Query-Centric Inverse Reinforcement Learning for Motion Forecasting in Autonomous Driving

Muleilan Pei and Shaojie Shen
Hong Kong University of Science and Technology
{mpei, eeshaojie}@ust.hk

***Abstract*—Motion forecasting for surrounding on-road agents is challenging and essential in safety-critical autonomous driving systems. In contrast to conventional data-driven approaches that primarily adopt the supervised learning paradigm, our focus is to explore the potential of inverse reinforcement learning (IRL) for autonomous vehicles, which holds promise due to its learning from interaction mechanisms. In this work, we present a novel Query-centric Inverse Reinforcement Learning framework for motion forecasting, termed QIRL. First, we encode the traffic agents and scene elements in a vectorized manner and aggregate the context features utilizing a query-centric paradigm. Subsequently, we employ the maximum entropy IRL method to infer the reward distribution and derive the policy capable of inducing multiple plausible plans. Finally, conditioned on the sampled plans, we introduce a DETR-like decoder with a refinement module to generate accurate future trajectories. Experimental results on the large-scale Argoverse motion forecasting dataset demonstrate our proposed IRL-based predictor exhibits highly competitive performance compared to existing supervised models.**

I. Introduction

Trajectory prediction plays a crucial role in bridging the upstream perception and downstream planning modules in safe autonomous driving. However, accurate motion forecasting of surrounding traffic agents poses significant challenges due to its inherent uncertainty and underlying multi-modality: an agent can have multiple plausible future trajectories given its past observations and the available scene information [22].

Existing works primarily adopt learning-based frameworks, which leverage deep neural networks to encode the historical motion profiles of agents, as well as topological and semantic information of high-definition (HD) maps in either rasterized or vectorized context representation. Recently, the Transformer-based architecture has been extensively explored for feature extraction and fusion [21, 29], because of its notable improvement in overall prediction performance.

In general, the data-driven predictor can be regarded as imitating the behaviors of human drivers from a large amount of recorded data in real-world driving scenarios. This typical imitation learning task in robotics can be approached mainly in two ways: behavior cloning (BC) and inverse reinforcement learning (IRL). Most advanced methods in motion forecasting predominantly employ the BC framework, which involves directly learning the distribution of trajectories from datasets in a supervised manner. On the other hand, the IRL architecture models the agent's behavior as a sequential decision-making process [17] and aims to infer the underlying reward function that is considered the most parsimonious and robust representation of the expert demonstrations [15].

Despite achieving impressive performance in motion forecasting benchmarks, the BC-based learning fashion still possesses inherent challenges such as the covariate shift issue, whereas IRL offers a promising pathway to alleviate them, thanks to its learning from interaction mechanisms. Another critical concern associated with the supervised approach is the modality collapse problem. As only one ground-truth future trajectory is provided as supervision, the predictor has to generate diverse plausible predictions via learning one-to-many mappings [20]. In contrast, the IRL framework holds the potential to address uncertainty by integrating the maximum entropy (MaxEnt) principle [30]. The MaxEnt IRL paradigm intends to derive the reward distribution with the highest entropy [10], which can better capture the intrinsic multi-modality of demonstrations. Furthermore, the learned reward, as an interpretable intermediate representation [25], can also benefit downstream decision-making and behavior planning in highly complicated and interactive scenarios [8, 18].

In light of its superior properties, the MaxEnt IRL framework has garnered significant attention in recent research [5, 24, 27]. However, to the best of our knowledge, most traditional IRL algorithms typically operate efficiently in grid-shaped environments, which prompts existing work to utilize rasterized context representations, rendering scene elements into bird's-eye-view (BEV) images as input [9, 3, 7]. Consequently, the performance of IRL-based predictors is hampered by scene information loss and inefficient feature extraction. This contrasts with state-of-the-art supervised models, which typically employ vectorized representations [29, 21, 14, 26]. To bridge this gap, we propose adopting a query-centric paradigm that can effectively aggregate vectorized features into spatial grid tokens, thereby empowering the IRL-based predictor with exceptional motion forecasting capabilities.

Overall, the main contributions of this work can be summarized as follows: (1) We present a novel **Q**uery-centric **I**nverse **R**einforcement **L**earning (QIRL) framework for the motion forecasting task, which is the first to integrate the MaxEnt IRL paradigm with vectorized context representation through the query-centric fashion. (2) We introduce a hierarchical DETR-like decoder with a refinement module to improve prediction accuracy. (3) Our approach achieves highly competitive performance on the large-scale Argoverse motion forecasting benchmarks [2] compared to existing supervised models.

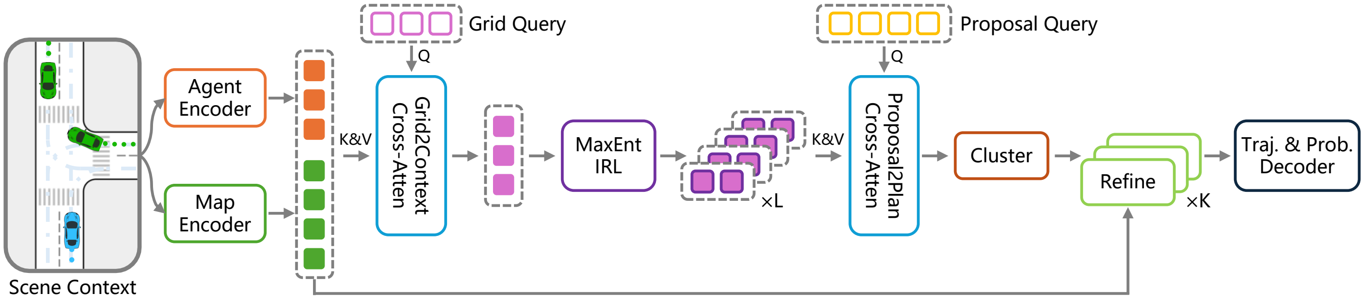

Fig. 1: Overview of the QIRL framework. We integrate the MaxEnt IRL paradigm with the query-centric motion forecasting pipeline to showcase its effectiveness. The scene context features are extracted using simple encoders and then aggregated to the grid queries. The sampled tokens derived from the MaxEnt IRL process are further employed to generate trajectory proposals. After clustering and refinement, multi-modal future trajectories along with their corresponding confidences are finally obtained by the trajectory and probability decoder.

II. Methodology

A. Framework Overview

An overview of the proposed QIRL framework is presented in Fig. 1, demonstrating its seamless integration of the MaxEnt IRL paradigm with the query-centric trajectory prediction pipeline. Firstly, we represent the driving context in a vectorized manner and leverage the agent encoder and map encoder to extract scene features. These fused features are then aggregated into spatial grid tokens through cross-attention mechanisms. Subsequently, the grid-based MaxEnt IRL algorithm is employed to infer the reward distribution, thereby obtaining the optimal policy that can be sampled to induce multiple plausible plans or traversals over the 2-D grid map. Finally, we introduce a DETR-like [1] trajectory decoder accompanied by a refinement module to generate multi-modal future trajectories in a hierarchical fashion.

B. Query-Centric Context Encoder

Considering context rasterization suffers from information loss, we represent agent trajectories and corresponding HD maps in a vectorized manner. Herein, we adopt the target-centric coordinate system, where all context instances and sequences are normalized to the current state of the target agent through translation and rotation operations. Subsequently, we leverage an agent encoder which is a simple 1-D CNN with a feature pyramid network (FPN) [11, 12] to encode the kinematic profiles of all agents in the scene, encompassing historical positions, velocities, and other relevant attributes. Additionally, we devise a PointNet-like network [19, 6] as the map decoder to extract static map features. The resulting agent and map features are then concatenated to form the context tokens. Unlike most IRL-based methods that commonly depend on image-like or rasterized features as input, we introduce learnable grid-shaped queries with 2-D spatial relative positional embeddings to aggregate the vectorized context features using the Grid2Context cross-attention module. The updated grid tokens are further reshaped into a grid-shaped map, thereby seamlessly adapting to the IRL framework.

C. MaxEnt IRL-based Policy Generator

Using the grid tokens as input, we employ a neural network to construct a nonlinear mapping from the context features to the reward function $\mathcal{R}$, which serves as a succinct representation of the driving context. Following the MaxEnt IRL framework [30], the probability of a state sequence (or plan), denoted $P(\tau)$, is directly proportional to the exponential of the total reward accumulated over the planning horizon $\mathcal{H}$:

$$P(\tau) = \frac{1}{Z}\exp\left(\mathcal{R}(\tau)\right) = \frac{1}{Z}\exp\left(\sum_{i=1}^{\mathcal{H}} \mathcal{R}(s_i)\right), \tag{1}$$

where $\tau = [s_1, \dots, s_{\mathcal{H}}]$ represents any given plan, s_i indicates the i-th state, and Z denotes the partition function. Further, we convert the continuous-valued future trajectories from the dataset into discrete state sequences using a simple uniform quantization technique with a specific resolution, constituting a set of demonstrations denoted as $\mathcal{D} = \{\tau_1, \dots, \tau_{|\mathcal{D}|}\}$. The objective is to maximize the log-likelihood of the demonstration data $\mathcal{L}_{\mathcal{D}}$ under the MaxEnt distribution. This optimization problem can be solved by employing gradient-based approaches, and the gradient is given by

$$\nabla\mathcal{L}_{\mathcal{D}} = \left(\mu_{\mathcal{D}} - \mathbb{E}[\mu]\right)\nabla\mathcal{R}, \tag{2}$$

where $\mu_{\mathcal{D}}$ represents the average state visitation frequencies (SVFs) from the demonstrations and $\mathbb{E}[\mu]$ refers to the expected SVFs under the policy [24], which can be derived via a forward RL process given the current reward distribution. Here, we leverage the approximate value iteration algorithm to obtain the policy $\pi(a|s)$ with the following expression:

$$\pi(a|s) = \exp\left(Q(s,a) - V(s)\right), \tag{3}$$

where $Q(s,a)$ represents the action-value function and $V(s)$ refers to the state-value function. Upon convergence of the reward distribution, we can acquire the optimal MaxEnt policy π^*, which enables the generations of multiple plans over the 2-D grid, acting as trajectory generation priors.

TABLE I: Quantitative results on the Argoverse 1 motion forecasting benchmark. The best and the second-best results are in **bold** and underlined, respectively. All metrics follow a lower-the-better criterion. brier-minFDE$_6$ is the official ranking metric.

Method	MR$_1$	minADE$_1$	minFDE$_1$	MR$_6$	minADE$_6$	minFDE$_6$	brier-minFDE$_6$
mmTransformer [13]	0.6178	1.7737	4.0033	0.1540	0.8436	1.3383	2.0328
SceneTransformer [16]	0.5921	1.8108	4.0551	0.1255	0.8026	1.2321	1.8868
HiVT [28]	0.5473	1.5984	3.5328	0.1267	0.7735	1.1693	1.8422
MultiPath++ [23]	0.5645	1.6235	3.6141	0.1324	0.7897	1.2144	1.7932
SIMPL [26]	0.5796	1.7501	3.9668	0.1165	0.7693	1.1545	1.7469
Wayformer [14]	0.5716	1.6360	3.6559	0.1186	0.7676	1.1616	1.7408
QCNet [29]	**0.5257**	**1.5234**	**3.3420**	**0.1056**	**0.7340**	**1.0666**	**1.6934**
QIRL (Ours)	0.5453	1.5824	3.4300	0.1209	0.7977	1.1652	1.7363

D. Hierarchical DETR Trajectory Decoder

Based on the converged reward, the MaxEnt policy, updated grid tokens, and context features, we introduce a DETR-like decoder with refinement to generate K multimodal predictions and their corresponding probabilities in a hierarchical fashion.

In the initial proposal stage, we generate multiple plausible trajectory proposals conditioned on the sampled grid tokens. To achieve this, we first employ the learned MaxEnt policy to induce plans over the 2-D grid map using the Markov chain Monte Carlo (MCMC) sampling strategy. Each sampled plan is then utilized to extract the associated grid tokens and relative position coordinates. After concatenating them as the plan feature embedding, we leverage trajectory proposal queries to aggregate these features with a Proposal2Plan cross-attention module. Recognizing the inefficiency and redundancy of sampling only K plans, as a small number of samples tend to produce similar outputs [4], we oversample L ($L \gg K$) plans in parallel, thereby inducing L trajectory proposals to better capture the trajectory distribution. Consequently, we derive K possible future trajectories from the set of L candidates through a clustering module.

In the second refinement stage, we predict trajectory offsets conditioned on the K initial trajectory proposals acting as anchors. Each trajectory proposal serves as a query and retrieves its nearby context features using a DETR-like decoder similar to the one used in the trajectory proposal module. The fused features are then fed into an MLP with residual connections, which comprises a regression head for producing predicted trajectory offsets, and a classification head followed by a softmax function for generating probabilities. Eventually, the predicted trajectory can be derived by summing the trajectory proposal and its corresponding offset.

E. Training Objectives

For the MaxEnt IRL process, our objective is to maximize the log-likelihood $\mathcal{L}_{\mathcal{D}}$ through stochastic gradient descent to derive the reward model and optimal policy, as explained in Section II-C. As for trajectory generation, the overall learning objective is composed of both regression loss and classification loss. Specifically, for regression, we apply the Huber loss to the predicted trajectory proposal $\mathcal{L}^{\mathrm{P}}_{reg}$, the refined trajectory $\mathcal{L}^{\mathrm{T}}_{reg}$, and its corresponding goal point $\mathcal{L}^{\mathrm{G}}_{reg}$. The winner-takes-all (WTA) training strategy is employed to mitigate the modality collapse issue, which only considers the best candidate with the minimum error in comparison to the ground truth. As for classification, we adopt the Hinge loss $\mathcal{L}_{cls}$ to distinguish the positive modality from the others, following the approach outlined in [11]. The total loss $\mathcal{L}$ in the trajectory decoder process can be expressed as follows:

$$\mathcal{L} = \mathcal{L}^{\mathrm{P}}_{reg} + \alpha\mathcal{L}^{\mathrm{T}}_{reg} + \beta\mathcal{L}^{\mathrm{G}}_{reg} + \gamma\mathcal{L}_{cls}, \tag{4}$$

where α, β, and γ are hyperparameters for balancing each loss component. In practice, we set $\alpha = \beta = 1$ and $\gamma = 3$.

III. Experiments and Results

A. Experimental Settings

1) Dataset: We train and evaluate the proposed approach on the large-scale Argoverse 1 motion forecasting dataset [2], which provides trajectory sequences collected from real-world urban driving scenarios, along with HD maps that encompass rich geometric and semantic information. Specifically, it consists of 205,942 training, 39,472 validation, and 78,143 testing sequences. Each sequence spans 5 seconds and is sampled at 10 Hz. The task involves forecasting the subsequent 3-second trajectories based on the preceding 2 seconds of observations.

2) Metrics: We evaluate the performance of trajectory prediction using the widely accepted metrics, including miss rate (MR$_K$), minimum average displacement error (minADE$_K$), minimum final displacement error (minFDE$_K$), and the Brier minimum final displacement error (brier-minFDE$_K$). Concretely, the MR$_K$ measures the proportion of scenarios where none of the K predicted endpoints fall within a 2.0-meter range of the ground truth. The minADE$_K$ calculates the average pointwise ℓ_2 distance between the best forecast among the K candidates and the ground truth, while the minFDE$_K$ solely focuses on the endpoint error. Furthermore, the brier-minFDE$_K$ incorporates prediction confidence by adding the Brier score $(1.0 - p)^2$ to minFDE$_K$, where p corresponds to the probability of the best forecast.

B. Quantitative Results

We conduct a comprehensive comparison of our approach with other state-of-the-art methods on the Argoverse motion forecasting benchmarks. The quantitative results on the Argoverse 1 test split are presented in Table I. As far as our knowledge extends, there is currently no publicly available IRL-based predictor on the Argoverse leaderboard. Therefore,

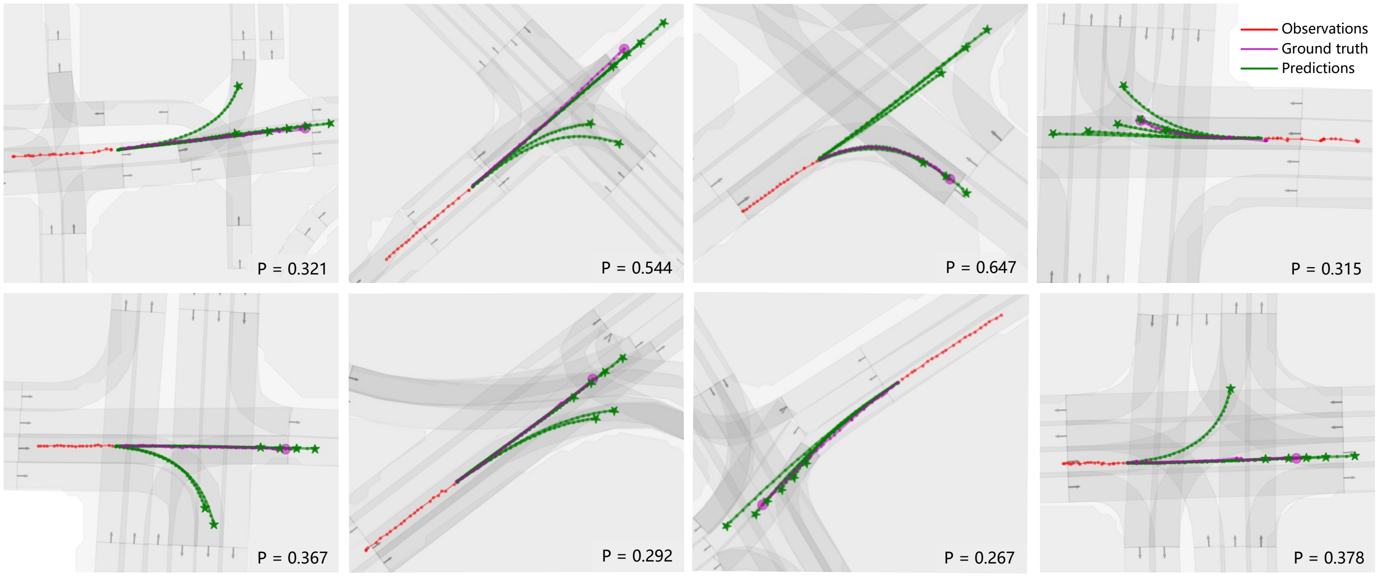

Fig. 2: Qualitative results of QIRL on the Argoverse 1 validation set. The historical trajectory, ground-truth future trajectory, and multi-modal predictions are depicted in red, magenta, and green, respectively. The lower-right corner showcases the probability associated with the best forecast in terms of the endpoint.

we compare our QIRL framework against several state-of-the-art supervised models utilizing transformer-based network architectures. Note that the current official ranking metric, brier-minFDE$_6$, takes both the precision of forecasted endpoints and the prediction confidence into account. It is evident from the results that our proposed method achieves highly competitive performance among the listed supervised models. In particular, QIRL outperforms strong baselines such as MultiPath++ [23], SIMPL [26], and Wayformer [14] in various evaluation metrics. Although our proposed QIRL falls slightly short compared to the top-ranked supervised model, QCNet [29], the findings unmistakably showcase the competitive performance of the IRL-based predictor and indicate its potential for further improvement in motion forecasting tasks.

C. Qualitative Results

We present visualizations of our proposed QIRL framework under diverse traffic scenarios from the Argoverse 1 validation set, as illustrated in Fig. 2. The qualitative results highlight the exceptional capability of our approach in accurately anticipating and generating feasible multi-modal future trajectories that align with the scene layout across a range of scenarios. This includes complex intersections and long-range situations, which demonstrates the effectiveness of our IRL-based predictor.

IV. Conclusion

In this work, we introduce QIRL, a query-centric inverse reinforcement learning framework for motion forecasting in autonomous driving. To the best of our knowledge, QIRL is the first trajectory predictor that combines the MaxEnt IRL paradigm with vectorized context representations through the query-centric pipeline. Additionally, the hierarchical DETR-like trajectory decoder significantly enhances prediction accuracy. Experimental results showcase that QIRL excels in generating scene-compliant multi-modal future trajectories and achieves highly competitive performance when compared to state-of-the-art supervised methods. Moreover, our work underscores the effectiveness of IRL-based predictors and provides a promising baseline for further investigations. Future work will involve evaluating the proposed QIRL on diverse datasets to assess its generalization abilities and extending its application to joint multi-agent motion forecasting scenarios.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Ming-Fang Chang, John Lambert, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

[3] Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020.

[4] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pages 203–212, 2021.

[5] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[6] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vec-

torized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

[7] Ke Guo, Wenxi Liu, and Jia Pan. End-to-end trajectory distribution prediction based on occupancy grid maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2242–2251, 2022.

[8] Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Conditional predictive behavior planning with inverse reinforcement learning for human-like autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[9] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 201–214. Springer, 2012.

[10] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11):1289–1307, 2016.

[11] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556, 2020.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[13] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021.

[14] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, et al. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 International Conference on Robotics and Automation (ICRA)*, pages 2980–2987. IEEE, 2023.

[15] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *International conference on machine learning*, volume 1, page 2, 2000.

[16] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.

[17] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018.

[18] Tung Phan-Minh, Forbes Howington, et al. Driveirl: Drive in real life with inverse reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1544–1550. IEEE, 2023.

[19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[20] Daniela Ridel, Nachiket Deo, Denis Wolf, and Mohan Trivedi. Scene compliant trajectory forecast with agent-centric spatio-temporal grids. *IEEE Robotics and Automation Letters*, 5(2):2816–2823, 2020.

[21] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[22] Haoran Song, Di Luan, Wenchao Ding, Michael Y Wang, and Qifeng Chen. Learning to predict vehicle trajectories with model-based planning. In *Conference on Robot Learning*, pages 1035–1045, 2021.

[23] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

[24] Markus Wulfmeier, Dushyant Rao, Dominic Zeng Wang, Peter Ondruska, and Ingmar Posner. Large-scale cost function learning for path planning using deep inverse reinforcement learning. *The International Journal of Robotics Research*, 36(10):1073–1087, 2017.

[25] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8660–8669, 2019.

[26] Lu Zhang, Peiliang Li, Sikang Liu, and Shaojie Shen. Simpl: A simple and efficient multi-agent motion prediction baseline for autonomous driving. *IEEE Robotics and Automation Letters*, 2024.

[27] Yanfu Zhang, Wenshan Wang, Rogerio Bonatti, Daniel Maturana, and Sebastian Scherer. Integrating kinematics and environment context into deep inverse reinforcement learning for predicting off-road vehicle trajectories. In *Conference on Robot Learning*, pages 894–905, 2018.

[28] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022.

[29] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023.

[30] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.