Emerging Cutting-Edge Developments in Intelligent Traffic and Transportation Systems
M. Shafik (Ed.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/ATDE240027

On the Way to Reliable Trajectory Prediction in Urban Traffic

Rock TERESA [a,1], Bleher THOMAS [a] and Dr. Bahram MOHAMMAD [a] and Prof. Dr. rer. nat. Marker STEFANIE [b]

[a] *BMW Group, Research and Technology, D-80788 Munich, Germany*
[b] *Technical University of Berlin, Chair of Naturalistic Driving Observation for Energetic Optimisation and Accident Avoidance, D-13355 Berlin, Germany*

Abstract. Predicting the intentions of other vehicles in traffic is a frequently addressed challenge in autonomous driving. Due to the complexity and diversity of urban traffic, it is a major challenge to develop prediction models that are able to generate reasonable predictions for a broad range of situations. Commonly employed data-driven approaches encounter problems related to the lack of transparency of black-box approaches and poor generalizability due to overfitting. Meanwhile, most of the publications to date have focused on the modeling part, but investigations that provide transparency into the transferability of learned patterns and the effect of different settings on generalizability are rarely addressed. This paper addresses these challenges by presenting an advanced evaluation method providing insight into the ability of models to create plausible predictions even in exceptional situations. The proposed method is applied to investigate variations in the provided input information, varying diversity in training data, and different model parameters. Among other things, our results show that providing semantic contextual information and enriching real training data with synthetic samples contributes to better generalizability. Furthermore, the evaluation revealed weaknesses of commonly used metrics, as the exclusive use of displacement errors can be misleading in terms of generalizability and plausibility of results. In summary, this contribution paves the way for reliable predictions in urban traffic by providing valuable insights and a methodology for a critical evaluation of prediction models.

Keywords. Trajectory Prediction, Neural Networks, Intelligent Vehicles, Urban Traffic, Human Factors, Big Data and Naturalistic Datasets

1. Introduction

Anticipating the intention of other vehicles in traffic is a common challenge in autonomous driving, as understanding and incorporating the future movements of other traffic participants is the basis for safe driving strategies and reasonable decision-making, especially in urban traffic. Compared to the well-researched highway traffic, new missions arise in urban traffic, as vehicle movements strongly depend on the situational con-

[1] Corresponding Author: Teresa Rock, is with BMW Group, Research and Technology, D-80788 Munich, Germany and with the Technical University of Berlin, Chair of Naturalistic Driving Observation for Energetic Optimisation and Accident Avoidance, D-13355 Berlin, Germany.; E-mail: Teresa.Rock@bmw.de; t.rock@campus.tu-berlin.de.

text, e.g., right-of-way regulations, reactions to other road users, or road network-specific dependencies. This results in new challenges for representing and structuring scenario information. The common objective in developing a prediction model is to generate accurate predictions across the wide range of situations encountered in traffic. The number of possibilities for addressing this challenge with common data-driven models is immense, starting from how and which situational information is provided, among a wide range of model architecture options, and finally the choice of learning parameters such as loss functions, regularizations, and optimizers. This richness of opportunities meets difficulties associated with data-driven approaches, namely overfitting and lack of transparency. On one hand, black-box models provide little transparency, so that insights into accuracy can only be gained in explicitly tested situations. In addition, data-driven models risk overfitting the training data, resulting in poor results in unknown situations, i.e., low generalizability. Furthermore, the representativeness of situations that appear in a database is always limited compared to all potentially occurring scenarios in urban traffic. Taking these facts together, one usually does not know what relationships the model has actually learned, and both the evaluation and training of such models depend strongly on the available datasets. Meanwhile, most state-of-the-art publications focus on problem-solving and introduce new concepts of how to generate accurate predictions on individual datasets, but rarely address the question of how various conceptual choices affect the generalizability of the model and rarely provide detailed evaluations. To address these challenges, this paper presents an advanced evaluation method that provides insight into the ability of models to generalize and generate plausible predictions even in exceptional situations. The multi-level evaluation method aims to provide more transparency about learned patterns and allow for more reliable and efficient model development. The evaluation method is applied to investigate the effects of differences in provided input information, varying diversity in training data, and different learning parameters for a simple exemplary prediction model. Accordingly, the following research questions are formulated:

R1: How to measure the generalizability of data-driven prediction models?

R2: How and to what extent is the generalizability of a data-driven prediction model affected by differences in the input information, variety of training data, and various learning parameters?

R3: Is it possible to combine real and synthetic traffic data samples to compensate for underrepresented situations in the future?

The rest of the paper is organized as follows. Section 3 describes the evaluation methodology and the concept of variations in training data and input features. The subsequent section describing the implementation (Section 4) contains all the necessary information about the problem, data acquisition, and data processing to obtain the described features. In addition, the applied metric is presented. All evaluation results are presented in Section 5, discussing the impact of the training data, input features, and tuning parameters on generalizability. Finally, a conclusion, future attempts, and specific limitations of the methods are presented in Section 6.

2. Related Work

2.1. Prediction Concepts

As vehicle trajectory prediction is a commonly addressed problem in autonomous driving, various concepts exist. Providing the entire range of currently published approaches would exceed the scope of this paper, therefore an overview of the main concepts used in state-of-the-art is given. One can categorize approaches by the model architecture, provided input information, and levels of behavioral discretization, i.e. the model output. Table 1 provides an overview of the named categories with corresponding references. Regarding the model architecture, as trajectory prediction is mostly formulated as a sequential problem, several approaches utilize recurrent network structures, such as Recurrent Neural Networks (RNN) or Long short-term memories (LSTM) such as proposed by Xia et al. [1]. Other promising concepts evolve Graph-neural networks (GNNs), as these structures offer great potential in representing spatial dependencies between road users, offer the possibility of handling dynamic input sizes, and are suitable to predict the entire scene development instead of predicting each road user individually as presented by Li et al. [2]. Other popular structures involve transformer networks [3] or variational autoencoder (VAE) [4]. Some approaches combine different model architectures into one prediction network, such as the *SCALE-Net* proposed by Jeon et al. combining a graph approach with LSTM and multilayer perceptron (MLP) layers [5]. The choice of architecture strongly depends on the representation of the input data (e.g. object lists or image data) and the concept of feature encoding.
Especially in urban traffic, driving behavior is affected by various influences. Consequently, there are several approaches for incorporating contextual information into the prediction. Concepts vary in terms of the information provided (e.g., static environment information of the map or neighboring road users) and the format in which this information is provided (on a semantic level [6], as raw data transformed into embeddings [7]). Finally, the model output can vary in the level of discretization. Some approaches predict the intention of other road users as maneuvers [8] or actions [9], while others directly predict a trajectory deterministically [10] or in a probabilistic manner [11].

Table 1. Overview state-of-the-art prediction models

Model architecture	CNN	[12]
	RNN, LSTM	[13,14,1,15,16,17]
	GNN, GCN	[10,8,2]
	Transformer	[18,19,20,21,22,3]
	VAE, CVAE	[4,23]
	RNN + GNN , GCN + LSTM + MLP	[24,25,5,26,27]
Input information	semantic representations for static environment	[28,2,12,23,6]
	raw representations of static environment or embeddings	[4,13,1,11,20,26,21,29,15,7]
	raw context information (position, dynamics of other road users)	[24,10,25,8,18,30,19,3]
	semantic context information (interaction partners, relationship)	[31,5,16,17]
Model output	manoeuvre prediction	[8,32]
	action prediction	[9]
	deterministic trajectory prediction	[10,25,8,18,19,26,21,16,17,3]
	probabilistic trajectory prediction	[4,13,28,1,11,30,20,5,29,12,15,23]

2.2. Evaluation Strategies

The evaluation of black-box models is a crucial part of the development, as it is the only chance to gain insight into the accuracy and reliability of the model. Table 2 shows a selection of employed state-of-the-art metrics that are used for evaluating trajectory prediction. The most common metrics for evaluating trajectory prediction are displacement errors, average displacement error (ADE), and final displacement error (FDE), measured as the L2 distance between the true and the predicted trajectory. Some approaches use variations of ADE and FDE or root mean square error (RMSE) as a metric. These metrics indicate how accurately the predicted trajectory matches the individual human-driven trajectory. However, the use of displacement errors cannot provide information on how functional or plausible the predicted trajectory was. Therefore, in some individual cases, more sophisticated evaluation strategies are applied, e.g., taking into account functional errors such as road violations [29] or unrealistic headways [8], as summarized in Table 2. Next to the applied metric, another key element of the evaluation strategy is the choice of test scenarios or test data. Most state-of-the-art approaches test their models on a retained test split of the training data. Only, a few approaches test their models on different datasets [18]. Furthermore, information about how close the used test data and training data are is rarely discussed in most publications, leading to insufficient findings regarding generalizability.

Table 2. Summary of Metrics for Trajectory Prediction

Metric	Explanation	Reference
ADE & FDE	Average Displacement Error & Final Displacement Error	[24,4,13,14,28,1,33,34,35,30,19,20,26,21,12,15,27,16,17,3,23]
variations of ADE & FDE	normalized ADE & FDE, minimum ADE & FDE	[24,1,2,18,29]
RMSE	Root Mean Square Error	[14,33,10,25,8,5,15]
Negative headway distance occurrence	Occurrence of unrealistic states due to poor decision-making	[8]
Jerk sign inversion	Quantifies oscillations in model's acceleration predictions	[8]
Miss Rate (MR)	Proportion of unacceptable trajectories measured by a region of interest.	[29,23,36]
Off-road rate	The ratio of predicted trajectories laying not entirely in the driveable area of the map to the total number of predicted trajectories	[29]
EMD distance	Quantifies amount of probability mass that has to be moved from the predicted distribution to match the true distribution.	[35]
Hard Off-road Rate (HOR)	The percentage of scenarios that have at least one off-road prediction in the trajectory points	[35]
Soft Off-road Rate (SOR)	The percentage of off-road prediction points over all prediction points and the average over all scenarios.	[35]
Drivable Area Compliance (DAC)	Count of future trajectories within the drivable area divided by the number of all possible trajectories.	[36]
TCC	Temporal Correlation Coefficient (high TCC, meaning predictions cover the time-varying motion patterns well)	[13]

3. Method

This paper addresses the question of how different conceptual choices affect the generalizability of a data-driven model for predicting trajectories. For this purpose, a multi-level evaluation method is introduced to assess the generalizability of models by providing detailed insights into the accuracy and plausibility of predictions at different levels of test data. The focus lies on examining how different information categories of features representing the situational context and the diversity of the training data contribute to generalizability. Therefore, models with different combinations of training data and input features are trained and analyzed. To gain insight into the magnitude of the effect compared to *simple* parameter tuning, one of the combination models is trained with different sets of learning parameters and evaluated using the same method.

3.1. Evaluation Methodology

As stated in Section 2.2, the two key aspects for evaluating a data-driven predictive model involve the used metrics and the choice of test data. Most publications use spatio-temporal error measurements, ignoring the situational context when evaluating a trajectory. However, using exclusively displacement errors cannot provide information on how functional or plausible the predicted trajectory was. According to our previous work [31], cases occur in which behavior deviates from the real trajectory but is still plausible, e.g., a slightly longer time gap when turning without becoming critical, while other predictions with similar error values enter non-driveable areas. Consequently, common error measurements are not able to distinguish between *false-bad* and *plausible-bad* trajectories. Meanwhile, it would be crucial for developing or tuning prediction models to identify the situations in which the model produces non-plausible or non-functional results. In order to address this scientific gap, a simple plausibility metric is formulated consisting of two categories to evaluate the plausibility of predictions adapting and further developing sophisticated metric approaches from literature [29,35,36,13]. The metric incorporates the following parts:

- spatial evaluation: maximal path deviation $max\_path\_violation$, road violation, maximal road violation $max\_road\_violation$
- temporal evaluation: collision check, minimum distance to other traffic participants $min\_dist2others$

To measure plausibility, a percentage score S_P is calculated based on these five components. Binary aspects, such as collision or road violation checks, return True or False, interpreted as 0 or 1. All distance measures are divided into bins and mapped to plausibility values between 0 and 1. The final plausibility score S_P is the average of the individual components. To further investigate spatial accuracy, the percentage of predictions with path deviations greater than 5 meters is measured. Based on the plausibility metric in conjunction with the common metrics ADE and FDE, model performance, involving accuracy and plausibility, can be evaluated against test data. Since data-driven models are based on black-box approaches, transparency in terms of model generalizability and reliability is achieved by applying a trained model to test data. Accordingly, the selection of test data is crucial for the significance of the evaluation. In this paper, a multi-level method is proposed for critical evaluation, involving four levels of test data, illustrated in Figure 1. The four levels present different challenges in terms of generalizability, as they include situations that are further apart from the training data. Starting with unknown situations at locations shown during training (L1), new locations from real traffic data (L2a), new locations from synthetic traffic data (L2b), and ending with testing in an exceptional situation (L3), in which the ego vehicle has to pass a static obstacle with oncoming traffic. For the L2a, L2b, and L3 levels, the proposed plausibility metric is applied. As the objective is to gain insight into how different settings affect model performance and which concepts contribute to improved generalizability, the following variants are investigated: Three variants of training data are trained over six variants of provided input feature settings. In addition, one of these models is trained with four different variants of learning parameter sets. In total, this results in 22 models to be evaluated.

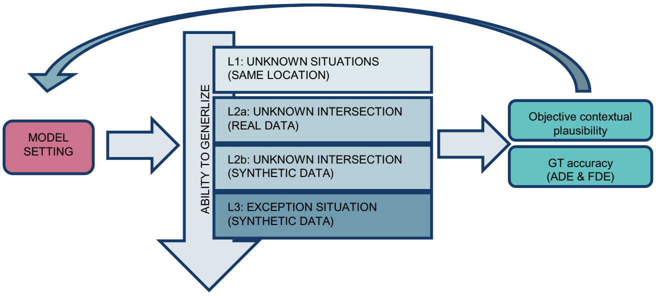
MODEL SETTING
ABILITY TO GENERLIZE
L1: UNKNOWN SITUATIONS (SAME LOCATION)
L2a: UNKNOWN INTERSECTION (REAL DATA)
L2b: UNKNOWN INTERSECTION (SYNTHETIC DATA)
L3: EXCEPTION SITUATION (SYNTHETIC DATA)
Objective contextual plausibility
GT accuracy (ADE & FDE)

Figure 1. Illustration of the evaluation concept measuring accuracy and plausibility of predictions on four different levels of test data

3.2. *Influence of Contextual Information*

Given the high complexity and variability of urban traffic, it is important to ensure that meaningful patterns are learned that can be applied to new situations. However, a model can only learn such patterns from accessible information, which creates challenges in terms of how and in which format the situational context can be represented for a model. The present paper aims at investigating which types of information contribute to better generalizability of a data-driven prediction model and to what extent model performance is affected. As shown in Table 1, there are different types of information categories that may influence driving behavior. Since temporal information is mostly generated by assembling situational information into a time series, the present approach focuses on examining the provided situational context itself.

As presented in our previous work [31], the key idea is to use prior knowledge to generate a scene representation incorporating raw and semantic information describing the context of interactive traffic situations. This should provide the basis for a data-driven model to apply transfer learning and allow for coping with the high variability of urban traffic. Four information categories were defined: Ego-information (E), Map-information (M), Partner-information (P), and Interaction-information (I), according to the specifications in Table 3. In the following, ego always refers to the vehicle being predicted. Partners are all road users potentially affecting the behavior of the ego vehicle. The scene

Table 3. Feature categories describing the driving situation.

Category	Related information occurring in features
Ego (E)	position, velocity, acceleration, heading, type, dimensions
Map (M)	turn direction, lane curvature, lane direction, center line coordinates ahead
Partner (P)	position, velocity, acceleration, heading, type, dimensions of all identified interaction partners
Interaction (I)	relationship, positioning and relative movement regarding the individual conflict zone for all identified interactions

representation includes both easily accessible information such as positions or classifications of road users and information that requires prior interpretation. Such information, obtained through interpretation, relies on heuristic recognition algorithms that attempt to identify potential interactions between road users and describe the resulting interaction at a feature level, e.g., relationships between road users. A key element of the heuristic recognition algorithms is the fusion of time series and map data with the objective of extracting additional contextual information such as right-of-way regulations or the location of conflict zones. A detailed description of these algorithms can be found in our previous work [31]. Based on this, models are trained with six different compositions of situational information: **EMPI, EMP, EMI, EI, EM, E**.

3.3. *Influence of Variety in Training Data*

The performance of data-driven prediction models strongly depends on the data presented during training. On the one hand, it is crucial that the training data represent the application domain as thoroughly as possible. On the other hand, the degree of variability can influence the learning process. Furthermore, in data-driven modeling, we often face the problem that some exceptional situations are underrepresented for adequate training. However, the creation of new data, especially in exceptional situations, is either costly or not possible due to the rarity or criticality of events in everyday traffic. As a result, it would be beneficial if it were possible to augment existing datasets with manually defined situations that are known to be underrepresented. The present paper investigates how the variability of the training data affects the model's ability to generalize. In addition, real traffic data is combined with synthetic traffic data from simulation to investigate the possibility of augmenting existing real datasets through synthetic samples from simulation. The following three levels of training data are investigated:

T1: Low variability: synthetic traffic data for training.

A simulation framework is used to create synthetic traffic data. Due to the limited possibility of individualizing a driver model, less diversity in behavior occurs.

T2: Medium variability: real traffic data for training.

For real traffic data an open-source dataset covering typical interactive urban situations is selected.

T3: High variability: a combination of real and synthetic traffic data.

4. Implementation

4.1. *Problem Formulation and Model Architecture*

The problem of trajectory prediction is formulated as follows. At time t the model predicts the future trajectory $Y_t^i = Y_{t+1}^i, Y_{t+2}^i, ..., Y_{t+T_{pred}}^i$ for the next T_{pred} seconds for one vehicle i based on the current scene X_i^t from the individual perspective of vehicle i. The future motion Y_t^i is a sequence of positions in a two-dimensional space: $Y_t^i = (x_t^i, y_t^i)$. The prediction horizon T_{pred} was set to 5 seconds. For labeling, the position of the respective vehicle after 1,2,3,4,5 second in global XY coordinates is used. The current situation from an ego perspective of vehicle i is represented as a concatenated feature vector X_i^t at time t. The vector X_i^t includes features describing the ego vehicle $X_{i_{ego}}^t$, and, depending on the setting to investigate, all features describing the map, potential conflict partners with their state, and the individual interaction with the ego:

$$X_i^t = (X_{i_{ego}}^t, X_{i_P}^t, X_{i_I}^t, X_{i_M}^t) \tag{1}$$

Since this work aims at investigating different levels of provided situational information, the training samples do not contain any past information, and consequently, no recurrent structures were used in the model architecture. For prediction, an MLP with four hidden layers consisting of 512, 256, 128, and 64 neurons and 10 neurons representing the output layer for the respective positions in XY for the next 5 seconds is created.

4.2. *Data for Training and Testing*

For training and evaluating the open-source drone dataset inD [2] [37] is used for representing real traffic. The dataset shows all key characteristics of interactive urban traffic (e.g. shared spaces, non-deterministic regulations, and interactions with Vulnerable Road Users (VRUs).) The dataset includes recordings of four German unsignalized intersections called Aseag, Bendplatz, Frankenburg, and Heckstrasse displayed in Figure 2. For the medium level of variability in training data (T2), models are trained only on real data, and recordings from Aseag, Bendplatz, and Frankenburg were selected for training, resulting in 900.000 training samples. For the evaluation at level L1 (unknown situations), one recording from each location was retained for testing and one as a validation set. For representing a low level of variety in training data (T1), data on four synthetic intersections was created with the help of the simulation framework *Spider* at BMW [38]. The choice of intersections intends to represent similar intersections compared to the ones represented in the inD dataset involving different complex intersections and merging topologies. Since all drivers in the simulation are based on the same heuristic agent model, but using different parameters, overall behavior shows less variety. For the low level of variability in training data (T1), models are trained on intersections 1, 2, and 3. Randomly selected vehicle IDs were chosen and retained for the validation set during training and for testing on L1 (unseen situation). It has to be noted that the synthetic data only contains vehicles and no VRUs. The synthetic data is recorded at the same sampling rate and shows the same characteristics as the inD dataset. The high degree of diversity (T3) in the training data is achieved by combining synthetic and real data. For this purpose, the real traffic data of Bendplatz and Aseag are combined with the synthetic data of intersections 1 and 3. The data is combined in such a way that synthetic and real data have a distribution of 50:50. For comparison, all training sets were set to have a similar number of samples. For evaluating at the L2 level, data from one real (L2a) and one synthetic (L2b) location were used. For L2a, the models were tested on recording 30 of the inD dataset, which represents a new intersection from reality (Heckstrasse). For L2b, data from a different four-armed intersection was created (isec 4). To test model performance in an exceptional situation (L3), the simulation framework was used to create a special scenario in which the path of the ego is occupied by an obstacle on a two-lane road with oncoming traffic. For data collection on L3, the vehicle is controlled by a real human in simulation. Pictures of all test and training locations are shown in Figure 2.

[2] https://ind-dataset.com/

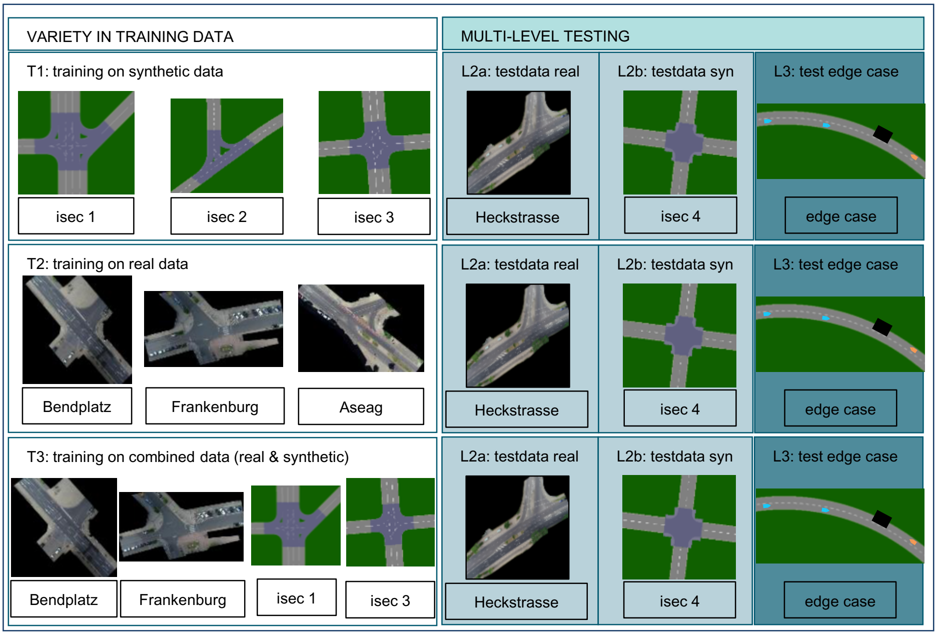

Figure 2. Overview of all training and test locations

4.3. Data Processing for Feature Calculation

Following our previous published work, time-series and map data are fused, in order to first identify potential interaction partners, and subsequently calculate interaction features to describe complex situations on a semantic level [31]. The algorithm is initialized with a maximal number of interaction partners, namely five vehicles and four VRUs. If fewer partners occur in the scene, features are represented as -1 aiming at training the model to ignore those since -1 is out of the normal feature range of $[0-1]$. When more partners occur only the closest ones are considered. The static environment of the ego vehicle is represented by semantic and raw features describing the lane the vehicle is following by turn direction, curvature, and lane center 5m, 10m, and 15m ahead. The different feature spaces show the following dimensions: EMPI: 209 features, EMP: 144 features, EMI: 96 features, EI: 78 features, EM: 31 features, and E: 13 features.

4.4. Training and Model Parameters

For training of all models regarding different feature settings and varying training data, the Adam optimizer with a default learning rate of 0.001 was employed, mean squared error (MSE) as loss function, and *relu* for activation by using Keras for model building [39]. In order to investigate the influence of parameter tuning relative to changes in features and training data, some variations were investigated, namely the choice of the loss function, activation function, optimizer, and batch normalization shown in Table 4. All models were trained with a batch size of 50 for maximal 80 epochs using early stopping with a minimum delta of 0.00001 and patience of 15 epochs.

Table 4. Tuning parameter-sets [39]

Setting ID	Batch Norm.	Loss Function	Optimizer	Activation
1	True	Mean Absolute Error	*sgd*	*sigmoid*
2	True	Mean Squared Error	*sgd*	*sigmoid*
3	True	Mean Squared Error	*adam*	*relu*
4	False	Mean Absolute Error	*adam*	*relu*

4.5. Metric Calculation

The evaluation method presented earlier returns the following measures for each model:

- Overall score S_O, overall accuracy score S_{ACC}, overall plausibility score S_P, overall ADE & FDE calculated across all test levels
- ADE & FDE individually on test data of L1, L2a, L2b, and L3
- S_P individually on test data L2a, L2b, and L3

In order to evaluate the accuracy of the proposed models, ADE and FDE for accuracy are calculated using L2 distance according to the general state-of-the-art [40]. Model performance is measured as a combined measure considering accuracy and plausibility, resulting in S_O. For model accuracy evaluation S_{ACC}, ADE and FDE are converted to a score under consideration of benchmark results according to Equation (2), where $ADE_B = 2m$ and $FDE_B = 5m$ [41].

$$S_{ACC} = (\frac{ADE_B}{ADE} + \frac{FDE_B}{FDE}) \cdot \frac{1}{2}) \cdot 100 \tag{2}$$

For calculation of the overall accuracy S_{ACC}, all displacement errors are combined, while the displacement errors of L2a, L2b, and L3 are weighted double to assign a higher priority to results on data further away from training. The score is calculated according to Equation (3). Total FDE is calculated accordingly.

$$\overline{ADE} = \frac{ADE_{L1} + 2 \cdot (ADE_{L2a} + ADE_{L2b} + ADE_{L3})}{7} \tag{3}$$

The score for the final model performance S_O is the calculated mean of plausibility S_P and accuracy score S_{ACC}.

5. Results

All results for all model variants are provided in Tables 5 and 6, whereby Table 5 shows the results for all different feature settings and levels of variability in training data. While Table 6 provides the evaluation of different learning parameters for models trained on the full feature space EMPI on the real dataset (T2).

5.1. Influence of Variability in Training Data and Provided Situational Information

The results show a clear benefit of more variability in training, as models trained on T3 provide the best results for S_O, S_P, and S_{ACC} across all test levels, illustrated in Figure 3 (right). In terms of plausibility and accuracy at different test levels, T3 either outperforms the other data variation settings or shows similarly accurate results. Models trained on

T1 show significantly higher error values when testing on unseen real data (L2a) and in the edge case (L3), illustrated in Figure 3 (left) and Table 5. Meanwhile, the low level of variability in the behavior data of T1 is beneficial when predicting *clean* synthetic behavior on test level L2b compared to the models trained on real data (T2). However, even on the *clean* synthetic behavior data of L2b, the models trained on T3 outperform models trained on T1. This indicates potential in combining real and synthetic data for handling underrepresented situations in the future. When considering the overall plausibility scores S_P of all models, a model trained on T3 shows the best overall plausibility score of 73%, followed by a model trained on T2 with 72%, while models trained on T1 only reach a maximum of 66% shown in Table 5. Regarding variations in the input information provided, results show that contextual features provide a clear benefit in terms of generalizability when training on T3. The best-performing model includes all contextual features (EMPI), as shown in Figure 4 and Table 5. Models trained on T1 and T3, show the overall best plausibility on the feature setting EMI but the differences in S_O and S_{ACC} when comparing the feature settings, do not show a clear tendency. When considering the ability of models to generate reasonable predictions in exceptional situations (L3), the feature setting EMI clearly outperforms the others when training on T1 or T2, while models trained on T3 show the best results when all features are included (EMPI). The models trained only on synthetic data (T1) show the poorest results overall. Considering different feature settings, no clear tendency could be found. Overall plausibility S_P, shows the best results on feature setting EMI, and overall accuracy S_{ACC} is best on feature setting E. But when it comes to the edge case scenario (L3), one can see a clear advantage of including context features during learning (up to 20% more accuracy and plausibility). The fact that models trained on synthetic data show less benefit from the inclusion of contextual features can be explained by the driver model used to create the synthetic data. The driver models are not able to interact and rarely respond to the behavior of others but follow predefined heuristic rules. Consequently, driver behavior in this dataset is less context-dependent compared to real traffic data. In addition, the interaction and partner feature spaces contain features for VRUs that are not present in the synthetic data. The *empty* features might hinder the training process. In general, a high dependency between the training data and the role of provided input information can be observed.

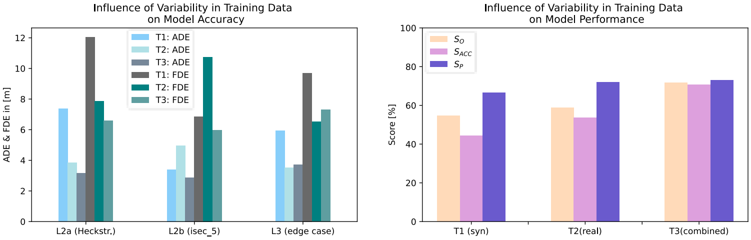

Figure 3. Influence of Variability in Training Data on Model Performance. Left: Accuracy measured by ADE and FDE on different test-levels with best feature setting of training data category. Right: Scores for accuracy, plausibility, and overall for different training data

5.2. Influence of Individual Feature Categories

The effect of map and interaction features on spatial, and temporal performance is analyzed to gain further insight into the impact of individual feature categories. In Figure 5 (left), it can be observed that interaction and partner features contribute on average to better temporal plausibility of the results, measured by the frequency of collisions. In addition, the best feature settings with and without map features at all training levels show an advantage of including map features regarding spatial plausibility of predictions, measured by the frequency of road violations and the percentage of path deviations over 5 m. However, the positive effect of including map features on better spatial perspective results is smaller than expected. Considering the individual values presented in Table 5 it can be observed that the spatial plausibility, measured on synthetic test data, partly shows better values without map features. This aspect should be investigated further. Semantic features are associated with a higher computational effort in data pre-processing. Therefore, the influence of interaction features representing situational context is analyzed in more detail. When testing on L3, a clear advantage of semantic interaction features can be observed. The same tendency, but with a smaller effect, is observed when testing on real (L2a) and synthetic data (L2b). In general, the inclusion of partner and interaction features contributes to better temporal accuracy in prediction as shown in Figure 5 (right). Again, a strong dependency between training data and the utility of each feature category can be observed. In addition, a lower benefit of contextual features is observed for models trained on synthetic data (T1), which can again be explained by a lower context dependence of behavior due to heuristic model strategies for artificial drivers.

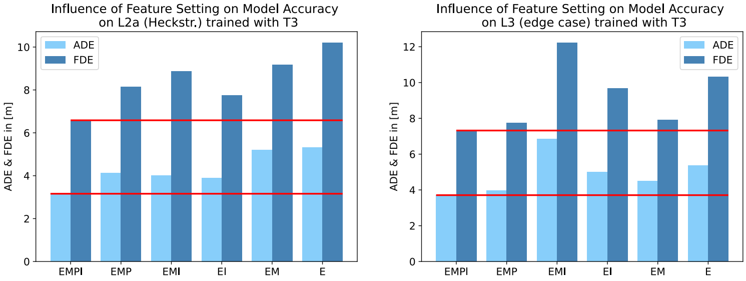

Figure 4. Influence of Different Feature Settings on Model Accuracy

5.3. Impact of Tuning Parameters

The results of the exemplary parameter tuning variants are provided in Table 6 and show similar effect sizes on accuracy and plausibility as differences in the provided input information, varying in a range of ±10%. Looking at the effects of variability in the training data, one can observe a much larger effect up to ±25%. Results show that the parameter tuning has a large impact on the generalizability of the model since variant ID 4, for example, shows the best results on test level L1 while providing weak generalizability. Meanwhile, variant ID 2 provides the lowest accuracy on test level L1 but outperforms variant ID 4 on all other test levels.

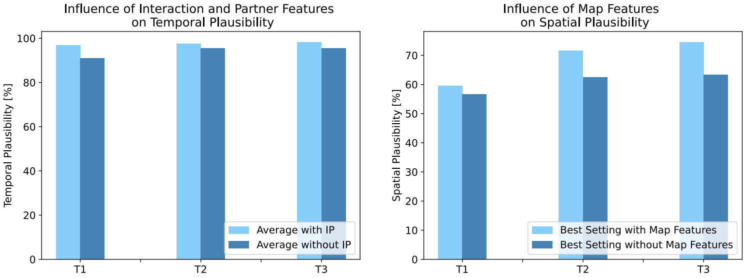

Figure 5. Influence of Interaction features on Temporal Plausibility (left) and Influence of Map Features on Spatial Plausibility (right)

5.4. Measuring Generalizability and Plausibility

Generalizability is assumed to be measurable by testing model performance on different traffic situations and scenarios in varying distances (test-level L1 - L3) from situations shown during training [42]. For the various settings shown in Table 5 and Table 6, the model showing the best performance on test level L1 is usually not the model providing the best performance on L2 or L3, emphasizing the necessity of a critical evaluation strategy. In particular, for the parameter settings, the model with the best results on data close to the training (L1) showed the weakest plausibility and accuracy overall, indicating poor generalizability. Consequently, an evaluation strategy that considers a wide range of test data is crucial even in an early stage of development. When considering plausibility in relation to accuracy, the results with the best accuracy do not necessarily show the best plausibility. In particular, when testing with real data (L2a), one can see large differences between accuracy and plausibility. Therefore, in Figure 6, two situations with three different model variants (EMPI, EMI, EM) trained on T3 are illustrated to provide some qualitative results. Considering ADE and FDE, the feature setting EMI and EM show similar poor accuracy. However, when measuring the plausibility score S_P or on a subjective qualitative basis, the EMI model shows way more reasonable learned patterns. This indicates that the plausibility metric is a more appropriate method to identify situations in which a model shows weak performance.

6. Conclusion, Limitations and Future Work

This method presents a multi-level evaluation method providing detailed insights into the generalizability of data-driven trajectory prediction models addressing research question R1. Testing at different levels highlights the criticality of selected test data with respect to the validity and significance of evaluation results. Since not only the accuracy but also the plausibility of results is considered, the proposed methodology allows the identification of samples showing inconsistent predictions. Such insights are crucial during the development process to develop reliable solutions. Two phenomena were observed: firstly, the plausibility of results does not necessarily correlate with accuracy; secondly, the best-performing setting on test data that is close to the training data is not necessarily the best setting in terms of generalization. Taking those facts together, a multi-dimensional evaluation involving a broad range of test data is crucial for determining the best model setting

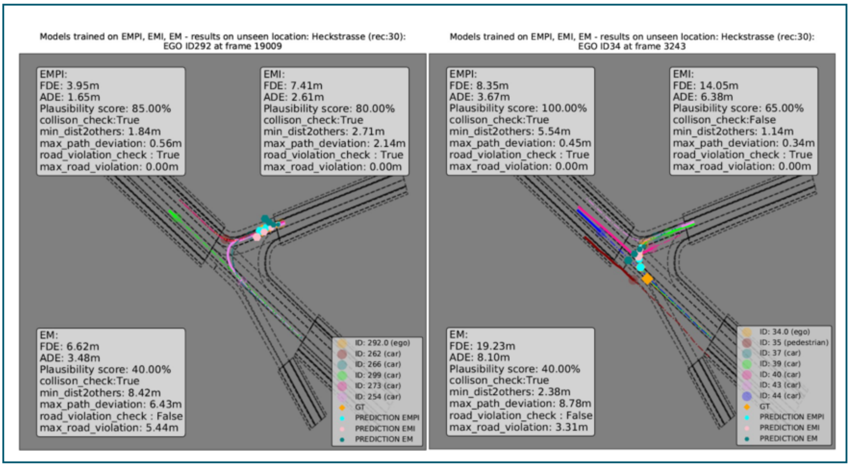

Figure 6. Qualitative Evaluation of Plausibility vs. Accuracy on L2a (real data: Heckstrasse) from models trained on T3

and should be considered in early stages of development. Considering research questions R2 and R3, the evaluation showed a large impact of the variability in training data on model performance and at the same time the potential to augment existing real datasets with synthetic samples. This is a highly valuable insight as it demonstrates the possibility of using simulation to create specific situations to compensate for those that are under-represented in the real training data. Of course, the extent to which human-like behavior can be generated by simulation in such situations depends strongly on the quality of the driver models in use. When investigating individual feature settings, advantages of providing features describing the situational context were identified. Next to interesting insights regarding the effect of training data, input information, and learning parameters on generalizability, the evaluation has shown that such model aspects can not be considered independently. There are strong inter-dependencies between data, model structure, and learning parameters, which make it challenging to derive general valid conclusions. This fact highlights the necessity of a complex and critical evaluation method to provide more transparency and reliable solutions when using black-box models. However, as the complexity of the evaluation method increases, so does the interpretation of results. Therefore, scores have been introduced to allow for easy assessment and comparison. However, these scores combine and average the individual results, which can lead to smoothing effects. A *simple* model approach for prediction is employed, dispensing on the consideration of temporal context or probabilistic outputs. However, such aspects are commonly addressed in state-of-the-art approaches and should be combined with the proposed methods in the future. Since the L3 test data contains only a small number of samples (600) to exemplify what such a level of testing might look like, more exceptional situations should be designed and included in testing at L3 to provide extensive insights for evaluating model performance. Furthermore, the plausibility metric employed is based on simple functional indicators. A more sophisticated plausibility metric is planned for the future, which will include additional parameters to investigate human similarity and situational plausibility of the results by considering parameters such as Post Encroachment Time (PET) and dynamic motion values such as ranges of accelerations driven by humans in similar situations.

References

[1] Xia B, Wong C, Peng Q, Yuan W, You X. CSCNet: Contextual semantic consistency network for trajectory prediction in crowded spaces. Pattern Recognition. 2022;126:108552.

[2] Li J, Yang F, Tomizuka M, Choi C. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. Advances in neural information processing systems. 2020;33:19783-94.

[3] Zou B, Li W, Hou X, Tang L, Yuan Q. A Framework for Trajectory Prediction of Preceding Target Vehicles in Urban Scenario Using Multi-Sensor Fusion. Sensors. 2022;22(13).

[4] Cheng H, Liao W, Yang MY, Sester M, Rosenhahn B. MCENET: Multi-Context Encoder Network for Homogeneous Agent Trajectory Prediction in Mixed Traffic. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC); 2020. p. 1-8.

[5] Jeon H, Choi J, Kum D. SCALE-Net: Scalable Vehicle Trajectory Prediction Network under Random Number of Interacting Vehicles via Edge-enhanced Graph Convolutional Neural Network. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2020. p. 2095-102.

[6] Schulz J, Hubmann C, Morin N, Lochner J, Burschka D. Learning Interaction-Aware Probabilistic Driver Behavior Models from Urban Scenarios. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2019. p. 1326-33.

[7] Mo X, Xing Y. ReCoG: A Deep Learning Framework with Heterogeneous Graph for Interaction-Aware Trajectory Prediction. ArXiv. 2020.

[8] Su J, Beling PA, Guo R, Han K. Graph convolution networks for probabilistic modeling of driving acceleration. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2020. p. 1-8.

[9] Kuefler A, Morton J, Wheeler T, Kochenderfer M. Imitating driver behavior with generative adversarial networks. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2017. p. 204-11.

[10] Li X, Ying X, Chuah MC. Grip: Graph-based interaction-aware trajectory prediction. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE; 2019. p. 3960-6.

[11] Liang J, Jiang L, Hauptmann A Alexander" editor="Vedaldi, Bischof H, Brox T, Frahm JM. SimAug: Learning Robust Representations from Simulation for Trajectory Prediction. In: Computer Vision – ECCV 2020. Cham: Springer International Publishing; 2020. p. 275-92.

[12] Su Z, Wang C, Bradley D, Vallespi-Gonzalez C, Wellington C, Djuric N. Convolutions for Spatial Interaction Modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 6573-82.

[13] Tao C, Jiang Q, Duan L, Luo P. Dynamic and static context-aware lstm for multi-agent motion prediction. In: European Conference on Computer Vision. Springer; 2020. p. 547-63.

[14] Wang Y, Chen S. Multi-Agent Trajectory Prediction With Spatio-Temporal Sequence Fusion. IEEE Transactions on Multimedia. 2023;25:13-23.

[15] Zhao T, Xu Y, Monfort M, Choi W, Baker C, Zhao Y, et al. Multi-agent tensor fusion for contextual trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 12126-34.

[16] Ma Y, Zhu X, Zhang S, Yang R, Wang W, Manocha D. TrafficPredict: Trajectory Prediction for Heterogeneous Traffic-Agents. ArXiv. 2018;abs/1811.02146.

[17] Chandra R, Bhattacharya U, Bera A, Manocha D. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 8483-92.

[18] Quintanar A, Fernández-Llorca D, Parra I, Izquierdo R, Sotelo M. Predicting vehicles trajectories in urban scenarios with transformer networks and augmented information. In: 2021 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2021. p. 1051-6.

[19] Yu C, Ma X, Ren J, Zhao H, Yi S. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer; 2020. p. 507-23.

[20] Li J, Ma H, Zhang Z, Tomizuka M. Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network. ArXiv. 2020;abs/2002.06241.

[21] Mo X, Huang Z, Xing Y, Lv C. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. IEEE Transactions on Intelligent Transportation Systems. 2022;23(7):9554-67.

[22] Guo S, Lin Y, Feng N, Song C, Wan H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence.

2019 jul;33:922-9.

[23] Lee N, Choi W, Vernaza P, Choy CB, Torr PH, Chandraker M. Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 336-45.

[24] Rossi L, Paolanti M, Pierdicca R, Frontoni E. Human trajectory prediction and generation using LSTM models and GANs. Pattern recognition. 2021 dec;120:108136.

[25] Mo X, Xing Y, Lv C. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE; 2021. p. 1934-9.

[26] Mo X, Xing Y, Lv C. Recog: A deep learning framework with heterogeneous graph for interaction-aware trajectory prediction. arXiv preprint arXiv:201205032. 2020.

[27] Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:2255-64.

[28] Syed A, Morris BT. Semantic scene upgrades for trajectory prediction. Machine vision and applications. 2023;34(2):23.

[29] Schäfer M, Zhao K, Bühren M, Kummert A. Context-Aware Scene Prediction Network (CASPNet). In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE Press; 2022. p. 3970–3977.

[30] Zhao H, Gao J, Lan T, Sun C, Sapp B, Varadarajan B, et al. Tnt: Target-driven trajectory prediction. In: Conference on Robot Learning. PMLR; 2021. p. 895-904.

[31] Rock T, Marker S, Bleher T, Bahram M. Data-Driven Prediction of Other Road Users' Intention for Better Scene Understanding in Traffic Agents. In: Kemeny A, Chardonnet JR, Colombet F, editors. Proceedings of the Driving Simulation Conference 2022 Europe VR. Strasbourg, France: Driving Simulation Association; 2022. p. 9-16.

[32] Deo N, Rangesh A, Trivedi MM. How would surrounding vehicles move? A unified framework for maneuver classification and motion prediction. IEEE Transactions on Intelligent Vehicles. 2018;3.

[33] Sighencea BI, Stanciu RI, Căleanu CD. A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction. Sensors. 2021;21(22).

[34] Gomes I, Wolf D. A Review on Intention-aware and Interaction-aware Trajectory Prediction for Autonomous Vehicles. TechRxiv. 2022 mar.

[35] Bahari M, Saadatnejad S, Rahimi A, Shaverdikondori M, Shahidzadeh AH, Moosavi-Dezfooli SM, et al. Vehicle trajectory prediction works, but not everywhere. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. p. 17102-12.

[36] Liu J, Mao X, Fang Y, Zhu D, Meng MQA Survey on Deep-Learning Approaches for Vehicle Trajectory Prediction in Autonomous Driving. 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO). 2021:978-85.

[37] Bock J, Krajewski R, Moers T, Runde S, Vater L, Eckstein L. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In: 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2020. p. 1929-34.

[38] Strobl MH. Spider-das innovative software-framework der bmw fahrsimulation/spider-the innovative software framework of the bmw driving simulation. 1745; 2003. .

[39] Chollet F, et al.. Keras; 2015. https://keras.io.

[40] Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A. Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 2255-64.

[41] Bahari M, Nejjar I, Alahi A. Injecting knowledge in data-driven vehicle trajectory predictors. Transportation Research Part C: Emerging Technologies. 2021;128.

[42] Lu J, Zhan W, Tomizuka M, Hu Y. Generalizability analysis of graph-based trajectory predictor with vectorized representation. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2022. p. 13430-7.

7. APPENDIX

Table 5. Results for all model variants across different training data and varying feature settings on all test-levels. *First five columns provide results summarized across all test levels. The rest provides results for the individual test levels. Best results within the training dataset in bolt, best result per column bolt and underlined.*

		S_O [%]	S_{ACC} [%]	S_P [%]	$\overline{FDE}$ [m]	$\overline{ADE}$ [m]	L1 FDE [m]	L1 ADE [m]	L2a FDE [m]	L2a ADE [m]	L2a S_P [%]	L2b FDE [m]	L2b ADE [m]	L2a S_P [%]	L3 FDE [m]	L3 ADE [m]	L3 S_P [%]
T1	EMPI	49.34	40.46	58.23	9.93	6.55	2.38	1.34	12.26	8.27	51.79	7.87	4.40	72.92	13.43	9.58	49.97
	EMP	50.31	40.87	59.74	9.96	6.34	**2.10**	**1.20**	13.41	8.42	58.23	7.82	4.04	73.25	12.57	9.13	47.75
	EMI	**54.57**	42.56	**66.57**	9.72	5.94	2.39	1.38	14.59	9.70	56.57	8.52	4.47	71.22	**9.70**	**5.93**	**71.92**
	EI	50.56	34.62	66.50	11.73	7.51	2.40	1.26	20.88	16.31	**60.71**	**6.86**	**3.40**	78.27	12.12	5.95	60.52
	EM	43.96	34.63	53.30	11.93	7.31	2.19	1.26	13.60	9.48	47.58	9.86	5.83	64.50	17.22	9.65	47.81
	E	54.02	**44.42**	63.61	**9.55**	**5.48**	2.60	1.45	**12.04**	**7.36**	57.41	7.05	3.47	**81.72**	13.03	7.63	51.70
T2	EMPI	55.84	48.93	62.75	8.87	4.82	4.72	2.88	**7.87**	**3.85**	72.47	11.05	5.92	61.54	9.75	5.68	54.26
	EMP	**58.72**	**53.68**	63.75	**8.62**	**4.05**	**3.90**	**1.95**	8.79	4.17	72.42	**10.75**	**4.97**	**64.10**	8.70	4.06	54.73
	EMI	57.65	43.37	**71.93**	10.60	5.05	3.97	2.09	9.70	4.55	70.52	18.91	8.58	62.08	**6.52**	**3.52**	**83.19**
	EI	53.66	47.74	59.59	9.56	4.64	4.60	2.20	10.84	4.77	61.75	11.02	5.68	57.33	9.28	4.67	59.70
	EM	55.96	45.58	66.34	8.83	5.80	3.94	2.20	9.17	4.91	67.28	12.73	9.78	59.39	7.02	4.50	72.34
	E	58.10	47.42	68.78	8.64	5.41	4.51	2.26	9.10	4.30	**73.12**	11.37	9.88	60.88	7.51	3.63	72.35
T3	EMPI	**71.83**	**70.63**	**73.03**	**6.50**	**3.11**	**2.70**	**1.33**	**6.58**	**3.16**	67.40	7.50	3.35	74.49	**7.32**	**3.71**	**77.21**
	EMP	66.24	62.11	70.37	7.24	3.62	3.12	1.73	8.13	4.12	65.20	7.91	3.72	69.29	7.74	3.98	76.62
	EMI	61.79	52.81	70.77	8.48	4.29	2.81	1.54	8.87	4.01	66.85	7.18	3.38	75.72	12.23	6.84	69.73
	EI	66.75	63.05	70.46	7.08	3.60	2.85	1.65	7.74	3.90	63.73	**5.97**	**2.88**	**82.70**	9.67	5.00	64.96
	EM	61.27	52.98	69.56	8.13	4.50	2.91	1.64	9.18	5.20	**68.96**	9.94	5.21	70.21	7.90	4.50	69.51
	E	60.79	54.56	67.02	8.19	4.16	2.74	1.62	10.21	5.33	59.00	6.76	3.05	81.74	10.31	5.38	60.32

Table 6. Results for different learning parameters on all test-levels according to Table 4

ID	S_O [%]	S_{ACC} [%]	S_P [%]	$\overline{FDE}$ [m]	$\overline{ADE}$ [m]	L1 FDE [m]	L1 ADE [m]	L2a FDE [m]	L2a ADE [m]	L2a S_P [%]	L2b FDE [m]	L2b ADE [m]	L2a S_P [%]	L3 FDE [m]	L3 ADE [m]	L3 S_P [%]
1	**55.87**	**45.42**	66.31	9.26	5.43	5.22	2.54	**7.07**	**3.65**	**74.27**	11.38	5.70	62.76	11.35	8.39	**61.91**
2	53.67	44.76	62.58	9.27	5.62	5.73	2.94	7.96	4.41	64.55	**9.68**	**5.20**	62.67	11.94	8.59	60.51
3	55.84	48.93	62.75	**8.87**	**4.82**	4.72	2.88	7.87	3.85	72.47	11.05	5.92	61.54	**9.75**	**5.68**	54.26
4	45.92	35.60	56.25	11.64	7.08	**3.91**	**1.89**	12.45	7.75	59.06	15.89	8.27	57.90	10.44	7.83	51.79