Injecting knowledge in data-driven vehicle trajectory predictors

Mohammadhossein Bahari [*], Ismail Nejjar , Alexandre Alahi

Ecole polytechnique fédérale de lausanne (EPFL), Switzerland

ABSTRACT

Vehicle trajectory prediction tasks have been commonly tackled from two distinct perspectives: either with knowledge-driven methods or more recently with data-driven ones. On the one hand, we can explicitly implement domain-knowledge or physical priors such as anticipating that vehicles will follow the middle of the roads. While this perspective leads to feasible outputs, it has limited performance due to the difficulty to hand-craft complex interactions in urban environments. On the other hand, recent works use data-driven approaches which can learn complex interactions from the data leading to superior performance. However, generalization, *i.e.*, having accurate predictions on unseen data, is an issue leading to unrealistic outputs. In this paper, we propose to learn a "Realistic Residual Block" (RRB), which effectively connects these two perspectives. Our RRB takes any off-the-shelf knowledge-driven model and finds the required residuals to add to the knowledge-aware trajectory. Our proposed method outputs realistic predictions by confining the residual range and taking into account its uncertainty. We also constrain our output with Model Predictive Control (MPC) to satisfy kinematic constraints. Using a publicly available dataset, we show that our method outperforms previous works in terms of accuracy and generalization to new scenes. Code is available at: https://github.com/vita-epfl/RRB.

1. Introduction

While driving, humans have this powerful capability to anticipate other drivers' decisions. Similarly, an autonomous vehicle should have the same prediction capability to safely navigate alongside human drivers. Some researchers addressed the vehicle trajectory prediction task, also known as microscopic traffic modeling, by building hand-crafted functions based on the available domain-knowledge to model average driving behaviors (Coscia et al., 2018; Cosgun et al., 2017; Keyvan-Ekbatani et al., 2016; Treiber et al., 2000; Xu et al., 2015; Coscia et al., 2016). These methods are interpretable and usually lead to a set of feasible predictions. However, they have limited performance since they not only miss non-average behaviors, but also are not able to model complex interactions. Conversely, recent works solely rely on experience, *i.e.*, learning from data, mostly using neural networks (Alahi et al., 2017; Xie et al., 2019; Zhang et al., 2019; Tang and Salakhutdinov, 2019; Kothari et al., 2020). Using large amount of data helps these methods to achieve accurate predictions without explicitly modeling the domain-knowledge. However, their predictions are not essentially realistic (on-road) and in some cases, even counter-intuitive. Moreover, they are prone to overfitting on the training data or poor performance on out-of-distribution data. Combining the domain-knowledge and the data will benefit from the strengths of both approaches and avoid their shortcomings.
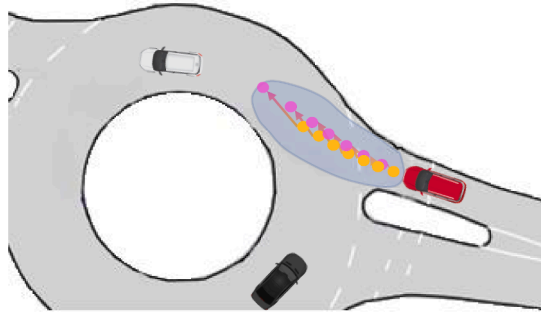
Fig. 1. Illustration of our Realistic Residual Block (RRB) model. The knowledge-based model generates a scene-compliant trajectory (yellow dots) which could not effectively account for other agents, hence, is too conservative. Our data-driven RRB improves the prediction (purple dots) by adding confined residuals conditioned on other agents in the scene (illustrated by the arrows). The blue region shows the physically-constrained output space for our RRB predictions.

We argue that driving is a skill learned from domain-knowledge and experience. The former is typically driven by physical constraints such as respecting road constraints or avoiding collisions. The later is driven by social conventions *e.g.*, the way drivers interact with each other or the safe way to enter a roundabout. Consequently, learning to predict vehicle trajectories can be re-framed as learning to combine both knowledge-driven and data-driven methods. Human's ability to employ both knowledge and data is not specific to driving. Researchers believe that humans learn rich representations (we refer to as knowledge) as well as patterns from observed examples in everyday life (Lake et al., 2017). This allows them to learn from fewer examples and to generalize to new conditions (Lake et al., 2015).

A popular way of adding domain-knowledge to the neural network is by adding constraints to the problem and optimizing the network under these constraints. Authors in (Niedoba et al., 2019) proposed a road-loss which can be interpreted as an approximation of the scene constraint in order to avoid off-road predictions. However, directly optimizing the model under constraints makes the optimization difficult and leads to sub-optimal results (Ganchev et al., 2010; Pathak et al., 2015).

There exist two main challenges towards creating a knowledge-driven and data-driven model. First, the combined model should be differentiable so that the data-driven part can be trained. This means that the integration needs careful design as the knowledge-driven part of the model usually is not differentiable. The second challenge is to preserve the benefits of both worlds after merging them *i.e.*, the final output should be realistic despite the fact that neural network might generate unrealistic outputs.

In this work, we address the aforementioned challenges. In the proposed approach shown in Fig. 1, the knowledge-driven (KD) trajectory is achieved by any knowledge-driven model. The KD output is then taken as input by our Realistic Residual Block (RRB) and the residuals required to be added to KD trajectory are found. In other words, the KD trajectory estimates coarse-grained behaviors based on the common driving performances while the residuals address fine-grained behaviors coming from non-modeled social interactions in KD trajectory as well as the long-tail of performances. This structure allows imposing knowledge by any function while allowing the residual block to be trainable. Moreover, in order not to diminish the feasibility of KD prediction, we can physically constrain the output of our RRB (*i.e.*, the allowed feasible offset) and leverage its uncertainty in the combination. Hence, our RRB uses a physically-constrained Inverse-variance weighting approach to add feasible and confident residuals to the KD trajectory. We finally constrain the final output with Model Predictive Control (MPC) to ensure kinematic-feasibility of predictions.

The contributions of this work can be summarized as: (1) proposing a Realistic Residual Block (RRB) as an effective way of leveraging both knowledge and data in vehicle trajectory prediction. Our RRB complements the knowledge-driven output with realistic and confident outputs. (2) Using Model Predictive Control (MPC) to bring kinematic constraints to the vehicle trajectory prediction task with latent control variables. (3) Demonstrate the generalization of our approach to new scenes and the multimodal case.

2. Related work

Vehicle trajectory prediction: Pioneering works addressed vehicle trajectory prediction problem by means of knowledge-driven methods. Researchers in (Cosgun et al., 2017) used Kalman filter (Kalman, 1960) to predict vehicle future trajectory. In order to incorporate scene information, (Ziegler et al., 2014) proposed associating vehicle's positions with the lanes of the road. Vehicle-vehicle interaction is addressed in (Treiber et al., 2000) to predict the longitudinal motion of a target vehicle. In Xu et al. (2015), an asymmetric optimal velocity model is presented to capture the asymmetry between acceleration and deceleration and (Keyvan-Ekbatani et al., 2016) studied lane change decision empirically. On the other hand, many researchers tackle the problem by leveraging data-driven models. A deep belief network is proposed in (Xie et al., 2019) to model lane-changing behavior. Authors in (Zhang et al., 2019; Deo and Trivedi, 2018) model lane-changing and car following behaviors simultaneously by employing long short-term memory (LSTM) neural networks and convolutional social pooling respectively. Authors in (Sadeghian et al., 2018) used an attention module to incorporate scene features into an LSTM model. An inverse optimal control (IOC) ranking module is used in (Lee et al., 2017) to determine the most likely hypotheses incorporating scene context and interactions. Moreover, some researchers addressed the
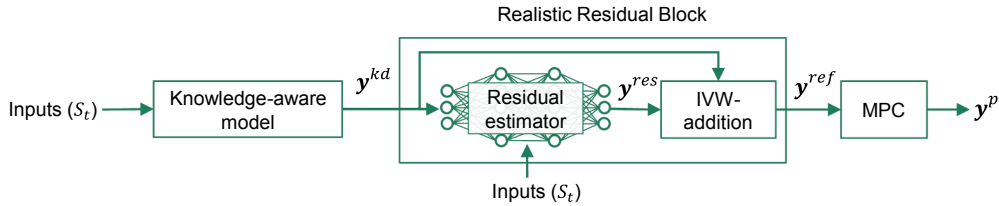
Fig. 2. Our proposed Realistic Residual Block (RRB) takes as input (i) the state of the scene, S_t, and (ii) the knowledge-driven prediction, $\mathbf{y}^{kd}$. The Residual estimator block builds physically-constrained residuals $\mathbf{y}^{res}$. Then, the IVW-addition block merges $\mathbf{y}^{kd}$ and $\mathbf{y}^{res}$ by Inverse-Variance Weighted sum and forms $\mathbf{y}^{ref}$ according to Eq. 3. Finally, Model Predictive Control (MPC) satisfies the kinema.tic constraints.

multimodal nature of human behavior prediction by using Winner-Takes-All (WTA) loss method (Cui et al., 2019a; Makansi et al., 2019). While all mentioned works rely on knowledge or data exclusively, our solution benefits from both approaches.

Injecting domain-knowledge to neural networks: Knowledge has been incorporated into the non-data-driven models to preserve realisticity and improve generalization. In An et al. (2015), Tang et al. (2016), space–time constraints were employed to confine the outputs to the feasible range. Researchers have proposed different means for incorporating domain-knowledge into neural networks (Willard et al., 2020; Borghesi et al., 2020; von Rueden et al., 2020). It can be injected into neural networks by designing specialized network structures (França et al., 2014; Garcez et al., 2019). As an example, Convolutional neural network (CNN) was created by changing multilayer perceptron (MLP) networks, considering image specifications (LeCun et al., 1998). Another approach for fusing neural networks with domain knowledge is to modify the learning algorithm. This can be done by exposing knowledge-driven negative samples to the model (Liu et al., 2020) or adding constraints to the outputs of the network (von Rueden et al., 2020). Other approaches are adding knowledge to the training data (Karpatne et al., 2017) and post-processing (Fang et al., 2017). From another perspective, some researchers used neural networks to address the imperfection of knowledge-driven models. They used knowledge-driven models to find biases of the data and the neural network compensates the error by outputting required residuals. Zeng et al. (2020), Silver et al. (2018) predict residuals on top of a physics-based robotic controller and (Kani and Elsheikh, 2017) finds residual minimiser of numerically discretized differential equations. A key limitation in residual modeling is by adding neural network-based residuals, the physics-based constraints, which are required for a realistic prediction, are hard to preserve (Willard et al., 2020). In this paper, we mitigate this limitation by confining the residual values. Moreover, we replaced naive addition of residuals by variance-based integration methods.

In the context of vehicle trajectory prediction, previous works attempt to add scene knowledge by adding penalty terms to the loss function. Researchers in (Bansal et al., 2019) defined an on-road loss to keep the predictions inside the road. The proposed loss requires the output to be an occupancy heatmap, which is computationally expensive to achieve and also not compatible with most of the trajectory prediction works (Niedoba et al., 2019). Authors in (Niedoba et al., 2019) proposed an off-road loss defined as the euclidean distance between each predicted waypoint and the nearest drivable point to penalize off-road predictions. Their experiments show that although off-road predictions are reduced, the performance deteriorates in terms of ℓ_2 loss. This is due to the fact that direct optimization of the model with constraints which are non-linear with respect to model parameters is difficult and leads to sub-optimal solutions (Ganchev et al., 2010; Pathak et al., 2015). In this paper, we propose a new method for adding scene knowledge to the model.

Kinematically-feasible predictions: Kinematic constraints are physical rules that need to be satisfied for a realistic vehicle prediction. Authors in (Cui et al., 2019b) showed that the predictions of the neural network model were not essentially kinematically-possible. Researchers in (Cui et al., 2019b; Li et al., 2019) solved the problem by adding the bicycle model (Kong et al., 2015) as a layer to the model. The network estimates control commands instead of coordinates and the kinematic layer converts them to a feasible trajectory. Despite the effectiveness of this approach, in many cases, it cannot be employed as most of the off-the-shelf models predict coordinates rather than control commands. Moreover, having control commands as the outputs prevents any further knowledge injection to the model's output, since any changes to the predicted trajectory can lead to a non-kinematically-feasible trajectory. In contrast, we propose using a Model Predictive Controller (MPC) to satisfy kinematic constraints. This allows the outputs to be in the coordinate format while the control commands are latent variables and mitigates the mentioned problems. MPC is commonly used for planning (Ziegler et al., 2014; Chen et al., 2019; Kazemi et al., 2018; Jeong et al., 2020; Bae et al., 2019). In this work, we show how it can also be used for the vehicle trajectory prediction task.

3. Proposed method

Humans have a clear understanding of the domain-knowledge while driving *e.g.*, where the drivable and non-drivable areas are. Moreover, they learn specific aspects of driving by experience *e.g.*, interacting with other agents. However, benefiting from the domain-knowledge and learning from experience simultaneously is challenging. We address the problem by proposing a Realistic Residual Block (RRB). RRB finds data-driven residuals conditioned on the knowledge-aware prediction. The output is a physically-constrained Inverse-Variance Weighted (IVW) sum of the knowledge-driven (KD) trajectory with the residuals. Finally, Model Predictive Control (MPC) is incorporated to satisfy kinematic constraints. Fig. 2 shows a high level picture of our proposed model. We will explain each part of the model in the following subsections.
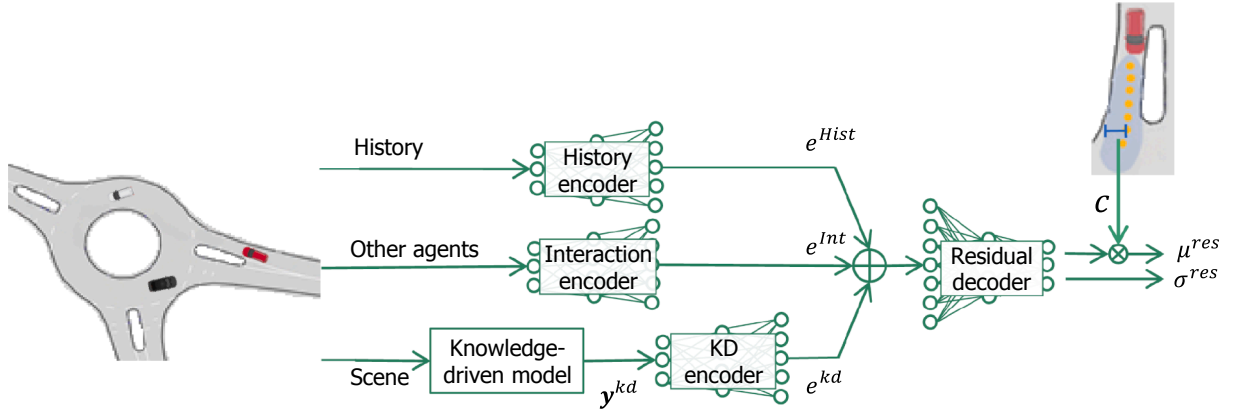
Fig. 3. The 'Residual estimator' block inside our RRB. Three encoders are employed to embed the history, the interactions and the KD trajectory. The residual decoder integrates the features and estimates the residual distribution. The mean of the distribution is confined to C, a real-word extracted parameter.

3.1. Problem formulation

The goal of our model is to predict future positions of a vehicle given its history and surroundings. Therefore, the state of the scene input to the model at time t, S_t, consists of an image of the scene and histories of ego-vehicle and other vehicles in the scene. By ego-vehicle, we refer to the vehicle whose future is predicted. The observation vectors for ego-vehicle and other vehicles' history comprises of T_o observed 2D coordinates of them in the world space. Without loss of generality, each time, one of the agents in the scene is taken as the ego-vehicle. At time t, the model predicts $\mathbf{y}_t^{ref}$ for the next T_p prediction frames as a Gaussian distribution with mean μ and variance σ as:

$$p\left(\mathbf{y}_t^{ref}|S_t, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t\right) = \prod_{j=1}^{T_p} \mathcal{N}\left(y_{t+j}^{ref}|S_t, \mu_{t+j}, \sigma_{t+j}\right). \tag{1}$$

Note that we indicate the sequence of values in a bold text. We omit the index t in the rest of the paper for simplicity.

3.2. Knowledge-aware prediction

Our proposed method is flexible in the choice of the knowledge-aware model. Hence, any off-the-shelf knowledge-driven prediction model can be employed to create the KD trajectory. In Section 4.5, we show the robustness of our RRB to different knowledge-driven models. In this work, we make the KD trajectory by utilizing the scene knowledge, the most influential domain-knowledge in vehicle prediction task. Inspired by (Ziegler et al., 2014), we use the lanes of the road to form the scene-compliant trajectory $\mathbf{y}^{kd}$. To have a probabilistic framework, we take KD trajectory as the mean of a Gaussian distribution with fixed variances which are approximated by the statistics of the training data.

3.3. Realistic residual block

While KD trajectory reflects average behaviors with respect to the scene, our proposed data-driven Realistic Residual Block learns the missing complex interactions. The architecture of RRB is depicted in Fig. 3. The learned residuals are first confined and then merged with the KD prediction by the IVW-addition block. This will lead to feasible and confident predictions. Both Residual estimator and integration mechanism are explained in the following subsections.

3.3.1. Residual estimator

Our Residual estimator's structure is shown in Fig. 3. The inputs are the KD trajectory and S_t consisting of the image of the scene and history of agents in coordinate format. The history of ego-vehicle is processed by the history encoder to find the driver's intention leading to the feature $\mathbf{e}^{Hist}$. We model the interaction among agents in a simple yet effective approach. Inspired by how vehicles interact in the real-word, our preprocessing eliminates non-interacting agents in the scene. In contrast with most previous works (Alahi et al., 2016; Tang and Salakhutdinov, 2019) which consider surrounding agents, it excludes all the agents behind the vehicle as usually, vehicles interact with the agents in front. Moreover, as only the closest vehicles can impact the driver's behavior, it only preserves a set of closest vehicles and passes them to the interaction encoder model to get the interaction features $\mathbf{e}^{Int}$. Our experiments show the effectiveness of the model in learning interactions between agents. The KD trajectory $\mathbf{y}^{kd}$, is encoded into the vector $\mathbf{e}^{kd}$ by the KD encoder network. All the encoders are feed-forward networks with ReLU non-linearities. Finally, the residual decoder estimates the residuals given the concatenated features as:

$$
\begin{aligned}
(\widehat{\mu}^{res}, \sigma^{res}) &= MLP\big(\big[\mathbf{e}^{Hist}, \mathbf{e}^{Int}, \mathbf{e}^{kd}\big]; W^{res}\big), \\
\mu^{res} &= C\widehat{\mu}^{res},
\end{aligned}
\tag{2}
$$

where $(\mu^{res}, \sigma^{res})$ are parameters of the Gaussian distribution $\mathbf{y}^{res}$ and MLP is a multilayer perceptron network parameterized by W^{res}. Note that we bound the range of $\widehat{\mu}^{res}$ to $(-1, 1)$ by using Tanh activation function in the last layer. μ^{res} is the scaled version of $\widehat{\mu}^{res}$ by C to adjust the max feasible deviation from the center of the road. C is a real-word extracted parameter which can be adapted to each scene. In our experiments, we chose C equal to the half of the minimum width of the road in each scene.

3.3.2. IVW-addition

The merged trajectory $\mathbf{y}^{ref}$ can be achieved by simply adding residual Gaussians to the KD Gaussian prediction, referred to "A-RRB" baseline in Section 4.3. However, adding uncertain residuals to the KD trajectory can impair the KD predictions. To mitigate that, we utilize Inverse-Variance Weighting (IVW) (Cochran, 1954) to scale residuals according to their uncertainties which results in the most certain output (in terms of output variance) among all weighted averages (Hartung et al., 2011). Hence, our goal is to find the weights $w, \widetilde{w}$ such that each point on the merged trajectory shown in Eq. 3 has the minimum variance.

$$
\mathbf{y}^{ref} = w\mathbf{y}^{kd} + \widetilde{w}\big(\mathbf{y}^{kd} + \mathbf{y}^{res}\big).
\tag{3}
$$

Let's represent $\mathbf{y}^{kd} + \mathbf{y}^{res}$ by $\mathbf{y}^{ad}$ for brevity. The points on each trajectory are assumed temporally-independent. Weight matrices have the form of $w = diag(w_1, w_2), \widetilde{w} = diag(\widetilde{w}_1, \widetilde{w}_2)$ where $w_i, \widetilde{w}_i$ are scalar values. Note that $\mathbf{y}^{kd}$ is a prior for $\mathbf{y}^{res}$ and hence they are not independent. The problem formulation is as follows:

$$
argmin_{w,\widetilde{w}}(\sigma_{11}^{ref}, \sigma_{22}^{ref}), \text{subject to}: \sigma_{ref} = w\sigma^{kd}w^T + \widetilde{w}\sigma^{ad}\widetilde{w}^T + 2w\sigma^{kd,ad}\widetilde{w}^T, w + \widetilde{w} = I_{2\times 2},
\tag{4}
$$

where σ^{kd}, σ^{ad} are the covariance matrices and $\sigma^{kd,ad}$ is the cross-covariance matrix. Solving the constraint problem using Lagrangian multipliers leads to the following solution:

$$
w = \begin{bmatrix} \dfrac{\sigma_{11}^{ad} - \sigma_{11}^{kd,ad}}{\sigma_{11}^{ad} + \sigma_{11}^{kd} - 2\sigma_{11}^{kd,ad}} \\[2em] \dfrac{\sigma_{22}^{ad} - \sigma_{22}^{kd,ad}}{\sigma_{22}^{ad} + \sigma_{22}^{kd} - 2\sigma_{22}^{kd,ad}} \end{bmatrix}, \quad \widetilde{w} = \begin{bmatrix} \dfrac{\sigma_{11}^{kd} - \sigma_{11}^{kd,ad}}{\sigma_{11}^{ad} + \sigma_{11}^{kd} - 2\sigma_{11}^{kd,ad}} \\[2em] \dfrac{\sigma_{22}^{kd} - \sigma_{22}^{kd,ad}}{\sigma_{22}^{ad} + \sigma_{22}^{kd} - 2\sigma_{22}^{kd,ad}} \end{bmatrix}.
\tag{5}
$$

3.4. Multimodal prediction and loss function

Our approach can easily be extended to a multimodal predictor using Winner-Takes-All (WTA) loss (also known as oracle loss) approach (Rupprecht et al., 2017). To do so, the knowledge-aware model should generate multiple plausible future trajectories. RRB takes the KD predictions and finds the associated required residuals. Finally, the loss will be calculated, considering the closest mode to the ground truth. Minimizing the log-likelihood will lead to the following loss function:

$$
l(\theta) = -\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbf{1}(m = m^*)\Big[\log p\Big(\mathbf{x}_{n,m}^p \Big| S_n, \mu_{n,m}^{ref}(\theta), \sigma_{n,m}^{ref}(\theta)\Big)\Big],
\tag{6}
$$

where N and M are number of samples and modes respectively, $\mathbf{1}$ is the indicator function, and m^* is the closest output to the ground truth in terms of ℓ_2 distance. Note that for the single modal case, the equation holds with $M = 1$.

3.5. MPC

To add kinematic feasibility, in contrast with the previous works (Cui et al., 2019b; Li et al., 2019) which utilize a kinematic layer after computing control commands, we employ Model Predictive Control (MPC). This gives the model the flexibility to estimate the positions instead of finding the control commands which is beneficial in adding the domain-knowledge. MPC minimizes its cost function subject to a set of constraints. The state parameters for agent i in time t is $s_t = [x_t, y_t, \phi_t, v_t]$ which consists of coordinates, orientation and speed. We denote control parameters acceleration and the steering angle by $u_t = [a_t, \gamma_t]$. The dynamics of the system is formulated using bicycle model F_{bic} (Kong et al., 2015) which is shown to be sufficient for normal manoeuvres (Kong et al., 2015). Then, the MPC solves the following optimization problem:

$$
\mathbf{y}_t^p = argmin_{s_{t:t+T_p}, u_{t:t+T_p}} \sum_{j=1}^{T_p} \| s_{t+j}[0:1] - \mathbf{y}_{t+j}^{ref} \|_2^2 + \lambda \| u_{t+j} - u_{t+j-1} \|_2^2
\tag{7}
$$

$$
\text{subject to}: s_{t+1} = F_{bic}(s_t, u_t), \quad s_0 = s_{init}, \quad u_{min} < u_t < u_{max},
$$

where $\mathbf{y}^{ref}$ is the reference trajectory, λ is a hyper-parameter and u_{min}, u_{max} are minimum and maximum feasible control values

respectively.

4. Experiments

We evaluate our RRB against other baselines to test the following hypotheses: (1) RRB brings the advantages of both knowledge-driven and data-driven models thus, outperforms both types of models, (2) RRB structure merges knowledge-driven and data-driven models better than other fusion techniques, (3) adding MPC makes outputs more realistic by satisfying kinematic constraints, (4) our trained RRB can improve performance of different knowledge-driven models without the need for fine-tuning, (5) IVW-addition and residual confinement are essential parts for more realistic outputs.

4.1. Dataset

We evaluate the proposed method on Interaction dataset (Zhan et al., 2019). It is a large-scale real-world dataset which consists of top-down scenes from intersections, highways, and roundabouts. The data is collected from three different continents (North America, Asia and Europe). It includes locations of dynamic agents such as vehicles and pedestrians for each frame of the 10 Hz downsampled video. The dataset also provides the static context of the scene. Interaction dataset is challenging as it includes interactions between vehicles, different environments, and potentially multiple plausible predictions. We used the same settings as (Zhan et al., 2019). The observation and prediction lengths indicate the number of frames used to represent the past states and to be predicted, respectively. We have set observation lengths equal to 5 frames (2.5 s) and and prediction length as 10 frames (5 s). The dataset includes scenes with different numbers of samples. Hence, in order not to be overwhelmed by scenes with large number of samples, we report the average of performance of a model on the three categories regardless of the number of samples in each category. Moreover, to study the model's generalization power, we consider two scenarios: the first scenario is scene-overfitting, in which all scenes exist in the training set, but 20% of data is kept for the test set. In this scenario the model should be able to overfit on the scene and generalize on the interactions. We argue that to assess model generalization in terms of scene perception, another scenario named scene-generalization is required. In this scenario, the three following scenes, 'DR_USA_Intersection_MA', 'DR_USA_Roundabout_SR', 'DR_CHN_Merging_ZS' are kept for the test set and the rest scenes are used in training. This challenging scenario reveals the performance of the models in a new environment.

4.2. Implementation details

The models are trained for 50 epochs with batch size of 32. We employed Adam optimizer (Kingma and Ba, 2014) with the initial learning rate of 0.001, which is decreased by half every 10 epochs. The model is implemented using PyTorch (Paszke et al., 2017). Network's building blocks are MLP networks with the following hidden layers: History and interaction encoders have (32,32,64), KD encoder has (32,64) and the decoder has (256,128,128,64) hidden layers.

4.3. Metrics and baselines

The following metrics were used for the evaluation:

1. Average/final displacement error **(ADE/FDE)**. Average displacement error (ADE) and Final displacement error (FDE) are adopted as two common evaluation metrics. In the multimodal case, similar to the previous works (Gupta et al., 2018; Lee et al., 2017), the closest mode to the ground truth is chosen.
2. Road violation **(RV)**. Inspired by (Niedoba et al., 2019), we define this metric as the percentage of average number of points predicted in the off-road area. For the multimodal case, we average RVs for the modes weighted by their probability. RV measures the feasibility of predictions concerning the scene.
3. Cross track **(CT)**. Cross track is the distance between the actual destination and the final point on the retimed predicted sequence by the ground truth speed profile(Gong and McNally, 2004). In other words, cross track is the distance between the actual destination and the final point on the retimed predicted sequence. Cross track metric is able to express the spatial effectiveness of the model, excluding temporal aspects.

To demonstrate the effectiveness of RRB, we compare the results with the following baselines:

Naive baselines: We used Kalman filter as a linear prediction **(Lin)**. We also report Constant velocity **(CV)** (Schöller et al., 2020) as another naive baseline.

Knowledge-driven baselines: In order to form the KD trajectory explained in Section 3.2, we utilize different strategies to predict vehicle velocity. We name Constant velocity strategy (Schöller et al., 2020) as **KD1**, and Leader follower (Treiber et al., 2000) which tackles interaction between agents as **KD2**.

Data-driven baselines: For the data-driven models that address agent-agent interactions, we report Social LSTM **(S-LSTM)** (Alahi et al., 2016), Social Attention **(S-ATT)** (Vemula et al., 2018) (numbers reported from (Li et al., 2020)), and Social GAN **(S-GAN_M)** (Gupta et al., 2018). We report Social WaGDAT **(S-WaGDAT_M)** (Li et al., 2020) (numbers reported from their paper) as a state-of-the-

Table 1
Quantitative results of baselines in scene-generalization scenario. ADE/FDE and CT are reported in meters. The lower the better for all metrics.

Models		Scene-generalization		
		ADE/FDE	RV	CT
Naive	Lin	4.13/ 8.77	34	3.6
	CV (Schöller et al., 2020)	3.12/ 7.34	24	3.19
Knowledge-driven	KD1 (Schöller et al., 2020; Ziegler et al., 2014)	2.92/ 6.62	**0**	1.87
	KD2 (Treiber et al., 2000; Ziegler et al., 2014)	2.85/ 6.55	**0**	**1.74**
Data-driven	S-LSTM (Alahi et al., 2016)	2.85/ 7.17	72	4.26
	S-GAN_M (Gupta et al., 2018)	2.34/ 5.82	52	5.71
	EDN	2.78/ 6.7	7	3.4
Mixed approaches	VI1	2.52/ 6.27	3	2.9
	VI2	2.62/ 6.38	2	2.32
	RL (Niedoba et al., 2019)	2.56/ 6.40	9	3.15
	RRB	2.44/ 6.04	0	1.98
	RRB_M	2.15/ 5.08	0	1.81
	RRB_M + MPC	**2.13/ 5.02**	0	1.81

Table 2
Quantitative results of baselines in scene-overfitting scenario. ADE/FDE are in meters.

Models	ADE/FDE
CV (Schöller et al., 2020)	2.80/ 6.59
KD1 (Schöller et al., 2020; Ziegler et al., 2014)	2.59/ 6.00
KD2 (Treiber et al., 2000; Ziegler et al., 2014)	2.53/ 5.83
S-LSTM (Alahi et al., 2016)	2.33/ 4.52
EDN	1.81/ 4.05
S-ATT (Vemula et al., 2018)	2.29/ 4.25
S-GAN_M (Gupta et al., 2018)	2.12/ 4.20
S-WaGDAT_M (Li et al., 2020)	1.62/ **3.35**
RRB_M	**1.49**/ 3.68

art work that models both scene and interactions. We also report the performance of our encoder-decoder neural network **(EDN)** similar to Fig. 3 trained with Eq. 6 as the loss function. We replaced the knowledge-driven model in Fig. 3 with a convolutional neural network to represent a fully data-driven model. Note that the multimodal baselines' names end with '_M'.

Mixed baselines: We implemented the road loss **(RL)** (Niedoba et al., 2019) to impose the scene constraint to the model. It adds the road loss to the NN model to help it learn the drivable and non-drivable regions. We also consider two alternatives to the RRB model. The first one is to use a Variance-based Integration **(VI1)** between the KD1 and data-driven prediction EDN by employing IVW method (Cochran, 1954). Note that this is equivalent to using Bayes rule while KD output is taken as the prior and EDN as the likelihood. The second approach is to replace the KD variance in VI1 by a fixed hyperparameter **(VI2)**.

Our proposed solution is RRB model **(RRB)** explained in Section 3.3 which uses KD1 method. We express the effectiveness of our method on other KD predictions in the ablation study. We also demonstrate the performance of our model in the multimodal case **(RRB_M)** with 2 modes and when constrained by the MPC **(RRB_M + MPC)**. Moreover, in the ablation study, we report the performance of RRB while IVW-addition is replaced by simple addition **(A-RRB)** and also a non-confined RRB model **(NC-RRB)**.

4.4. Results

Table 1 provides the results in scene-generalization scenario. Knowledge-driven models outperform other models in terms of RV and CT metrics since they can perceive scene well and also generalize to new environment. The data-driven approaches have better ADE/FDE as they can learn interactions from data. Although EDN performs well in scene-overfitting scenario (shown in Table 2), it fails in generalizing to the new environment leading to high RV and CT values. The large gap between the performance of data-driven methods in scene-overfitting and the scene-generalization scenarios reveals the limited generalization power of solely data-driven approaches, which can be mitigated by leveraging domain-knowledge. The proposed RRB model outperforms knowledge-driven models as it learns vehicle-vehicle interactions from data and goes beyond average behavior. Compared to the data-driven models, RRB can better generalize to new environments due to utilization of the scene knowledge. Compared with the mixed approaches, RRB outperforms both VI1 and VI2 because as opposed to them that merge two independent models, RRB generates residuals conditioned on KD output. Our method has zero RV because of the realistic residual scale which prevents off-road predictions. Note that RL could not improve EDM performance as it is very sensitive to the choice of hyperparameters and has many local minima.

The possibility of having multiple outputs is investigated by increasing the number of modes of RRB model to 2. The results approve
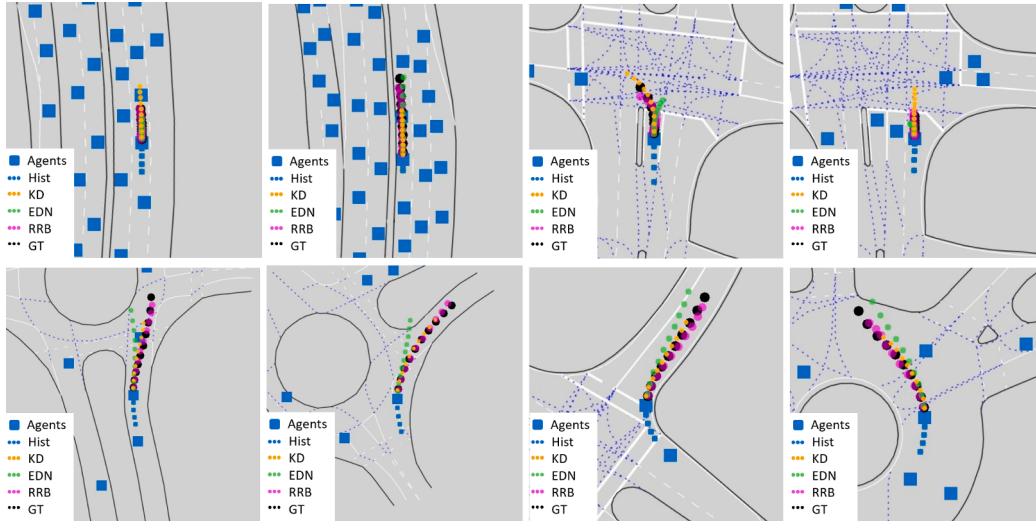
Fig. 4. Qualitative results of different baselines. Given the scene and history (Hist), the models predict the future positions. The ground truth (GT) is shown in black. The encoder-decoder neural network (EDN) captures the interactions while is prone to create unrealistic outputs shown in first row. Knowledge-driven (KD) model has realistic predictions as it uses the scene knowledge but cannot reason about the interactions among vehicles as shown in the second row. Our proposed RRB predicts realistic and interaction-aware outputs.

Table 3
Results of the two ablation studies on robustness and realistic elements of the proposed RRB.

(a) Assessing the robustness of RRB with respect to different knowledge-driven models. RRB is not trained in this experiment but we used the fixed model in Table 1.				(b) Ablation study on the two realistic elements of the proposed RRB, residual confinement and IVW-addition block.			
Models	Scene-generalization			Models	Scene-generalization		
	ADE/FDE	RV	CT		ADE/FDE	RV	CT
Lin	4.13/ 8.77	34	3.6	NC-RRB	2.28/ 5.59	10	1.94
Lin + RRB	3.11/ 7.50	2	3.59	A-RRB	2.47/ 6.08	3	2.07
KD2	2.85/ 6.55	0	1.74	RRB	2.44/ 6.04	0	1.98
KD2 + RRB	2.49/ 6.14	0	2.04				

that RRB_M is able to successfully capture multiple modes while it provides the required residuals for each mode individually leading to a multimodal knowledge-aware model. MPC is employed in RRB_M + MPC in order to ensure the kinematic-feasibility of the outputs. The results show that kinematic constraints are satisfied without loss of performance. Hence, the model better approximates human behavior.

The results of previous works in scene-overfitting scenario are reported in Table 2. Our RRB_M model outperforms previous works and especially the recent S-WaGDAT_M (Li et al., 2020) model which expresses the ability of the model to learn from data. Note that RRB_M could successfully improve performance of KD1 by adding the residuals.

We visualize the outputs of different models in Fig. 4. The first row images visualize the cases where KD prediction is not accurate because of the missing interactions. RRB complements KD trajectory by accounting for interactions in the residuals. The second row images emphasize on realisticity of the predictions. Although EDN can reason about interactions, it can have unrealistic predictions with respect to the scene especially in a new environment. However, the realistic elements of RRB preserve the feasibility of the final output as shown in all images.

4.5. Ablation study

In this section, a set of ablation studies are performed to shed light on the effectiveness of different parts of the model. First, we want to study the robustness of RRB concerning the different knowledge-driven models. We report two knowledge-driven methods in Table 1 and report the RRB performance while KD1 was employed. Without retraining the model, we replaced KD1 with Lin (Lin + RRB) and KD2 (KD2 + RRB) models. The results are expressed in Table 3a, where the performance of the knowledge-driven model is reminded in the parenthesis. It shows that RRB could successfully improve the performance of the models in all metrics. This experiment shows the robustness and effectiveness of the proposed RRB for different KD trajectories.

The second ablation study aims at assessing realistic elements of RRB which are the realistic parameter **c** and IVW-addition. We trained the model without the limitation on the range of RRB as non-confined RRB (NC-RRB). In addition, IVW-addition can be

replaced with simply adding the residuals to the KD prediction which is reported as A-RRB. This is equivalent to setting $w = diag(0, 0), \widetilde{w} = diag(1, 1)$ in Eq. 3. The results are expressed in Table 3b. Non-confined model has more freedom in improving the KD trajectory hence achieves a better ADE/FDE. But it is prone to unrealistic predictions with large RV error. Also, neglecting uncertainties in A-RRB will increase RV and CT as the network utilized uncertain residuals. These experiments approve that the added elements play important roles in having realistic predictions.

5. Conclusions and future work

In this paper, we addressed the safety–critical task of vehicle trajectory prediction also known as microscopic traffic modeling. We argue that neither solely knowledge-driven nor purely data-driven models can provide generalizable and accurate predictions. Given a set of experiments on a real-world dataset, our experiments demonstrate that knowledge-driven models have inferior performance but can generalize to new scenes. On the other hand, although data-driven models provide more accurate predictions, they tend to have unrealistic predictions in new scenes. Our proposed solution, named RRB, effectively merges knowledge-driven with data-driven models by finding residuals required to be added to the knowledge-driven prediction in order to model human stochastic behavior. It leverages physically-constrained Inverse-variance weighting approach to build realistic and confident residuals. To further make the predictions realistic, we employed a Model Predictive Control (MPC) to bring kinematic constraints to the final output. RRB outperforms all other counterparts in terms of accuracy and generalizability. As future work, we will study the effectiveness of RRB when employing more complex knowledge-driven methods on a variety of agents, including pedestrians. Pedestrians are not constrained to the roads but respect specific social rules. We can take the existing knowledge-driven models such as "Social force" (Helbing and Molnar, 1998) as prior predictions and learn residuals required to better capture pedestrians dynamics. Moreover, our approach can help in complex multi-agent environments where there exist vehicles, pedestrians and cyclists each one with its own constraints and dynamics. In such scenarios, RRB can benefit from using strong knowledge-driven priors for each category and provide safe and accurate data-driven residuals. We hope that our work will pave the way to more methods combining the best of knowledge and data driven approaches.

CRediT authorship contribution statement

Mohammadhossein Bahari: Conceptualization, Software, Methodology, Investigation, Validation. **Ismail Nejjar:** Software, Methodology, Investigation, Validation. **Alexandre Alahi:** Supervision, Resources, Project administration.

Acknowledgements

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., June 2017. Social lstm: Human trajectory prediction in crowded spaces. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Alahi, Alexandre, Ramanathan, Vignesh, Goel, Kratarth, Robicquet, Alexandre, Sadeghian, Amir, Fei-Fei, Li, Savarese, Silvio, 2017. In: Learning to Predict Human Behavior in Crowded Scenes. Academic Press, pp. 183–207. https://doi.org/10.1016/B978-0-12-809276-7.00011-4. ISBN 9780128092767 (Chapter 9).
An, L., Tsou, M.H., Crook, S.E., Chun, Y., Spitzberg, B., Gawron, J.M., Gupta, D.K., 2015. Space–time analysis: Concepts, quantitative methods, and future directions. Ann. Assoc. Am. Geogr. 105 (5), 891–914.
Bae, S., Saxena, D., Nakhaei, A., Choi, C., Fujimura, K., Moura, S.J., 2019. Cooperation-aware lane change maneuver in dense traffic based on model predictive control with recurrent neural network. arXiv abs/1909.05665.
Bansal, M., Krizhevsky, A., Ogale, A.S., 2019. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In: Robotics: Science and Systems XV. vol. 15.
Borghesi, A., Baldo, F., Milano, M., 2020. Improving deep learning models via constraint-based domain knowledge: a brief survey. arXiv preprint arXiv:2005.10691.
Chen, Y., Hu, C., Wang, J., 2019. Human-centered trajectory tracking control for autonomous vehicles with driver cut-in behavior prediction. IEEE Trans. Veh. Technol. 68 (9), 8461–8471.
Cochran, W.G., 1954. The combination of estimates from different experiments. Biometrics 10 (1), 101–129.
Coscia, P., Castaldo, F., Palmieri, F.A.N., Ballan, L., Alahi, A., Savarese, S., 2016. Point-based path prediction from polar histograms. In: 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, Germany, pp. 1961-1967.
Coscia, P., Castaldo, F., Palmieri, F.A., Alahi, A., Savarese, S., Ballan, L., 2018. Long-term path prediction in urban scenarios using circular distributions. Image Vis. Comput. 69, 81–91.
Cosgun, A., Ma, L., Chiu, J., Huang, J., Demir, M., Anon, A.M., Lian, T., Tafish, H., Al-Stouhi, S., 2017. Towards full automated drive in urban environments: A demonstration in gomentum station, california. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1811–1818.
Cui, H., Radosavljevic, V., Chou, F.C., Lin, T.H., Nguyen, T., Huang, T.K., Schneider, J., Djuric, N., 2019a. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE. pp. 2090–2096.
Cui, H., Nguyen, T., Chou, F.C., Lin, T.H., Schneider, J., Bradley, D., Djuric, N., 2019b. Deep kinematic models for physically realistic prediction of vehicle trajectories. arXiv preprint arXiv:1908.00219.
Deo, N., Trivedi, M.M., 2018. Convolutional social pooling for vehicle trajectory prediction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1549–15498.
Fang, Y., Kuan, K., Lin, J., Tan, C., Chandrasekhar, V., 2017. Object detection meets knowledge graphs.
França, M.V., Zaverucha, G., Garcez, A.S.d., 2014. Fast relational learning using bottom clause propositionalization with artificial neural networks. Machine Learning 94 (1), 81–104.
Ganchev, K., Gillenwater, J., Taskar, B., et al., 2010. Posterior regularization for structured latent variable models. J. Machine Learn. Res. 11 (Jul), 2001–2049.

Garcez, A.d., Gori, M., Lamb, L.C., Serafini, L., Spranger, M., Tran, S.N., 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv:1905.06088.

Gong, C., McNally, D., 2004. A methodology for automated trajectory prediction analysis. In: AIAA Guidance, Navigation, and Control Conference and Exhibit. p. 4788.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A., 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2255–2264.

Hartung, J., Knapp, G., Sinha, B.K., 2011. Statistical Meta-analysis with Applications, vol. 738. John Wiley & Sons.

Helbing, D., Molnar, P., 1998. Social force model for pedestrian dynamics. Phys. Rev. E 51. doi:10.1103/PhysRevE.51.4282.

Jeong, Y., Kim, S., Yi, K., 2020. Surround vehicle motion prediction using lstm-rnn for motion planning of autonomous vehicles at multi-lane turn intersections. IEEE Open J. Intell. Transport. Syst. 1, 2–14. https://doi.org/10.1109/OJITS.2020.2965969.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems.

Kani, J.N., Elsheikh, A.H., 2017. Dr-rnn: A deep residual recurrent neural network for model reduction. arXiv preprint arXiv:1709.00939.

Karpatne, A., Watkins, W., Read, J., Kumar, V., 2017. Physics-guided neural networks (pgnn): An application in lake temperature modeling. arXiv preprint arXiv: 1710.11431.

Kazemi, H., Mahjoub, H.N., Tahmasbi-Sarvestani, A., Fallah, Y.P., 2018. A learning-based stochastic mpc design for cooperative adaptive cruise control to handle interfering vehicles. IEEE Trans. Intell. Vehicles 3 (3), 266–275.

Keyvan-Ekbatani, M., Knoop, V.L., Daamen, W., 2016. Categorization of the lane change decision process on freeways. Transport. Res. Part C: Emerg. Technol. 69, 515–526.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kong, J., Pfeiffer, M., Schildbach, G., Borrelli, F., 2015. Kinematic and dynamic vehicle models for autonomous driving control design. In: 2015 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1094–1099.

Kong, J., Pfeiffer, M., Schildbach, G., Borrelli, F., 2015. Kinematic and dynamic vehicle models for autonomous driving control design. In: 2015 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1094–1099.

Kothari, Parth, Kreiss, Sven, Alahi, Alexandre, 2020. Human trajectory forecasting in crowds: A deep learning perspective, arXiv preprint arXiv:2007.03639.

Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B., 2015. Human-level concept learning through probabilistic program induction. Science 350 (6266), 1332–1338.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J., 2017. Building machines that learn and think like people. Behav. Brain Sci. 40.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, N., Choi, W., Vernaza, P., Choy, C., H.S. Torr, P., Chandraker, M., 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. pp. 2165–2174. doi:10.1109/CVPR.2017.233.

Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M., 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 336–345.

Li, J., Ma, H., Tomizuka, M., 2019. Conditional generative neural system for probabilistic trajectory prediction. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6150–6156.

Li, J., Ma, H., Zhang, Z., Tomizuka, M., 2020. Social-wagdat: Interaction-aware trajectory prediction via wasserstein graph double-attention network. arXiv preprint arXiv:2002.06241.

Liu, Yuejiang, Yan, Qi, Alahi, Alexandre, 2020. Social NCE: Contrastive Learning of Socially-aware Motion Representations, arXiv preprint arXiv:2012.11717.

Makansi, O., Ilg, E., Cicek, O., Brox, T., 2019. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7144–7153.

Niedoba, M., Cui, H., Luo, K., Hegde, D., Chou, F.C., Djuric, N., 2019. Improving movement prediction of traffic actors using off-road loss and bias mitigation. In: Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch (2017 NiPS Talk).

Pathak, D., Krahenbuhl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1796–1804.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D., 2017. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3591–3600.

Sadeghian, A., Legros, F., Voisin, M., Vesel, R., Alahi, A., Savarese, S., 2018. Car-net: Clairvoyant attentive recurrent network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 162–180.

Schöller, C., Aravantinos, V., Lay, F., Knoll, A., 2020. What the constant velocity model can teach us about pedestrian motion prediction. IEEE Robot. Automat. Lett. 5, 1696–1703.

Silver, T., Allen, K., Tenenbaum, J., Kaelbling, L., 2018. Residual policy learning. arXiv preprint arXiv:1812.06298.

Tang, C., Salakhutdinov, R.R., 2019. Multiple futures prediction. In: Advances in Neural Information Processing Systems. pp. 15398–15408.

Tang, J., Song, Y., Miller, H.J., Zhou, X., 2016. Estimating the most likely space–time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method. Transport. Res. Part C: Emerg. Technol. 66, 176–194.

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 62 (2), 1805.

Vemula, A., Muelling, K., Oh, J., 2018. Social attention: Modeling attention in human crowds. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1–7.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al., 2020. Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems. arXiv preprint arXiv:1903.12394.

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2020. Integrating physics-based modeling with machine learning: A survey. arXiv preprint arXiv:2003.04919.

Xie, D.F., Fang, Z.Z., Jia, B., He, Z., 2019. A data-driven lane-changing model based on deep learning. Transport. Res. Part C: Emerg. Technol. 106, 41–60.

Xu, X., Pang, J., Monterola, C., 2015. Asymmetric optimal-velocity car-following model. Physica A 436, 565–571.

Zeng, A., Song, S., Lee, J., Rodriguez, A., Funkhouser, T., 2020. Tossingbot: Learning to throw arbitrary objects with residual physics. IEEE Trans. Rob.

Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kümmerle, J., Königshof, H., Stiller, C., de La Fortelle, A., Tomizuka, M., 2019. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. arXiv:1910.03088 [cs, eess].

Zhang, X., Sun, J., Qi, X., Sun, J., 2019. Simultaneous modeling of car-following and lane-changing behaviors using deep learning. Transport. Res. Part C: Emerg. Technol. 104, 287–304.

Ziegler, J., Bender, P., Schreiber, M., Lategahn, H., Strauss, T., Stiller, C., Dang, T., Franke, U., Appenrodt, N., Keller, C.G., et al., 2014. Making bertha drive–an autonomous journey on a historic route. IEEE Intell. Transport. Syst. Mag. 6 (2), 8–20.