jing, and Z. Ding, "Psi: A pedestrian behavior dataset for socially intelligent autonomous car," *arXiv preprint arXiv:2112.02604*, 2021.

[104] J. Lorenzo, I. Parra, and M. Sotelo, "Intformer: Predicting pedestrian intention with the aid of the transformer architecture," *arXiv preprint arXiv:2105.08647*, 2021.

[105] A. Singh and U. Suddamalla, "Multi-input fusion for practical pedestrian intention prediction," in *ICCV Workshops*, 2021, pp. 2304–2311.

[106] T. Yau, S. Malekmohammadi, A. Rasouli, P. Lakner, M. Rohani, and J. Luo, "Graph-sim: A graph-based spatiotemporal interaction modelling for pedestrian action prediction," in *ICRA*, 2021, pp. 8580–8586.

[107] S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels, "Context model for pedestrian intention prediction using factored latent-dynamic conditional random fields," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 6821–6832, 2021.

[108] Y. Yao, E. M. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "Coupling intent and action for pedestrian crossing behavior prediction," in *IJCAI*, 2021, pp. 1238–1244.

[109] H. Razali, T. Mordan, and A. Alahi, "Pedestrian intention prediction: A convolutional bottom-up multi-task approach," *Transport. Res C: Emer.*, vol. 130, 2021.

[110] J. A. Abbasi, N. M. Imran, and M. Won, "Watchped: Pedestrian crossing intention prediction using embedded sensors of smartwatch," *CoRR*, vol. abs/2208.07441, 2022.

[111] C. Xu, W. Mao, W. Zhang, and S. Chen, "Remember intentions: Retrospective-memory-based trajectory prediction," in *CVPR*, 2022.

[112] S. Zhao, H. Li, Q. Ke, L. Liu, and R. Zhang, "Action-vit: Pedestrian intent prediction in traffic scenes," *IEEE Signal Processing Letters*, vol. 29, 2022.

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

[114] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Trans. Intell. Transp. Syst.*, pp. 1 – 12, 2022.

[115] X. Zhang, P. Angeloudis, and Y. Demiris, "St crossingpose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction," *IEEE Trans. Intell. Transp. Syst.*, pp. 1 – 10, 2022.

[116] D. Yang, H. Zhang, E. Y. andf Keith A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 221–230, 2022.

[117] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" in *IV*, 2022, pp. 895–902.

[118] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-modal hybrid architecture for pedestrian action prediction," in *IV*, 2022, pp. 91–97.

[119] A. Y. Naik, A. Bighashdel, P. Jancura, and G. Dubbelman, "Scene spatio-temporal graph convolutional network for pedestrian intention estimation," in *IV*, 2022, pp. 874–881.

[120] H. Q. Dang, J. Fürnkranz, A. Biedermann, and M. Hoepfl, "Time-to-lane-change prediction with deep learning," in *ITSC*, 2017, pp. 1–7.

[121] Y. Hu, W. Zhan, and M. Tomizuka, "Probabilistic prediction of vehicle semantic intention and motion," in *IV*, 2018, pp. 307–313.

[122] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Syst. Appl.*, vol. 91, pp. 452–463, 2018.

[123] O. Scheel, L. A. Schwarz, N. Navab, and F. Tombari, "Situation assessment for planning lane changes: Combining recurrent models and prediction," in *ICRA*, 2018, pp. 2082–2088.

[124] 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, 2019.

[125] T. Han, J. Jing, and Ü. Özgüner, "Driving intention recognition and lane change prediction on the highway," in *IV*, 2019, pp. 957–962.

[126] R. Izquierdo, Á. Quintanar, I. Parra, D. F. Llorca, and M. Á. Sotelo, "Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM," in *ITSC*, 2019, pp. 3657–3662.

[127] A. Zyner, S. Worrall, and E. M. Nebot, "Naturalistic driver intention and path prediction using recurrent neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1584–1594, 2020.

[128] V. Mahajan, C. Katrakazas, and C. Antoniou, "Prediction of lane-changing maneuvers with automatic labeling and deep learning," *Transport. Res. Rec.*, vol. 2674, no. 7, pp. 336–347, 2020.

[129] A. Girma, S. B. Amsalu, A. Workineh, M. A. Khan, and A. Homaifar, "Deep learning with attention mechanism for predicting driver intention at intersection," in *IV*, 2020, pp. 1183–1188.

[130] L. Li, W. Zhao, C. Xu, C. Wang, Q. Chen, and S. Dai, "Lane-change intention inference based on RNN for autonomous driving on highways," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5499–5510, 2021.

[131] K. Griesbach, M. Beggiato, and K. H. Hoffmann, "Lane change prediction with an echo state network and recurrent neural network in the urban area," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6473–6479, 2022.

[132] X. Chen, H. Zhang, F. Zhao, Y. Hu, C. Tan, and J. Yang, "Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles," *IEEE Trans. Intell. Transp. Syst.*, 2022.

[133] Y. Hu, X. Jia, M. Tomizuka, and W. Zhan, "Causal-based time series domain generalization for vehicle intention prediction," in *ICRA*, 2022, pp. 7806–7813.

[134] W. Wang, T. Qie, C. Yang, W. Liu, C. Xiang, and K. Huang, "An intelligent lane-changing behavior prediction and decision-making strategy for an autonomous vehicle," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2927–2937, 2022.

[135] G. Xie, K. Yang, C. Xu, R. Li, and S. Hu, "Digital twinning based adaptive development environment for automotive cyber-physical systems," *IEEE Trans. Ind. Informatics*, vol. 18, no. 2, pp. 1387–1396, 2021.

[136] F.-Y. Wang, X. Wang, L. Li, and L. Li, "Steps toward parallel intelligence," *IEEE CAA J. Autom. Sinica*, vol. 3, no. 4, pp. 345–348, 2016.

[137] W. Zhou, Z. Cao, N. Deng, X. Liu, K. Jiang, and D. Yang, "Long-tail prediction uncertainty aware trajectory planning for self-driving vehicles," *arXiv preprint arXiv:2207.00788*, 2022.

[138] C. Zang, M. Pei, and Y. Kong, "Few-shot human motion prediction via learning novel motion dynamics," in *IJCAI*, 2021, pp. 846–852.

[139] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," in *EMNLP*, 2018, pp. 3090–3099.

[140] O. Styles, A. Ross, and V. Sanchez, "Forecasting pedestrian trajectory with machine-annotated training data," in *IV*, 2019, pp. 716–721.

[141] J. Wang, W. Chai, A. Venkatachalapathy, K. L. Tan, A. Haghighat, S. Velipasalar, Y. Adu-Gyamfi, and A. Sharma, "A survey on driver behavior analysis from in-vehicle cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10 186–10 209, 2022.

[142] Z. Hu, S. Lou, Y. Xing, X. Wang, D. Cao, and C. Lv, "Review and perspectives on driver digital twin and its enabling technologies for intelligent vehicles," *IEEE Trans. Intell. Veh.*, 2022.

[143] J.-Y. Kwak, B. C. Ko, and J.-Y. Nam, "Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime," *Infrared Phys.Technol.*, vol. 81, pp. 41–51, 2017.

[144] Y. Xing, C. Lv, H. Wang, H. Wang, Y. Ai, D. Cao, E. Velenis, and F. Wang, "Driver lane change intention inference for intelligent vehicles: Framework, survey, and challenges," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4377–4390, 2019.

[145] C. Dong and J. M. Dolan, "Continuous behavioral prediction in lane-change for autonomous driving cars in dynamic environments," in *ITSC*, 2018, pp. 3706–3711.

[146] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," *arXiv preprint arXiv:2204.00102*, 2022.

[147] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[148] S. Cai, Y. Shu, and W. Wang, "Dynamic routing networks," in *WACV*, 2021, pp. 3588–3597.

[149] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, 2021.

[150] Y. Chen, B. Ivanovic, and M. Pavone, "Scept: Scene-consistent, policy-based trajectory predictions for planning," in *CVPR*, 2022, pp. 17 103–17 112.

[151] D. Xie, T. Shu, S. Todorovic, and S. Zhu, "Learning and inferring "dark matter" and predicting human intents and trajectories in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1639–1652, 2018.

[152] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," in *ACM SIGKDD Explorations*, vol. 22, no. 1, 2020, pp. 18–33.

[153] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference," in *IROS*, 2020, pp. 10 711–10 718.

[154] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: steering counterfactual explanations with semantics," *CoRR*, vol. abs/2111.09094, 2021.

[155] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *CVPR*, 2022, pp. 15 159–15 168.

[156] Y. Cao, C. Xiao, A. Anandkumar, D. Xu, and M. Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction," *arXiv preprint arXiv:2209.08744*, 2022.

[157] R. Jiao, X. Liu, T. Sato, Q. A. Chen, and Q. Zhu, "Semi-supervised semantics-guided adversarial training for trajectory prediction," *arXiv preprint arXiv:2205.14230*, 2022.
[158] W. Ding, H. Lin, B. Li, and D. Zhao, "Causalaf: Causal autoregressive flow for safety-critical driving scenario generation," in *CoRL*, 2022.
[159] N. Muscholl, M. Klusch, P. Gebhard, and T. Schneeberger, "Emidas: explainable social interaction-based pedestrian intention detection across street," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021, pp. 107–115.
[160] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "Trafficsim: Learning to simulate realistic multi-agent behaviors," in *CVPR*, 2021, pp. 10 400–10 409.
[161] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *CVPR*, 2021, pp. 9909–9918.
[162] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation - A methodological perspective," *CoRR*, vol. abs/2202.02215, 2022.
[163] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, "CARLA: an open urban driving simulator," in *CoRL*, vol. 78, 2017, pp. 1–16.
[164] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *the 11th International Conference on Field and Service Robotics*, vol. 5, 2017, pp. 621–635.
[165] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *ICCV*, 2017, pp. 2232–2241.
[166] Q. Li, Z. Peng, Z. Xue, Q. Zhang, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *CoRR*, vol. abs/2109.12674, 2021.
[167] F. Wang, N. Zheng, D. Cao, C. M. Martinez, L. Li, and T. Liu, "Parallel driving in CPSS: a unified approach for transport automation and vehicle intelligence," *IEEE CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 577–587, 2017.
[168] J. Wang, X. Wang, T. Shen, Y. Wang, L. Li, Y. Tian, H. Yu, L. Chen, J. Xin, X. Wu, N. Zheng, and F. Wang, "Parallel vision for long-tail regularization: Initial results from IVFC autonomous driving testing," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 286–299, 2022.
[169] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *CVPR*, 2022, pp. 17 222–17 231.
[170] J. Bai, X. Fang, J. Fang, J. Xue, and C. Yuan, "Deep virtual-to-real distillation for pedestrian crossing prediction," in *ITSC*, 2022.
[171] K. Kim, Y. K. Lee, H. Ahn, S. Hahn, and S. Oh, "Pedestrian intention prediction for autonomous driving using a multiple stakeholder perspective model," in *IROS*, 2020, pp. 7957–7962.
[172] W. Chen, Z. Yu, Z. Wang, and A. Anandkumar, "Automated synthetic-to-real generalization," in *ICML*, vol. 119, 2020, pp. 1746–1756.
[173] W. Chen, Z. Yu, S. D. Mello, S. Liu, J. M. Alvarez, Z. Wang, and A. Anandkumar, "Contrastive syn-to-real generalization," in *ICLR*, 2021.
[174] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
[175] T. J. Schoonbeek, F. J. Piva, H. R. Abdolhay, and G. Dubbelman, "Learning to predict collision risk from simulated video data," in *IV*, 2022, pp. 943–951.
[176] H. Kim, K. Lee, G. Hwang, and C. Suh, "Crash to not crash: Learn to identify dangerous vehicles using a simulator," in *AAAI*, vol. 33, no. 01, 2019, pp. 978–985.
[177] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Viena2: A driving anticipation dataset," in *ACCV*, 2018, pp. 449–466.
[178] W. Tian, S. Wang, Z. Wang, M. Wu, S. Zhou, and X. Bi, "Multi-modal vehicle trajectory prediction by collaborative learning of lane orientation, vehicle interaction, and intention," *Sensors*, vol. 22, no. 11, p. 4295, 2022.
[179] Z. Su, G. Huang, S. Zhang, and W. Hua, "Crossmodal transformer based generative framework for pedestrian trajectory prediction," in *ICRA*, 2022, pp. 2337–2343.
[180] C. Choi, S. Malla, A. Patil, and J. H. Choi, "Drogon: A trajectory prediction model based on intention-conditioned behavior reasoning," in *CoRL*, 2020.
[181] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. Ni, "Joint intention and trajectory prediction based on transformer," in *IROS*, 2021, pp. 7082–7088.
[182] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, "LOKI: long term and key intentions for trajectory prediction," in *ICCV*, 2021, pp. 9783–9792.
[183] L. Li, J. Yao, L. Wenliang, T. He, T. Xiao, J. Yan, D. Wipf, and Z. Zhang, "Grin: Generative relation and intention network for multi-agent trajectory prediction," vol. 34, 2021, pp. 27 107–27 118.
[184] J. Liu, Y. Luo, H. Xiong, T. Wang, H. Huang, and Z. Zhong, "An integrated approach to probabilistic vehicle trajectory prediction via driver characteristic and intention estimation," in *ITSC*, 2019, pp. 3526–3532.
[185] Z. Huang, A. Hasan, K. Shin, R. Li, and K. Driggs-Campbell, "Long-term pedestrian trajectory prediction using mutable intention filter and warp lstm," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 542–549, 2020.
[186] R. Lopez, P. Boyeau, N. Yosef, M. Jordan, and J. Regier, "Decision-making with auto-encoding variational bayes," vol. 33, 2020, pp. 5081–5092.
[187] A. Bhattacharyya, D. O. Reino, M. Fritz, and B. Schiele, "Europvi: Pedestrian vehicle interactions in dense urban centers," in *CVPR*, 2021, pp. 6408–6417.
[188] L. Sun, C. Tang, Y. Niu, E. Sachdeva, C. Cho, T. Misu, M. Tomizuka, and W. Zhan, "Domain knowledge driven pseudo labels for interpretable goal-conditioned interactive trajectory prediction," in *IROS*, 2022.
[189] J. Schulz, C. Hubmann, J. Löchner, and D. Burschka, "Multiple model unscented kalman filtering in dynamic bayesian networks for intention estimation and trajectory prediction," in *ITSC*, 2018, pp. 1467–1474.
[190] H. Wu, L. Wang, S. Zheng, Q. Xu, and J. Wang, "Crossing-road pedestrian trajectory prediction based on intention and behavior identification," in *ITSC*, 2020, pp. 1–6.
[191] B. Tang, S. Khokhar, and R. Gupta, "Turn prediction at generalized intersections," in *IV*, 2015, pp. 1399–1404.
[192] Q. Xu, H. Wu, J. Wang, H. Xiong, J. Liu, and K. Li, "Roadside pedestrian motion prediction using bayesian methods and particle filter," *IET Intell. Transport Syst.*, vol. 15, no. 9, pp. 1167–1182, 2021.
[193] X. Shen, M. Lacayo, N. Guggilla, and F. Borrelli, "Parkpredict+: Multimodal intent and motion prediction for vehicles in parking lots with cnn and transformer," *arXiv preprint arXiv:2204.10777*, 2022.
[194] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of vulnerable road users from motion trajectories using stacked lstm network," in *ITSC*, 2017, pp. 327–332.
[195] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCV Workshops*, 2017, pp. 206–213.
[196] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo, "The prevention dataset: a novel benchmark for prediction of vehicles intentions," in *ITSC*, 2019, pp. 3114–3121.
[197] J. Xue, J. Fang, T. Li, B. Zhang, P. Zhang, Z. Ye, and J. Dou, "Blvd: Building a large-scale 5d semantics benchmark for autonomous driving," in *ICRA*, 2019, pp. 6685–6691.
[198] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.
[199] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019, pp. 6262–6271.
[200] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *CVPR*, 2020, pp. 11 186–11 196.
[201] A. Rasouli, T. Yau, P. Lakner, S. Malekmohammadi, M. Rohani, and J. Luo, "Pepscenes: A novel dataset and baseline for pedestrian action prediction in 3d," *arXiv preprint arXiv:2012.07773*, 2020.
[202] 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, 2022.
[203] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, "Multi-modal probabilistic prediction of interactive behavior via an interpretable model," in *IV*, 2019, pp. 557–563.
[204] H. Ma, Y. Sun, J. Li, M. Tomizuka, and C. Choi, "Continual multi-agent interaction behavior prediction with conditional generative

memory," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8410–8417, 2021.

[205] E. Banijamali, M. Rohani, E. Amirloo, J. Luo, and P. Poupart, "Prediction by anticipation: An action-conditional prediction method based on interaction learning," in *ICCV*, 2021, pp. 15 621–15 630.

[206] 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 *ITSC*, 2020, pp. 1–8.

[207] T. Bandyopadhyay, C. Z. Jie, D. Hsu, M. H. Ang, D. Rus, and E. Frazzoli, "Intention-aware pedestrian avoidance," in *International Symposium on Experimental Robotics*, 2013, pp. 963–977.

[208] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Congestion-aware multi-agent trajectory prediction for collision avoidance," in *ICRA*, 2021, pp. 13 693–13 700.

[209] S. Ferguson, B. Luders, R. C. Grande, and J. P. How, "Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions," in *WAFR*, vol. 107, 2014, pp. 161–177.

[210] K. Hirayama, K. Miyake, T. Shiotani, and T. Okimoto, "Dssa+: Distributed collision avoidance algorithm in an environment where both course and speed changes are allowed," *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 13, no. 1, 2019.

[211] Y. Tamura, P. Dai Le, K. Hitomi, N. P. Chandrasiri, T. Bando, A. Yamashita, and H. Asama, "Development of pedestrian behavior model taking account of intention," in *IROS*, 2012, pp. 382–387.

[212] D. Osipychev, "Collision avoidance for autonomous cars based on human intention," Ph.D. dissertation, Oklahoma State University, 2015.

[213] F. Chan, Y. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *ACCV*, vol. 10114, 2016, pp. 136–153.

[214] W. Bao, Q. Yu, and Y. Kong, "DRIVE: deep reinforced accident anticipation with visual explanation," in *ICCV*, 2021, pp. 7599–7608.

[215] A. V. Malawade, S. Yu, B. Hsu, D. Muthirayan, P. P. Khargonekar, and M. A. A. Faruque, "Spatio-temporal scene-graph embedding for autonomous vehicle collision prediction," *CoRR*, vol. abs/2111.06123, 2021.

[216] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident DB," in *CVPR*, 2018, pp. 3521–3529.

[217] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9590–9600, 2022.

[218] M. M. Karim, Y. Li, and R. Qin, "Towards explainable artificial intelligence (XAI) for early anticipation of traffic accidents," *Transport. Res. Rec.*, vol. abs/2108.00273, 2021.

[219] Z. Zheng, N. Zheng, and Y. Yang, "Parameter-efficient person re-identification in the 3d space," *IEEE Trans. Neural Networks Learn. Syst.*, 2022.

[220] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, 2019, pp. 8748–8757.

[221] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11 621–11 631.

[222] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *WACV*, 2021, pp. 1258–1268.

[223] J. Gesnouin, S. Pechberti, B. Stanciulescu, and F. Moutarde, "Assessing cross-dataset generalization of pedestrian crossing predictors," in *IV*, 2022, pp. 419–426.

[224] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *AAAI*, 2015, pp. 2901–2907.

[225] D. Widmann, F. Lindsten, and D. Zachariah, "Calibration tests in multi-class classification: A unifying framework," in *NeurIPS*, 2019, pp. 12 236–12 246.

[226] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *CVPR Workshops*, 2019, pp. 38–41.

[227] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[228] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *CVPR*, 2021, pp. 5579–5588.

[229] J. Zhang, Y. Dai, M. Xiang, D.-P. Fan, P. Moghadam, M. He, C. Walder, K. Zhang, M. Harandi, and N. Barnes, "Dense uncertainty estimation," *arXiv preprint arXiv:2110.06427*, 2021.

[230] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017, pp. 6402–6413.

[231] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *ICML*, 2018, pp. 1184–1193.

[232] 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.*, 2022.

[233] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *ACM MM*, 2020, pp. 1395–1403.

[234] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 2683–2693, 2020.

[235] S. Kang, H. Guo, and Y. Xue, "A reinforcement learning framework with description language for critical driving scenario generation," *arXiv preprint arXiv:2209.10078*, 2022.

[236] S. Han, D. Cao, L. Li, L. Li, S. E. Li, N.-N. Zheng, and F.-Y. Wang, "From software-defined vehicles to self-driving vehicles: A report on cpss-based parallel driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 1, pp. 6–14, 2018.

[237] Q. Zhang, Y. Gao, Y. Zhang, Y. Guo, D. Ding, Y. Wang, P. Sun, and D. Zhao, "Trajgen: Generating realistic and diverse trajectories with reactive and feasible agent behaviors for autonomous driving," *arXiv preprint arXiv:2203.16792*, 2022.

[238] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, 2018.

[239] B. Pang, K. Zha, H. Cao, J. Tang, M. Yu, and C. Lu, "Complex sequential understanding through the awareness of spatial and temporal concepts," *Nat. Mach. Intell.*, vol. 2, no. 5, pp. 245–253, 2020.

[240] P. Zhang, L. Bai, J. Xue, J. Fang, N. Zheng, and W. Ouyang, "Trajectory forecasting from detection with uncertainty-aware motion encoding," *CoRR*, vol. abs/2202.01478, 2022.

[241] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Vip3d: End-to-end visual trajectory prediction via 3d agent queries," *CoRR*, vol. abs/2208.01582, 2022.

[242] P. Gao, R. Guo, H. Lu, and H. Zhang, "Correspondence identification for collaborative multi-robot perception under uncertainty," *Auton. Robots*, vol. 46, no. 1, pp. 5–20, 2022.

[243] H. Wang, X. Yuan, Y. Cai, L. Chen, and Y. Li, "V2I-CARLA: A novel dataset and a method for vehicle reidentification-based V2I environment," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[244] U. Kim, D. Ka, H. Yeo, and J. Kim, "A real-time predictive pedestrian collision warning service for cooperative intelligent transportation systems using 3d pose estimation," *CoRR*, vol. abs/2009.10868, 2020.

[245] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *ECCV*, 2020, pp. 605–621.

[246] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, and Q. Ye, "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[247] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[248] S. Ren, S. Chen, and W. Zhang, "Collaborative perception for autonomous driving: Current status and future trend," *CoRR*, vol. abs/2208.10371, 2022.