Research Proposal

1 Introduction

Autonomous driving technology (such as lane departure warning system, adaptive cruise, automatic parking system, etc.) is considered to be an important technology to ensure car safety. Active safety performance is the core issue of autonomous driving. In order to assess safety risks, the surrounding environment and agents (such as vehicles, pedestrians, cyclists, and so on) must be monitored in real-time. Compared with static environmental information, it is often more difficult to monitor the behavior of moving agents. As the surrounding agents are in motion, in order to realize the accurate safety assessment of the self-driving vehicle, the monitoring of the agents needs to pay attention not only to its current position and behavior mode but also to its possible position in the future. Therefore, it is necessary to predict the trajectory of the surrounding agents.

Recently, deep learning methods have been gradually introduced into the field of trajectory prediction. Due to the powerful expressive ability of neural networks, it is helpful for the model to find laws that cannot be obtained by general machine learning from large-scale data, thereby achieving more accurate trajectory prediction. However, some issues remain to be resolved. It is a topic worthy of research on how to learn robust scene representation and model the spatial interaction between agents under complex traffic conditions, give full play to the expressive ability of advanced deep learning models, and realize a highly accurate and highly interpretable trajectory prediction model.

2 Related Works and Research Gap

The development of trajectory prediction algorithms in the past few decades has reflected in two aspects: One is to introduce a model with stronger expressive ability, which has gradually evolved from the kinematics and dynamics model in the early years to the current popular deep learning model. The second is to introduce more abundant information, from only paying attention to the instantaneous position and speed information of the target agent in the early years to introducing the historical information of the target agent to explore the rules in time sequence, to starting to pay attention to the spatial impact of surrounding agents on the movement of the target agent, and even to the game of vehicle driving intention. These two improvement directions are not irrelevant but complement each other. The effective application of complex information depends on the strong expressive ability of the model, and the design of complex models depends on our understanding and simplification of driving behavior.

Most of the state-of-the-arts for autonomous driving trajectory prediction are deep learning approaches. With 4 key components as input representation, output representation, interaction representation, and prediction modeling, existing approaches can be classified as follows:

- Input Representations

  Concerning input representations for trajectory prediction, most approaches take advantage of either graph-based representations [1]–[5] or rasterization-based representations [6]–[10]. Casas, Luo, and Urtasun [6] propose to rasterize static information such as roads, lanes, crossing, and traffic signs together with dynamic information that keeps changing like traffic lights. Gao, Sun, Zhao, *et al.* [2] utilize local graph networks to encode the representation of the entity such as agents and road structures first and then use a global graph network to model the interactions.

Although rasterized maps provide richer geometric and semantic features, it's easier to build interaction models based on graph representations as structures of graph models have a strong inductive bias for modeling entities and their interactions [11].

- Interaction Representations

  Correctly modeling the interaction between agents is important for predicting the trajectory with a longer time horizon. Many early studies only focused on simple traffic scenes with sparse agents and interactions such as through highways. Relying only on the historical information of the target agent is usually enough to make reliable predictions in such traffic scenes. However, with the increasing attention to dense and complex traffic scenes, the spatial constraints between agents have a increasingly greater impact on the future trajectory. Therefore, many recent studies attempt to model the spatial interaction between agents and integrate it into the trajectory prediction model.

  Deo and Trivedi [12] and Altché and La Fortelle [13] extend the input of the encoder from the historical track of a single agent to multiple historical tracks of the target agent and its adjacent agents, and use the LSTM network to learn the impact of other agents on the target agent track. [14], [15] model social influence between adjacent pedestrians by social pooling mechanism and predicts their interactive trajectories. [4], [16]–[18] take advantage of Graph Neural Network (GNN) to model agent-to-agent interactions. [19]–[22] adopt attention mechanism to model interaction relationships between multiple agents.

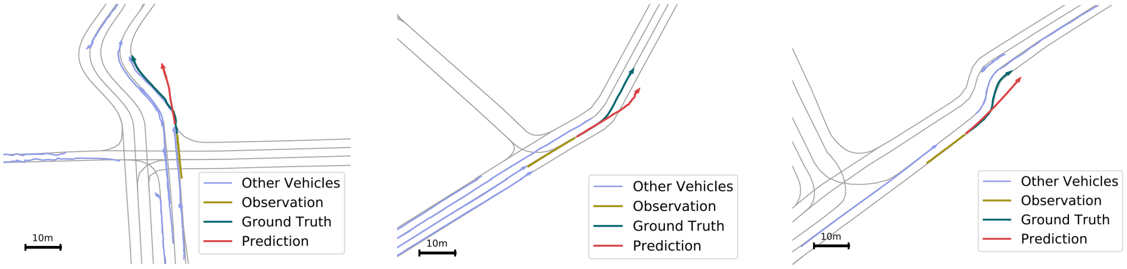
  However, since all these approaches build implicit interaction models and learn dependencies between agents in an end-to-end manner, they offer little interpretability and often fail to generate scene-compliant trajectories, especially when faced with unseen data. Therefore, auxiliary collision loss [20] or a critic built by inverse reinforcement learning [23] are utilized to discourage colliding trajectories. Moreover, Liu, Yan, and Alahi [24] proposes to incorporate contrastive learning to learn socially-aware representations and avoid undesirable events such as collisions. Bahari, Saadatnejad, Rahimi, *et al.* [25] report poor generalization of learning-based approaches on unseen traffic scenes and propose a adversarial scene generation method to deal with the problem, which is claimed to be helpful in 30%-40% off-road rate reduction.

  In addition, conditional trajectory predictors identify interaction relations explicitly between agents. [4], [5], [26], [27] condition trajectories of agents on the future motion of another agent to predict correlated future trajectories. However, these approaches rely heavily on the knowledge of the future trajectory of an agent such as the autonomous vehicle or a robot, whose motion plan is available to the prediction model. Sun, Huang, Gu, *et al.* [28] go beyond by identifying influencers and reactors between the agents and predicts the future trajectories of influences first and future trajectories of reactors conditioned on influencers' trajectories then. But this work is still limited to 1-on-1 interaction relations.

- Output Representations

  Most papers directly predicts trajectories [4], [8] or occupancy maps [10], [29]–[31]. However, some propose to utilize intermediate results to predict conditional trajectories. Zhao, Gao, Lan, *et al.* [32] and Gu, Sun, and Zhao [33] predict the targets in the first stage and then refines the motion between the target and current position. Casas, Luo, and Urtasun [6] makes trajectory predictions based on drivers' intents, which are high-level features learned end-to-end.

- Prediction modeling



(a) DATF[42]

(b) WIMP[26]

(c) LaneGCN[3]

Figure 1: The predictions of different models in some generated scenes. All models are challenged by the generated scenes and failed in predicting in the drivable area. Images from [25]

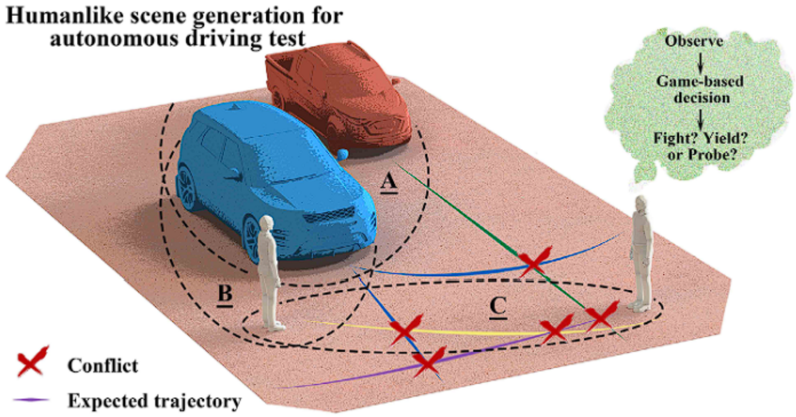


Figure 2: Schematic illustration of driving conflicts among road users.A) Vehicle–vehicle interaction; B) pedestrian–vehicle interaction;C) pedestrian–pedestrian interaction. Image from [43]

Previous research has used either discriminative models [34]–[37] or generative models [38]–[40]. For discriminative models, either a single MAP trajectory per agent is predicted via supervised regression, or distribution over multiple possible trajectories is generated using multi-modal loss like multiple-trajectory prediction loss [9], [38]. Generative models leverage random sampling to model future uncertainty. However, generative models may suffer from poor interpretability, low inference efficiency, and the problem of modal collapse [41].

The existing research has the following two deficiencies:

(1) Agent trajectories depend on road structure and are closely related to road traffic elements such as lane width and lane alignment, lane user type and speed limit, lane driving direction, and traffic lights, and vary greatly in different road sections. However, most of the existing trajectory models either focus on simple cases such as highways, where trajectory predictions are acceptable even without road structures, or the representation of road elements and structure only concentrate on spatial information. When the road structure is complex and changeable, the prediction accuracy and robustness of prediction models are often limited. Moreover, when faced with unseen data with the covariate shift, learning-based approaches show poor scene compliance, as shown in Figure 1.

(2) The behavior of multiple adjacent agents affects each other. As shown in Figure 2, There are usually game relations between adjacent agents, such as yield, fight, blend-in and avoidance,

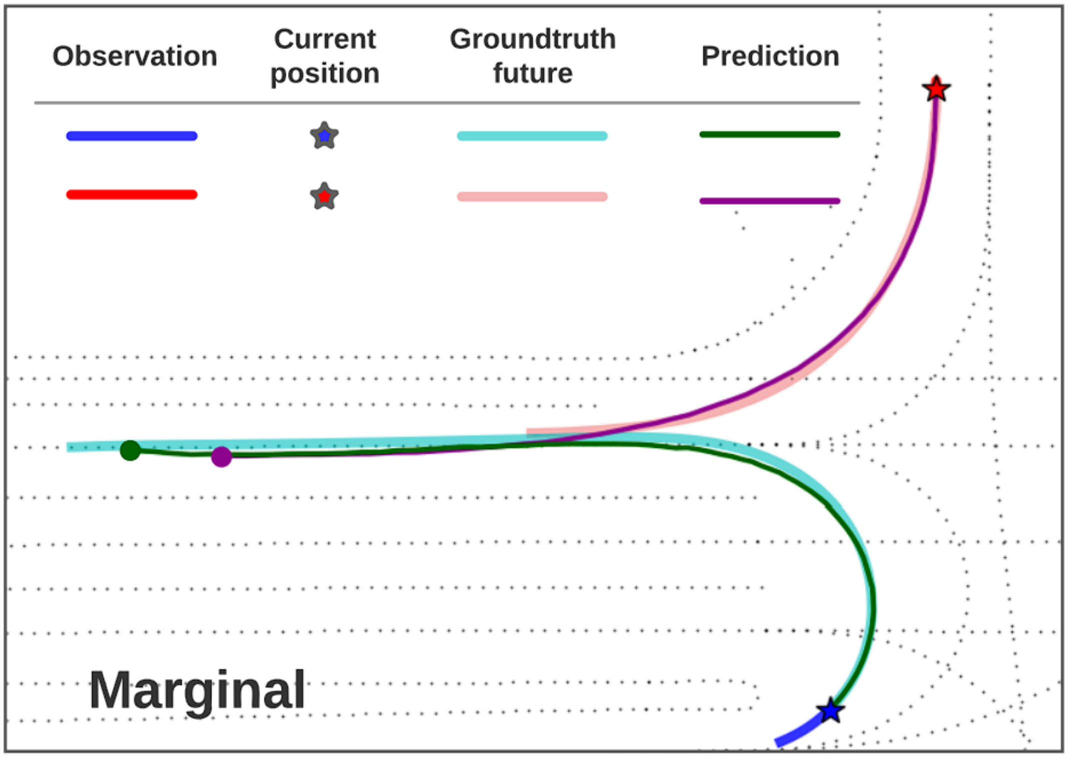


Figure 3: An example in which the marginal predictor produces overlapping and inaccurate predictions. Image from [28]

and the interaction between agents is generally two-way. However, the existing prediction methods usually focus on marginal trajectory prediction and rarely explicitly model the interaction between multiple agents, which makes it difficult to deal with complex road scenes that often include multi-agent game interactions, as shown in Figure 3. Even if the interaction of multiple agents is considered, the model is often limited to the unidirectional impact of other agents on the target agent. As a result, the predicted trajectory might be socially unacceptable or scene incompliant. Moreover, downstream modules like the planning module require scene-compliant predictions for easier cost/risk assessment of planned trajectories.

3 Research Aims and Questions

Problem 1: How to learn robust representation on complex traffic scenes?

It's worth exploring the representation of road traffic elements suitable for trajectory prediction in the multi-agent interaction scene and the modeling method of multi-agent interaction relationship so as to realize the long-time interaction domain trajectory prediction of surrounding agents in complex traffic scenes. Another focus is on solving the problem of low accuracy of trajectory prediction caused by the lack of assistance with more information on road structures and elements, especially when faced with unseen data, and incorporating more regulation on what's not socially appropriate and scene-compliant.

Problem 2: How to model interactions between agents (and traffic elements) to improve scene compliance?

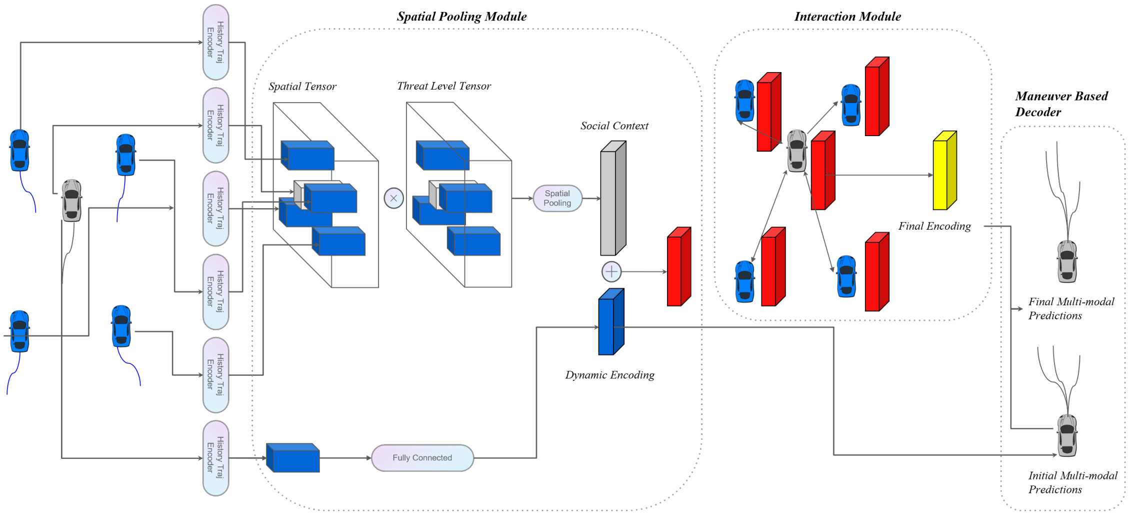


Figure 4: Proposed Model: First historical trajectory information of agents are encoded with LSTMs. Encoded agent dynamic state are concatenated with HDMap feature map to obtain the spatial tensor. Then spatial tensor is multiplied by threat level tensor and used as input to social pooling layers. Output of the social pooling layers are viewed as the social context (grey vector) for the target agent and is concatenated with the dynamic encoding (blue vector underneath) in order to obtain the agent states (red vector) for the interaction module. Dynamic encoding of the target agent is used to calculate an initial trajectory prediction. Finally the final encoding (yellow vector) of the target agent obtained from the interaction module is used to refine the prediction.

The respective influences of the historical states and the spatial interactions of surrounding agents need to be distinguished. On this basis, a trajectory prediction framework of "prediction correction" can be constructed. The trajectory is estimated according to the historical states, and the modeling of the spatial interaction between agents is used to output the correct sequence for the estimated result, and the threat degree of the surrounding agents to the target agent is evaluated as the weight of the correction. Efficient use of spatiotemporal information can simulate the driving behavior of human drivers, that is, the process of fine-tuning the driving trajectory according to surrounding agents (and scene elements), so as to avoid risky behaviors as much as possible, thus helping to greatly improve the scene compliance and effectiveness of trajectory prediction.

4 Methodology

In order to solve the above two research problems, we plan to propose a scene-interaction-aware trajectory prediction method for autonomous driving. The core idea of the prediction framework is "understanding-prediction," that is, to predict the trajectory of surrounding agents after understanding the road traffic scene. Specifically, it includes two core procedures: scene modeling and trajectory prediction as follows.

4.1 Scene Modeling

Scene modeling refers to the establishment of a general state expression and feature extraction method for the road structure, obstacles, traffic rules, and other elements in the road traffic scene so as to provide a standardized input form and rich feature information for the follow-up automatic driving tasks such as trajectory prediction, behavior decision-making, path planning and so on.

For scene modeling, the specific workflow we envision is as follows: (1)the coordinate system centered on the target agent is determined, the spatial grid within a certain range of the center of the coordinate system is meshed, and different grid sizes are set according to the distance from the center; (2)assign different values to the grid occupied by road structures such as driveable areas, lane lines and road edges according to the lane direction and other geometric features the to obtain the road structure feature map, and similarly obtain the obstacle feature map; (3)mark regions under different traffic situation and obtain the traffic rule feature map; (4)concatenate all feature maps and feed it into CNN to extract the spatial features and obtain the global feature map of the road structure and traffic flow; (5) agent states are obtained from historical information by RNNs; (6) The global feature map and agent states are feed into convolutional spatial pooling layer to further extract features of different levels and model social interdependencies.

Besides, considering the huge difference in scene distribution, we may consider building a meta-scene offline model based on the meta-learning method. The model can adopt Transformers [44] as the basic structure. The autonomous vehicle obtains scene information online and performs online fine-tuning on the meta-scene offline model. This approach has the potential to improve the representational power of the scene model.

4.2 Trajectory Prediction Framework with Coupled Spatiotemporal Information

The key to trajectory prediction is to fully consider the difference in the time series information and spatial interaction information and treat them differently in the model. The proposed framework takes a coarse-to-fine manner.

Specifically, as the time-series information contains rich and comprehensive knowledge about the current motion of the agent, it is concatenated with scene representations to predict the initial trajectory. The spatial interaction information is then used to correct the agent trajectory. In other words, the historical movement of the predicted target and surrounding scene is considered for coarse predictions. Then interactions between agents (and traffic elements) are considered to refine the coarse predictions and make them socially acceptable and scene complaint.

Furthermore, since the proposed model will focus on multiple surrounding agents at the same time, multiple trajectory correction sequences need to be effectively superimposed to form a global correction sequence and generate the final prediction result. The model will evaluate the security threat level between the surrounding agents and the target agent so as to measure the influence of each surrounding agent on the prediction result of the target agent and dynamically modify the weights in the above superposition process so as to get more reasonable (less security risk) predictions. In addition, this process based on the superposition of dynamic weights is also consistent with our intuition. In fact, human drivers also estimate whether the surrounding agents will pose a safety threat to themselves while driving and actively avoid risks. This also gives some interpretability to the prediction process of this model.

For evaluation of Safety Distance and Calculation of Dynamic Weights, firstly, based on the relative position and relative speed, the safety distance of the two agents is calculated, and the safety distance is compared with the actual distance of the two agents to obtain a variable representing the safety threat level between the two agents. Finally, all the global "security threat levels" are normalized so as to obtain the weight required for the superposition of the correction sequence. The entire calculation process above is based on the instantaneous state of the two agents, and in the long run, the calculated weights are therefore dynamically changing.

In addition, the normalization makes the algorithm adequate for situations where there are no surrounding agent detection results, which also increases the flexibility of the algorithm.

4.3 Model Structure

In order to realize the calculation process of ”prediction correction”, we would refer to the composite structure of Recurrent Neural Network (RNN) in [45] and divide the LSTM models involved in the composite into two categories: Component LSTM (ComponentLSTM, C-LSTM) and Interaction LSTM (Interaction LSTM, I-LSTM) to simultaneously model temporal and spatial interactions. When designing the model, the output and stacking method of each substructure can be clearly defined so as to incorporate the core ideas of hierarchical processing of information and dynamic weight construction mentioned above. In addition, the model can adopt a step-by-step training process during the training phase to adapt to the semantics of the substructure. Moreover, the framework of contrastive learning can be adopted to discourage trajectory predictions that are not scene-compliant (e.g. off-road waypoints, colliding trajectories)

5 Timeline and Expected Outcome

The above research contents will be completed within 3-4 years. The specific research plans are as Table 1 shows:

Table 1: Timeline and expected outcome

	Expected Outcome
1st Year	The grid representation of road traffic elements will be studied.We will analyze the relationship between the road traffic environment and the prediction of the trajectories of the surrounding agents, and design a prediction framework of the trajectories of the surrounding agents that integrates the complex road traffic environment.
2nd Year	Interaction-aware trajectory prediction methods will be studied. We will design interaction-aware trajectory prediction models and conduct simulation and experimental studies.
3rd Year	Performance enhancements and method improvements will be carried out. We will further improve the research work and complete the main work of the dissertation.

Through this study, it is expected to solve the problem that the complex road structure is difficult to describe. At the same time, it can also solve the problem that the prediction accuracy drops when the traffic is dense, and improve the accuracy of the prediction trajectory. The research work will be published in top journals/conferences.

References

[1] N. Homayounfar, W.-C. Ma, J. Liang, X. Wu, J. Fan, and R. Urtasun, "Dagmapper: Learning to map by discovering lane topology," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2911–2920.

[2] J. Gao, C. Sun, H. Zhao, *et al.*, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.

[3] M. Liang, B. Yang, R. Hu, *et al.*, "Learning lane graph representations for motion forecasting," in *European Conference on Computer Vision*, Springer, 2020, pp. 541–556.

[4] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision*, Springer, 2020, pp. 683–700.

[5] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *European Conference on Computer Vision*, Springer, 2020, pp. 598–614.

[6] S. Casas, W. Luo, and R. Urtasun, "Intentnet: Learning to predict intention from raw sensor data," in *Conference on Robot Learning*, PMLR, 2018, pp. 947–956.

[7] M. Bansal, A. Krizhevsky, and A. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," *arXiv preprint arXiv:1812.03079*, 2018.

[8] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[9] H. Cui, V. Radosavljevic, F.-C. Chou, *et al.*, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 2090–2096.

[10] X. Ren, T. Yang, L. E. Li, A. Alahi, and Q. Chen, "Safety-aware motion prediction with unseen vehicles for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 731–15 740.

[11] P. W. Battaglia, J. B. Hamrick, V. Bapst, *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[12] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2018, pp. 1179–1184.

[13] F. Altché and A. de La Fortelle, "An lstm network for highway trajectory prediction," in *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, IEEE, 2017, pp. 353–359.

[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 conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[15] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255–2264.

[16] S. Casas, C. Gulino, R. Liao, and R. Urtasun, "Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 9491–9497.

[17] S. Casas, C. Gulino, S. Suo, K. Luo, R. Liao, and R. Urtasun, "Implicit latent variable model for scene-consistent motion forecasting," in *European Conference on Computer Vision*, Springer, 2020, pp. 624–641.

[18] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 424–14 432.

[19] V. Kosaraju, A. Sadeghian, R. Martın-Martın, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] L. L. Li, B. Yang, M. Liang, *et al.*, "End-to-end contextual perception and prediction with interaction transformer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 5784–5791.

[21] N. Kamra, H. Zhu, D. K. Trivedi, M. Zhang, and Y. Liu, "Multi-agent trajectory prediction with fuzzy query attention," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 530–22 541, 2020.

[22] J. Ngiam, B. Caine, V. Vasudevan, *et al.*, "Scene transformer: A unified architecture for predicting multiple agent trajectories," *arXiv preprint arXiv:2106.08417*, 2021.

[23] T. van der Heiden, N. S. Nagaraja, C. Weiss, and E. Gavves, "Safecritic: Collision-aware trajectory prediction," *arXiv preprint arXiv:1910.06673*, 2019.

[24] Y. Liu, Q. Yan, and A. Alahi, "Social nce: Contrastive learning of socially-aware motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 118–15 129.

[25] M. Bahari, S. Saadatnejad, A. Rahimi, *et al.*, "Vehicle trajectory prediction works, but not everywhere," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 123–17 133.

[26] S. Khandelwal, W. Qi, J. Singh, A. Hartnett, and D. Ramanan, "What-if motion prediction for autonomous driving," *arXiv preprint arXiv:2008.10587*, 2020.

[27] E. Tolstaya, R. Mahjourian, C. Downey, B. Vadarajan, B. Sapp, and D. Anguelov, "Identifying driver interactions via conditional behavior prediction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 3473–3479.

[28] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao, "M2i: From factored marginal trajectory prediction to interactive prediction," *arXiv preprint arXiv:2202.11884*, 2022.

[29] S. Hoermann, M. Bach, and K. Dietmayer, "Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018, pp. 2056–2063.

[30] D. Ridel, N. Deo, D. Wolf, and M. Trivedi, "Scene compliant trajectory forecast with agent-centric spatio-temporal grids," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2816–2823, 2020.

[31] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *European Conference on Computer Vision*, Springer, 2020, pp. 414–430.

[32] H. Zhao, J. Gao, T. Lan, *et al.*, "Tnt: Target-driven trajectory prediction," *arXiv preprint arXiv:2008.08294*, 2020.

[33] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 303–15 312.

[34] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[35] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 3569–3577.

[36] N. Deo and M. M. Trivedi, “Convolutional social pooling for vehicle trajectory prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1468–1476.

[37] T. Yang, Z. Nan, H. Zhang, S. Chen, and N. Zheng, “Traffic agent trajectory prediction using social convolution and attention mechanism,” in *2020 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2020, pp. 278–283.

[38] T. Zhao, Y. Xu, M. Monfort, *et al.*, “Multi-agent tensor fusion for contextual trajectory prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 126–12 134.

[39] B. Ivanovic and M. Pavone, “The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2375–2384.

[40] X. Weng, Y. Yuan, and K. Kitani, “Ptp: Parallelized tracking and prediction with graph neural networks and diversity sampling,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4640–4647, 2021.

[41] H. Thanh-Tung and T. Tran, “Catastrophic forgetting and mode collapse in gans,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–10.

[42] S. H. Park, G. Lee, J. Seo, *et al.*, “Diverse and admissible trajectory forecasting through multimodal context understanding,” in *European Conference on Computer Vision*, Springer, 2020, pp. 282–298.

[43] Y. Zhang, P. Hang, C. Huang, and C. Lv, “Human-like interactive behavior generation for autonomous vehicles: A bayesian game-theoretic approach with turing test,” *Advanced Intelligent Systems*, vol. 4, no. 5, p. 2 100 211, 2022.

[44] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[45] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 5308–5317.