Predicting Vehicles Trajectories in Urban Scenarios with Transformer Networks and Augmented Information

A. Quintanar[1], D. Fernández-Llorca[1,2], I. Parra[1], R. Izquierdo[1] and M. A. Sotelo[1]

***Abstract*— Understanding the behavior of road users is of vital importance for the development of trajectory prediction systems. In this context, the latest advances have focused on recurrent structures, establishing the social interaction between the agents involved in the scene. More recently, simpler structures have also been introduced for predicting pedestrian trajectories, based on Transformer Networks, and using positional information [1]. They allow the individual modelling of each agent's trajectory separately without any complex interaction terms. Our model exploits these simple structures by adding augmented data (position and heading), and adapting their use to the problem of vehicle trajectory prediction in urban scenarios in prediction horizons up to 5 seconds. In addition, a cross-performance analysis is performed between different types of scenarios, including highways, intersections and roundabouts, using recent datasets (inD, rounD, highD and INTERACTION). Our model achieves state-of-the-art results and proves to be flexible and adaptable to different types of urban contexts.**

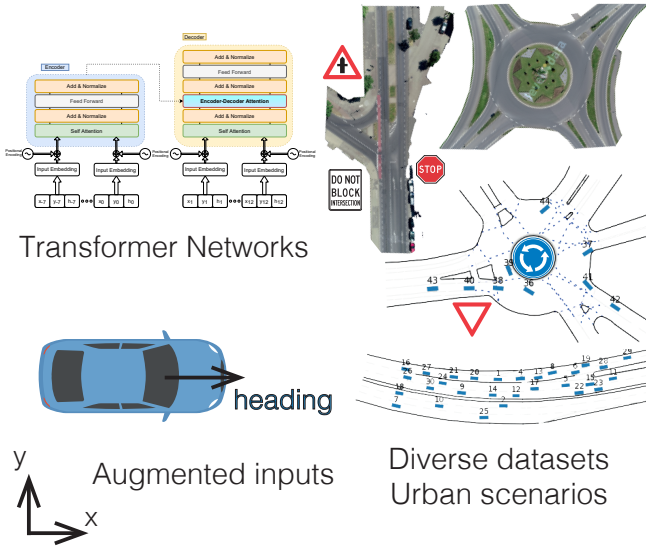

Fig. 1. System overview.

I. Introduction

Predicting road users trajectories is essential for autonomous driving. It enables path planning taking into account future states of dynamic agents, resulting in safer and more comfortable driving. It is reasonable to think that agents are affected in their behavior by traffic conditions and road structure, so any potential solution must be flexible enough to be applicable to various scene contexts. In addition, although recent approaches model the behaviors of multiple agent types within a single model (vehicles, cyclists and pedestrians) [2], [3], having specific models for each agent type simplifies the problem, and facilitates the use of simple and effective architectures, such as Transformer (TF) networks. These have been proposed for Natural Language Processing (NLP) to deal with word sequences, using attention instead of sequential processing [4].

TF networks have been recently applied to predict pedestrians trajectories [1], by using positional information. These are considered as "simple" models because each agent is modelled separately without considering complex interactions such as social recurrent networks [5] and graph neural networks [6] approaches. In this paper, we explore, for the first time, the applicability of TF networks to predict vehicles trajectories in multiple scenarios. We study the effect of augmenting the positional information with additional variables (i.e., velocity and orientation) for the context of vehicles. We evaluate the proposed model using four recent datasets (highD [7], inD [8], rounD [9] and INTERACTION [10]) which include different scenarios of different complexity (intersections, roundabouts, and straight road segments). In this way, we validate and assess the flexibility and learning transferability of the TF networks when dealing with vehicle trajectory prediction in multiple urban scenarios.

II. Related Work

In the early stages of trajectory prediction, classical approaches relied essentially on linear regression, Bayesian filtering or Markov decision process. These methods performed properly, but since they are based on a explicit physical model of the agents' behavior, their scaling and generalization capabilities are limited. Data-driven approaches have become more predominant to address these limitations. Recently, Deep Learning based methods have emerged for vehicle maneuvers [11], [12] and trajectories [13] prediction. More specifically, Recurrent Neural Networks (RNNs), such as GRUs and LSTMs, have been widely used in the field. In order to account for interactions, these approaches were adapted by including a social pooling layer (Social-LSTM) for pedestrians [14] and also for vehicles [5]. In order to overcome some limitations of the social pooling layer, we can find approaches based on occupancy grids [15], (Scene-LSTM) [16], message passing (SR-LSTM) [17], Generative Adversarial Networks (Social GAN) [18], (SoPhie) [19] and multi-agent tensors [20].

Another interesting approach to model spatial interactions for trajectory forecast is through Graph Convolutional (GNN)

[1]Computer Engineering Department, Universidad de Alcalá, Alcalá de Henares, Spain. alvaro.quintanar@uah.es
[2]European Commission, Joint Research Center, Seville, Spain.

or Graph Attention (GAT) Networks. They use a graph to represent each agent (nodes) and their interactions (edges), and update each node state and implement a weighted message passing mechanism by using convolutional or feed-forward layers, or attention mechanisms. They have been applied for modeling traffic participant interactions [6]. In order to integrate temporal information, graph representations are usually combined with recurrent-based ensembles such as Social-BiGAT [21], Social-STGCNN [22], GRIP++ [23], or adapted to allow learning temporal patterns (ST-GCN) [24]. Recently, the STAR model [25] proposed to combine GATs to model spatial interactions, with Transformers to model temporal interactions. Finally, we can find recent proposals based on the combination of some encoder-decoder architecture with Conditional Variational Auto-Encoders (CVAE) such as AMENet [2] or DCENet [3]. CVAEs are used to encode spatial-temporal information into a latent space. Future trajectories of the agents are then predicted by repeatedly sampling from the learned latent space. Most of the aforementioned approaches focused on pedestrian trajectories.

This paper is mainly inspired by [1] which adapted Transformer Networks (TF) to predict pedestrian trajectories in crowded spaces. They achieved state-of-the-art results in TrajNet benchmark [26], by relying only on self positional information without explicitly modeling interactions. TF models overcome the limitations of RNN-based models which suffer when modeling data in long temporal sequences, or in cases in which there is a lack of input data in observations (very common in real systems involving physical sensors), being more parallelizable and requiring significantly less time to train. Moreover, its main weakness, i.e. the absence of explicit modeling of spatial interactions (which is explicitly addressed by graphical-based approaches), also represents its main strength, i.e. the simplicity of the model, which also facilitates explainability. Spatial interactions and context can be easily incorporated into the input embedding without increasing the model complexity.

To the best of our knowledge, this is the first attempt to use TF models in the specific context of vehicle trajectory prediction. We evaluate the effect on performance of adding the heading to positional information, as well as the effect of the sampling frequency. Another important contribution is the evaluation of the system on four different datasets, which include different types of urban environments such as roundabouts, different types of intersections, straight road sections, etc. Different cross experiments are performed to validate the flexibility and generalization capability of this approach.

III. Methodology

This section describes the methodology employed to deploy the model: defining input and output data, pre-processing BEV datasets and modifying a Vanilla-Transformer to introduce new inputs and process them adequately.

A. Addressing the problem

As stated before, to predict a trajectory the objective is to forecast future positions of agent i by observing its current and previous positions, being defined an observation window (seen sequence) and a prediction horizon. The objective is to provide predictions about the position of the agent at the future κ steps.

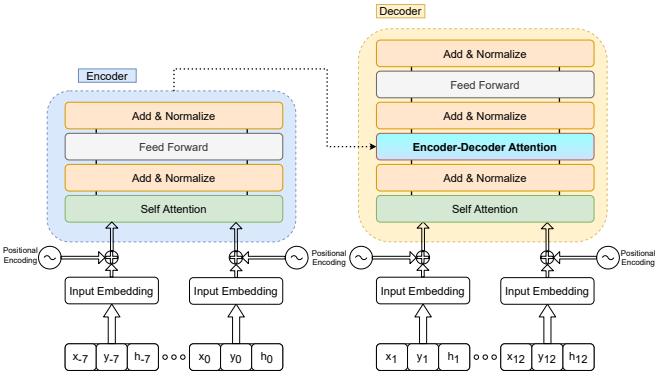

Fig. 2. Architecture overview: adding new inputs.

1) Positional Information and Encoding: Compared to an LSTM, TF does not process the input sequentially, so it needs a way to encode the temporal information. For this it makes use of positional encoding, where each input embedding has its corresponding timestamp, calculated through sine and cosine functions, as in [4]. The input embedding is concatenated with the positional encoding vector. This feature of the TF makes it possible to learn in parallel from all the time instants, while the LSTM needs to be processed sequentially, performing back-propagation. If a certain element is missing from the input, the positional encoding will consider it, which is not as problematic as in a LSTM, where this information could be lost. This claim was proven by the original authors of the model, showing that the architecture can perform satisfactorily even when missing data, degrading performance in inverse proportion to the age of the lost samples. Normalization of the input is vital for the performance, so the inputs are normalized by subtracting the mean and dividing by SD of the train set. This model is not multimodal, although it would be possible to classify the inputs into different classes by representing the inputs as vectors (i.e., through classification). According to its original developers, regression works better than classification-based approaches, so this approach is the one chosen.

We kept the original architecture of [1], adopting an L2 loss in which position increments (to enhance the independence of each given position) and normalized heading are configured. $d_{model} = 512$, using 6 layers and 8 attention heads. Warmup have been setup to 10 epochs, using a decaying learning rate in the subsequent epochs.

The memory of the model is based on the attention modules, where in the encoding stage a representation for the observation sequence is produced. Thus, this creates two vectors to be handed to the decoder stage, as seen in the figure 2. TF keeps the memory separate from the

decoded sequence, unlike LSTM, which keeps everything in the hidden state. For this reason, TF is known to work better with longer prediction horizon.

The additional input to the Transformer is the heading normalized between 0 and 1. This is then combined with the training loss calculation to complement the velocity (position increments) at each time instant.

2) Preprocessing the data: As data contained in every dataset is enclosed in a different way, as will be detailed in the corresponding section, it is necessary to preprocess them. Firstly, it is required to take into consideration the framerate, even though in this case the input data will be kept in order to enable possible studies to be carried out, considering this matter directly in the data loader. The sets are separated by classes, according to the tests to be performed. After analyzing part of the recordings of one of the datasets, it was noticed that there were static cars parked steadily, which data could affect the result of the inference. Thus, a filtering is applied to remove this specific information from the datasets, resulting the structure $frame, track, x, y, vx, vy, heading$. Some tests have been developed with vx and vy, but finally they have not been employed in this work. Increments of x and y have been calculated directly as velocity value, considering the time step. Units are expressed in SI. The sliding window step for each valid full trajectory capture has been set to 1, in order to analyze all possible trajectories in each split.

IV. Results

A. Datasets

The use of bird's eye view (BEV) datasets has been remarkably extended in the recent works to develop trajectory prediction systems, emphasizing the TrajNet [26] for pedestrian trajectory prediction. This dataset offers a Challenge that constitutes a solid multi-scenario forecasting benchmark, observing 8 values of position ground-truth (3.2 seconds) and predicting the following 12 positions (4.8 seconds). All positions are given in world plane coordinates, being the 8-12 protocol a consistent fashion for diverse datasets, as explained in the following section.

Beyond purely pedestrian-based approaches, NGSIM datasets [27] [28] pioneered in providing coverage of a highway area, offering information taken from cameras mounted at a skyscraper. Other multi-agent focused datasets have been developed in the past few years, some focusing on highway scenarios, such as highD [7] for highway vehicle trajectory prediction, offering aerial images obtained with a drone located over various locations of german *Autobahn*, Vehicle labeling ensures that the error is below 10 centimeters, providing a combined total of 147 hours of drive time on more than 100,000 vehicles. Moreover, the authors of this dataset went further with the concept, moving to urban scenarios: inD [8] and rounD [9] record different intersections and roundabouts, respectively. In addition to previously mentioned, the INTERACTION Dataset [10] combines all these scenarios, including ramp merging, signalized intersections and roundabouts. This dataset also provides diverse material in driving behavior, showing multiple critical maneuvers, including an accident. These are the situations that add value to a trajectory prediction solution, and should be evaluated here in a qualitative way. Table I offers an overview of datasets employed to develop the experiments. Furthermore, while the use of 2D datasets taken from drones or fixed locations in bird's eye view enables a relatively simple creation and labeling process, the ultimate purpose of such datasets would be to train models that can later be ported to vehicles with onboard sensors, which can be tested in datasets like the PREVENTION Dataset [11].

B. Evaluation metrics

TrajNet performance is measured in terms of Mean Average Displacement Error (MAD/ADE) and Final Average Displacement (FAD/FDE). ADE measures the aligned Euclidean distance from the prediction w.r.t. the ground truth, making an average of the error at every time step. That is, ADE reports a mean value of the general fit of the forecast in the predicted trajectory. FDE measures the Euclidean distance at the very last step, comparing the prediction to the corresponding ground truth position.

1) inD: Comparative results: In order to make a fair comparison, the model execution for this section has been performed with the same data split used by the DCENet authors to carry out their quantitative analysis. This includes all types of agents, which are loaded and analyzed globally in the results without differentiation, which may affect the results of the TF models.

Beyond this aspect, the data split proposed for the table II includes in the training recordings of intersections of the same location that will be analyzed later in the test, but in any case a recording has been included in both training and test. In addition, it would be interesting to propose an alternative that compares temporal horizons, instead of setting the 12 prediction frames (4.8 seconds) as the only horizon.

As we can see, the Vanilla-TF model is behind AMENet and DCENet in this test, while Oriented-TF improves the results to some extent, without outperforming those described. Thus, in this test it is not possible to conclude categorically whether the inclusion of heading improves trajectory forecasting.

For this reason, in the following comparative tests of generalization of the models, different splits will be selected, depending on the type of test to be performed, which avoid the visualization of equivalent scenes by the model in the training set. Furthermore, from now on only vehicles (cars, trucks, vans, trailers, etc.) will be evaluated.

C. Testing in different datasets

1) Single dataset tests: The aim of this section is to perform tests with different data splits within each dataset, in order to analyze the performance of the model for different scenes, keeping completely separate the data with which the model is trained and the test. It is also possible to compare the performance of the original system and the one that includes the heading. As shown in Table III, the Oriented-TF

TABLE I
DATASETS USED IN THIS WORK

Dataset	Country	Locations	# Tracks	Road User Types	Data Frequency	Maps
inD	Germany	urban intersections (4)	11500	pedestrian, bicycle, car, truck, bus	25 Hz	yes
rounD	Germany	(sub-)urban roundabouts (3)	13746	pedestrian, bicycle, motorcycle, car, van, truck, bus, trailer	25 Hz	yes
highD	Germany	highway (6)	110000	car, truck	25 Hz	no
INTERACTION	USA Germany China	roundabout (5), intersection (4), highway (2)	40054	pedestrian/bicycle, car, truck	10 Hz	yes

TABLE II
GENERAL PERFORMANCE

InD	Average
S-LSTM	1.88/4.47
S-GAN	2.38/4.66
AMENet	0.73/1.59
DCENET	0.69/1.52
Vanilla-TF	1.07/2.65
Oriented-TF	1.02/2.57

takes advantage in the INTERACTION recordings, improving the FDE by more than one meter in all scenarios.

As can be appreciated in the table, the results of split 4 of the inD are notably weaker in all metrics. Analyzing the recordings, it is possible to think that in this intersection the network does not have other previous references, since the only one that could be similar is 3, slightly less complex and with lower occurrence in the dataset. As expected, the results in the highD are remarkably favorable, due to the strong linear component that exists in this highway dataset. Note that in the highD the authors do not provide the heading as it, so a careful selection of another included metric, the minimal distance headway (in meters), is introduced directly instead of the heading (it is not normalized in this example).

As tested in the comparative analysis against other architectures, it was also planned to carry out a data split including similar video sequences, to evaluate the performance of the two architectures in similar conditions to the ones studied in the inD. Thus, the results obtained are as expected, with a decrease of about 4 times the error in the rounD case. This may be due to the marked imbalance of the data per scenario in this dataset. In the INTERACTION, the error is also lower, but to a minor extent, and the results are slightly better for the Oriented-TF model.

2) Mixing datasets: similar scenarios of different datasets: In order to assess the generalization potential of the system in terms of coordinate prediction independently of the inputs, this analysis will test the model in similar scenarios to those already known, but from a completely different dataset.

As can be seen in table IV, the results are quite satisfactory for the intersections, obtaining similar figures to those obtained by performing the train on the dataset itself. Something specific can be observed in the case of roundabouts, where a lower error is obtained when testing on roundabouts of a dataset different from the dataset on which the model has been trained. In the case of the Oriented-TF, no improvement is seen here, being marginal only for the highD. The generalization of the model in this case is excellent, obtaining results that are even better than its own.

3) Generalization between different scenarios: After testing comparable scenarios, the performance of vehicle dynamics learning will now be assessed, independently of the trajectories observed in the training videos. So, for example, could it know how a vehicle will act at a junction if it has been trained with roundabouts? As observed in Table V, the generalization in this case is also fairly adequate, highlighting an improvement to the single results in the inD-rounD and inD-INT-round test. It seems quite significant that the model has improved the results in roundabouts training with intersections, and it is also remarkably the performance improvement of the Oriented-TF in the training and test cases in the INTERACTION.

4) Changing frame rate of input data - Vanilla TF: The frequency of data input also seems to be vital in the performance of a prediction system. In this section, Vanilla-TF has been tested on the set of inD vehicles, with training strategy on 3 scenes and test on the remaining one. The tests have been carried out with the original layout of 2.5 fps (8-12) and on 5 fps (16-24). For the test sets on scenarios 1 and 2, the choice of doubling the framerate is the winning option, with FDE improvements of 0.43 m and 0.57 m, respectively. However, for scenarios 3 and 4, the 2.5 fps framerate is still the better option, with FDE improvements in favor of 1.49 m and 3.5 m, respectively. Junctions 1 and 2 coincide with crossroads at perpendicular intersections where the vehicle dynamics are different to the others, possibly extracting more information and taking advantage of the additional framerate.

D. Qualitative results

Beyond the quantitative results, it is always convenient to have an approach closer to reality by directly representing the input and output data of the system. Figure 3 shows three prediction situations that were observed in one of the cross experiments, specifically in the rounD - INTERACTION Roundabouts. In one of them the system has correctly predicted a linear trajectory, in another it is forecasting quite correctly a moderately tight turn, while in the last one it has chosen a turn in the wrong direction, making a very significant error according to the established metrics. The selected figures include one of the sections of the USA_EP map, where there is a junction stretch adjacent to the roundabout. Thus, it is also possible to appreciate the model's generalization capacity, which has been solely trained with European roundabouts from the rounD dataset. The figure 4 shows the histogram of FDE for various forecast

TABLE III

SINGLE DATASET TESTS

Training // Test	Vanilla-TF ADE / FDE	Oriented-TF ADE / FDE	Training // Test	Vanilla-TF ADE / FDE	Oriented-TF ADE / FDE
inD: 123 // 4	**7.67** / 17.22	7.71 / **16.83**	**rounD: 01 // 2**	**6.59 / 16.87**	6.62 / 17.09
inD: 124 // 3	**1.46 / 3.85**	1.56 / 4.08	**rounD: 02 // 1**	**6.64 / 17.04**	6.88 / 17.53
inD: 134 // 2	**2.80 / 7.46**	3.47 / 9.02	**rounD: 12 // 0**	**6.68 / 16.71**	7.98 / 19.82
inD: 234 // 1	1.91 / 5.18	**1.89 / 5.14**	**rounD: mixed**	**1.88 / 4.85**	1.94 / 5.10
inD: mixed	1.07 / 2.65	**1.02 / 2.57**	**highD**	**1.19 / 2.96**	2.20 / 3.75
INT - intersection: EP0-EP1-MA // GL	2.54 / 6.95	**2.10 / 5.66**	**INT - roundabout: SR-FR-EP-OF // LN**	4.46 / 11.65	**3.81 / 9.51**
INT - intersection: MA-GL-EP0 // EP1	3.27 / 8.17	**2.80 / 7.16**	**INT - roundabout: LN-SR-FT-EP // OF**	4.27 / 11.63	**3.68 / 10.11**
INT - intersection: mixed	2.09 / 5.85	**1.81 / 4.98**	**INT - roundabout: mixed**	2.75 / 7.78	**2.31 / 6.38**

TABLE IV

EQUIVALENT SCENARIO TESTS (TRAINING ON ENTIRE DATASET)

Training // Test	Vanilla-TF ADE / FDE	Oriented-TF ADE / FDE
inD // INT-int	**3.12 / 8.10**	4.89 / 10.87
INT-int // inD	**4.04 / 10.10**	4.24 / 10.32
rounD // INT-round	**3.19 / 8.34**	5.18 / 11.72
INT-round // rounD	**5.30 / 14.13**	6.99 / 16.54
highD // INT-merg	2.45 / 5.14	**2.35 / 4.77**

TABLE V

DIFFERENT SCENARIOS TESTS (TRAINING ON ENTIRE DATASET)

Training // Test	Vanilla-TF ADE / FDE	Oriented-TF ADE / FDE
inD // rounD	**5.87 / 15.08**	5.97 / 15.37
rounD // inD	**3.27 / 8.35**	3.40 / 8.59
INT-int // INT-round	5.04 / 12.84	**4.51 / 11.68**
INT-round // INT-int	2.99 / 8.21	**2.67 / 7.28**
inD // INT-round	**3.34 / 8.58**	5.26 / 11.46
INT-round // inD	**3.36 / 8.83**	4.44 / 10.08

horizons in the qualitative scenario, considering in all cases 8 frames observed. It is visible that the errors increase as the time horizon is extended, showing slopes similar to those of a normal distribution.

As seen in this section, BEV datasets used with the TF can deliver surprising results, performing better in some situations when they have been trained with foreign scenes. This raises the question of whether these scenarios really contribute anything to model learning, opening the debate as to whether it is better to have a large amount of data or whether more variability in the scenes, coupled with detailed labeling, is more valuable. An interesting phenomenon has also been noticed with the inclusion of heading, working better in the INTERACTION dataset, while in the others it has hardly improved. This opens the door to a modification of the normalization method and processing of heading in these datasets.

V. CONCLUSIONS AND FUTURE WORK

Based on the experiments performed, it is possible to conclude that the Oriented-TF model, as well as Vanilla-TF, are fully competent among the state-of-the-art models for the datasets analyzed in this work, confirming its good performance in TrajNet by its original authors, considering that it is a single agent approach, where no context variables or interaction with other agents are included. Thus, a first approach to the analysis of its generalization ability has also been carried out, by conducting multiple cross-tests between similar scenarios of diverse datasets, analyzing the obtained results. A novel use of the Transformer is also proposed, by adding the agent's orientation as an input variable to improve the trajectory prediction, observing interesting results, depending on the dataset analyzed. As future work, the core task is to further develop the model, measuring its possibilities for single agent input data processing, as well as exploring the social architectures already proposed based on graphs. In addition, the direct use of this model could involve other datasets that also contain data that can be expressed in 2D, as is the case of the information that we can obtain from PREVENTION through the radars. This will enable testing, for example, the inference time in a real situation by obtaining information from the radars of an instrumented vehicle.

ACKNOWLEDGMENT

This work was funded by Research Grants S2018/EMT-4362 (Community Reg. Madrid), DPI2017-90035-R (Spanish Min. of Science and Innovation), BRAVE Project, H2020, Contract #723021 and PRE2018-084256 (Spanish Min. of Education) via a predoctoral grant to the first author.

REFERENCES

[1] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer Networks for Trajectory Forecasting," pp. 1–18, 2020. [Online]. Available: http://arxiv.org/abs/2003.08111

[2] H. Cheng, W. Liao, M. Y. Yang, B. Rosenhahn, and M. Sester, "AMENet: Attentive Maps Encoder Network for trajectory prediction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 253–266, feb 2021. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2020.12.004

[3] H. Cheng, W. Liao, X. Tang, M. Y. Yang, M. Sester, and B. Rosenhahn, "Exploring Dynamic Context for Multi-path Trajectory Prediction," Tech. Rep. [Online]. Available: https://github.com/wtliao/DCENet.

[4] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," Tech. Rep.

[5] N. Deo and M. M. Trivedi, "Convolutional Social Pooling for Vehicle Trajectory Prediction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1549–15 498.

[6] F. Diehl, T. Brunner, M. T. Le, and A. Knoll, "Graph Neural Networks for Modelling Traffic Participant Interaction," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 695–701.

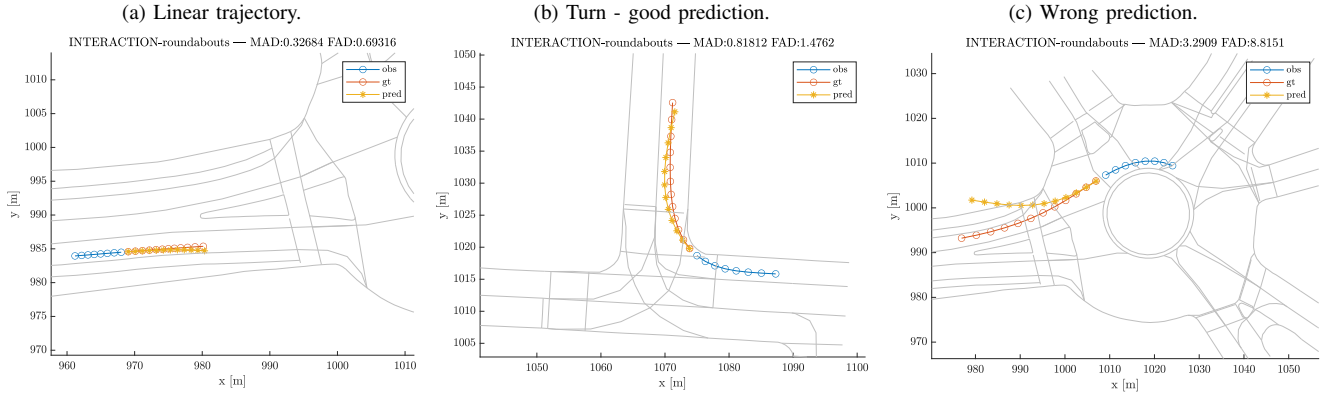

Fig. 3. Sample outputs for cross experiment (trained in rounD - tested in INTERACTION Roundabouts set). Observed trajectory is depicted in blue, ground truth in red and predicted trajectory in yellow.

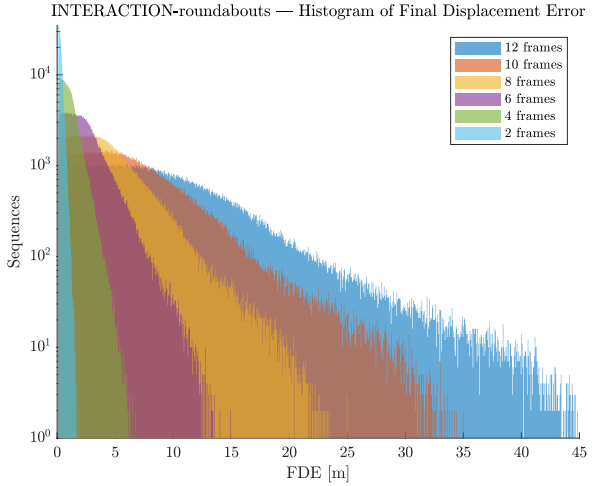

Fig. 4. Histogram of FDE.

[7] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2018-Novem. Institute of Electrical and Electronics Engineers Inc., dec 2018, pp. 2118–2125.

[8] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, "The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections," *arXiv*, nov 2019. [Online]. Available: http://arxiv.org/abs/1911.07602

[9] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany." Institute of Electrical and Electronics Engineers (IEEE), dec 2020, pp. 1–6.

[10] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv*, sep 2019. [Online]. Available: http://arxiv.org/abs/1910.03088

[11] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 3657–3662.

[12] M. Biparva, D. Fernández-Llorca, R. Izquierdo-Gonzalo, and J. K. Tsotsos, "Video action recognition for lane-change classification and prediction of surrounding vehicles," 2021.

[13] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Vehicle Trajectory Prediction in Crowded Highway Scenarios Using Bird Eye View Representations and CNNs," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020*. Institute of Electrical and Electronics Engineers Inc., sep 2020.

[14] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem. IEEE Computer Society, dec 2016, pp. 961–971.

[15] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena, "A Data-driven Model for Interaction-Aware Pedestrian Motion Prediction in Object Cluttered Environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5921–5928.

[16] H. Manh and G. Alaghband, "Scene-LSTM: A Model for Human Trajectory Prediction," 2019.

[17] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 12 077–12 086, mar 2019. [Online]. Available: http://arxiv.org/abs/1903.02793

[18] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," Tech. Rep.

[19] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, S. H. Rezatofighi, and S. Savarese, "SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints," 2018.

[20] Y. N. W. Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, "Multi-Agent Tensor Fusion for Contextual Trajectory Prediction," pp. 12 126–12 134, 2019.

[21] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, S. H. Rezatofighi, and S. Savarese, "Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks," *arXiv*, jul 2019. [Online]. Available: http://arxiv.org/abs/1907.03395

[22] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14 412–14 420, feb 2020. [Online]. Available: http://arxiv.org/abs/2002.11927

[23] X. Li, X. Ying, and M. C. Chuah, "Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving," 2020.

[24] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," Tech. Rep. [Online]. Available: www.aaai.org

[25] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction," Tech. Rep. [Online]. Available: https://github.com/Majiker/STAR

[26] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi, "Trajnet: Towards a benchmark for human trajectory prediction," *arXiv preprint*, 2018.

[27] J. Colyar and J. Halkias, "NGSIM - US Highway 101 Dataset," 2007. [Online]. Available: https://www.fhwa.dot.gov/publications/research/operations/07030/07030.pdf

[28] J. Halkias and J. Colyar, "NGSIM - Interstate 80 Freeway Dataset," 2006. [Online]. Available: https://www.fhwa.dot.gov/publications/research/operations/06137/06137.pdf