Vision-based Intention and Trajectory Prediction in Autonomous Vehicles: A Survey

Izzeddin Teeti[1,2] , **Salman Khan**[1] , **Ajmal Shahbaz**[1] , **Andrew Bradley**[2] and **Fabio Cuzzolin**[1]
[1]Visual Artificial Intelligence Laboratory, Oxford Brookes University, Oxford, UK
[2]Autonomous Driving & Intelligent Transport, Oxford Brookes University, Oxford, UK
{iteeti, salmankhan, ashahbaz, abradley, fabio.cuzzolin}@brookes.ac.uk

Abstract

This survey targets intention and trajectory prediction in Autonomous Vehicles (AV), as AV companies compete to create dedicated prediction pipelines to avoid collisions. The survey starts with a formal definition of the prediction problem and highlights its challenges, to then critically compare the models proposed in the last 2-3 years in terms of how they overcome these challenges. Further, it lists the latest methodological and technical trends in the field and comments on the efficacy of different machine learning blocks in modelling various aspects of the prediction problem. It also summarises the popular datasets and metrics used to evaluate prediction models, before concluding with the possible research gaps and future directions.

1 Introduction

Prediction is about forecasting future events before they occur, which is beneficial for human-machine interaction systems, including surgical and industrial robots and autonomous vehicles (AVs). This survey will focus on prediction in autonomous vehicles. In this domain, prediction refers to the ability of the AV to forecast either the intention of other road users (e.g., they will turn left, right), including other cars, pedestrians and cyclists, or their future trajectory (when using video as input, their location in the image or map plane over a horizon of N future video frames). The accurate prediction of the intention and/or the trajectory of other road users can enable the 'planning' module of the autonomous (ego) vehicle or assistive system to avoid paths potentially leading to accidents and propose safe and reliable courses of action. Furthermore, it can foster social intelligence, as the first step towards understanding the thinking of other road users. For all these reasons, many companies have recently started to create a dedicated pipeline for AV prediction.

Multiple approaches have been suggested to tackle the prediction problem, including both physics-based models and machine learning-based models. The former depends on dynamic equations that govern the vehicle's motion to create an evolution model that predicts the final point and time of the current motion [Brännström *et al.*, 2010]. Such models do not capture contextual information of the scene; therefore, they cannot capture the high-level intentions of road users, and their ability to model the uncertainty of future trajectories is limited [Lin *et al.*, 2000]. Machine learning approaches, on the other hand, can distil high-level understanding from driving scenes by learning from data. They have arguably proven more successful in handling the above challenges in terms of accuracy and generalisation [Rasouli *et al.*, 2019], as they can model both temporal dynamics and the social features of road users and produce suitably multi-modal predictions [Yuan *et al.*, 2021].

Machine Learning (ML)-based prediction methods have explored different approaches to the problem. Some focus on predicting LiDAR point clouds in the future frames [Weng *et al.*, 2020], while others predict occupancy maps [Hoermann *et al.*, 2018]. This survey will focus on ML-based methods outputting visual predictions for road agents using input camera frames, either from an ego-vehicle perspective or from a *bird's eye viewpoint* (BEV).

Isolated surveys do exist on both the general prediction problem [Rasouli, 2020] and prediction in AVs [Mozaffari *et al.*, 2020]. The latter classifies prediction models based on input, method used and output, doing a good job highlighting the pros and cons of each class. However, neither papers systematically compares the contributions of the models' constituents blocks in tackling different aspects of the prediction problem. Also, existing surveys only list datasets without a detailed comparison between their attributes, failing to indicate which dataset to use to tackle a particular problem.

To the best of the authors' knowledge, this is the first survey to provide such an in-depth analysis of prediction models and datasets. Its main contributions are: (1) a formulation of a generic AV prediction pipeline; (2) a critical review of the ability of the various models' building blocks to model the social, temporal and generative dimensions of the prediction task, as a guide for researchers to select the correct ML technique for their problem; (3) a comparison of the relevant datasets in terms of size, input modalities, target agent and problem; (4) a summary of the latest methodological trends and a reasoned analysis of future research directions.

The survey is structured as follows. Section 2 introduces the prediction pipeline and formalises the prediction problem for both intention and trajectory. Section 3 highlights the main challenges of the prediction task, while Section 4 compares the most popular prediction datasets and lists the associated performance metrics. State-of-the-art (SOTA) prediction approaches are recalled in Tables in Section 5. Based on that, Section 6 focuses on the current research trends and comments on the efficacy of different ML techniques. Research gaps and future directions are discussed in Section 7.

2 Problem Formulation

Hereafter, the vehicle on which the sensors are mounted is called the *ego-vehicle*, while all the other road agents (including vehicles, pedestrians, cyclists and traffic lights) are termed *target agents*, as the prediction model aims to anticipate their future behaviour.

The input to the task is, given a current time instant, a sequence of features $\mathbf{X} = \{X_t^a\}_{t=1:T_{obs};a=1:A}$ related to A target agents extracted over the previous T_{obs} time instants. These features might include the target agent's location (either local, in image plane, or global, from a BEV), velocity, heading angle, or pose, as well as context information I in the form of RGB images, LiDAR point clouds, HD maps, semantic segmentation maps, etc. capturing road structure, traffic conditions or other environmental factors.

The problem is to calculate one or both of the following.

1. **Intention**. A set of K possible future intentions for the A agents in the next T_{pred} (future) frames, each with an associated probability. In most cases this is considered a classification problem, since intention is assumed to be annotated, either as a mind state or as a future action.
2. **Trajectory**. A set of K possible future locations for the A agents in the next T_{pred} future frames, formally $\mathbf{S} = \{S_k^a\}_{k=1:K,a=1:A}$ where each S_k^a contains a trajectory $s = \{s_t\}_{t=1:T_{pred}}$, and a scalar confidence c which captures the probability of this trajectory. Trajectory coordinates are measured either in the image plane or in the map plan (BEV).

To address this problem, a typical prediction pipeline will encode the input related to the agents in the scene, to then model their temporal and social dimensions before decoding their uni-modal or multi-modal intention or trajectory. Figure 1 shows the workflow of a typical AV prediction stack.

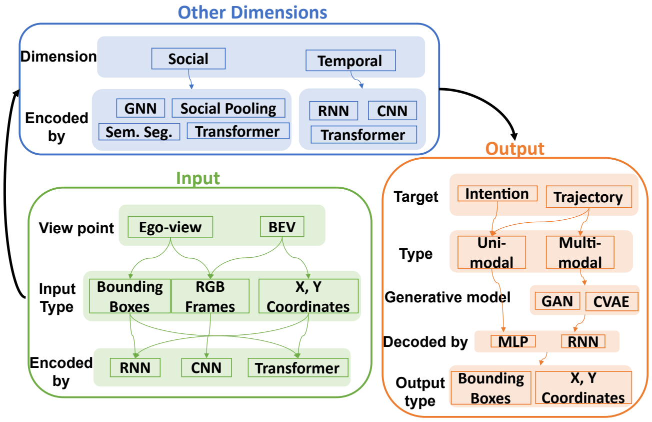

Figure 1. Prediction stack components and flow.

3 Challenges

Prediction in the AV domain is a complex problem due to the following characteristics of the driving environment.

1. **Dynamic**: Both the ego-vehicle (observer) and the target agents are moving. Therefore, the future trajectory of the target agent depends on its motion as well as the ego-vehicle's motion. Furthermore, in a dynamic environment the captured data is available sequentially, requiring that the model makes predictions in an *online*, rather than batch, fashion.
2. **Multi-agent**: Being multi-agent environments, roads host agents characterised by different goals and features. In particular, pedestrians possess more diverse and unpredictable trajectories than cars, which generally operate within a finite number of predefined lanes in manners restricted by road rules and geometry. Furthermore, the actions of one agent may influence the actions of other agents and vice versa in a *social interaction* process.
3. **Stochastic**: There is inherent uncertainty about a user's future intention or trajectory and the multimodality of agents' behaviours, as one past trajectory can have multiple possible future trajectories.
4. **Partially observable**: The surrounding environment and target agents are only partially observable by the ego vehicle due to occlusions etc; that is, the location, speed, heading angle, latent beliefs of other agents are not accessible/known by the ego-vehicle at all times.
5. **Real-time requirement**: In such a critical environment in which agents are moving at relatively high speeds, prediction algorithms need to perform in real-time, allowing time for the ego vehicle to react. This, in turn, increases the computational burden on the on-board PC.

In the light of the challenges mentioned above, an ideal prediction algorithm must model the social and temporal features of the ego vehicle and its surrounding road users (including cars and pedestrians) in an online processing fashion. Furthermore, it must generate multiple future trajectories (each with an associated probability), while running in real-time and assuming partial observability.

4 Training and Evaluation

Prediction approaches are tested on various standard datasets, and their performance evaluated using suitable metrics.

4.1 Datasets

While older datasets were limited in terms of environments and agent categories, modern benchmarks have significantly boosted progress in the AV prediction field.

For example, NGSIM-180 [Coifman and Li, 2017] and the highD Dataset [Krajewski *et al.*, 2018] used drones and surveillance cameras to capture cars on highways, featuring a single type of agent with a limited set of possible actions (move left/right and keep straight). On the other hand, ETH [Pellegrini *et al.*, 2009] contemplated trajectory prediction for off-road pedestrians only. PIE [Rasouli *et al.*, 2019], and its predecessor JAAD [Kotseruba *et al.*, 2016] also targeted the

	ETH	KITTI	PIE	Lyft	Waymo	nuScenes	Argoverse	LOKI
Year of release	2009	2012	2019	2019	2019	2019	2019	2021
Type of Prediction	Trajectory	Intention, Trajectory	Intention, Trajectory	Trajectory	Trajectory	Trajectory	Trajectory	Intention, Trajectory
# Classes	1	3	1	3	4	10	26	8
Night/Rain	✗/✗	✗/✗	✗/✗	✗/✗	✓/✓	✓/✓	✓/✓	✓/✗
# Cities	1	1	1	1	6	2	6	-
# Scenes	8	22	-	366	1k	1k	113	644
Annotated RGB frames	-	15k	293k	46k	200k	40k	22k	41k
LiDAR Ptc	✗	15k	✗	46k	200k	400k	44k	-
3D-BB	✗	✓	✗	✓	✓	✓	✓	✓
HD Map	✗	✗	✗	✓	✓	✓	✓	✗
Multi-Agent Evaluation	✗	✗	✗	✓	✓	✗	✓	✓
Sampling Rate	2.5Hz	10 Hz	30 Hz	10 Hz	10 Hz	2 Hz	10 Hz	5 Hz
Total Time	-	1.5h	6h	1118h	574h	5.5h	305h	2.3h
Trajectory Duration	8s	5s	2s	25s	9.1s	8s	11s	8s
Forecast Horizon	4.8s	3s	1.5s	5s	8s	6s	6s	5s
Download Size	-	354 GB	70 GB	22 GB	1.4 TB	48 GB	16 GB	-
Annotation	Manual	Manual	Manual	Auto, noisy	Auto, high quality	Manual	Auto, noisy	Manual

Table 1. Features of the most widely used AV prediction datasets. Hyphens "-" indicate that information is not available.

Method/Year	Input	Encoding	Temporal Encoding	Social Encoding	Generative model	Decoding	Loss/training	Target Agent	Uni/Multi-modal	view	Observe	Predict	Dataset/Metric/Performance
[Zaech *et al.*, 2020]	Past trajectory, velocity, HD map	CNN, GNN	3D-CNN	-	-	HMM	ADAM optimiser with binary cross-entropy for goal generation, smooth l1 loss for trajectory	Cars	Uni	BEV	2s	3s	Argoverse: mAP=0.61
[Liu *et al.*, 2020]	RGB frames, 2D-BB	CNN (ResNet-18)	GRU	GNN	-	GRU	-	Pedestrians	Uni	Ego	1s	1s, 2s, 3s	JAAD: Acc@1s=0.77
[Kotseruba *et al.*, 2021]	2D-BB, Pose, Ego-speed, RGB images.	RNN for sequences 3D-CNN for images	Temporal Attention followed by Modality Attention	-	-	FC layer	ADAM optimiser with cross entropy loss.	Pedestrians	Uni	Ego	0.5s	1 frame	PIE: Acc=0.87, AUC=0.86, F1=0.77 JAD: Acc=0.85, AUC=0.86, F1=0.68
[Benterki *et al.*, 2021]	Velocity, Acceleration, Heading angle	4-layer DNN	-	-	-	4-layer DNN	ADAM optimiser, categorical cross entropy loss.	Cars	Uni	BEV	5s	1 frame	The highD Dataset: Acc@3s=0.93
[Fang *et al.*, 2021]	RGB images, Semantic segmentation maps.	3D-CNN	3D-CNN	GCN	-	ConvLSTM	ADAM optimizer, KLDiv term, linear correlation coefficient, and normalized scanpath saliency	Cars, Cyclists, Pedestrians	Uni	Ego	-	1 frame	DADA-2000: AUC-J=0.92
[Kosaraju *et al.*, 2019]	RGB frames, Past locations.	MLP for past trajectory. VGG for images	LSTM	Graph Attention Network	GAN	LSTM	GAN loss, noise-reconstruction loss, L2 loss for trajectory, KL loss on the generated noise.	Pedestrians	Multi	BEV	3.2s	4.8s	ETH: ADE=0.48m, FDE=1m
[Salzmann *et al.*, 2020]	RGB frames, BB, Past location, HD maps.	LSTM for trajectory. CNN for HD map	LSTM and Dynamic model	GNN with Attention	CVAE	GRU with GMM	infoVAE loss	Pedestrians, Cars	Multi	BEV	2s 3.2s	6s 4.8s	nuScenes: FDE@3s=1.14 ETH: ADE=0.21, FDE=0.41
[Neumann and Vedaldi, 2021]	Camera intrinsics, RGB frames, 2D-BB, Odometry.	ResNet, self-supervised depth model which output ego-motion prediction	-	-	-	FC layer	Adam Optimiser. photometric loss for ego-motion predictor. L2 for bounding box locations.	Pedestrians	Uni	Ego	0.5 s	1.5s	PIE: MSE@0.5s=42, MSE@1.0s=153, MSE@1.5s=453.
[Bhattacharyya *et al.*, 2021]	RGB frames, 2D-BB	LSTM for trajectory. VGG for images	LSTM	Condition each agent's latent space on other agent's space	CVAE	LSTM	Adam optimiser with β-VAE loss	Pedestrians	Multi	Ego	5s	3s	EURO-PVI: FDE@3s=0.13m. nuScenes: FDE@3s=0.51m.
[Mangalam *et al.*, 2021]	RGB images, Past trajectory, Segmentations, Heatmaps.	Unet with ResNet backbone and ReLU	-	Semantic segmentation	Estimated distribution through Unet	Unet with Sigmoid	Adam optimiser with cross-entropy loss for Goal, Waypoints, and Trajectory distributions.	Pedestrians	Multi	BEV	3.2s	4.8s	ETH: ADE=0.18m, FDE=0.27m
[Pang *et al.*, 2021]	RGB frames, Past locations.	MLP	MLP	Attention-based Social Pooling	Latent space estimated by MLP	MLP	Adam optimiser with KL loss	Pedestrians	Multi	BEV	3.2s	4.8s	ETH: ADE=0.21s, FDE=0.38s
[Gilles *et al.*, 2021]	Past trajectory, Velocity, Angle, HD map.	GRU for trajectory. GNN for HD map	GRU and Cross Attention	GNN	Head map with probability distribution	-	Adam optimiser with pixel-wise focal loss	Cars	Multi	BEV	2s 2s	3s 6s	ArgoVerse: minADE@K5=1.59 nuScenes: minADE@K6=0.94
[Weng *et al.*, 2021]	Lidar, 3D-BB, RGB images,	LSTM for past trajectory. MLP for current detection	LSTM	GNN	CVAE	MLP	Ourentropy, cross entropy for multi-object detection. diversity, and L2 losses for prediction	Pedestrians, Cars	Multi	Ego	2s	3s	nuScenes: ADE@3s=1.02, FDE@3s=1.53. KITTI: ADE@3s=1.32, FDE@3s=2.3.
[Yuan *et al.*, 2021]	Past trajectory, Velocity, Angle, HD map.	Transformer for trajectory. CNN for map	Transformer	Transformer	CVAE	MLP	Adam optimiser with ELBO loss, L2 trajectory loss, diversity loss.	Pedestrians, Cars	Multi	BEV	2s 3.2s	6s 4.8s	nuScenes: ADE@K5=1.86, FDE@K5=3.89. ETH: ADE=0.23, FDE=0.39.
[Gu *et al.*, 2021]	Past trajectory, HD map	Attention	-	GNN (VectorNet)	Probability estimation using MLP	MLP	ADAM optimiser with binary cross-entropy for goal generation, smooth l1 loss for trajectory	Cars	Multi	BEV	2s	3s, 5s, 8s	Argoverse: minADE=0.88, minFDE=1.28 nuScenes: minADE=1.04, minFDE=1.55
[Hu *et al.*, 2021]	RGB frames, Past trajectory Ego-speed	CNN (EfficientNet)	Spatial Transformer, 3D-CNN	-	CVAE	ConvRNN	KL divergence loss, cross-entropy, L2 loss	Cars	Multi	BEV	1s	2s	nuScenes: IoU@short=59.4 Lyft: IoU@short=57.8
[Varadarajan *et al.*, 2021]	Past trajectory, HD map	LSTM, MLP-based fusion unit	LSTM	LSTM of other agent's relative features	GMM	MLP	ADAM optimiser with log likelihood	Cars	Multi	BEV	2s 1s	3s 8s	Argoverse: minADE@K6=0.79 Waymo: minADE@K6=0.56
[Schmidt *et al.*, 2022]	Series of discrete past displacements	LSTM	LSTM	GNN, attention	Using multiple decoders	Linear residual layer	ADAM optimiser with smooth L1 loss and Winner-takes-all loss	Cars	Multi	BEV	2s	3s	Argoverse: minADE@K6=0.85 minFDE@K6=1.44
[Deo *et al.*, 2022]	Past trajectory, speed, angle, HD map	GRU for past observations GNN for HD map	GRU	Attention	Probability estimation using MLP	MLP	ADAM optimiser with Negative Log Likelihood loss, winner takes all average displacement loss	Cars	Multi	BEV	2s	6s	nuScenes: minADE@K5=1.30, minADE@K10=1.00
[Cai *et al.*, 2020]	Lateral position, Lateral velocity	-	HMM with Dynamics model	-	-	-	Trained using Expectation-Maximization (EM) algorithm	Cars	Uni	BEV	2s	1 frame	NGSIM-180
[Rasouli *et al.*, 2021]	2D-BB, Ego-speed, RGB frames	LSTM	LSTM	Semantic segmentation, CNN, LSTM, Attention	-	LSTM	RMSProp optimiser with log(cosh), cross entropy, and multi-class entropy losses.	Pedestrians	Uni	Ego	3.2s	4.8s	PIE-intention: Acc=0.91, F1=0.85 PIE-Traj: ADE=15.21, FDE=35.03
[Girase *et al.*, 2021]	Actions, 2D-BB RGB images.	RNN	RNN	GNN, MLP, Attention	CVAE	MLP for intention. GRU for trajectory.	ADAM optimiser with KL term, L2-goal, L2-trajectory, and cross entropy for intention.	Pedestrians, Cars	Multi	BEV	3s	5s	LOKI: ADE= 0.79, FDE= 2.28.

Table 2. Comparison of SOTA models' attributes. Hyphens "-" indicate that information is either not applicable, or not available

trajectories of pedestrians only. Unlike ETH, they used an ego-vehicle camera to capture the scenes, and added an intention label for crossing or not crossing, allowing the dataset to be used for intention and trajectory prediction. All such datasets would capture the scene using camera sensor only.

KITTI [Geiger *et al.*, 2012] was one of the first multimodality datasets to provide, in addition to camera frames, LiDAR pointclouds for the input scenes. This, in turn, gave rise to the recent interest in detecting objects using 3D bounding boxes [Chen *et al.*, 2017]. Furthermore, KITTI provides annotations for both cars and pedestrians.

The deeper the AI model, the more images it needs to generalise efficiently. Recent datasets such as Lyft [Houston *et al.*, 2020], Waymo [Ettinger *et al.*, 2021], nuScenes [Caesar *et al.*, 2020] and Argoverse [Wilson *et al.*, 2021] have brought about a significant increase in terms of number of annotated frames, paving the road for the training of deep models. In addition to camera and LiDAR, these datasets provide High Definition (HD) maps that capture the topology of the road [Seif and Hu, 2016]. Adding HD maps has made possible to investigate global navigation abilities, enabling in turn the training of models upon a longer prediction horizon. Furthermore, unlike previous datasets, the datasets listed above contemplate more classes, recorded ego-vehicle odometry data, a variety of different cities, various weather and lighting conditions (including rain and night), and labels for other agents including traffic lights and road rules. Nevertheless, they still lack intention prediction-related labels.

LOKI [Girase *et al.*, 2021] has addressed this problem by providing action labels that can be harnessed by intention prediction models. Furthermore, the creators considered additional attributes that may influence the prediction model (e.g. age and weather). E.g., the probability of a pedestrian crossing the street not on a zebra crossing differs if the pedestrian is a teenager or an elderly. Finally, LOKI has an average of 21.6 agents per frame, useful for multi-agent prediction.

In summary, modern datasets helped addressing most of the prediction challenges by providing a huge amount of diverse, multi-agent, multi-modal data which can be used to train models able to predict the behaviour of different types of interacting agents in different weather conditions. Moreover, they provide annotations useful for high-level understanding of the driving scene including location, action, events. Commonly used datasets are outlined in Table 1.

4.2 Evaluation Metrics

Intention prediction. As a classification task, it is evaluated using traditional classification metrics including *accuracy*, i.e. the percentage of correctly classified samples over the total number of samples. Since it deals only with true positives, accuracy does not quite reflect the true performance in case of class-imbalance. The model will learn to focus on the more populated class in order to output many true positives, while generating a large number of false predictions for the other classes. Therefore, it is recommended to use other complimentary metrics including *precision* and *recall*, which consider the false positives and the false negatives, respectively. Precision and recall can also be utilised to calculate other metrics such as the *F1 score*, *mean Average Precision* (mAP) and the *Area Under the Curve* (AUC) [Rasouli, 2020].

Trajectory prediction. It is considered a regression problem in which the output is either a sequence of bounding boxes around the object of interest or its x - y locations. The *Average Displacement Error* (ADE) is widely used to evaluate the output. It is the mean l_2 distance between all the locations forming the predicted trajectory and the corresponding ground truth. Another common metric is the *Final Displacement Error* (FDE), which performs the same calculation but only for the final predicted location in the trajectory and its ground truth at the prediction horizon T_{pred}.

Models which use probabilistic generative methods to generate multi-modal prediction employ additional metrics. *Best of N* calculates ADE and FDE for the best N samples out of all the generated trajectories. When N equals 1, only the generated trajectory closest to the ground truth is chosen: this method is called *minADE* and *minFDE*, respectively. Other generative model-related measures are the different versions of *Negative Log Likelihood* (NLL), comparing the distribution of the generated trajectories against the ground truth.

5 Overview of the Recent Literature

Behaviour prediction can be either implicit, in the form of future trajectories, or explicit in terms of predicting future actions or events. An agent's intentions can be affected by: a) the agent's *own belief* or will (challenging to model, since it is usually not observed); b) its *social interactions*, which can be modelled using different approaches like social pooling, Graph Neural Networks, Attention, etc.; c) *environmental constraints*, such as the road layout, which can be encoded by High Definition (HD) maps; d) *contextual information*, in the form of RGB frames, LiDAR point clouds, optical flow, segmentation maps, etc.

On the other hand trajectory prediction is inherently more challenging - unlike intention (which is typically considered as a classification (discrete) problem), trajectory prediction is a regression (continuous) problem. Table 2 outlines state-of-the-art models targeting intention prediction (first three rows), trajectory prediction (remainder of the table), and prediction of both intention and trajectory (last three rows). The table analyses the type of input modalities, along with their encoding techniques. It then compares different methods used to model the social, temporal, and generative dimensions of the prediction problem. It also contrasts various loss functions, optimisation methods, and datasets used to train the models, before listing their output type and measured metrics.

The different classes of techniques used by the methods listed in the above tables are discussed and compared next to highlight the emerging trends in AV prediction.

6 Take-home Messages

6.1 Ego-Motion and Dynamics

In order to tackle the aforementioned (Section 3) challenge that accompanies a dynamic environment, methods have included a dedicated processing unit to model the ego-vehicle motion [Kotseruba *et al.*, 2021; Neumann and Vedaldi, 2021; Rasouli *et al.*, 2021; Salzmann *et al.*, 2020] to generate more accurate trajectories. In opposition, some methods model the target agents' motion, using either deep networks or dynamical models. [Cai *et al.*, 2020; Neumann and Vedaldi, 2021; Salzmann *et al.*, 2020] use dynamic equations at different levels of their algorithms to constrain the generated trajectory, making sure that they are dynamically feasible. In a related effort, researchers have leveraged additional quantities computed from the inputs directly provided by the dataset, such as pose, optic flow, semantic maps and heat maps.

6.2 Ego-Camera vs Bird's Eye View

In terms of the view in which predictions are formulated, methods can be categorised into two classes. In *BEV methods* the algorithm processes data coming from a top, map-like view [Salzmann *et al.*, 2020; Varadarajan *et al.*, 2021]. In *ego-camera prediction*, the algorithm perceives the world from the ego-vehicle perspective [Neumann and Vedaldi, 2021; Bhattacharyya *et al.*, 2021]. The latter is generally more challenging than the former due to a number of reasons.

Firstly, perceiving the world from a BEV allows for a wider field of view and more informative predictions. By contrast, ego-camera views have a shorter visual range that limits the prediction horizon, because the car cannot plan further than it sees. Also, ego-cameras are occluded more commonly than BEVs. Thus, BEV methods suffer less from the 'partial-observability' challenge than those based on an ego-camera. Secondly, unless LiDAR data is available, monocular vision makes it difficult for the algorithm to infer target agents' depth - which is an important clue in predicting their behaviour. Finally, as mentioned above, the ego-camera is moving - which requires processing both target agent's motion and the ego-vehicle motion, unlike the static BEV.

Despite inherent challenges in processing an ego-camera view, it remains more practical than BEV. Cars rarely have access to a camera showing BEVs of the road and target agents' locations. The conclusion is that a prediction system should be capable of viewing the world from the ego-vehicle's perspective. It may be advantageous to include LiDAR and / or stereoscopic camera data to perceive the world in 3D.

Another important related point is that, whenever the target agent's location has to be included for prediction, it is desirable to employ bounding box locations rather than purely centre points, as the former coordinates implicitly capture the changes in relative distance between ego-vehicle and pedestrians as well as the camera ego-motion [Rasouli *et al.*, 2019; Rasouli *et al.*, 2021]. In other words, bounding boxes become larger as the agent approaches the ego-vehicle, providing an additional (albeit rudimentary) estimation of depth.

6.3 Temporal Encoding

Since the driving environment is dynamic with many moving agents, it is critical to encode the temporal dimension of the agents to build a better prediction system connecting what happened in the past to what will happen in the future through the present. Knowing where an agent is coming from can help in guessing where it may go next. Most ego-camera-based models [Fang *et al.*, 2021; Kotseruba *et al.*, 2021] deal with shorter time horizons, and use *3D-Convolutional Neural Networks* 3D-CNN models to capture the temporal dimension of video data. 3D-CNNs stack the 2D video frames along the third dimension (time), then apply 3D convolutions. This can capture temporal evolution for short-horizon tasks such as intention, in which the last few frames before the event are the most relevant and important. Other works use *Hidden Markov Model* (HMM) [Cai *et al.*, 2020]. For longer horizons, however, a more complex structure is arguably required including Transformers [Yuan *et al.*, 2021] and variants of *Recurrent Neural Networks* (RNNs) [Girase *et al.*, 2021] such as *Long Short-term Memory* (LSTMs) [Bhattacharyya *et al.*, 2021; Fang *et al.*, 2021; Kosaraju *et al.*, 2019; Rasouli *et al.*, 2021; Salzmann *et al.*, 2020; Weng *et al.*, 2021; Yuan *et al.*, 2021] and *Gated Recurrent Units* (GRUs) [Gilles *et al.*, 2021]. It is noted that Transformer- and RNN-based models outperformed others in achieving SOTA results.

6.4 Social Encoding

To tackle the 'multi-agent' challenge, most top-performing algorithms use different types of *graph neural networks* (GNNs) to encode the social interaction between agents. [Gilles *et al.*, 2021; Kosaraju *et al.*, 2019] use a fully connected undirected graph in which all agents in the scene are connected. This method is clearly exponential in the number of connections. In opposition, [Fang *et al.*, 2021; Girase *et al.*, 2021; Salzmann *et al.*, 2020; Weng *et al.*, 2021] use a sparsely connected graph so that only agents that are located within a certain distance to each other are connected together - significantly lowering the number of links. Similarly, [Kosaraju *et al.*, 2019; Yuan *et al.*, 2021] implement sparse graphs, assuming that different types of agents have different perception ranges - and thus making the connections in the graph directional. The unique feature of [Girase *et al.*, 2021] is its use of a *heterogeneous graph* that does not only include road agents but also significant road elements (e.g. exit, entrance) thus modelling the relationship between agents and environment. As for the optimal number of GNN layers, [Addanki *et al.*, 2021] suggest that the deeper the graph the better the performance - in stark contrast, [Weng *et al.*, 2021] and [Liu *et al.*, 2020] claim that the optimal number is just 2.

[Rasouli *et al.*, 2021; Mangalam *et al.*, 2021] use semantic segmentation to extract the visual features of different classes, then find the relationship between them using attention. Unless panoptic segmentation is used (which distinguishes between semantic instances), all surrounding agents of the same class are treated the same whether they are interacting with the ego-vehicle or not. A less efficient method is *social pooling*, which was used along with attention in [Pang *et al.*, 2021]. [Bhattacharyya *et al.*, 2021] model the social dimension in the latent space of each agent, conditioned on that of previous agents. The limitation of this method is that it uses a human-biased ordering to decide on which agent the modelled agent depends, which may be incorrect.

Most approaches encode the temporal and social dimension separately - either starting with the temporal aspect to then consider the social one [Gilles *et al.*, 2021; Girase *et al.*, 2021; Kosaraju *et al.*, 2019; Rasouli *et al.*, 2021; Weng *et al.*, 2021], or by doing the opposite [Salzmann *et al.*, 2020]. [Yuan *et al.*, 2021] argues that such systems are bound to be suboptimal, and proposed a transformer-based model that can simultaneously encode both dimensions.

6.5 Goal-Conditioning

Studies in neuroscience [Valentin *et al.*, 2007] and computer vision [Gilles *et al.*, 2021; Mangalam *et al.*, 2021] suggest that humans are goal-directed agents. Furthermore, while making decisions, humans follow a series of successive levels of reasoning that eventually create their short or long-term plans. Based upon this, various algorithms have followed a strategy that appears to be beneficial to prediction performance: breaking down the prediction problem into two sub problems. The first is termed *epistemic*, and seeks the goals of each agent or answers the question 'where are the agents going?' The second is called *aleatoric* and solves for the trajectory that will carry the agent to the calculated goal, answering the question 'how is this agent going to reach its goal?' Methods may do that explicitly [Girase *et al.*, 2021; Mangalam *et al.*, 2021; Pang *et al.*, 2021] or implicitly [Gilles *et al.*, 2021], often conditioning the generated trajectories on the calculated goals. Those methods have achieved SOTA results in different driving datasets, reflecting the importance of goal-conditioning in improving prediction performance.

6.6 Multimodality

As the road environment is stochastic, one prior trajectory can unfold different future trajectories. Thus, a practical prediction system, that addresses 'stocasticity' challenge, is one that models the uncertainty of the problem. *Conditional variational auto-encoders* (CVAEs) achieve this by learning the latent space for the agents, then sampling from them during inference. [Girase *et al.*, 2021; Weng *et al.*, 2021; Yuan *et al.*, 2021] model agents independently and use the variational lower bound of the log-likelihood (ELBO) function, while [Bhattacharyya *et al.*, 2021] employ a CVAE to jointly model agents using a β-VAE loss function, arguing that ELBO suffers from posterior collapse. [Salzmann *et al.*, 2020] also model agents independently using a CVAE, but use an infoVAE loss instead. In their ablation studies, however, none of these authors investigate the effect of using different CVAE loss functions on the final prediction.

[Kosaraju *et al.*, 2019] use a *generative adversarial network* (GAN)-based method that uses adversarial losses for multimodal prediction. [Pang *et al.*, 2021], instead, use a latent belief energy-based method that is close to generative adversarial imitation learning (GAIL) - arguing that GAIL is preferable to inverse reinforcement learning (which is from the same family of models). The latter calculates an agent's policy through a two loop process (an outer one for the cost and an inner one for the policy, which is highly computationally expensive) while the former directly optimises a policy network [Pang *et al.*, 2021]. Interestingly, other authors have created datasets contemplating multiple future trajectories for a same observed trajectory [Liang *et al.*, 2020].

An important observation is that, in the current literature, multimodality *is only applied to trajectories*, completely neglecting its potential for intention prediction, despite the existence of ways to model the latent space of a discrete variable.

A final shared trend among prediction systems is the implementation of *attention* mechanisms to calculate the weighted importance of features or input modalities [Fang *et al.*, 2021; Gilles *et al.*, 2021; Girase *et al.*, 2021; Kosaraju *et al.*, 2019; Kotseruba *et al.*, 2021; Pang *et al.*, 2021; Rasouli *et al.*, 2021].

7 Research Gaps and Future Directions

Most researchers did not address the 'real-time' challenge, and did not touch on the computational demand of their systems. For the minority who did, including [Salzmann *et al.*, 2020], their system was not running in real-time. Therefore, a promising future direction is to design online prediction systems.

Several successful multi-modal approaches to trajectory prediction using different generative losses have been brought forward. However, the ideal loss function is still to be identified as none of the existing studies have performed suitable ablation studies in this sense. Further, all existing multimodal systems exclusively deal with trajectory prediction, leaving as a clear future research direction the creation of systems for joint multi-modal intent and trajectory prediction.

Current state of the art algorithms struggle to make accurate predictions in the presence of scenes different from those they were trained upon. This is an inherent limitation of prediction methods that learn from examples. A promising (yet unexplored) approach to modelling the social interaction of road users, and directly generalising goal-directed methods, is *theory of mind* (ToM) [Singh *et al.*, 2021; Cuzzolin *et al.*, 2020]. A notion from cognitive psychology, ToM is the ability to understand and anticipate other agents by simulating their mental states. Applied to autonomous driving, ToM could allow an ego vehicle to think from the perspective of the target vehicle to produce more accurate predictions of its future behaviour, especially in rare situations or cases in which humans change course without visible cues.

Finally, future prediction methods should not be limited by the availability of annotation. The AV industry is generating huge amount of visual data, which can only be leveraged by moving toward *self-supervised learning* methods.

8 Conclusions

The ability to accurately predict intentions and trajectories of other road agents can significantly enhance AV capabilities. Recent developments leveraging large, diverse and multi-modal datasets have significantly boosted research in this area. SOTA prediction models have harnessed powerful models such as attention, GNNs, LSTMs and transformers, although many outstanding challenges remain. Prediction in autonomous driving provides the key to a higher level of understanding of road scenes, thereby offering the potential improve the safety of future autonomous driving technology.

References

[Addanki *et al.*, 2021] Ravichandra Addanki, Peter W Battaglia, David Budden, Andreea Deac, Jonathan Godwin, Thomas Keck, Wai Lok Sibon Li, Alvaro Sanchez-Gonzalez, Jacklynn Stott, Shantanu Thakoor, et al. Large-scale graph representation learning with very deep gnns and self-supervision. *arXiv preprint arXiv:2107.09422*, 2021.

[Benterki *et al.*, 2021] Abdelmoudjib Benterki, Moussa Boukhnifer, Vincent Judalet, and Choubeila Maaoui. Driving intention prediction and state recognition on highway. *2021 29th Mediterranean Conference on Control and Automation, MED 2021*, pages 566–571, 6 2021.

[Bhattacharyya *et al.*, 2021] Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6404–6413, 2021.

[Brännström *et al.*, 2010] Mattias Brännström, Erik Coelingh, and Jonas Sjöberg. Model-based threat assessment for avoiding arbitrary vehicle collisions. *IEEE Transactions on Intelligent Transportation Systems*, 11:658–669, 9 2010.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Cai *et al.*, 2020] Wenqi Cai, Ganglei He, Jianlong Hu, Haiyan Zhao, Yuhai Wang, and Bingzhao Gao. A comprehensive intention prediction method considering vehicle interaction. *2020 4th CAA International Conference on Vehicular Control and Intelligence, CVCI 2020*, pages 204–209, 12 2020.

[Chen *et al.*, 2017] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[Coifman and Li, 2017] Benjamin Coifman and Lizhe Li. A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, 105:362–377, 2017.

[Cuzzolin *et al.*, 2020] Fabio Cuzzolin, Alice Morelli, B Cîrstea, and Barbara Jacquelyn Sahakian. Knowing me, knowing you: theory of mind in ai. *Psychological Medicine*, 50:1057 – 1061, 2020.

[Deo *et al.*, 2022] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pages 203–212. PMLR, 2022.

[Ettinger *et al.*, 2021] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[Fang *et al.*, 2021] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[Gilles *et al.*, 2021] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. *arXiv preprint arXiv:2109.01827*, 2021.

[Girase *et al.*, 2021] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[Gu *et al.*, 2021] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.

[Hoermann *et al.*, 2018] Stefan Hoermann, Martin Bach, and Klaus Dietmayer. Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling. pages 2056–2063, 2018.

[Houston *et al.*, 2020] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020.

[Hu *et al.*, 2021] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras, 2021.

[Kosaraju *et al.*, 2019] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian D. Reid, Seyed Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 2019.

[Kotseruba *et al.*, 2016] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016.

[Kotseruba *et al.*, 2021] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Benchmark for evaluating pedestrian action prediction. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1257–1267, 2021.

[Krajewski *et al.*, 2018] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018.

[Liang *et al.*, 2020] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2020.

[Lin *et al.*, 2000] Chiu Feng Lin, A. Galip Ulsoy, and David J. LeBlanc. Vehicle dynamics and external disturbance estimation for vehicle path prediction. *IEEE Transactions on Control Systems Technology*, 8:508–518, 2000.

[Liu *et al.*, 2020] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[Mangalam *et al.*, 2021] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

[Mozaffari *et al.*, 2020] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):33–47, 2020.

[Neumann and Vedaldi, 2021] Lukas Neumann and Andrea Vedaldi. Pedestrian and ego-vehicle trajectory prediction from monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10204–10212, June 2021.

[Pang *et al.*, 2021] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11819, 2021.

[Pellegrini *et al.*, 2009] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[Rasouli *et al.*, 2019] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6261–6270, 2019.

[Rasouli *et al.*, 2021] Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021.

[Rasouli, 2020] Amir Rasouli. Deep learning for vision-based prediction: A survey. *arXiv:2007.00095*, 2020.

[Salzmann *et al.*, 2020] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

[Schmidt *et al.*, 2022] Julian Schmidt, Julian Jordan, Franz Gritschneder, and Klaus Dietmayer. Crat-pred: Vehicle trajectory prediction with crystal graph convolutional neural networks and multi-head self-attention, 2022.

[Seif and Hu, 2016] Heiko G. Seif and Xiaolong Hu. Autonomous driving in the icity—hd maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016.

[Singh *et al.*, 2021] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, et al. Road: The road event awareness dataset for autonomous driving. *arXiv preprint arXiv:2102.11585*, 2021.

[Valentin *et al.*, 2007] Vivian V. Valentin, Anthony Dickinson, and John P. O'Doherty. Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, 27:4019 – 4026, 2007.

[Varadarajan *et al.*, 2021] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. *arXiv preprint arXiv:2111.14973*, 2021.

[Weng *et al.*, 2020] Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, and Nicholas Rhinehart. Inverting the pose forecasting pipeline with spf2: Sequential pointcloud forecasting for sequential pose forecasting. In *CoRL*, 2020.

[Weng *et al.*, 2021] Xinshuo Weng, Ye Yuan, and Kris Kitani. Ptp: Parallelized tracking and prediction with graph neural networks and diversity sampling. *IEEE Robotics and Automation Letters*, 6:4640–4647, 7 2021.

[Wilson *et al.*, 2021] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[Yuan *et al.*, 2021] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[Zaech *et al.*, 2020] Jan-Nico Zaech, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Action sequence predictions of vehicles in urban environments using map and social context. In *2020 IEEE/RSJ IROS*, pages 8982–8989. IEEE, 2020.